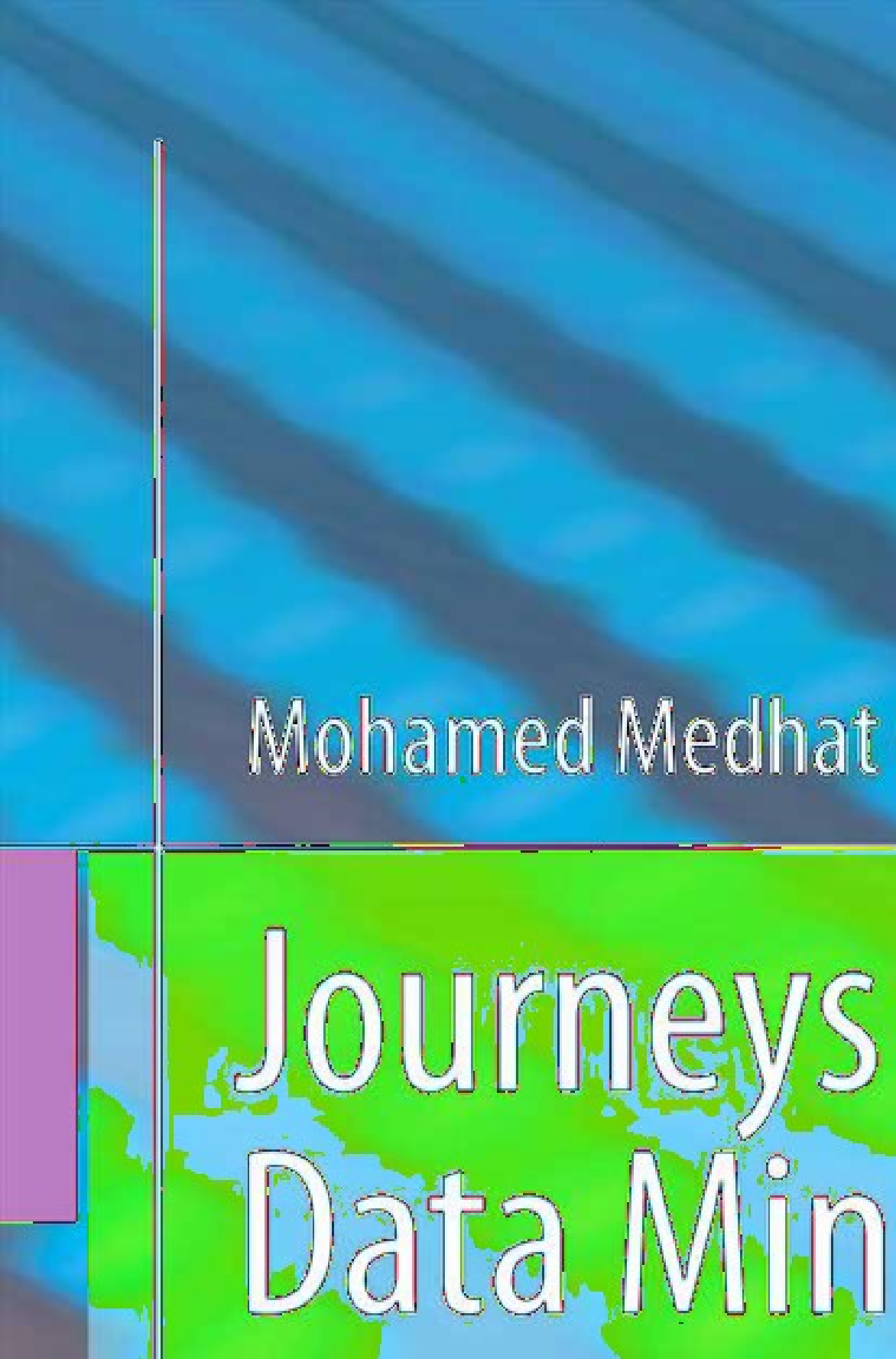

Mohamed Medhat
Journeys
Data Min

Journeys to Data Mining

Mohamed Medhat Gaber
Editor

Journeys to Data Mining

Experiences from 15 Renowned Researchers

Springer

Editor
Mohamed Medhat Gaber
School of Computing
University of Portsmouth
Portsmouth
United Kingdom

ISBN 978-3-642-28046-7 ISBN 978-3-642-28047-4 (eBook)
DOI 10.1007/978-3-642-28047-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012942594

ACM Computing Classification (1998): H.3, I.2, I.7, G.3, K.7

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

List of Contributors

Dean W. Abbott Abbott Analytics, Inc., San Diego, CA, USA, dean@abbottanalytics.com

Charu C. Aggarwal IBM T.J. Watson Research Center, Hawthorne, NY, USA, charu@us.ibm.com

Michael R. Berthold Department of Computer and Information Science, University of Konstanz, Konstanz, Germany, michael.berthold@uni-konstanz.de

Christopher W. Clifton Department of Computer Sciences, Purdue University, West Lafayette, IN, USA, clifton@cs.purdue.edu

John F. Elder IV Elder Research, Inc., Charlottesville, VA, USA, elder@dataminginglab.com

David J. Hand Department of Mathematics, Imperial College, London, UK, d.j.hand@imperial.ac.uk

Cheryl G. Howard IBM Corporation, Washington, DC, USA, cghoward@us.ibm.com

J. Dustin Hux VP Analytics, Elder Research, Inc., Charlottesville, VA, USA, dustin@dataminginglab.com

Hillol Kargupta Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, MD, USA; Agnik, LLC, Columbia, MD, USA, hillol@cs.umbc.edu; hillol@agnik.com

Colleen McLaughlin McCue GeoEye, Herndon, VA, USA, mccue.colleen@geoeye.com

Geoff McLachlan Department of Mathematics, University of Queensland, St. Lucia, Brisbane, QLD, Australia, gjm@maths.uq.edu.au

Gregory Piatetsky-Shapiro KDnuggets, Brookline, MA, USA, Gregory@kdnuggets.com

Shusaku Tsumoto Department of Medical Informatics, Faculty of Medicine, Shimane University, Shimane, Japan, tsumoto@computer.org

Graham J. Williams Togaware Pty Ltd., Canberra, ACT, Australia, Graham.Williams@togaware.com

Mohammed J. Zaki Rensselaer Polytechnic Institute, Troy, NY, USA, zaki@cs.rpi.edu

Introduction

Mohamed Medhat Gaber

"If I have seen further it is only by standing on the shoulders of giants"
by Sir Isaac Newton (1643–1727)

1 Preamble

It has been a great honour to have been given the opportunity to edit this book and a great pleasure to work with such a respected group of data mining scientists and professionals. It is our belief that the knowledge provided by studying the journeys these respected and recognised individuals took through the area of data mining is as important as simply gaining the required knowledge in the field. The contributors to this volume are successful scientists and professionals within the field of data analytics. All the authors in this volume have helped to shape the field of data analytics through their many valuable contributions.

It all began with a workshop co-organised by one of the contributors to this book, namely, Dr. Gregory Piatetsky-Shapiro in conjunction with the International Joint Conference on Artificial Intelligence (IJCAI) in 1989. Today, the number of publication venues and dedicated data analytics companies reflects the fast growing interest in the data mining field.

My own journey while editing this book has been quite remarkable. Invitations were sent to a number of renowned researchers and practitioners in the field. The feedback received from the invitees was very positive. However, other commitments made it difficult for some of these great researchers to contribute. Despite not being able to contribute, many of them were very supportive of the

M.M. Gaber (✉)
School of Computing, University of Portsmouth, Buckingham Building, BK1.41, Lion Terrace, Portsmouth, Hampshire PO1 3HE, UK
e-mail: mohamed.gaber@port.ac.uk

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_1,

project. During my journey through the editing of this book it was a great pleasure to receive chapter after chapter from the contributors and to read the amazing journeys they took through data analytics. My own journey took 18 months to complete, being the longest time I ever needed to edit a book. Nonetheless, it has probably been the most enjoyable editing experience I have had. I have very much enjoyed communicating with the renowned scientists that contributed to this book.

The rationale behind editing this book has been to teach young researchers how they can proceed in the data mining area and gain recognition. Some of our author's journeys, as the reader will see later in the book, are very interesting showing how talented individuals changed their careers and were then able to achieve recognition. The book is not only targeted at young researchers, but also established data mining scientists and practitioners will find it very useful to read. Learning how fields are related to each other and how to build a portfolio of skills in order to be recognised as a data scientist are needed by both early career researchers and the more established ones.

The contributors were asked to describe their personal journeys through the field of data mining whilst answering the questions listed below. Although a chapter structure was suggested we thought it is important to give our authors the freedom to organise their chapters in the way that best fits their own personal journey. This has in fact resulted in an interesting collection of chapter structures, each suiting the journey the chapter narrates.

1. What are your motives for conducting research in the data mining field?
2. Describe the milestones of your research in this field.
3. What are your notable success stories?
4. What did you learn from your failures?
5. Have you encountered unexpected results?
6. What are the current research issues and challenges in your area?
7. Describe your research tools and techniques.
8. What would you advise a young researcher to make an impact?
9. What do you predict for the next 2 years in your area?
10. What are your expectations in the long term?

Below we will introduce the reader to the distinguished contributors to this book. Please note that the chapter order is alphabetical according to the author's surname.

2 Dean Abbott

Dean Abbott is President of Abbott Analytics, Inc. in San Diego, California. Mr. Abbott has over 21 years of experience applying advanced data mining, data preparation, and data visualisation methods in real-world data-intensive problems, including fraud detection, response modelling, survey analysis, planned giving, predictive toxicology, signal process, and missile guidance. In addition, he has

developed and evaluated algorithms for use in commercial data mining and pattern-recognition products, including polynomial networks, neural networks, radial basis functions, and clustering algorithms, and has consulted with data mining software companies to provide critiques and assessments of their current features and future enhancements.

Mr. Abbott is a seasoned instructor, having taught a wide range of data mining tutorials and seminars for a decade to audiences of up to 400, including DAMA, KDD, AAAI, and IEEE conferences. He is the instructor of well-regarded data mining courses, explaining concepts in language readily understood by a wide range of audiences, including analytics novices, data analysts, statisticians, and business professionals. Mr. Abbott has also taught both applied and hands-on data mining courses for major software vendors, including Clementine (SPSS, an IBM Company), Affinium Model (Unica Corporation), Statistica (StatSoft, Inc.), S-Plus and Insightful Miner (Insightful Corporation), Enterprise Miner (SAS), Tibco Spotfire Miner (Tibco), and CART (Salford Systems).

3 Charu Aggarwal

Charu Aggarwal is a Research Scientist at the IBM T.J. Watson Research Center in Yorktown Heights, New York. He completed his BS from IIT Kanpur in 1993 and his Ph.D. from Massachusetts Institute of Technology in 1996. He has since worked in the field of performance analysis, databases, and data mining. He has published over 155 papers in refereed conferences and journals, and has been granted over 60 patents. Because of the commercial value of the above-mentioned patents, he has received several invention achievement awards and has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bio-terrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of an IBM Research Division Award (2008) for his scientific contributions to data stream research.

He has served on the program committees of most major database/data mining conferences, and served as program vice-chairs of the SIAM Conference on Data Mining, 2007, the IEEE ICDM Conference, 2007, the WWW Conference, 2009, and the IEEE ICDM Conference, 2009. He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering Journal from 2004 to 2008. He is an associate editor of the ACM TKDD Journal, an action editor of the Data Mining and Knowledge Discovery Journal, an associate editor of the ACM SIGKDD Explorations Journal, and an associate editor of the Knowledge and Information Systems Journal. He is a fellow of the IEEE for “contributions to knowledge discovery and data mining techniques”, and a life member of the ACM.

4 Michael Berthold

After receiving his Ph.D. from Karlsruhe University, Germany, Michael Berthold spent over 7 years in the US, among others at Carnegie Mellon University, Intel Corporation, the University of California at Berkeley and—most recently—as director of an industrial think tank in South San Francisco. Since August 2003 he has held the Nycomed Chair for Bioinformatics and Information Mining at Konstanz University, Germany, where his research focuses on using machine learning methods for the interactive analysis of large information repositories in the Life Sciences. Most of the research results are made available to the public via the open source data mining platform KNIME. In 2008 M. Berthold co-founded KNIME.com AG, located in Zurich, Switzerland. KNIME.com offers consulting and training for the KNIME platform in addition to an increasing range of enterprise products.

M. Berthold is a Past President of the North American Fuzzy Information Processing Society, Associate Editor of several journals and the President of the IEEE System, Man, and Cybernetics Society. He has been involved in the organisation of various conferences, most notably the IDA-series of symposia on Intelligent Data Analysis and the conference series on Computational Life Science. Together with David Hand he co-edited the successful textbook "Intelligent Data Analysis: An Introduction", which has recently appeared in a completely revised, second edition. He is also coauthor of the brand new "Guide to Intelligent Data Analysis" (Springer Verlag), which appeared in summer 2010.

5 John Elder

Dr. John Elder heads the USA's leading data mining consulting team—with offices in Charlottesville Virginia and Washington, DC (http://www.datamininglab.com). Founded in 1995, Elder Research, Inc. focuses on investment, commercial and security applications of advanced analytics, including text mining, image recognition, process optimisation, cross-selling, biometrics, drug efficacy, credit scoring, market sector timing, and fraud detection.

John obtained a BS and MEE in Electrical Engineering from Rice University, and a Ph.D. in Systems Engineering from the University of Virginia, where he is an adjunct professor teaching Optimization or Data Mining. Prior to 17 years at ERI, he spent 5 years in aerospace defence consulting, four heading research at an investment management firm, and two in Rice's Computational and Applied Mathematics department. Dr. Elder has authored innovative data mining tools, is a frequent keynote speaker, and was co-chair of the 2009 Knowledge Discovery and Data Mining conference in Paris. John was honoured to serve for 5 years on a panel appointed by the President to guide technology for National Security. His book with Bob Nisbet and Gary Miner, Handbook of Statistical Analysis and Data Mining Applications, won the PROSE award for Mathematics in 2009. His book

with Giovanni Seni, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions, was published in February 2010. A book on Practical Text Mining was published in January 2012.

6 Chris Clifton

Chris Clifton is an Associate Professor of Computer Science and (by courtesy) Statistics at Purdue University, and director of the Indiana Center for Database Systems. His primary research is on technology ensuring privacy in the analysis and management of data. He also works on challenges posed by novel uses of data mining technology, including data mining of text and data mining techniques applied to interoperation of heterogeneous information sources. Prior to joining Purdue, Dr. Clifton was a principal scientist in the Information Technology Division at the MITRE Corporation. Before joining MITRE in 1995, he was an assistant professor of computer science at Northwestern University. He has a Ph.D. from Princeton University, and Bachelor's and Master's degrees from the Massachusetts Institute of Technology, all in Computer Science.

7 David Hand

David Hand studied mathematics at Oxford University and statistics and pattern recognition at the University of Southampton. He has been Professor of Statistics at Imperial College, London, since 1999, and before that, from 1988 to 1999 was Professor of Statistics at the Open University, where he made a number of television programmes about statistics. He is currently on leave, working as Chief Scientific Advisor to Winton Capital Management, one of Europe's leading hedge funds. He was a member of Lord Oxburgh's enquiry panel into the UEA's Climategate affair in 2010, and has served in many other public and private advisory roles, including serving on the AstraZeneca Expert Statistics Panel, the GlaxoSmithKline Biometrics Advisory Board, the Office for National Statistics Methodology Advisory Committee, and the Technical Opportunities Panel of the Engineering and Physical Sciences Research Council. He served a term of office as president of the Royal Statistical Society for 2008 and 2009, and a second term for 2010.

He has published 26 books, including Principles of Data Mining, and over 300 papers. In 1999 he was elected an Honorary Fellow of the Institute of Actuaries, and in 2003 a Fellow of the British Academy, the UK's National Academy for the Humanities and Social Sciences. He won the Royal Statistical Society's Guy Medal in Silver in 2002, and the IEEE-ICDM Outstanding Contributions Award in 2004. In 2006 he was awarded a Wolfson Research Merit Award from the Royal Society, the UK's national academy for the natural sciences.

8 Cheryl Howard

Dr. Cheryl Howard has worked in the fields of systems engineering, machine learning, and predictive analytics for over 25 years. After graduating from The University of Rochester, NY, she began her career at the US Army Center for Night Vision and Electro-Optics in Fort Belvoir, VA; there she became interested in the application of intelligent image analysis to the challenges of automated target recognition. Her concurrent doctoral research implemented a module for extracting geometric concepts from texture map images; this module formed part of a general-purpose machine learning system for image and signal analysis (Bock et al. 1993). The resulting system was applied to a wide range of industrial challenges under the sponsorship of Robert Bosch, GmbH at the Research Institute for Applied Knowledge Processing (FAW) in Ulm, Germany. After receiving her doctoral degree from The George Washington University in Washington, DC, she joined the research laboratories of the Thomson Corporation where she applied data mining and machine learning to problems in the information publishing and financial markets.

Dr. Howard spent 10 years as Vice President of Research at Elder Research, Inc. where she specialised in fraud and insider threat detection for the public sector. She was responsible for the development and deployment of several highly successful fraud detection applications. She is currently a Senior Managing Consultant at IBM Corporation in the Washington, DC area.

9 Hillol Kargupta

Hillol Kargupta is a Professor of Computer Science at the University of Maryland, Baltimore County. He is also a co-founder of AGNIK LLC, a data analytics company for mobile, distributed, and embedded environments. He received his Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign in 1996. His research interests include mobile and distributed data mining.

Dr. Kargupta is an IEEE Fellow. His work received the 2010 Frost and Sullivan Enabling Technology of the Year Award. He won the IBM Innovation Award in 2008 and a National Science Foundation CAREER award in 2001 for his research on ubiquitous and distributed data mining. He and his team received the 2010 Frost and Sullivan Enabling Technology of the Year Award for the MineFleet vehicle performance data mining product. His other awards include the 2010 IEEE Top-10 Data Mining Case Studies Award for his work at Agnik, the best paper award for the 2003 IEEE International Conference on Data Mining for a paper on privacy-preserving data mining, the 2000 TRW Foundation Award, and the 1997 Los Alamos Award for Outstanding Technical Achievement. His dissertation earned him the 1996 Society for Industrial and Applied Mathematics annual best student paper prize.

He has published more than one hundred peer-reviewed articles. His research has been funded by the US National Science Foundation, US Air Force, Department of Homeland Security, NASA and various other organisations. He has co-edited

several books. He serve(s/d) as an associate editor of the IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Systems, Man, and Cybernetics, Part B and Statistical Analysis and Data Mining Journal. He is/was the Program co-chair of the 2009 IEEE International Data Mining Conference, general chair of 2007 NSF Next Generation Data Mining Symposium, Program co-chair of 2005 SIAM Data Mining Conference and Associate general chair of the 2003 ACM SIGKDD Conference, among others. For more information please visit: http://www.cs.umbc.edu/hillol

10 Dustin Hux

Dustin Hux has 15 years of applied statistical modelling and data mining experience. He has led teams solving problems in the commercial, financial, and scientific domains including fraud detection, cross-selling, customer profiling, product bundling, direct marketing, biometric identification, stock selection, market timing, text mining, and atmospheric modelling. Dustin is particularly honoured to be a part of projects that apply data mining and advanced analytics to challenges facing the intelligence community.

Mr. Hux is expert with leading data mining software tools and advises vendors on enhancements. For KDD, he was on the 2004 Data Mining Standards, Services, and Platforms Program Committee and was 2006 Sponsorship Committee co-chair. Dustin earned a Master's degree in Environmental Sciences from the University of Virginia and Bachelor's degrees in Biology and Economics from Emory and Henry College.

11 Colleen McCue

Dr. Colleen McLaughlin McCue is the Senior Director, Social Science and Quantitative Methods at GeoEye. In this role, she supports a variety of public safety, national security, and commercial clients; bringing more science and less fiction to the field of operational security analytics and helping her clients gain the insight necessary to prevent crime and improve public safety outcomes. Dr. McCue brings over 18 years of experience in advanced analytics and the development of actionable solutions to complex information processing problems in the applied public safety and national security environment. Her areas of expertise include the application of data mining and predictive analytics to the analysis of crime and intelligence data, with particular emphasis on deployment strategies, surveillance detection, threat and vulnerability assessment, and the behavioural analysis of violent crime.

Dr. McCue's experience in the applied law enforcement setting and pioneering work in operationally relevant and actionable analytical strategies has been used to support a wide array of national security and public safety clients. In her free time she

enjoys reading books on science, medicine, nature, and business process in an effort to identify novel approaches to security analytics and advance the science. Dr. McCue has published her research findings in journals and book chapters, and has authored a book on the use of advanced analytics in the applied public safety environment entitled, Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis. She earned her undergraduate degree from the University of Illinois at Chicago and Doctorate in Psychology from Dartmouth College, and completed a 5-year postdoctoral fellowship in the Department of Pharmacology and Toxicology at the Medical College of Virginia where she received additional training in pharmacology and molecular biology. Dr. McCue lives with her husband, NCIS Supervisory Special Agent (ret.) Richard J. McCue, and their children in Richmond, Virginia.

12 Geoff McLachlan

Geoff McLachlan is Professor of Statistics in the Department of Mathematics and a Professorial Research Fellow in the Institute for Molecular Bioscience at the University of Queensland. He is also a chief investigator in the ARC (Australian Research Council) Centre of Excellence in Bioinformatics. He currently holds an ARC Professorial Fellowship. He has written numerous research articles and six monographs, the last five in the Wiley series in Probability and Statistics. They include "Discriminant Analysis and Statistical Pattern Recognition", "Finite Mixture Models" (coauthored with David Peel), and the "EM Algorithm" and Extensions (with Thriyambakam Krishnan). His current research interests are focussed on the fields of machine learning, multivariate analysis, and bioinformatics.

In 1994, he was awarded a Doctor of Science by the University of Queensland on the basis of his publications in Statistics and in 1998 was made a fellow of the American Statistical Association. He is an ISI Highly Cited Author in the category of Mathematics and was recently awarded the Pitman gold medal of the Statistical Society of Australia in recognition of his contributions to the discipline of Statistics. Also, he won the IEEE-ICDM Outstanding Contributions Award in 2011. He is currently President of IFCS (the International Federation of Classification Societies). He was head of the Mathematics Department at the University of Queensland from 2007 to 2009 and was a panel member of the College of Experts of the Australian Research Council for 2008–2010. He has served on the editorial boards of numerous journals and is currently an associate editor for BMC Bioinformatics, Journal of Classification, Statistics and Computing, Statistical Modelling, and Statistics Surveys.

13 Gregory Piatetsky-Shapiro

Gregory Piatetsky-Shapiro, Ph.D. is the President of KDnuggets, which provides research and consulting services in business analytics and data mining. Previously, he led data mining groups at GTE Laboratories, Knowledge Stream Partners, and Xchange.

He has extensive experience in applying analytic methods to many areas including customer modelling, healthcare data analysis, fraud detection, bioinformatics, and Web analytics, and analyzed data for many leading companies in banking, e-commerce, insurance, telecom, and pharma fields.

Gregory is also the Editor of KDnuggets News, the leading newsletter on analytics and data mining, and the Editor of the http://www.KDnuggets.com site, a top-ranked site for analytics and data mining, covering news, software, jobs, companies, courses, education, publications and more.

Gregory coined the term Knowledge Discovery in Data (KDD) when he organised and chaired the first three workshops on KDD in 1989, 1991, and 1993. These workshops later grew into KDD Conferences (http://www.kdd.org), currently the leading conference in the field. Gregory was also a founding editor of the Data Mining and Knowledge Discovery Journal.

Gregory is a co-founder of ACM SIGKDD, the leading professional organisation for Knowledge Discovery and Data Mining and served as the Chair of SIGKDD (2005–2009). He also serves on the Steering Committee of the IEEE International Conference on Data Mining.

As a visiting professor at Connecticut College (2003) Gregory taught a course on Data Mining and developed teaching materials which are freely available on the Web.

Gregory received the ACM SIGKDD Service Award (2000) and IEEE ICDM Outstanding Service Award (2007).

Gregory has over 60 publications, including two best-selling books and several edited collections on topics related to data mining and knowledge discovery.

He was born in Moscow, Russia and received his MS and Ph.D. from New York University. He is married and has two children.

14 Shusaku Tsumoto

Shusaku Tsumoto graduated from Osaka University, School of Medicine in 1989, during which time he was involved in developing a medical expert system. After his time as a resident of neurology at Chiba University Hospital, he worked in the emergency division (ER room) at Matsudo Municipal Hospital from 1989 to 1991. He then moved to the Division of Medical Informatics in Chiba University Hospital and was involved in developing a hospital information system from 1991 to 1993. He moved to Tokyo Medical and Dental University in 1993 and started his research on rough sets and data mining in biomedicine. He received his Ph.D. (Computer Science) on application of rough sets to medical data mining from Tokyo Institute of Technology in 1997. He became a Professor at the Department of Medical Informatics, Shimane University in 2000. From this year he has been in charge of the network system on the Izumo Campus of Shimane University and of the hospital information system in Shimane University Hospital. In 2008, he became a visiting professor of the Institute of Statistics and Mathematics. His research interests

include approximate reasoning, contingency matrix theory, data mining, fuzzy sets, granular computing, knowledge acquisition, intelligent decision support, mathematical theory of data mining, medical informatics, rough sets, risk sciences, and service-oriented computing. He served as a president of the International Rough Set Society from 2000 to 2005 and served as a co-chair of the Technical Committee on Granular Computing in the IEEE SMC society from 2008. He served as a PC chair of RSCTC2000, IEEE ICDM2002, RSCTC2004, ISMIS2005, and IEEE GrC2007 and as a Conference chair of PAKDD 2008 and IEEE GrC 2009. He also served as a workshop chair of IEEE ICDM2006 and as a publicity chair of SIAM DM2007, 2008 and CIKM2010.

15 Graham Williams

Dr. Graham Williams is Chief Data Miner with the Australian Taxation Office. Previously he was Principal Computer Scientist for Data Mining with CSIRO and Lecturer in Computer Science, Australian National University. He is now an Adjunct Professor University of Canberra and Australian National University, and International Expert and Visiting Professor of the Chinese Academy of Sciences.

Graham has been involved in many data mining projects for organisations including the Health Insurance Commission, the Australian Taxation Office, the Commonwealth Bank, NRMA Insurance Limited, Department of Health, and the Australian Customs Service. His significant achievements include Multiple Decision Tree Induction (1989), HotSpots for identifying target areas in very large data collections (1992), WebDM for the delivery of data mining services over the Web using XML (1995), and Rattle (2005), a simple to use Graphical User Interface for data mining. His text book on Data Mining with Rattle and R was published by Springer in 2011.

Graham is involved in numerous international artificial intelligence and data mining research activities and conferences, as chair of the steering committees for the Australasian Data Mining Conference and the Pacific Asia Knowledge Discovery and Data Mining conference. He is also a member of the steering committee of the Australian Artificial Intelligence conference.

Graham's Ph.D. (Australian National University, 1991) introduced the then new idea of combining multiple predictive models for the better understanding of data and predictive capability. The thesis explored algorithms for building and combining multiple decision trees. Similar approaches are now widely used as ensembles, boosting, and bagging, and provide significant gains for modelling.

Graham has worked for a number of organisations including: CSIRO Land and Water in Canberra, Australia, developing award-winning spatial expert systems (using Prolog); BBJ Computers as Research and Development and then Marketing Manager, overseeing the implementation of a data mining tool for integration with a 4GL database environment; and was involved in developing one of the first and longest deployed Expert Systems in Australia, for Esanda Finance, Melbourne, Australia.

16 Mohammed J. Zaki

Mohammed J. Zaki is a Professor of Computer Science at RPI. He received his Ph.D. in computer science from the University of Rochester in 1998. His research interests focus on developing novel data mining techniques, especially in bioinformatics. He has published over 200 papers and book chapters on data mining and bioinformatics.

He is the founding co-chair for the BIOKDD series of workshops. He is currently Area Editor for Statistical Analysis and Data Mining, and an Associate Editor for Data Mining and Knowledge Discovery, ACM Transactions on Knowledge Discovery from Data, Knowledge and Information Systems, ACM Transactions on Intelligent Systems and Technology, Social Networks and Mining, and International Journal of Knowledge Discovery in Bioinformatics. He was/is the Program co-chair for SDM'08, SIGKDD'09 and PAKDD'10, BIBM'11, CIKM'12 and ICDM'12. He received the National Science Foundation CAREER Award in 2001 and the Department of Energy Early Career Principal Investigator Award in 2002. He received an HP Innovation Research Award in 2010 and 2011. He is a senior member of the IEEE, and an ACM Distinguished Scientist.

17 Remarks

The reader will now be left to enjoy following the journeys of some of the great data miners that we all admire. As mentioned previously, the list of contributors that have kindly devoted their precious time to share their journeys with us represent only some of the notable data mining experts active in the field. There are many other successful experts in this area and we would be grateful if they would consider sharing their journeys of success with us in possible future volumes.

Data Mining: A Lifetime Passion

Dean Abbott

1 Early Data Mining

I loved watching and playing sports as a boy, and I began playing Farm League at age 8 and then Little League from age 10 to 12. I loved pitching because of the cognitive battle: nothing was more fun than figuring out a way to make the hitter look foolish by throwing a pitch that was unexpected. Over time, I began to notice patterns in individual hitters' approaches that could be exploited. My father and I would spend hours discussing hitters and how to get them out, including which pitch sequences would work best for each hitter. I did not think about it at that time, but in retrospect, this was data collection and predictive modeling in a rudimentary (but exciting!) form. The only thing I was missing was storing the transactional data in an Oracle database.

For me, an aspect of playing baseball that was almost as fun as pitching was computing the statistics after the game. I still have my game-by-game box scores complete with cumulative batting average and earned run average (ERA). I thank my father for teaching me how to calculate these statistics. ERA is not an easy thing for a 9 year old to compute, though it is a very nice statistic because it normalizes the runs allowed over a typical game. Tracking cumulative statistics helped give me a sense of how important the denominator is in computing statistics: the longer the time period for calculating rates, the more difficult it is to change their value.

This led to another childhood pastime that involved summary statistics: Strat-O-matic Baseball. For those unfamiliar with this game, their web site describes the game like this:

> Since 1961, Strat-O-Matic has created the most realistic simulation of statistically accurate baseball. No other game combines Strat-O-Matic's realism, statistical accuracy, and ease of play. That's because no other game matches Strat-O-Matic's in-depth research—up to

D. Abbott (✉)
Abbott Analytics, Inc., P.O. Box 22536, San Diego, CA 92192-2536, USA
e-mail: dean@abbottanalytics.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_2,

> 1,500 hours to recreate a season in supremely realistic detail. Easy enough to learn swiftly, Strat-O-Matic's ability to capture all the variety of real baseball keeps gamers coming back for more—for decades.

I would play hundreds of games and track the batting and pitching statistics for hundreds of players. Remember, these are the days of pencil and paper, so this required having sheet after sheet of box scores and stats, all calculated by hand. I still believe that calculating by hand deepens our understanding of the data's meaning in a way that calculators and computers cannot.

When the time came to apply to college, I was convinced of only three things: I wanted to be a mathematician, this new "computer" stuff was going to be valuable, and I did not want to take any more English. That narrowed me down to engineering schools (and one Ivy-Dartmouth), but in the end, I chose Rensselaer Polytechnic Institute in Troy, NY, over Georgia Tech and Rose-Hulman Institute of Technology.

2 Undergraduate Education

For undergraduate education, I went to Rensselaer Polytechnic Institute (RPI) in Troy, NY, where I had a choice of three mathematics majors, two of which were of interest to me: mathematics of computation and operations research. The most enjoyable of the courses I took were differential equations (ordinary and partial) and any computational course, such as numerical analysis, programming languages, numeric ODEs, and numeric PDEs. Unfortunately, I never took a probability or statistics course, which meant I had to acquire knowledge in these areas during my professional career.

When I arrived at RPI, the state of computing was as follows: Fortran was big, Pascal was an instructional language, punch cards were still in use, and green-text terminals were really, really cool. I had very little programming experience before arriving, only having dabbled in BASIC on paper tape, so learning WATFIV (Waterloo Fortran IV) was a challenge. The mathematics of computation degree also required taking courses in multiple computing languages, including APL, SNOBOL, LISP, and Pascal (but not yet C), and numerical analysis courses.

I was convinced by a friend to minor in electrical engineering. After taking the basic circuits and electronics courses, I focused on the "systems" courses: linear systems, discrete time systems, etc. These courses influenced my thinking throughout my career and led to my often-used phrase "all problems can be described by block diagrams."

In my senior year, I did the usual interviewing with companies on campus while looking for a job, and it became brutally clear that while my undergraduate degree in mathematics was foundational, it did not provide me with the practical skills employers were looking for. I was very interested in some of the projects the interviewers were describing to me, but I realized that I could provide no immediate, practical value to those projects—yet. At that point, I knew I had to

go to graduate school to get at least a master's degree. I ended up at the University of Virginia (UVA) in their Applied Mathematics Department, which was a part of the School of Engineering.

3 Graduate School

I continued down the path of differential equations and controls, taking every control systems course in applied math and engineering as much as I could while minimizing real analysis, a course that was difficult for me and not at all enjoyable (I think we can all agree here: it is pretty bad!). The courses that resonated most with me were optimal control and nonlinear control, where the latter relied heavily on computer methods in solving nonlinear equations, in contrast to older, pencil-and-paper methods of linearization and drawing isoclines.

Graduate school was a time of philosophical reflection: how would I really apply Lyapunov theory and Lebesgue integration theory in the business world? Two key episodes helped clarify why I should continue doing mathematics. The first was conveyed to me by the chairman of the Mathematics Department at UVA, Dr. James Taylor, who was also a friend and an elder at Trinity Presbyterian Church in Charlottesville. He told me that whenever he proves a theorem, he discovers something more about how a rational God designed the universe we live in. There is a great joy that results from uncovering a new piece of the huge mosaic that is mathematics. Dr. Taylor frequently took time to remind me with his British accent and dry sense of humor that as an *applied* mathematician, I was not truly a *proper* mathematician. I could only nod my head in humble agreement.

The second event unfolded as part of my master's project (a short version of a master's thesis). My advisor suggested that I investigate properties of systems of equations with periodic coefficients. It was interesting, but I was not sure what value this project would provide. Since I was right there in the EE Department, I asked three or four professors there about applications that might benefit from this kind of formulation. Each of them thought for a while and replied that they knew of none.

4 An Introduction to Data Mining: Barron Associates, Inc.

At about that time, I saw on the e-mail system a solicitation for consulting jobs at Barron Associates, Inc. (BAI), which intrigued me because it provided an outlet for using optimal control. I still had no experience in pattern recognition at this point. It became evident from the onset of my experience there that this was exactly what I had been missing: a clear way to apply what I had been learning as a student to real-world problems.

The president of BAI, Roger Barron, was familiar with the master's project requirement and suggested that I combine my interest in optimal control with

a practical need of his (i.e., a contract) to solve guidance problems through optimization. My advisor accepted the offer for Roger to be a technical supervisor of my master's project, while he would give feedback and provide the final approval.

Roger Barron was truly a pioneer in the use of statistical learning methods as applied to guidance and control problems. His first company, Adaptronic Corp., was described by Robert Hecht-Nielsen as the only commercially successful "neural network" company in existence [1]. I very much enjoyed the intellectual stimulation and work at BAI. Roger gave his young employees tremendous responsibility, and he expected a lot out of us. Yet no one worked harder or longer than Roger himself. Although it may not have been clear to me at the time, my experience at BAI formed and launched the trajectory of my career in data mining.

The young men who worked under Roger at BAI in the first year I was there included John Elder (now president of Elder Research, Inc., and author of the chapter "Driving Full Speed, Eyes on the Rear-View Mirror"), Paul Hess and Gerry Montgomery (the two went on to found Abtech Corp., a data mining software and services company, which later became Marketminer, Inc.), and Richard Cellucci, a fantastic computer scientist. Gene Parker joined BAI just before I left and eventually became president, and he is running the company today. These individuals were not only top-notch researchers, they were also personally enjoyable to work with. John, in particular, became a great friend as we went to the same church, played on the same softball teams, served on a missions committee together, though interestingly, we never coauthored a paper while at BAI; that would finally come about in 1998.

In my first project, I assisted Roger in developing a new version of the optimum path-to-go (OPTG) guidance algorithm as applied to an MK82 glide bomb retrofitted with an inertial guidance unit and fins. The first step of the process was to derive the equations for the Lagrange multipliers using the calculus of variations from the equations of motion and constraints. The optimum guidance commands were derived as functions of these Lagrange multipliers, and the system of equations (equations of motion, Lagrange multipliers, and control actions) were simulated to form the optimum trajectories.

The primary objective of the bomb was to hit the target with as much speed and as steeply as possible from any feasible launch position. One approach would be to simulate the launch from position X and try to hit target position Y. This requires an iterative search to find the optimum trajectory. John realized however that since every path is an optimum path, we do not need to search for the optimum solution from *each* particular X–Y pair; we can instead fly all trajectories from Y (the target position) backward to a family of initial starting positions. These became the set of optimum trajectories we could use in our solution.

However, these optimum trajectories cannot be utilized in real time. Instead, we began with the set of optimum trajectories stored as a data set (the "training data") and then used polynomial networks to predict the optimum guidance command action for any given position of the bomb in order to hit the specified target position. The predictions would be updated on a regular basis throughout the flight of the bomb (closed-loop guidance). It is the polynomial network, a statistical learning

network, that provided the means to implement the optimum guidance commands in real time [2] and [3].

I love that this work represented innovation in two distinct research areas: optimal control and data mining. It was also some of the most enjoyable work I have been a part of, despite the long hours it took to work through the formulations over and over again. But this hard work taught me what it takes to be at the top of your field: innovation, persistence, dedication, and lots of time.

Polynomial networks, sometimes called polynomial neural networks (PNNs), are statistical learning networks developed as an improvement to the group method of data handling (GMDH) algorithm. PNNs formed the technological basis for Roger's first company Adaptronics, Inc., in the early 1960s and remain my favorite data mining algorithm.

However, after my wife, Barbara, finished her Ph.D. at UVA, we had a difficult decision to make. While we loved Charlottesville, our respective parents live in New Hampshire and San Diego. In the interest of the family we planned on having, we decided to move closer to one side of the family or the other and that ended up being San Diego to join Martin Marietta.

5 The Employee Years

Martin Marietta was my one and only large company experience, and while I can certainly live without the bureaucracy, the time greatly enhanced my technical understanding of real-time systems and image processing.

When I was hired by Martin Marietta into their Advanced Development Operations (ADO) in 1991, they pushed hard to get me there as soon as possible because of the "needs" they had for my services. So I came promptly and had nothing to do. After writing a final report for a project I did not work on, I spent months without any real technical work. For someone who was used to working 50–60 h per week at BAI on cutting-edge research and development projects, the lack of tangible work at Martin Marietta was depressing.

However, what made this time valuable then and even today was that I taught myself programming in C, something I had only dabbled in before and did not have a good grasp of at that time. But learning in the abstract does not stick as well as learning with an application, so I decided to build a set of libraries for linear regression, where the key pieces were how to represent matrices and how to invert them. The libraries soon expanded beyond just linear regression to include neural networks as well.

Anyone who has developed algorithms knows that one can superficially understand algorithms by reading about them, but one can only truly understand algorithms by writing the code to implement them. There are always degeneracies, end conditions, and unforeseen circumstances that are finessed when describing the algorithm but have to be dealt with when writing code. After building the linear

regression libraries, I understood the "what" and "how" of regression far better than I did before.

Within a few months, I was picked up by the Optical Pattern Recognition (OPR) group. That group comprised a total of six or seven individuals; three of us had experience with neural networks and other pattern recognition algorithms. One of our first applications was part of a large bid to the US Post Office for a mail sorting system. The OPR group was tasked with building a postnet bar code reader that would outperform the current system. At that time, the read rate was about 90%. We had a time budget of 0.7 s per envelope to complete all of the image processing and pattern recognition algorithms.

My particular task was the interpretation of the bar code after that part of the image had been segmented out. Data representation of the bar code was key. At that time, many solutions for a postnet decoder used a frequency domain approach, looking for regular patterns of bars and identifying the heights of these bars. As part of our team, we had an expert in postnet bar codes who recommended treating the sequence of bars, two tall and three short per digit in the ZIP code, as a "word." From a pattern recognition perspective then, this could be translated into building an algorithm to detect the "words" using a classifier with ten possible outcomes: the numbers 0–9.

I built multiclass logistic regression models (using the "Class" software from Barron Associates) but had to incorporate any data preparation and classification steps in C code that would process all of this in under about 0.1 s. We spent weeks in Albuquerque, NM, at the Martin Marietta Postal Systems Division running our algorithms on chunks of mail, tweaking the algorithms, and improving efficiency. Our final product achieved a read rate of greater than 95%.

The contrast between R&D and actual implementation of predictive models in a real-time system is a critically important concept to grasp. Decisions about data preparation, feature creation, and algorithm types have great consequences in real-time systems. These do not come naturally to mathematicians or statisticians; they do to a great degree to engineers. In the barcode reader project, we could have built neural networks and achieved higher accuracy, but this would have required a longer timeline to compute the neural network squashing function on our i860 board or more expensive hardware. We opted for the compromise of a simpler algorithm and cheaper hardware. These lessons have stayed with me to this day.

A second project the OPR group worked on was a proposal to digitize the entire criminal record history of the British Home Office (Scotland Yard) stored on microfiche. Our competitors were offshore manual key entry services; only our solution contained an automated optical character recognition (OCR) solution. For expediency, we licensed an OCR classification box from AEG that contained polynomial networks as the core algorithm for classification. I actually met the developer of the algorithms in the AEG solution, one of the real pioneers of polynomial regression and classification in Germany, Jürgen Schürmann. I first met him at a postal systems conference around the time of this contract, and then later when I was at Elder Research, I visited him in Ulm, Germany, in 1997 at the Daimler-Benz research facility, where he was at that time. I only wish my German

was better so that I could have read his papers. Perhaps the most impressive part of the application of data mining technology was that it worked reliably and consistently, something we could not say for the competitor OCR products we evaluated. After his passing in 2001, the Frontiers in Handwriting Recognition dedicated their event to Dr. Schürmann in 2002.

We did not win the British Home Office contract, and despite other research by the OPR group, no other project emerged with that same creativity and innovation. Unfortunately, it also became obvious to me that ADO was struggling financially and would likely be shutting their doors before long, so I looked for more pattern recognition work in San Diego. The doors to Martin Marietta were in fact shut down within about a year.

The next stop was PAR Government Systems Corp. (PGSC), whose La Jolla, CA, office primarily built algorithms to detect objects from long-range radar. I had done some radar and active sonar work at Martin Marietta, as well as passive sonar work at BAI, so this seemed to be a good fit. Perhaps the most significant aspect of the time I spent there was implementing algorithms in Fortran and C for their pattern recognition software package called OLPARS (On-Line Pattern Analysis and Recognition System). The "online" hearkened back to the 1960s when online meant "by computer." The creator of OLPARS was John Sammon, perhaps best known for his Sammon projection, which projected multidimensional data to a lower dimensional space while maintaining relative distances between data points.

I added radial basis function (RBF) networks to OLPARS and made enhancements to their neural networks, perceptrons, and added the ability for OLPARS to read and write Matlab files. I was never able to add polynomial networks however, though I did write the code to integrate them if I ever got the go-ahead. This was the first time I had actively supported data mining software, including providing technical support and fixing bugs. I must say, designing the algorithms is much more enjoyable than fixing the bugs.

Eventually, however, all of the pattern recognition projects ended, and I was put on a project that required a background in physics and optics rather than pattern recognition. While I enjoyed the people at PGSC, it became apparent the window for work in pattern recognition was closed, and therefore it was time for me to move on.

6 The Data Mining Deep Dive

In 1996, John Elder of Elder Research (ER) asked me to help him on a data mining project while I was still employed at PGSC. We had stayed in touch since I left Charlottesville, and John had already established himself as an outstanding data mining course instructor and consultant. He had invited me to be a guest lecturer at one or two of his 5-day courses while I was at PGSC. After that initial project, John made a job offer to me to become ER employee #1, a risk for him and for me, but one that I gladly accepted.

Data mining at last was at the center of the work I would do. Interacting with John on a regular basis was a real pleasure. As a bonus, John valued going to data mining conferences, including Data Mining Summit, KDD, and The Symposium on the Interface.

At the Data Mining Summit in San Francisco, I was first introduced to the Clementine data mining software package (ISL, Ltd., now part of IBM as the tool IBM SPSS Modeler). The icon-based data flow interface was one I liked very much and was familiar with, having used on primarily HP-UX systems. Cantata was a graphical front end to the Khoros system, an information processing and visualization package which was essentially a collection of executables that ran on the unix command line [4]. Cantata's glyphs (icons) and configuration settings called the commands sequentially according to the direction of the connections of the glyphs. It was quite flexible and even extensible, and I used it whenever I could while at Martin Marietta, Corp. Clementine reminded me of this package, and it would not be long before I would be able to use Clementine myself on the "DFAS" project.

I also was introduced to some key individuals in the data mining world at the Summit, including Gordon Linoff who gave a very practical and informative talk on business applications of data mining. It was particularly interesting to me because prior to beginning at Elder Research in 1995, all of the data mining applications I worked were engineering applications: guidance and control, signal processing, or image processing. I had relatively little experience at this time on the focus of Gordon's talks: customer analytics.

The fact that data mining was still a young discipline was evident in a talk that described the perils of data sampling. Why? That is because one might not include "the needle in the haystack" pattern in the training data. Data mining fortunately has grown up since those days. As Brad Efron once said, "*Those who ignore statistics are condemned to reinvent it.*"

The consulting work at ER included a mix of financial modeling and data mining projects. The most influential of the consulting projects was with the Defense Finance and Accounting Services (DFAS), which provides payment services to the United States Department of Defense. DFAS was truly a forward-thinking organization. John partnered with Federal Data Corporation (FDC) to provide DFAS with services to (1) select a data mining tool, (2) build initial models to detect fraud in invoices submitted to DFAS, and (3) identify misuse of government credit cards (the use of the cards for personal purchases). We first surveyed the data mining tool landscape, interviewed our top ten tool vendors, selected the top five of these, and spent roughly a month using each of the tools on DFAS data before selecting the final tool. The ultimate winner was Clementine (SPSS).

The project also had significant visibility within DoD as representatives from other organizations came to each of our review meetings, including the Financial Crimes Enforcement Network (FinCEN), the Secret Service, and the National White Collar Crime Center. I also know that the data mining courses I taught in the 2000s included a steady flow of DoD personnel.

The visibility of the DFAS contracts continued beyond my time at ER, and they have become a long-lasting source of journal articles and conference talks.

They were the source of the KDD-98 Workshop [5], an IEEE Systems, Man, and Cybernetics conference talk [6], articles in Federal Computer Week, talks at the 2000 Federal Computing Conference [7] and SPSS Public Sector Users conference [8], case studies to demonstrate the power of data mining to the IRS, and, more recently, talks at Predictive Analytics Summit in 2010 and the 2011 Predictive Analytics World for Government Conference [9].

7 Going at Data Mining Alone

I started as an independent consultant on March 1, 1999. I was fortunate to have had great employers in the past and to have maintained good relations with them. My first three contracts were with former employers: Elder Research (to continue the DFAS work), PAR Government Systems, and Barron Associates. I picked up another contract after meeting a program manager at Orincon Corp. on the basketball court of the La Jolla YMCA.

I was never good at sales, so it was obvious from the beginning that I would be better served by being a technical resource than being responsible for the entire sales cycle; I was not very good at "selling the vision" but could deliver on the vision. For the next decade, with a few notable exceptions, I partnered with companies that needed consultants, and I have been able to keep fully engaged without interruption for more than a month or two from the beginning of Abbott Consulting. I am grateful to the leaders of these organizations who provided such rewarding relationships, including Eric King and The Modeling Agency, many managers at SPSS including Susan Stellmacher and James Cho, John Elder and Elder Research, and Mike Rote of NCR.

Another avenue for Abbott Consulting opened during the summer of 1999 when I received a call from Eric King, formerly of American Heuristics Corp.'s Gordian Institute. Eric ran the week-long courses John taught (and I assisted) while I was at ER, and we had stayed in touch over the years. He was visiting San Diego with his wife and just wanted to connect while he was here. We sat by the pool of the Town and Country Hotel, sipped ice teas, and talked about a range of topics, one of which was data mining courses. During the conversation, we outlined a new kind of data mining course series for business practitioners, named (unimaginatively, but descriptively) Data Mining Level I and Data Mining Level II. Eric knew another great data mining and instructor, Tim Graettinger, and we essentially flipped a coin that resulted in Tim teaching Level I and me teaching Level II. After working out all the administrative details and outlining the two courses, the first course offering was slated for June 2000.

I had no material for the course, but providentially, KDD-1999 was in San Diego. I volunteered to help out on-site to get in, and during the week, I struck up a conversation with Randy Kerber, then a consultant at NCR. Randy was one of the authors of CRISP-DM 1.0 and someone who just enjoyed shooting "the technical breeze." He was interested in getting my feedback on an idea he

had for a data mining software product, and we struck an agreement: I would spend time with him and his product, and he would introduce me to Mike Rote, the director of Data Mining at NCR/Teradata. Mike and I had a great conversation, and providentially, it turned out that they had a new customer who needed to learn data mining. Teradata planned for an 8-week training course for them but did not yet have an instructor. Since Teradata was in Rancho Bernardo, just a 20-min drive away for me, it was an ideal collaboration.

I frantically developed course material in the evenings for each section of the course. Teradata was a strong supporter of the CRISP-DM process model for data mining, and the course was to be designed around CRISP-DM. My idea was to deliver the lecture material and then create exercises for the attendees to work through over the next ½ week to a week based on the lecture material. The course was well received, and developing the materials was the boost I needed for Data Mining Level II.

Probably the most beneficial part of developing the material was researching the topics deeply. I knew already how to use PCA, neural networks, Bayes rule, etc., but teaching them is far more difficult than using them. However, my rule of thumb in teaching is to know at least ten times more material than I teach, including the history of the algorithms or techniques, variations of how they are used, and clear analogies and illustrations to describe in nontechnical terms how they work.

Teaching the courses was and continues to be humbling. After teaching data mining to more than 1,000 different individuals, fielding questions, and getting into fascinating side conversations with attendees, one quickly realizes that there is an abundance of sharp and talented analysts out there. The best advice I can give to anyone who teaches any subject is to know what you know, know what you do not know, and differentiate facts from opinions.

I continued to teach Data Mining Level II through The Modeling Agency from that 2000 start through April 2010, and in about 2005, a Data Mining Level III Hands-On course was added to the sequence. Eric also asked me to teach 1-day versions of the course at TDWI World Conferences on a few occasions. I used a variety of data mining software tools for Level III (and other hands-on courses I taught on-site with customers), including Clementine, Insightful Miner, Affinium Model, Statistica, and PolyAnalyst, and later SAS Enterprise Miner for the Predictive Analytics World Conference. Additionally, I demonstrated these and many other data mining software products in the Level II course, including CART, Random Forests, TreeNet, GGobi, JMP, S-Plus, Matlab, WizWhy, Neuralware Predict, MineSet, and OLPARS. They all were good for demonstrations and certainly helped attendees understand the data mining landscape better.

While some direct consulting work came through those courses, the greater value from a consulting standpoint was having the flexibility to use whatever tool the customer had and to be able to use it at an advanced or expert level. I have built models on data mining projects either as an employee or as a consultant using nearly every tool on the list above.

In 2003, I renamed my company Abbott Analytics to better reflect the nature of the business, which was focusing on analytics. I wish I could say that the use of the

term "analytics" was prescient, given the adoption of that term in all things related to statistical learning, such as predictive analytics, business analytics, customer analytics, and even web analytics. However, while I like the term analytics, it also provides alliteration and a great alphabetical listing for the company acronym (AA).

8 Visibility as a Data Miner

As an independent consultant, I continued to explore ways to demonstrate proficiency in data mining and keep current with the latest ideas in the field. Data mining conferences were an option, but I found them to be too academic, leaving practitioners with little they could use immediately in their work. In addition, they were not good venues for generating new consulting leads. At that time, the applied data mining conferences I found were Users conferences sponsored by data mining tool vendors. I gave talks at conferences sponsored by SPSS, Salford Systems, and Insightful Corp. and developed good contacts. Giving conference talks was a good way to summarize the most important findings in ongoing contracts.

From a technical standpoint, the points of the talks I was giving were always the same: pose the right problem to solve, make the results understandable to the stakeholders, and use model ensembles to improve accuracy whenever possible. Every one of the eight conference talks I coauthored or gave from 1999 to 2008 included ensembles.

The advent of Predictive Analytics World (PAW), a brainchild of Eric Siegel, provided a unique opportunity to convey practical data mining insights in a vendor-neutral conference forum. Eric and I became acquainted through The Modeling Agency (in the context of teaching courses for TMA) and hit it off well. Eric asked me to give a talk at the first PAW and then to teach the Hands-On Predictive Analytics workshop at each of the subsequent conferences. With the visibility of PAW, and the continued visibility of my blog, consulting work began to "come to me." I still could not "sell," but now I still did not need to! I trust this will provide some comfort to the many data miners out there who love the technical aspects of analytics, but do not like, nor have the inclination toward trolling and selling.

Keeping abreast of the rapidly changing field of data mining is a challenge for those of us who dedicate the vast majority of our time to solving problems in industry. Most often, in these engagements, we do not have the luxury of extensive experimentation; we must solve problems and generate ROI for the customer. When I have presented findings at conferences like Predictive Analytics World, I often get questions along the lines of "did you try support vector machines," or even if we did use SVMs, "did you try a Gaussian kernel." The answer is often "no," and not because it was not a good idea. We just did not have time or the need to try these approaches in order to be successful. In the industry, success is defined by delivering what the customer requests within a specified timeframe. However, I often present the possibility of additional experimentation to the customer for a subsequent solution.

9 Lessons Learned and Advice

Looking back, I could not have charted the course that was taken, nor could I have envisioned how data mining as a field would mature and become mainstream in the business world. Seeing data mining (and its sibling, Predictive Analytics) emerge as one of the top growth technical fields is astonishing.

Here is my advice:

1. Do what you enjoy, but keep your eye out for ways to make what you enjoy relevant to the business world. Relatedly, identify problems in the industry that are a good match for your skills and interests.
2. Take opportunities to learn; opportunities are far greater now for this with YouTube and the academic papers that are available online. In the 1990s, I would made regular visits to the UCSD library to find and read references in papers that I liked and just to browse abstracts of journals I did not receive or see on a regular basis. This is how I discovered the Machine Learning journal, for example. I also read everything several academics wrote who were clear writers and thought leaders.
3. Interact with thought leaders on a regular basis. Working day to day with them is, of course, easiest, but it does not have to be that way. I have had interactions with other thought leaders in data mining, sometimes without ever having met them in person. One example of that is Will Dwinnell, my co-blogger. Will is another practical data miner with a deep understanding of data mining concepts. It is not unusual for one of us to call the other to ask a question about something we have seen in building decision trees or using PCA for example. We would get rolling, an hour would pass, and the initiator of the call might say, "we haven't gotten to my real question yet—do you mind if I call you later in the week?" I think it was 5 years before I even met Will in person.
4. Write down your ideas. This is far easier now than it ever has been with the advent of blogs and also with linkedin.com groups. Just reading these provides a wealth of information about how practitioners deal with day-to-day issues. I do not always agree, but I usually learn from their perspective. Writing down my opinions and having them shot down (on occasion) helps me correct my errors and articulate more clearly when I am right.

We in data mining are no longer just a niche group in the business world. But as much science as there is in data mining, there is still significant "art." We are not yet at the point where we can write up a series of recipes for an untrained analyst to follow mindlessly. While the "art" is there, I will stay in the world of data mining and will teach others to do so as well.

References

1. R. Hecht-Nielsen, *Neurocomputing* (Addison-Wesley, Reading, MA, 1990)
2. R.L. Barron, D.W. Abbott, et al. Trajectory Optimization and Optimum Path-to-Go Guidance of Tactical Weapons: Vol. I—Theory and AIWS Application, August, 1988; Vol. II—Closed-Loop OPTG Guidance of Mk 82 Glide Weapon, September 1987; Vol. III—Open-Loop Trajectory Optimization of Skipper Boost-Glide Weapon, June 1988; Vol. IV—Calculation of Lagrange Multipliers of Vertical-Plane Maximum-Range Trajectories of 11:1 LID Boost-Glide AIWS, January 1989, Barron Associates, Inc. Final Technical Report for HR Textron Inc. under U.S. Naval Weapons Center contract N60530-88-C-0036
3. R.L. Barron, D.W. Abbott, Use of polynomial networks in optimum real-time, two-point boundary-value guidance of tactical weapons, in *Proceedings of the Military Computing Conference*, 3–5 May 1988
4. M. Young, D. Argiro, S. Kubica, Cantata: visual programming environment for the khoros system. Computer Graphics **29**(2), 22–24 (1995)
5. J.F. Elder, D.W. Abbott, A Comparison of Leading Data Mining Tools, in *4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, New York, NY, August 1998
6. D.W. Abbott, I.P. Matkovsky, J.F. Elder, An evaluation of high-end data mining tools for fraud detection, in *1998 IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA, October 1998
7. D.W. Abbott, H. Vafaie, M. Hutchins, D. Riney, Improper Payment Detection in Department of Defense Financial Transactions (320KB), Federal Data Mining Symposium, Washington, DC, March 28–29, 2000. http://www.abbott-consulting.com/pubs/afcea_2000.pdf
8. D.W. Abbott, Model Ensembles in Clementine, SPSS 2000 Public Sector Users' Exchange, Washington, DC, December 6, 2000
9. D.W. Abbott and D. Riney, Predictive Modeling to Detect Fraud at DFAS, Predictive Analytics World-Government, Washington, DC, September 12, 2011

From Combinatorial Optimization to Data Mining

Charu C. Aggarwal

1 Motivation

My work in data mining started almost immediately after my Ph.D. was completed and was initially unrelated to my Ph.D. thesis, until more recently, when I started working in the field of graph mining. At the time I started working in data mining, the field was still in its infancy. I had graduated in 1996 from MIT in the field of combinatorial optimization and network flows and was mostly interested in problems of a theoretical nature. I had joined IBM Research, which, at the time, contained some of the strongest researchers in the field of data mining, including some of its key founding fathers.

Along the way, during my Ph.D. years, I had picked up an interest in probability and statistics. My Ph.D. was in the field of operations research, which essentially contained two tracks. The first track was in the field of combinatorial optimization and mathematical programming, and the second track was in the field of probability and statistics. Students were expected to complete their Ph.D. thesis in one of these two tracks, though they were also expected to complete a strong course work and gain research knowledge of both tracks. While my Ph.D. was in the combinatorial optimization and mathematical programming track, I always retained a certain level of interest in the field of probability and statistics. In particular, I still vividly remember the entertaining lectures of the well-known Alvin Drake during my first semester at MIT, who made the rather dry field of probability a very interesting one with his unique style and delivery. This is one of the reasons that I have never considered the field of probability and statistics a dry one, though I would also understand why someone else might, especially if they focus too much on the algebra of the advanced stochastic courses. Alvin Drake's main contribution as a teacher was to make us think of problems in probability and statistics in an intuitive way first, and only then make

C.C. Aggarwal (✉)
IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA
e-mail: charu@us.ibm.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_3,

use of the algebraic formalisms. This general concept is true not just in the process of learning but also in the process of research. As we will discuss later, the relative importance of intuitive understanding and the complex techniques required to enable the ideas resulting from such an understanding share a relationship in which the former drives the latter, and not vice versa. A clear understanding of this is key to good research, teaching, and learning. My interest in the field of probability and statistics grew over the years, as I continued to complete more course work in this area.

The field of data mining was just starting at the time I joined IBM in 1996. The first KDD Conference had just been completed in the year 1995. I had already developed an interest in probability theory and statistics during my student years and was immediately interested in the opportunity to work with large amounts of data in an area of my interest. Other than this, my combinatorial optimization and mathematical programming background ensured that I had an interest in algorithmic methods for high-dimensional data. Consequently, many of my first years were focussed on subspace analysis of high-dimensional algorithms and related problems such as dimensionality reduction and principal component analysis.

Other than this, my own thesis work in graph theory and network optimization has had some influence on my recent work in data mining, though graph mining has been a relatively recent evolution of this area. Along the way, my advisor James Orlin introduced me to several unconventional optimization techniques (such as genetic algorithms), which were not directly related to my Ph.D. thesis, but which I have found immensely useful in some of the unstructured data mining problems I have encountered over the years. I must use this opportunity to say that I was unable to fully appreciate the value of such methodologies at the time, especially since I was focussed on completing my Ph.D. quickly. It was only later that I was able to understand the practical implications of many of these techniques, and their value as tools for a variety of problems. I was also able to appreciate my advisor better for the breadth of the exposure he provided me during my formative years. I am grateful to him today for these technical digressions. One lesson from this to young Ph.D. students is to avoid the tendency to become "tunnel-visioned" during their Ph.D. years, in which they often focus only on completing their Ph.D. thesis, without trying to gain knowledge about related areas. A researcher's area of work evolves over time, and the greatest level of exposure to related fields is essential in helping define and redefine oneself, as the field evolves over time. In my particular case, my interest in the field of data mining came out of its relationship to the non-thesis-related courses and the non-thesis-related research which I had performed in the early years.

2 Milestones and Success Stories

I have worked in a wide variety of research topics during my data mining years, though most of my work can be categorized into one of five topics:

- High-dimensional data mining
- Data streams

- Privacy-preserving data mining
- Uncertain data mining
- Graphs and social networks

Other than these, I have also done some work on frequent pattern mining, visual data mining, and text mining, though the vast majority of my work has been in the topics listed above. Of the five topics listed above, the last two are relatively recent. Therefore, for the case of these two, I will provide some idea of the parts which I think may be significant, though the research community may eventually have a different judgment. In the case of the first three topics, the impact on the community is clear on the basis of citations and other metrics; therefore, I will discuss their contributions in that context. I will discuss my key milestones in the different topics by organizing them into different subsections.

2.1 High-Dimensional Data Mining

The "curse of dimensionality" was a well-known phenomenon in the field over many years. The key issue with the dimensionality curse was that high-dimensional data leads to the *sparsity property which causes a number of challenges*:

- Many problems such as indexing were known to show poor performance in high dimensionality because of the performance degradation which was associated with the use of such methods. As a result, there was a tremendous focus on the design of effective high-dimensional index structures [1]. It was well known that with increasing dimensionality, the performance of most index structures degrades to the point that most of the data needs to be accessed.
- Many data mining algorithms such as clustering did not yield meaningful results [2], when they were applied to algorithms in high-dimensional space. In particular, the sparsity of high-dimensional data ensured that the clusters were often poorly knit.

We note that one of the above two issues is about *performance*, and the other is about *quality* of the results. While the fact that the curse of dimensionality caused a significant number of problems for data mining and management applications was clear, it was not clear to me that the direction of work in the area was very insightful, because it remained trapped within traditional definitions of what problems such as clustering or nearest neighbor were *traditionally*, rather than what they *should be*. For example, if the results of many data mining problems such as clustering (based on proximity) were not meaningful, it meant that the results of problems such as nearest neighbor were also not of high quality. In this light, it was not clear to me why so much research effort should be focussed on designing index structures for doing so *efficiently*. Even more importantly, recent data sets continued to increase in dimensionality because of advancements in data collection techniques. Logically, the access to greater number of features of the

same entity should lead to higher quality results; yet, it seemed that algorithms such as clustering would perform worse with more information. From these observations, it seemed clear to me that *the curse of dimensionality was often a direct result of trying to force traditional definitions of problems such as clustering and nearest neighbor search on the high-dimensional scenarios*. A higher dimensionality brings with it clear combinatorial challenges from an exploration point of view, yet it also provided unprecedented opportunities which remained unexploited by traditional definitions. There was little work on exploring the inherent blessings of having more information, and the focus was always on the curse.

Therefore, it seemed logical to focus on a proper design of these problems [3]. For example, the work in [4] asked the fundamental question of how a nearest neighbor really should be defined in high-dimensional space. Much of my recent work showed that high-dimensional problems should be defined in terms of *more general subspace variations*, rather than in full dimensionality. The latter is simply a special case, which works well for low-dimensional data, but not for high-dimensional data. In one of my subsequent papers in the area [5], we showed that one can leverage the additional information in high-dimensional data in order to improve *both* the performance and quality of a nearest neighbor indexing structure, simply by changing the definition of the distance function and the nearest neighbor. I consider this line of work and vision as a major milestone, which lead to several derivative discoveries, such as the design of fractional norms for distance function [6], application-specific distance function learning [7], qualitative improvements arising from dimensionality reduction [8], local dimensionality reduction [9], and outlier detection [10]. The key vision paper for this line of work may be found in [3]. As we will discuss later in the impact section, one of the common reasons why research in the field does not move forward is because of a natural human tendency to stick with the "safe" and "traditional" definitions while solving problems, without regard to the applicability of such definitions on new and evolving scenarios.

2.2 Data Stream Mining

Just as advances in data collection technology has lead to large amounts of data in terms of dimensionality (horizontal dimension), it has also lead to large amounts of data in terms of continuous collection and processing (vertical dimension). This leads to large volumes of continuous data being processed over time. Such continuously increasing collections of data are referred to as *data streams*. Data stream mining techniques bring with them a number of unique challenges which are often not encountered in the scenario of traditional data which can be stored on a disk and processed.

My major focus in this field is from the perspective of designing new analytical methods for problems such as clustering [11], classification [12], anomaly detection [13], and reservoir sampling [14]. I consider my work on summarization of streams in terms of clustering [11] and reservoir sampling [14] because the most challenging aspect about stream mining is to enable careful summarization for a variety of

mining tasks. The works in [11] and [14] are designed to calibrate the summarization process carefully, so that it can be used even when the stream evolves significantly over time. A discussion of the key summarization methods in the stream literature may be found in [15]. The developed summarization methods were leveraged for a variety of other tasks such as classification [12] and even privacy-preserving data mining of dynamic data sets [16]. The edited book on data streams was also my first in a series of edited books over the years and was therefore a milestone of sorts as it got me into the habit of creating subsequent edited books [17–20]. Much of my goal in creating these edited books is to provide educational value to the community in key topics. As I work in industry, I have little opportunity to teach; the creation of edited books is a small contribution to the community in that direction. Beyond this, I also do believe that spending time on writing of edited books and surveys is important in enabling thorough learning of an area. When you create a carefully organized edited book or a survey, it forces you to internally organize the material well and learn. In my view, researchers should spend at least 10% of their time in writing such educational material; the time is well spent from a learning perspective, and the by-products are useful to the community at large.

2.3 Privacy-Preserving Data Mining

The problem of privacy-preserving data mining was first explored in the data mining community in [21]. The goal in these techniques is to design methods for mining data which may be sensitive in nature. In such cases, it may be useful to reduce the granularity of representation, so that useful information can be mined without compromising the integrity of the underlying data. One key challenge at the time was to design a theoretical model for quantification of privacy-preserving data mining algorithms [22]. This subsequently formed a basis for future theoretical work in privacy-preserving data mining. I continued to explore several other theoretical aspects of the privacy area, including its applicability of high-dimensional domains [23, 24]. This work shows that privacy is elusive for the case of high-dimensional data, and it may often not be possible to implement methods which preserve privacy effectively, without a complete loss of utility. This is a fundamental result, in that it establishes practical limits on what may or may not be achieved in this area. Another practical issue which arises with privacy techniques is that the data is often not converted to a format which is friendly to the use of *existing data mining or management techniques*. For example, the method in [21] converts the data into an *aggregate probability distribution* in which there is no concept of individual records. Such data representations are hard to use directly with current databases, because they are not naturally designed to represent aggregate versions of records. In one of my subsequent work, I showed that the recent work on probabilistic databases can be leveraged quite effectively with certain privacy transformations [25]. This leverages existing (and developing) work on uncertain databases to the problem of privacy-preserving data mining.

2.4 Other Recent Topics

Some of my recent work has been in the field of uncertain data and social networks. For the case of uncertain data, I introduced a number of fundamental data mining problems such as clustering and outlier detection to this domain [26, 27]. The importance of using uncertainty in order to improve the quality of the data mining results had not been noticed at the time, and one of my goals was to design data mining algorithms which were not only efficient but also qualitatively effective. In order to achieve this goal, a density-based framework [28] was designed for uncertain data mining. This framework was initially used for the problem of classification but was subsequently extended to the problem of outlier detection [27].

Another area of research which has found increasing prominence in recent years is that of graphs and social networks. In particular, dynamic data streams which are generated by activity on social networks and graphs pose an enormous challenge to a wide variety of applications. Such streams are difficult to process because of their structural nature. Other challenging cases are those in which the graph cannot be stored in main memory, but it needs to be stored on disk. Some of our recent work makes a progress in this direction [29, 30], in which we design methods in order to mine massive graph streams.

3 Research Tools and Techniques

A research paper can have two kinds of contributions: (1) a modeling contribution in terms of new problem of model being proposed and (2) a technique contribution, in terms of the algorithm, model, or methodology which was used to solve the problem.

Both of these characteristics are important, and in some cases, very nicely proposed problems and models do not survive because of poor techniques which were used to solve the problem. This is often because of a poor grounding in the skills which are needed in order to solve a particular problem. In general, the skills required to solve the different data mining problems may require any one of a number of the following:

- A strong background in probability and statistics is essential in order to properly assimilate and solve problems which are related to exploration of large volumes of data. In addition, many areas of data mining such as privacy and uncertain data explicitly require a strong stochastic background. Similarly, many of the stochastic optimization techniques (e.g., EM methods), which are used for problems such as clustering require some level of understanding of probability theory.
- A strong background in linear algebra and matrix techniques is required in order to perform the subspace analysis which are required for problems of high dimensionality. In addition, many standard analytical techniques such as SVD,

principal component analysis, and matrix factorization require a deep understanding of such techniques.
- In addition to the above specific topics in mathematics, an overall strong background in mathematics may be useful in tackling a variety of optimization problems which arise in the context of data mining. Many standard data mining problems such as clustering and classification can be formulated as optimization problems, and therefore, it is essential to know all the mathematical tools which are available for such problems. This could include conventional methods such as integer programming or unconventional techniques such as genetic algorithms and simulated annealing.
- A strong understanding and knowledge in discrete algorithms, graph theory, and combinatorial algorithms can be very useful. Combinatorial algorithms are particularly important because of the recent increase in interest in problems involving graph mining and social networks.

Other than this, a fundamental understanding of the classification of different kinds of data mining problems is essential in order to approach the field in a meaningful way. For example, the "big four" problems in data mining are clustering, classification, frequent pattern mining, and outlier detection. Anyone who is working in the field of data mining should be thoroughly familiar with the problems which relate to this area as well as the classical solution techniques which are used for such problems.

Finally, I do believe that implementation of one's ideas is not the realm of students alone, and even senior researchers should actively explore the ideas on a "hands-on" basis. Such hands-on exploration is critical in developing research insights, since it provides an understanding of how different variations of a solution may affect the underlying quality. Data mining requires different kinds of software tools at the different stages of data parsing, algorithmic development, and hands-on exploration. In terms of software tools, the following are particularly useful:

- A strong tool for data parsing such as *Perl*, which can process large amounts of unstructured data and convert them into a structured format for further processing.
- One of the many programming languages such as *C*, *C++*, or *Java*, which may be used for core implementation. In general, *C* or *C++* implementations tend to be more efficient. When the volume of data is large and efficiency is a concern, it is best to work with the *C* language.
- A mathematical toolbox such as *MATLAB* can be very useful for intermediate exploration of the data mining results. One advantage of *MATLAB* as compared to a language such as C is that it is an interpreted language within an environment which provides full access to the underlying variables for the particular data mining problems. This makes ad hoc exploration very easy. Such exploration can be very useful at the initial stages of a research effort in order to gain valuable insights about the statistical behavior of a data set or a class of algorithms. In fact, it is usually helpful to precede the work effort with an open-ended exploration stage with an interpreted language in order to gain better insight of the problem, before writing the efficient version of the code in a language such as C.

4 Lessons in Learning from Failures

A researcher in the field of data mining can have a failure which is primarily of two types while pursuing a research idea or topic.

- A particular idea may not work as well as originally thought when it was conceived.
- A particular idea may have been thought of in some form by someone else, and one ends up discovering it only after a while.

In some cases, there is little one can do in providing an effective remedy. In other cases, it is possible to learn from the nature of the setback, and turn it into a success.

Let us take the first case, where an idea may not work as well as it was originally intended. A natural question which arises in this case is as to the *causality of why* this idea did not work well, especially if it was a natural idea to begin with and would also be thought of by other researchers as a natural way of approaching the problem. In such cases, the *insight* of why such a natural direction does not work is in of itself worthy of further investigation and may lead to other research ideas. A specific example of this situation was one in which we investigated the extension of FP-Tree frequent pattern mining to the case of uncertain data [31]. Initially, we expected the natural extension of the deterministic algorithm to the uncertain case to work well, but our tests continuously showed us otherwise. Later, we found that the core of the FP-Tree advantage was in sharing of common prefixes in the deterministic case; this idea did not work well for the uncertain case, because of need to explicitly store the different probability values specific to different transactions. We were able to use this insight to design an uncertain version of the H-mine algorithm, which retained many of the key qualities of the FP-Tree algorithm, but did not rely on sharing of prefixes explicitly. Furthermore, we retained the negative results on the FP-Tree algorithm within our work [31] in order to provide researchers the insight on why such a natural line of approach would not work well. I consider this case one in which we were able to learn from our setbacks in order to result in a successful outcome.

The other situation, where a particular idea may have been thought of by someone else is also quite common, especially in the fast-moving data mining community. This has happened to me a few times; in one case, I was unable to publish a paper in the field of privacy because of numerous papers which had appeared on the topic. In another case, one of my ideas on text indexing and representation seemed to have been explored earlier by the community. Such situations can be avoided in a number of ways:

- A more thorough literature review at the very beginning of commencing the work, which ensures that such situations do not arise.
- Unfortunately, it has become increasingly necessary to move fast after the conception of an idea. In the past, one could spend a while exploring the idea in a variety of different ways. It now becomes more important to seek quicker

completion and publication. This is simply a necessary evil which results from the direction in which the community has progressed.

- In some cases, it may be possible to examine whether the current work has any advantages over the previously proposed work. In such a case, a comparison can be explicitly made, and the advantages of the new proposed work can be presented. This is sometimes not a fully satisfying solution, as it tends to create a more incremental piece of work. However, in some other cases, the new piece of work may be able to provide different insights because it may have been conceived independently from a different point of view. Even negative results from one's own research can be sometimes useful to the community.

5 Making an Impact

The single largest reason why many papers which are written today are not influential has been the tendency to do incremental work and focus more on the specific technique used to solve a problem, rather than the problem which is being solved. I believe that the single most important decision while writing a paper is to ask oneself as to why a particular problem is important and what gap in the research community it might fill. Some examples of common answers (both good and not so good) to this question are as follows:

1. This problem has not been solved, because no one thought of it as important before. Here is why I think it is important, and here is a nice (or even reasonable) solution to do it. An example of such a paper would be [32].
2. This problem has been solved, but researchers are currently trying to solve the wrong problem. Here is why I think this is the case, and here is what they should be solving. An example of such a paper would [4].
3. This problem has been formulated before, but no one knows how to do it (or at least in polynomial time). Here is a method which provides a first solution to the problem. An example would be polynomial simplex algorithm for minimum cost flows discussed in [33].
4. This problem has been solved before, but the current solutions are slow or impractical. Here is a solution which can improve the results by one or two orders of magnitude. An example of such a paper would be in [34]. Alternatively, here is *new class of solutions*, which opens a new direction of work. An example of such a paper would be [35].
5. Here is a very simple and practical observation about a current problem or class of solutions, which no one has noticed before. Here is why the implications of this observation are important. Here are the theoretical and experimental reasons why this happens. An example would be the page-rank paper discussed in [36].
6. The basic problem has been solved before, but I am going to examine the special case, where the objective function has costs, the variables are correlated, and some of the parameters (e.g., number of clusters) need to be determined in an automated way. Here is a solution for this problem.

7. The basic problem has been solved before, but I have a solution which improves the current solutions by 10%.
8. Same as either of the last two bullet points, except that my solution also borrows a lot of fancy (though not really required) techniques from the literature, so the reviewer will get distracted from the insignificance of the overall impact and accept the paper on the basis of technical depth.

I have not given specific examples of the last three bullet points, since I do not mean them in a positive way, and at least 50% of the papers being published today fit one of the last three categories. The above classification is also not exhaustive; I have picked these specific cases intentionally in order to represent some important and commonly encountered cases, and there are many papers which fall somewhere in the middle in terms of importance and impact. For example, emerging areas provide a number of quick opportunities to make an impact, if they are spotted early enough. For example, the exploration of an old problem in a new scenario or new data domain may sometimes not be as exciting, but may be important nonetheless. In general, emerging areas provide a number of such opportunities, but such opportunities are available only in the early years of a field. For example, while fundamental problems such as clustering and classification were very important for the stream domain a few years back, it is hard to write an interesting or significant paper on this subject today. It is also expected that a researcher may have a wide spectrum of papers of different types, and a small percentage of their papers may fall into the first five categories.

However, the tendency in the past few years has been to focus on papers in one or more of the last three categories. Some of the reason for this is that the focus of many research papers has been to write a paper with the most fancy hammer in search of nail, rather than *formulate the most important nail* in need of a hammer. While elegance of technique is always a joy to have in a paper, it should not be an end in of itself. The technical depth of the paper should be a natural means to achieving a goal, rather than a goal in itself. In fact, I believe that simplicity of ideas always trumps technical depth in many scenarios. For example, it is an almost universal observation across all fields that the most influential and "classic" papers are often quite simple in terms of technical depth. For example, the importance of the original association rule paper [32] cannot be disputed, yet the actual methods in the paper are extremely simple. Similarly, the page-rank paper [36] is an extraordinarily influential paper, but the core of the paper is the simple observation that modeling page importance by a surfer's random walk probability provides high quality results. The actual mathematics of the paper is quite simple, and fairly standard linear algebra. In fact, most influential papers are simple, intuitive, and typically have a *single elegant formulation or observation* which makes the paper important. Even some of the technically detailed papers such as the FP-Tree [35] depend upon the *simple underlying concept of prefix-based representation*, which is more valuable than the actual details of exactly how it is done. This is evidenced by the fact that numerous subsequent papers have built upon the core idea with different (and in many cases simpler) implementations quite successfully. In my

opinion, *the most influential paper is the simplest idea, which is both valuable and currently remains unexplored*. One test for this is to ask oneself: *Can I summarize the key idea (or takeaway) of this paper in one or two brief sentences, without having to write a full paragraph or bulleted list?*

Some of the most influential ideas arise from personal experiences either in terms of facing a technical challenge while building a practical system or by reading about technical trends and directions of the industry. While reading academic papers also provides insights and ideas, the most fundamental insights arise out of our ability to assimilate and combine this academic knowledge with our day-to-day observations about technical advancements, practical experiences, and our readings of slightly less academic sources such as general technical news about industry directions. Such an approach provides the simple insights which create great papers. It is also sometimes the case that well-known simple and influential papers have achieved recognition only in hindsight. At the time they were written, they were not considered extraordinary either during the review process or in terms of the initial reception. In some cases, even the review process can pose a challenge to such research. Papers which are either somewhat simple in approach or different from the current issues in focus may often be considered too risky to accept from a reviewer perspective. This is to be expected, because reviewers are typically also authors, and the same biases present in an author are also present in reviewers. Such experiences should not discourage young authors from exploring and writing about what they feel is most important.

5.1 Educational Impact

Aside from the research impact, it is also the responsibility of researchers to make an educational impact to the community either by teaching or by writing books and surveys. In particular, for a busy researcher, edited books provide the best balance between time and impact. For a specialized area, it is hard for a single researcher or even a small group of researchers to write comprehensively in every subtopic of that area. A better plan is to write about 25% of the book in the areas of one's greatest expertise and farm out the remaining book to the most appropriate authors. While some edited books may be popular, it is often the case that such books are often notorious for that lack of impact to the community. This is particularly because such books are often not properly planned, and the content is sometimes driven by the need to be able to collect almost any paper from well-known researchers. The process of creating valuable edited books should follow some general guidelines:

- The book should revolve around a coherent theme which is of interest to the community at large. A theme which is too general will result in a bunch of unrelated papers and reduced impact for the book. For example, it would be hard to create an edited book simply on the rather generic topic of "algorithms," but easy enough for an authored book.

- The book should not be an unplanned collection of papers on a topic. Often, edited books serve as a conduit to publish individual research results which could not be published elsewhere. This defeats the purpose of creating a book on the topic with the use of what is essentially journal material.
- The content should be carefully planned, and authors should be invited with specific guidelines on content. The authors should be picked carefully based on expertise, and the book should try to be as comprehensive as possible in coverage of subtopics of the overall theme of the book.
- Individual chapters should be about comprehensive coverage of a topic in the form of a survey, rather than the individual's own results.

By using such an approach, it is possible to create edited books which are a valuable resource to the community. One observation is that prospective authors of book chapters are often much less willing to spend the time required to write a survey chapter from scratch, than simply contribute one of their old papers which they could not publish elsewhere. It is necessary to be firm on this point in order to maintain the quality of the book. It is better to pick a less well-known researcher who accepts the proper guidelines on chapter content, rather than a better-known researcher who provides nonrelevant material for the book. Edited books should be *planned*, not solicited as a collection of papers.

6 Future Insights

New problems or challenges in the field of data mining often arise because of advances in hardware and software technology which can create new kinds of data or enable significantly more efficient processing of data. Such advances can also be a result of increase in storage, communication bandwidth, and processing speeds. For example, the data stream domain was created as a result of advances in hardware technology which enable fast data collection, transmission, and processing. Similarly, the bioinformatics domain was created as a result of hardware and algorithm advances, which enabled the mapping of large sets of genomes, and thereby, the creation of vast repositories of data. In this respect, I will discuss short-term challenges corresponding to the new fields which have recently arisen in this context.

6.1 Short-Term Challenges

In the short term, a number of recent trends in hardware and software technology have lead to advancements, which require significant investment in terms of algorithmic research effort. In particular, the following fields may be immediately relevant.

6.1.1 Graph Mining and Social Networks

Advances in social networking have lead to an increase in interest in the field of graph mining. While algorithms for graph analysis have been around for many years [37], these methods are generally not very useful for the planetary scale graphs which are typically relevant in the field of social networks. Furthermore, such networks often have tremendous amount of activity which is built on top of these infrastructures such as chat messages, communications, and posts, which are rich in both content and structure. This leads to the following kinds of challenges:

- The classical algorithms for graph analysis are designed for graphs which are of modest size. For example, an algorithm which is designed for graphs which are memory resident cannot be applied to the disk resident case.
- Many of the graphs have activity which is built on top of them. For example, in social networks, there may be a considerable amount of activity between the different participants. These are referred to as *edge streams*. The design of algorithms for the streaming scenario is an enormous challenge because of the dynamic structural aspects which need to be captured and used.
- Many recent graph mining applications are also associated with content. This leads to the possibility of designing algorithms which work effectively with both content and structure. The presence of content can lead to massive constraints on the kinds of algorithms which are designed for structural analysis. Combining content and structural analysis is also a challenge because of the inherently different representation of these kinds of data.

6.1.2 Cloud Computing

A recent trend has been the use of centralized repositories to store and process data for a wide variety of applications. This has been made possible by the increases in bandwidth over different Internet, wireless, and mobile data networks, which allows the rapid transfer of data from repositories to users. The enabling of centralized repositories which allows the centralized storage of data and applications has been a trend which has posed unique challenges in the design of data mining algorithms. For example, the design of the *map-reduce* framework has been a result of the increasing amounts of data and applications which are present on the cloud. One of the challenges with this is that many data mining applications need to be fundamentally redesigned in order to enable their use with a wide variety of applications.

6.2 Long-Term Challenges

The long-term challenges to the field of data mining will arise from both *greater complexity of problem domains* because of advances in data collection technology

or new platforms/applications for data collection and *better enablement of algorithmic solutions* because of advances in hardware technology (such as storage, processing power, and bandwidth) which allow for greater complexity in *problem definitions*. Either way, most long-term challenges are likely to arise because of advances in hardware (and in some cases, software) technology. This trend has also been historically true for data mining. Some historical examples are as follows:

- Advances in data collection and software processing technology have lead to new domains of data such as streams, sensor data, and biological data. These have brought new and rich problem domains.
- A number of new platforms and applications such as the Web and social networks have resulted in new kinds of structural data, which are very challenging from the perspective of a number of mining applications.
- Advances in processing speeds, storage limits, and new flash-based storage devices have resulted in different kinds of processing trade-offs for data mining algorithms. These different kinds of processing trade-offs result in important changes in the algorithmic techniques needed to address the different problems.

While the first two kinds of advances are likely to lead to unforeseen changes in problem definitions and domains, the last needs some further discussion. While it would seem that the increase in processing power and storage should ease the development of data mining algorithms, the reality is that an increasing ability to process data increases our scientific appetite to explore problems more deeply and, in some cases, also creates more challenging scenarios. In this respect, the trade-offs between the resource constraints may also change over time, which may lead to a fundamental redesign of a variety of data mining algorithms. For example, if the relative trade-offs between storage constraints, disk I/O speeds, and processing rates change rapidly, this may lead to a complete redesign of existing algorithms. Even relatively modest changes in such trade-offs (as evidenced by the behavior of the flash drive) have lead to huge changes in algorithmic design for a variety of data management and mining algorithms. Similarly, fundamental leaps in processing rates with the use of new technologies such as quantum computers may doom many current efficiency-focussed research problems to irrelevance. On the other hand, advances in such computing power may lead to previously unimagined (and assumed to be unsolvable) problems to become much more solvable and relevant. It will be fascinating to see how these challenges unfold.

7 Summary

I will end this paper with some directions for young researchers in the field. At an early stage in their careers, researchers should have the broadest possible exposure to research problems, tools, and techniques. This is because the field continues to evolve over time, and having a wide exposure is critical in having the tools needed to solve different kinds of problems.

Young researchers should continue to be "hands-on" in the initial phase of their careers (at least the first 10 years after completion of their Ph.D.). This is critical in order to build the research maturity and insight needed for understanding the subtle impacts of different kinds of techniques. It is all too often, that we find that young professors rarely spend the "hands-on" time necessary to build deep insights. Some of this is unfortunately because of the tremendous load that a proposal-driven system has placed on researchers, especially within the US academic system.

Simplicity is the key to making an impact in research. Good research work often contains simple observations which are deeply insightful and lead to fundamental advances. While writing a paper, a researcher should focus on bringing value to the community and focus on the most relevant techniques for a given problem rather than the most "impressive" ones. The value of an elegant technique is limited only by the problem it addresses and the intuitive concepts that it enables. This requires an understanding not just of the solution techniques but also of the emerging problems that researchers may care about, and the (as yet) unaddressed problems in the space.

References

1. S. Berchtold, C. Bohm, H.-P. Kriegel, The pyramid-technique: towards breaking the curse of dimensionality, in *ACM SIGMOD Conference*, 1998
2. A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data* (Prentice Hall, Upper Saddle River, NJ, 1988)
3. C.C. Aggarwal, Re-designing distance functions and distance-based applications for high dimensional data, in *ACM SIGMOD Record*, March 2001
4. A. Hinneburg, C.C. Aggarwal, D.A. Keim, What is the nearest neighbor in high dimensional space? in *VLDB Conference*, 2000
5. C.C. Aggarwal, P.S. Yu, The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space, in *ACM KDD Conference*, 2000
6. C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, in *ICDT Conference*, 2010
7. C.C. Aggarwal, Towards systematic design of distance functions for data mining applications, in *ACM KDD Conference*, 2003
8. C.C. Aggarwal, On the effects of dimensionality reduction on high dimensional similarity search, in *ACM PODS Conference*, 2001
9. C.C. Aggarwal, P.S. Yu, Finding generalized projected clusters in high dimensional spaces, in *ACM SIGMOD Conference*, 2000
10. C.C. Aggarwal, P.S. Yu, Outlier detection for high dimensional data, in *ACM SIGMOD Conference*, 2001
11. C.C. Aggarwal, J. Han, J. Wang, P. Yu, A framework for clustering evolving data streams, in *VLDB Conference*, 2003
12. C.C. Aggarwal, J. Han, J. Wang, P. Yu, On demand classification of data streams, in *ACM KDD Conference*, 2004
13. C.C. Aggarwal, On abnormality detection in spuriously populated data streams, in *SDM Conference*, 2005
14. C.C. Aggarwal, On biased reservoir sampling in the presence of stream evolution, in *VLDB Conference*, 2006

15. C.C. Aggarwal, *Data Streams: Models and Algorithms* (Springer, New York, 2007)
16. C.C. Aggarwal, P.S. Yu, On static and dynamic methods for condensation-based privacy-preserving data mining. ACM Trans. Database Syst. **33**(1), 1–39 (2008)
17. C.C. Aggarwal, P.S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms* (Springer, New York, 2008)
18. C.C. Aggarwal, *Managing and Mining Uncertain Data* (Springer, New York, 2009)
19. C.C. Aggarwal, *Managing and Mining Graph Data* (Springer, New York, 2010)
20. C.C. Aggarwal, *Social Network Data Analytics* (Springer, New York, 2011)
21. R. Agrawal, R. Srikant, Privacy-preserving data mining, in *ACM SIGMOD Conference*, 2000
22. D. Agrawal, C.C. Aggarwal, On the design and quantification of privacy-preserving data mining algorithms, in *ACM PODS Conference*, 2001
23. C.C. Aggarwal, On *k*-anonymity and the curse of dimensionality, in *VLDB Conference*, 2005
24. C.C. Aggarwal, On randomization, public information, and the curse of dimensionality, in *ICDE Conference*, 2007
25. C.C. Aggarwal, On the design and quantification of privacy-preserving data mining algorithms, in *ICDE Conference*, 2009
26. C.C. Aggarwal, P.S. Yu, A framework for clustering uncertain data streams, in *ICDE Conference*, 2008
27. C.C. Aggarwal, P.S. Yu, Outlier detection with uncertain data, in *ICDE Conference*, 2008
28. C.C. Aggarwal, On density-based transforms for uncertain data mining, in *ICDE Conference*, 2007
29. C.C. Aggarwal, Y. Li, P. Yu, R. Jin, On dense pattern mining in graph streams, in *VLDB Conference*, 2010
30. C.C. Aggarwal, Y. Zhao, P. Yu, Outlier detection in graph streams, in *ICDE Conference*, 2010
31. C.C. Aggarwal, Y. Li, J. Wang, J. Wang, Frequent pattern mining with uncertain data, in *ACM KDD Conference*, 2009
32. R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in *SIGMOD Conference*, 1993
33. J.B. Orlin, A polynomial time primal network simplex algorithm for minimum cost flows. Math. Program **77**, 109–129 (1997)
34. R.J. Bayardo Jr., Efficiently mining long patterns from databases, in *ACM SIGMOD Conference*, 1998
35. J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in *ACM SIGMOD Conference*, 2000
36. S. Brin, L. Page, The anatomy of a large scale hypertextual engine, in *WWW Conference*, 1998
37. R.K. Ahuja, T.L. Magnanti, J.B. Orlin, *Network Flows: Theory, Algorithms and Applications* (Prentice Hall, Englewood Cliffs, NJ, 1992)

From Patterns to Discoveries

Michael R. Berthold

1 The Past

Over the past two and half decades or so, a typical data analyst, such as myself—even though I really did not even realize at start that I was doing data analysis—started off by learning descriptive statistics and getting truly excited about the promises of machine learning. The ability of those methods to match powerful models to—today ridiculous but back then huge amounts of data opened up endless opportunities. I dove into neural networks and other similar approaches, together with their algorithms to match complex, distributed models to large data sets [1].

From this, the step to trying to actually understand those models was a natural desire. Being able to predict sufficiently ahead of time that quality in a production line will be dropping is one aspect, but being able to explain why this is the case allows to actually fix the problem at the root. So in the 1990s, I started looking into rule models and decision trees to extract interpretable patterns from large data sets. Extensions of these methods based on imprecise logic allowed to inject (inherently imprecise) expert knowledge and extract more general but still performant interpretable models [2]. This shift made it possible to find interpretable models to somewhat larger data sets, but we quickly ran into walls nevertheless.

Due to one of those strange shifts in life, I started working on data sets from the life science industries—the company hiring me to head their data analysis think tank realized that there was a wealth of powerful methods in the data mining community with potential for their own applications and was looking for fresh insights—or as their CEO put it: someone unbiased by any actual knowledge about the underlying applications and previous work in the field. In me they found

M.R. Berthold (✉)
Nycomed-Chair for Bioinformatics and Information Mining, Department of Computer and Information Science, Graduate School on Chemical Biology (KoRS-CB), University of Konstanz, Konstanz, Germany
e-mail: Michael.Berthold@Uni-Konstanz.DE

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_4,

a perfect candidate; I could barely spell "biology." When my focus shifted to life science data analysis, it became apparent that researchers in this area struggled with much more pronounced problems: they often did not even know what questions to ask! It took me a few months to abandon the holy grail of data mining: performance. In the life science areas, nobody cared about that famous last percentage point of model accuracy—the experts in the field would not trust our models anyway. They wanted explanations. And much more importantly, they wanted to see something new, something interesting, and something that would trigger a truly new discovery! During one of the long discussions we had with our users, trying to find out what kind of patterns they were looking for, a chemist finally got really nervous and exclaimed:

> I don't know what I am looking for, but I know it when I see it!

Since then, this phrase has driven much of the research in my group; see, for example, [3–6].

2 Types of Miners

It is interesting to note that the different types of data analysis described informally in the section before can be grouped into three main categories and nicely matched to different phases of scientific research.[1] Figure 1 illustrates this analogy.

2.1 Parameterization

This phase of data mining concerns essentially the fine tuning of an answer we already know exists, but we are lacking a few parameter values to really understand the underlying system—statistical data analysis represents the core technology on the data mining end. This type of analysis corresponds to the third phase in scientific research: Formalization. The system is well understood and the remaining questions rotate around fitting parameters. Theory formation and systematic experimentation go hand in hand.

2.2 Pattern Detection

The classical data mining area focuses on finding patterns. The nature of the patterns (e.g., association rules) is clear, but we do not know yet where those

[1] Personal communication with Christian Borgelt who cited (from memory) a publication that we were unable to find. Please contact the author if you know the reference.

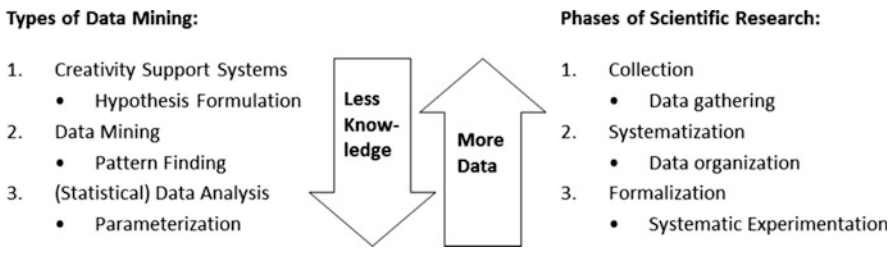


Fig. 1 The three phases of scientific research and the matching types of data mining

patterns occur. So we need highly efficient algorithms to dig through vast amounts of data to find the (often only few) local instantiations. The scientific research equivalent is conceptualization or systematization: we have an understanding of major parts of the system but lack some details. In a way, this phase relates nicely to question answering systems: can you find me more instances of this pattern? Can you determine the subset of features needed to explain this outcome? Finding a predictor is really a subdiscipline here: we may not care about the interpretability of the result, but we know (or at least we believe we do) which features may have an influence on the target and try to find some black box mimicking the underlying system's behavior.

2.3 Hypothesis Generation

The third phase of data mining activities corresponds to the early discovery phase of a new scientific area. Nothing really is known about the overall structure of the underlying system, and the goal is essentially to collect evidence wherever possible that can be used to poke holes into the fog and gradually build up a more complete understanding. This relates to the setup sketched above: the user cannot even really specify the question. Instead, we need to support users to form questions in the first place! This phase is often characterized by the ability to generate huge amounts of data and not really knowing what to do with it. The scientific research phase "Collection" stresses this aspect: we are collecting bits and pieces of information but do not (yet) know how those pieces fit into the overall puzzle. The goal is help users form new hypotheses. Quite a few new research areas arose around this topic in various fields: visual analytics in the visualization community, active learning in the machine learning sphere, and even data mining have a spin off: explorative data mining. But the overarching goal is hardly ever stressed explicitly: the aim is to build creativity support systems that help users to generate new, exciting questions.

One big problem when moving into this last phase becomes validation. Methods for parameter fitting and pattern mining can rather objectively be evaluated on standard benchmarks and using well-founded metrics. Obviously, there is a bit of a

risk of overfitting[2] but, with a bit of care, strong conclusions can be drawn from a well-done experimental analysis.

However, for the third type "hypothesis generation," this is far more complex. Finding a new hypothesis is not something that can be recreated in an isolated test bed. Measuring what is interesting is a difficult problem altogether. So the danger is that we will see—again—lots of papers presenting new approaches consisting of small modifications of existing methods and present their proclaimed superiority on one or two well-chosen examples. Better would be reports of successes on real life use cases, but are those really broad evaluation and can those really be used for comparisons? Fair validations of such systems will be an interesting challenge.

3 Tools

It is interesting to look at the evolution of development of tools for data processing, analysis, or mining over the past 20 years as well—they tend to follow the phases outlined above but typically lack behind a number of years. Maybe we can learn something from this trend for future developments.

Initially, tools were essentially table based, following the usual spreadsheet setup of VisiCalc or later mainly Excel. These tools allow the users to do increasingly complex operations but restrict those operations essentially to a single table. A parallel development was highly sophisticated statistical programming languages such as S and the open source version R. These languages allow highly skilled experts to run pretty much any analysis anyone can think of on a variety of data sets. However, the resulting code is far from being maintainable, and usually, research departments have a few experts that the analysis hinges upon. In a way, this development matched the first phase outlined above—at more or less advanced levels, of course.

A variety of other tools showed up briefly (such as SNNS, the Stuttgart Neural Network Simulator), but none really stuck around for long. One very notable exception is Weka, initially a machine learning toolkit which, when data mining research gained momentum, also became enormously popular among data miners. Weka, similar to R, managed to create a community of researchers adding new components to the platform and quickly became a reference platform among machine learning and data mining researchers. Weka still requires users to be pretty firm in the use of sophisticated algorithms and somehow exemplifies phase 2: if you are in search for a state-of-the-art data mining algorithm, you can likely find it in either Weka or—even more likely—also R. Of course, in parallel to those open source developments, also statistical analysis firms such as SPSS and SAS started to develop or buy tools similar to those resulting in tools such as Enterprise Miner and

[2] As Pat Langley once put it: "The entire Machine Learning community is kind of overfitting on the UCI Benchmark collection."

Clementine. In contrast to R and Weka, however, they did not foster community integrations and are not as heavily used by data mining researchers.

For phase 3, we are looking at a different picture: we need to enable users from other application areas to use our research prototypes instead of the usual benchmark evaluation. We can neither expect them to be aware of the latest bleeding edge development in our field nor can we expect them to make use of every little parameter twist in existing algorithms. They want to use tools, play with their data, and, by exploration, come up with new ideas and hypotheses. Those tools need to be interactive to allow exploration and—most importantly—intuitive to use also for expert as well as novice users!

This results in a dramatic shift in requirements for data mining researchers trying to trigger progress in this area: supporting complex, explorative analyses will require—even for research prototypes—fairly stable, professional grade tools. Otherwise, researchers in the application areas will simply not be able to use those in their live working environments and, in return, researchers in the data mining area will not be able to seriously evaluate their methods. This, of course, poses an entire set of new challenges for researches in the data mining field: in addition to publishing algorithms, we will now need to accompany these publications with the deployment of the new methods in environments that can be used in a professional context.

Obviously, we cannot require professional tool development from each and every research group. I personally do not see a way around highly modular frameworks that can be used by the broader research community. At the University of Konstanz, we invested over 2 years of time building up such a framework, guided by three main goals and resulting in KNIME, the Konstanz Information Miner [6][3]:

- Professional grade: from day one experienced, software engineers were part of the KNIME core team.
- Modularity: an integration platform can only be as good as its weakest piece. Therefore, KNIME essentially sandboxes all of the individual modules sanity checking as much of their operation and subsequent output as possible. This way, we can quickly determine why a workflow failed and isolate the misbehaving module.
- Community involvement: in order to allow others to (a) benefit from the work invested on our end and (b) integrate their own research results, KNIME has been designed to allow for simple integration of additional types of data, algorithms, and data handling routines.

It has taken time to convince the community (and the actual users working on real world problems!) that KNIME is not yet another one of those cool but not really useful open source projects that will die away when the responsible Ph.D. student graduates. But in the past years, KNIME has increasingly gained traction in both:

[3] http://www.knime.org.

the academic community and real-life applications. It has been exciting to see how research in my own group started benefiting from a common underlying framework after a few years, but it has been a real thrill to see how this also scales to industrial users and the academic community recently. As so often, the value of the whole is much greater than the sum of the individual contributions.

4 Sparking Ideas

As outlined above, triggering new insights can still be considered the holy grail of data mining. Now, however, we understand much better what this really means. Already Wilhelm Busch knew:

> Stets findet Überraschung statt. Da, wo man's nicht erwartet hat.

Translated roughly as: surprise happens where you least expect it. However, many of the existing data mining methods and especially the tools deploying them to normal users do not really support this quest. Often, what is called "data fusion" essentially results in a very early requirement for information (source) selection and the use of automated mechanisms for feature selection which aim to minimize a predefined metric which stears that process. However, in reality, we very often do not know which data source or which metric are most appropriate to guide us toward finding the unexpected. Prediction or some other measure of accuracy on a subsample of the available data cannot be the guiding force here.

Being able to push the information source and feature selection process as deeply into the analysis process as possible will be a requirement to actually support the user in making new discoveries. Ultimately, one should be able to keep all of the available data sources involved, not only the ones that one believes from the start could have some value. Truly novel discoveries will come from discovering connections between previously unconnected domains.

Our recently concluded European Project under the acronym BISON [7] launched a first attempt to address this issue. During project definition, we stumbled across Arthur Kostler's seminal work on creativity[4] where he defined the term *bisociation* as a description of a discovery that crosses domains. This term emphasizes the difference to a typical association which is defined within a domain (or at least a well-defined union of domains). The term *Bisociative Knowlegde Discovery* therefore nicely illustrates what we are after, contrasting this against methods for well-defined pattern discovery. The project did not find the ultimate solution for this daunting task, of course, but a number of very promising directions for future research, most notably in the discovery of bridging concepts of various types, were initiated. Lots of work still needs to be done before we can actually give

[4] Arthur Koestler: *The Act of Creation*, 1964.

users a system that does support the discovery of bisociations, but we took a big step into that direction.

In my opinion, true discoveries will arise when we combine sophisticated exploration with complex, heterogeneous data sources and a multitude of data mining algorithms. The resulting systems will support the discovery of new insights across domains—and require serious efforts on joint research among the various disciplines of data mining in conjunction with researchers in the applied domains. After more than 20 years, data mining has just started to be exciting all over again!

5 Conclusions

My discussions above are naturally highly biased by the challenges we face in our daily work with researchers in Chemical Biology at Konstanz University and by our interactions with biotechs and pharma companies. Research in the other areas of data mining, on methods and theories, is, of course, still of high importance and will continue to impact research in the other areas. However, I believe that in terms of grand challenges, the development of a complex data mining system that enables a true, domain bridging discovery has the potential of a huge impact on how scientific research will be done in the future.

References

1. Michael R. Berthold, Jay Diamond, in *Boosting the Performance of RBF Networks with Dynamic Decay Adjustment*. Advances in Neural Information Processing Systems, vol 7 (MIT, Cambridge, MA, 1995), pp. 521–528
2. Rosaria Silipo, Michael R. Berthold, Input features' impact on fuzzy decision processes. IEEE Trans. Syst. Man Cybern. B **30**(6), 821–834 (2000)
3. Christian Borgelt, Michael R. Berthold, in *Mining Molecular Fragments: Finding Relevant Substructures of Molecules*. Proceedings of the IEEE International Conference on Data Mining ICDM (IEEE, 2002), pp. 51–58
4. Michael R. Berthold, in *Data Analysis in the Life Sciences: Sparking Ideas*. Knowledge Discovery in Databases, Machine Learning: PKDD/ECML 2005, Lecture Notes in AI, no. 3720 (Springer, 2005), p. 1
5. Michael R. Berthold (ed.), *Bisociative Knowledge Discovery*. Lecture Notes in Computer Science, Springer, in press
6. N. Cebron, M.R. Berthold, Active learning for object classification. Data Min. Knowl. Discov. **18**(2), 283–299 (2009)
7. Michael R. Berthold, Fabian Dill, Tobias Kötter, Kilian Thiel, in *Supporting Creativity: Towards Associative Discovery of New Insights*. Proceedings of PAKDD 2008, LNCS 5012 (Springer, 2008), pp. 14–25
8. Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel, in *KNIME: The Konstanz Information Miner*. Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization (Springer, 2007), pp. 319–326

Discovering Privacy

Chris Clifton

1 Motivation

I like to say that I backed into data mining. I come from the database community, although with a dissertation in database support for hypertext (while at Princeton University) I have never really been a "core database" researcher. With 5 years (Bachelor's and Master's) at M.I.T., I had also had more machine learning than most. This led to my first real "data-mining" work (although it was not called that at the time): applying machine learning techniques to a problem in heterogeneous databases. Schema integration, or more specifically schema element matching, has all of the hallmarks we see in good knowledge discovery problems:

- Lots of data—by definition, multiple databases worth.
- A desire for the big picture—what types of data (columns) in one database match columns in others?
- No simple answers—even experts would disagree on the correct answer to a simple question: Do two columns contain the same type of data or not?
- Too little knowledge—even well-documented schemas often evolve into something new without the documentation keeping pace.

I left Northwestern University to join The MITRE Corporation to work with others doing database integration work. That same year, data mining hit the big time with the first KDD conference. MITRE needed data-mining experts, and with my background, I became one of them.

That said, I think the field is a natural fit. I remember a line that expresses how I feel: "Lies, secrets, how I hate not knowing things." (I attribute this to Oscar Hammerstein II but have not been able to find or remember exactly where.) Taking

C. Clifton (✉)
Department of Computer Science, Purdue University, 305 N. University Street, West Lafayette, IN 47907-210, USA
e-mail: clifton@cs.purdue.edu

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_5,

data and finding new knowledge in it is fun; finding new ways to do this with ever-larger data sets is even better. Text mining fits right in; I love to read, but there is no way I can read everything being written. But what if I could use a computer to analyze a corpus and give me knowledge that cannot be found in any single article?

Of course, we also have things we would prefer *not* be known. I recognized early that data-mining technology could provide the ability for others to learn things we would rather keep hidden—this led to an early position paper on the security and privacy implications of data mining [3]. I have since come to the conclusion that an even greater risk is through poorly protected data gathered to support data-mining projects (or even for the possibility of unknown, future data-mining projects.) Most data-mining technology is like research—the goal is *generalizable knowledge*, not specifics about individuals. The data, unfortunately, is often about individuals—and it is the unintended disclosure or misuse of that data that causes damage. Reconciling "I want to know" with "I want to control my own image" is what drives my current research—how do we mine data without looking at it? My hope is that through new privacy technology, we will continue to be able to enjoy the benefits of data mining but still have control over data about ourselves.

2 Milestones and Success Stories

The first real milestone in my progress as a data-mining researcher came before I really even knew the field. The SemInt system [5], developed with my Ph.D. student Wen-Syan Li at Northwestern University, used neural networks to solve what was essentially a multiclass classification problem without independent training data. Essentially what we did was use data describing the attributes in a database as features for a classifier and use the features of a database to train a classifier to recognize that database. While this seems like a recipe for overtraining, through limiting the capacity of the classifier, and using the fact that columns in different tables in the same database may reflect the same attribute (class), we obtained surprisingly effective results.

This idea that we could train on the very data we were trying to understand was useful in a very different application, detecting unusual changes in overhead imagery [1]. Again, the problem was a lack of training data—almost by definition, we do not know what an unusual change looks like. But we do know that most changes *are not* unusual, even if they are substantial (e.g., white changing to green in New England images taken in January and July). By training on the very data we wanted to understand, and seeing where the classifier prediction did not match the actual outcome, we could effectively use a classifier where we did not have a traditional training data set.

A second milestone was the realization that data mining really needed to be viewed in the context of a problem. I began a text-mining research project at The MITRE Corporation with the idea that through appropriate preprocessing and application of data-mining tools, we could learn things from large text corpuses

that would go well beyond what a human reader could see. The result was too much data (as *output* from the data-mining tools) and too little knowledge. It was only when I looked at the problems being addressed by the topic detection and tracking project [8] that I realized I could turn this data into useful knowledge. The TopCat system [2] achieved topic detection performance comparable to systems developed in that program, but through use of data-mining techniques such as the Apriori algorithm was able to scale to dramatically larger corpuses, and on smaller corpuses could achieve near-interactive response time. Combined with geotemporal visualization tools, this resulted in a system that provided useful and novel knowledge.

A third milestone was understanding that the real privacy issues in data mining were with the data, not the outcomes. By identifying problems where sharing the data was the issue, not the data-mining result, I (with my Purdue Ph.D. students Jaideep Vaidya, Murat Kantarcıoğlu, and others) developed numerous techniques to learn from distributed data sources without sharing the data [9]. This led to a new adversary model (with Ph.D. student Wei Jiang) [4] to prove that such techniques met real-world needs; most previous work assumed a semi-honest model where parties were assumed to have no desire to see data they were not supposed to. This led to new work in data anonymization; with Ph.D. student Mehmet Ercan Nergiz, we discovered that existing methods to measure quality of anonymization had little bearing on its quality of models learned from the anonymized data [7]. The outcome of this is privacy technology that really does give us the ability to have our cake and eat it too—in this case, to maintain individual privacy but still gain valuable knowledge.

3 Lessons in Learning from Failures

While privacy technology has in my view been a research success, I fear that it has been a market failure. More precisely, there has not even been the interest to bring the technology to market. We may all want our privacy, but we are not willing to pay much for it. Ask yourself—will you give out your age, gender, income, etc., to anyone who asks? Most will say no. But will you provide that information to get a discount on a purchase? How much of a discount? In general, not enough to pay for the technology required for the retailer to build their customer models without actually learning (and potentially accidentally disclosing) your private details.

This has lead to a new line of research: technology supporting collaboration using confidential data, targeted at corporate confidentiality. Companies place great value in their data (including private data they hold about individuals) and are willing to spend to protect it. At the same time, they realize the value others can provide through use of that data, posing a challenging trade-off between sharing data for the value received and protecting it to keep that value. I hope that companies will see the value in technology for learning from data without disclosing it and support the development of technology that also enhances privacy.

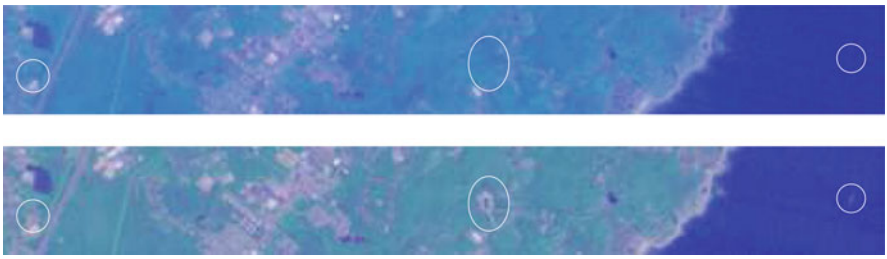


Fig. 1 Two images with unusual changes as detected by Clifton [1] *circled*

Actually, my experiences with data mining have lead to many other cases where I have had to learn from failure. Data-mining research inherently leads to failure, as the problems to be solved are rarely clearly defined, and even what it means to succeed or fail may not be known until the research is well underway. In Dr. Hand's essay, he quotes a definition of data mining as finding "unsuspected relationships"—as a researcher, this means we are trying to develop new techniques that will find knowledge we do not even know is there to be found. For example, Fig. 1 shows two images of the same location, taken 4 years apart. The system in [1] detected the circled changes—but are these interesting? Is the result complete? It depends on what the user wants to know—and they may not even know this until after they see the system in operation. This makes our life as researchers very challenging, as we often cannot tell success from failure.

That said, failure can give us insights that lead to success. My work in intrusion detection (as with my work in text and image mining) initially resulted in a flood of uninteresting results. With both the intrusion detection and image-mining work, it was the realization that the results *were* uninteresting that led to the breakthrough—what was really interesting was the data instances that did not match the flood of results. While there was no guarantee that these were interesting, it drastically reduced the amount of data a human needed to look at to determine what was interesting. (This insight has not been helpful in privacy-preserving data mining, where showing someone a few interesting data items would probably be the worst way to violate privacy).

4 Current Research Issues and Challenges

My current data-mining focus is on privacy and anonymity in textual data. This combines two very hard problems: text mining (in various guises, including information retrieval and ranking) with text de-identification and anonymization. Textual data contains a lot of information but also a lot of noise and redundancy. The redundancy means that simple de-identification techniques may fail; the noise means that straightforward anonymity measures will almost always tell us the text is not sufficiently anonymous. My Ph.D. student Mummoorthy Murugesan

and I explored ways to generate "cover queries" to hide a user's intent in Web search; the problem is that it is difficult to ensure that the real user's query does not stand out. Through use of data-mining techniques, we were able to generate queries sharing many characteristics with a user query, but on completely different topics, thus hiding the actual user intent [6]. This is not only a challenging problem to solve; it is an extremely challenging one to evaluate—how do we know if we have succeeded in preserving privacy?

To answer question such as this, I have recently started an NSF-funded project *Anonymizing Textual Data and its Impact on Utility*. The goal is not only to develop techniques to anonymize textual data but also to develop measures of both the real-world effectiveness of anonymization and the impact of using anonymized data on real-world research problems. Through collaboration with researchers from areas as diverse as health care and linguistics, we hope to derive new methods that will improve availability of data for research, while providing well-understood standards of privacy protection.

5 Research Tools and Techniques

So how do I proceed with my research? I first start with the passion—a general area of interest that excites me (e.g., privacy/security issues in data mining). Then, I try to understand the open problems—what is the real issue? What has and has not been done? Who cares? The last is critical to finding funding. Given a problem that someone else cares about, funding is just a matter of time and effort. While I generally have to seek out the funding, I have had occasions where simply making my interests and ideas known in the right community led sponsors to seek me out. And for those thinking that this is just a problem for academics, people in industrial research have it even tougher—they need to find not only someone who cares, but this someone also has to be in the company (or one of the company's customers.)

Once the ideas and funding are in place, the next steps can vary considerably. Much of the privacy work involves pencil and paper; the ideas must be proven to preserve privacy before it makes sense to implement anything. Other times, it is important to start exploring the data to gain understanding of the problem, using whatever tools happen to be available and seem to be the best match. Only after exploration of the data and some initial understanding is gained the really good ideas start to flow.

Most of my work stops at the proof-of-concept prototype stage. In general, I leave performance issues to other researchers and bulletproof code to the commercial (or increasingly, open source) world. I often implement on top of a relational database. For example, I developed a generalized version of Apriori in SQL on a commercial database. While not the fastest, it allows rapid exploration and validation of ideas, as determining *what* we want a system to do must come before figuring out the best way to do it.

One research tool I find invaluable is other researchers. While I occasionally work alone, I feel that collaborating with others results in much faster progress, fewer dead ends, and is generally the right way to do research. I am currently blessed with a great group of faculty and students to work with—Purdue has put together a machine learning/data-mining group spanning multiple departments (primarily computer science, statistics, and electrical and computer engineering) that enables me to contribute where I am strong and find others who take over where I lack background. This leads to results that any single one of us would probably not achieve.

6 Making an Impact

One way to make an impact is to carefully select a problem that has significant commercial value and solve problems that enable significant new value or cost savings. A second is to be in the right place at the right time. So far, I think my privacy impact has been more of the latter.

I really focused on privacy-preserving data mining when I started at Purdue in 2001, 2 years before concerns over potential invasions of privacy in attempts to combat terrorism lead to Senator Feingold's introduction in the US Senate of the "Data Mining Moratorium Act of 2003." This immediately raised awareness of privacy issues in the data-mining community, and my understanding of the issues and ability to demonstrate that we could reconcile them put me at the forefront of a new research area. Given the low value most people place on their private data, commercial development of the technology may well require government mandate, and (thankfully, in my opinion) such mandate has not come before the technology is ready (or it may well have been with the broad strokes of the 2003 act, which would have banned, among other things, any "data-mining program of the Department of Defense," including research and development.)

That said, my primary goal has never been research commercialization; if it had been, I would have joined a start-up in the mid-1980s boom. Even though I had seen my research prototypes tried out in real-world exercises while at MITRE, I found that the success of my students was a more important impact to me. Because of this, I rejoined academia, and the accomplishments of my students make me feel I have had a real impact, even when those accomplishments are not a direct result of my research.

7 Future Insights

So what would I suggest for a new researcher who wants to make an impact? First, be passionate about what you do. If you are not excited about it, why would someone else be? That passion, the willingness to push a good idea even when it is not clear it will succeed, is what will really result in changes.

Second, find good mentors. This is not just for students searching for a thesis advisor; my push into privacy would probably not have happened without the ideas and encouragement of Dr. Bhavani Thuraisingham, my mentor at MITRE. A senior person who will help you to develop and sell your ideas is invaluable, both for the experience (so you spend less time learning from failures and more making an impact) and for the visibility and credibility.

Third, pick the right problem. I put this third because I believe that without the first two, picking the right problem may result in a few papers, but it will not result in real impact on either the research community or the real world. That said, I will give my thoughts on the "right problem" for those working in data mining.

7.1 Short Term

In the next couple of years, I see growing interest in "big data" mining: dealing with petabyte and larger data sets. Both in research and commercially, there has been considerable investment in generating and collecting data—everyone will be looking to data mining to get a return on that investment.

This leads to three research challenges. The first is obvious—scaling existing techniques up to such large data sets. This may build on work such as mining data streams, using massively parallel architectures, or (best of all) new approaches not even thought of.

The second challenge is to show that for a lot of problems, we do not really need that much data. Does increasing the training set size really result in a better classifier? Beyond a certain point, the value is minor. Unfortunately, the people who have invested in producing huge data sets do not want to be told "just give me 1% of it, the rest is not useful anyway." Research in this area will need to prove that the results obtained from a (possibly carefully chosen) sample of the data are just as good as using the whole data set.

7.2 Long Term

I did say that large data sets give us three challenges, but only listed two. The third I consider a long-term challenge: data mining to find completely new types of knowledge that can only be found from such large quantities of data. Can such large data sets enable us to escape the "curse of dimensionality" and actually gain valuable knowledge that is truly high dimensional? Will "outlier detection" become instead "outlier summarization," enabling us to learn interesting rules that apply to only a small fraction of the data but are interesting in spite of that? What other things might we discover given petabyte-scale (and larger) data sets? Does the variety of types and relationships in such large data sets (e.g., social networks) allow us to answer questions nobody has yet thought to ask? There is more than enough here to keep the field advancing for many years.

Do I consider privacy to be a long-term challenge? The answer is yes, but a relatively limited one. I think we have developed a lot of good privacy technology in this community and have a good understanding of how to solve new problems as they arise. I expect we will see flurries of activity as regulations (or, hopefully, commercial demand) raise new challenges. I plan to continue working in this area, but then again, it is an area about which I am passionate (and as I said, I believe passion matters more than picking the "hot" topic.) Of course, there is always room for others who are as passionate about privacy as I am.

8 Summary

I have been active in the data-mining field for 15 years and have seen a field grow from developing tools (and companies to market those tools) to one that addresses real-world problems and is integral to operations of some of the world's biggest companies. Technology developed in this field permeates data and information management—sometimes so embedded as to be unrecognizable, but still driving how we use data.

I also see a field that continues to grow and expand, continuously identifying new challenges and moving beyond its borders to address them. I see this continuing; we will never develop a "general purpose data-mining tool suite" that addresses all challenges. The search for new knowledge will always need new ways to find that knowledge.

I will continue to push for privacy technology—the benefits of data mining need not come at the expense of losing control of our personal information. The ever-expanding nature of the field makes this a continuing challenge, and in my view, this means endless opportunity. My journey is just beginning.

References

1. C. Clifton, Change detection in overhead imagery using neural networks. Int. J. Appl. Intell. **18**(2), 215–234 (2003). http://www.kluweronline.com/issn/0924-669X
2. C. Clifton, R. Cooley, J. Rennie, TopCat: data mining for topic identification in a text corpus. IEEE Trans. Knowl. Data Eng. **16**(8), 949–964 (2004). http://csdl.computer.org/comp/trans/tk/2004/08/k0949abs.htm
3. C. Clifton, D. Marks, Security and privacy implications of data mining, in *Workshop on Data Mining and Knowledge Discovery* (ACM SIGMOD, Montreal, Canada, Jun 1996), pp. 15–19, http://www.cs.purdue.edu/homes/clifton/cv/pubs/dmkd.pdf
4. W. Jiang, C. Clifton, AC-framework for privacy-preserving collaboration, in *SIAM International Conference on Data Mining*, Minneapolis, MN, 26–28 Apr 2007, http://www.siam.org/proceedings/datamining/2007/dm07_005wjiang.pdf
5. W.S. Li, C. Clifton, SEMINT: a tool for identifying attribute correspondences in heterogeneous databases using neural networks. Data Knowl. Eng. **33**(1), 49–84 (2000). http://dx.doi.org/10.1016/S0169-023X(99)00044-0

6. M. Murugesan, C. Clifton, Providing privacy through plausibly deniable search, in *2009 SIAM International Conference on Data Mining*, Sparks, NV, 30 Apr–2 May 2009, http://www.siam.org/proceedings/datamining/2009/dm09_070_murugesanm.pdf
7. M.E. Nergiz, C. Clifton, Thoughts on k-anonymization. Data Knowl. Eng. **63**(3), 622–645 (2007). http://dx.doi.org/10.1016/j.datak.2007.03.009
8. Topic detection and tracking project TDT (25 Sep 2000), http://www.nist.gov/speech/tests/tdt/index.htm
9. J. Vaidya, C. Clifton, M. Zhu, *Privacy Preserving Data Mining*. Advances in Information Security, vol. 19 (Springer, New York, 2006). http://www.springeronline.com/sgw/cda/frontpage/0,11855,4-40356-72-52496494-0,00.html

Driving Full Speed, Eyes on the Rear-View Mirror

John F. Elder IV

1 Bright Lights

It is midnight and I am wide awake. My wife, Elizabeth, is sleeping soundly in the fancy Las Vegas hotel suite we have been lent, but I am stuck on the speech that I will give in the morning for the M2010 conference. There will be rock music, stage lights, a big introduction, and then I pop out to talk about. . . math. Or, rather, stories about analytics helping business. Unlike a visiting preacher, who can put all his best stories in one sermon, I need new material; I am back to keynote for the fourth time. The pressure is on to deliver a high-tech, but entertaining, talk (especially since Dick DeVeaux gave a great one yesterday. But he was a professional singer and dancer before becoming a great statistician, and I do not have that to fall back on!). I have, of course, been thinking about it for some time, but it has not gelled. As a frustrated perfectionist, it takes a tight deadline for hard things to get done.

I go out to walk the halls and let my mind wander over the topics I would like to touch: How I got to be a data miner. Why data mining is such great work, useful for almost everything! The importance of people—character, teamwork, shared goals, and connections. The crucial need to pay attention to human hopes and fears. The distinction between technical success, where we solve the defined problem in the lab, and business success, where the solution actually gets used.

For me, working in this industry is a joy. I labor alongside dedicated colleagues who want to contribute to something bigger than themselves—friendly people eager to help others. And with smart clients with whom I can swap expertise, we can meet their goal. The beauty of being a data miner is that your skills are so useful that wherever you go, people are happy to see you. You add value by helping others improve their processes and make sense of their data. And there are good

J.F. Elder IV (✉)
Elder Research, Inc., 300 W. Main Street, Suite 300, Charlottesville, VA 29901, USA
e-mail: elder@dataminginglab.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_6,

opportunities to publish your discoveries, since the field is still new and many industries (application areas) are just awakening to its potential.

2 Tom Swift and His Amazing Computing Device

I have had the good fortune to work in this emerging field since its beginning. Today, there are some schools with degrees in data mining, and those types of programs will certainly grow. I was trained as an engineer. I was a big fan of science fiction, having read every Tom Swift book I could as a young kid. I dreamed of becoming an inventor, solving technology puzzles. (My first program simulated blackjack, with all its fancy options, and I learned the thrill of seeing my creation used and enjoyed as classmates would sneak away to play it.) I went into Electrical Engineering—to design and build real things—at Rice University, in Houston, Texas, which was a great fit for me.

I focused on signal processing and computers. I loved programming; I (frankly) endured some of the engineering classes but thrived with algorithms and debugging. Nonetheless, I am glad I went through the discipline of engineering. In working with circuit parts, for instance, you learn how to make reliable devices from parts that are themselves very unreliable. That is, you gain an instinctive appreciation for noise and uncertainty—so central to understanding statistics—that you do not get in the pristine world of 1's and 0's of programming.

I put myself through Rice with scholarships, loans, and summer jobs.[1] While in high school, I won a National Fellowship for a summer job at the Smithsonian in Washington, DC.[2] It sounded like a great position, but the title and the reality were far apart. I was in the Museum's Electricity Department and found myself in what was essentially a windowless closet, cataloging obscure insulators (the glass and ceramic inverted cups that used to be affixed to telephone poles to insulate the lines). The only exciting aspect of the job was that I occasionally got to help repair electricity exhibits. Once, as I leaned down to unscrew the case of one of Ben Franklin's inventions, a guard challenged me—hand on gun, until I scrambled for my badge! I did create a temporary exhibit on unusual insulators, but mostly the job was a good lesson in what not to do with my life. I wanted to think, to create, to solve problems, and to interact with people, then I would look forward to work each day.

In the summer before Rice, I worked for the United States Geological Survey. I was charged with developing a way to predict how long the cartographic work for

[1] One could do that a generation ago, when the ratio of initial salary to tuition was probably four to five times better than it is today! College tuitions have grown so fast relative to other costs—partly due to being a "high touch" (not easily automated) commodity, but mostly due to subsidies—that college is increasingly becoming a money-losing investment for those not entering a profession valued by our economy. Fortunately, data mining has a very healthy return on learning investment.

[2] Then, the Museum of History and Technology, nicknamed by some "America's Attic."

a map should take. It took me a while to realize that this task put me at odds with the workforce, which did not appreciate management's view that the maps could be drawn more efficiently. I did not excel at the predictive task, but when it was concluded, I got to do some detailed cartography myself, which was surprisingly interesting. Yet as I picked up skills, the older guys stopped talking to me. One who had been something of a gruff mentor (who was an accomplished artist in his spare time) finally explained that my work was making them look bad.

This was my first experience with a discouraging environment where excellence was not the top goal. I have since understood how such behavior can be a rational (though regrettable) response to an organization too large, constrained, or impersonal to reward results well. Today, I love working with government agencies in the law enforcement and national security arenas as they are full of hardworking, patriotic individuals with the public interest in mind. They somehow accomplish great work while burdened with the procedures and constraints of a huge bureaucracy. I have it quite easy, in comparison, working with a handpicked team of bright colleagues. As a small, responsive company, we can get things done more quickly and inexpensively than the norm. The results of data mining analytics are very measurable and can stand up to the brutal bottom-line evaluation that successful commercial firms survive by. Thankfully, competence in analytics allows us to contribute in the very different cultures and communities of Wall Street, commercial firms, and government agencies.

That summer job at the USGS taught me a few things. First, not everyone may want the project to succeed! Second, it is important to know what the other people involved care about; interpersonal relationships are sometimes not the top priority for us science-types, but cannot be taken for granted. And third, I discovered more about the kind of organization in which I did not want to work.

One point about the predictive task that I undertook for the USGS: To determine how long it should take to draw a map, we were advised to sample random sections and extrapolate to the whole. There was heavy disagreement about the random part: would we need to estimate how much of the map was dense, etc., and use stratified sampling to proportionally weight the samples, or could we disregard being "fair" to different types of areas and assume it would all work out? None of us nonspecialists were equipped to answer this well enough to convince others—which virtually paralyzed the project—so I became interested in acquiring the tools to meet such needs.

3 Robots that Learn

I did have great summer jobs during my years at Rice. The father of a college friend ran a company near Washington DC called Adaptronics. I learned as much in those summers as I had during the school year. With about 50 employees, it was a good size; you could know everyone but still be given as much responsibility as you

could handle. I had great supervisors who gave me freedom and enough direction so that I was able to really contribute.

A growing business for the company was building robots that found cracks in nuclear cooling pipes. The radioactive cooling water in nuclear reactors makes the metal pipes brittle, which eventually becomes dangerous. Instead of the client spending a million dollars a day to drain the water and have people inspect the pipes, they could send in the robots. These would bounce an ultrasound signal through the pipes and record the reflected response. By "training" a computer model on examples of known flaws and nonflaw anomalies (like welds), the system could flag potentially dangerous findings for close evaluation by experts. The "training" took a long time with a large computer and needed as complete as possible a set of signal examples, but once built, the model with the "judgment" encoded was small and fast and could be loaded onboard the robot. In some cases, its judgment was better than a human's judgment as it was more consistent and the modeling algorithm could potentially find complex patterns not easily noticed by analysts.

Working at Adaptronics motivated me to stick with E-school because I could see that the end of the tough slog could hold a challenging and enjoyable job. One of the wonderful things about programming is that you can quickly make your thoughts become reality. As code is abstract, it becomes real much faster than the mechanical devices of engineering. And, as you build up a toolkit of reliable programs to do tasks, you can move on to higher and harder problems, and your productivity grows rapidly. Spending three summers immersed in nondestructive evaluation and similar tasks introduced me to inductive modeling, where one learns to extract general principles from specific cases—a key foundation for what would become data mining.

After graduating from Rice, I worked for a summer as an applications engineer at Texas Instruments with their new signal processing chip (the TMS320). We were to encourage sales by showing how the chip might be used. I wrote an article that included complete and debugged machine code to do a key task and learned how fun it is to see your work used, as TI employed it for training for years.

I was then supposed to head to Berkley for Ph.D. graduate work in robotics. But I had fallen in love with a girl, later my wife, who was leaving Rice to study theology and history at Emory University. Together, we instead decided to stay at Rice for master's degrees. This might not have been the best career move, but it was the right one personally. In his 2011 commencement speech at Rice University, the columnist David Brooks described marriage as "the most important decision you will face in your life. If you have a great career and a bad marriage, you will be miserable. If you have a great marriage and a bad career you can still be joyful." I felt this, instinctively, and I have tried to have my own business be a blessing to my colleagues' families, rather than being its chief competitor.

Near the end of my master's year, I paid a social visit to the president of Adaptronics. He had just recently sold the company and bought land in rural Virginia, where he would live as a gentleman rancher and start a new business doing high-tech aerospace consulting in guidance and flight control. He invited me

to join him as his right-hand man in the new company. What a great opportunity to learn from an experienced mentor! While the job turned out to be extremely demanding, it was a crucial step to my becoming a data miner.

4 High Finance and Higher Ed

Elizabeth and I moved to Charlottesville, Virginia, the nearest large town to the farm in Greene county, and I commuted 30 min out into the countryside. In 5 years, the company grew to about 15 employees, including one who would become a close friend, Dean Abbott (see chapter "Data Mining: A Lifetime Passion"), who years later would be a vital part of my own startup. I learned a lot about the aerospace industry, as well as very useful lessons in how to build a company. The best year of the job was an early project to write a data mining algorithm, using the technology of polynomial networks,[3] with which my mentor had been an early, though little-recognized, pioneer. It was exhilarating to create—or rather, improve upon earlier creations of—an algorithm to be like a "crystal ball," capable of peeking into the future.[4] A spin-off company, AbTech, helmed by Gerry Montgomery and Paul Hess, took my code and built around it a commercial product, and it is still a thrill today to see it on a client's shelf and find out how they use it.

Eventually, however, I had to leave. Expecting everyone to share his complete dedication, my boss pushed for incredibly long hours that were detrimental to family life. One week I put in 107! I looked forward to days when I had to get something to the FedEx office in Charlottesville, because it closed by 7:00 p.m., making it an "early" night for me. My originally close working relationship with my boss became very strained. I delayed leaving out of loyalty, and because having found Charlottesville to be a very pleasant place, I wanted to lay down roots here (I had moved 20 times in my Army family youth and wanted a stable, appealing place to live and raise my children). I also worried that there were very few other high-tech companies in Charlottesville, so I hung in there overlong.

In fact, I had to be essentially kicked out (through an ego-shocking reduction in pay after a small dispute), but this big setback proved to be a tremendous blessing. So, in 1989, I went back to graduate school, just down the road at the University of Virginia, where I studied with a great advisor, Don Brown. Two influences that sent me back to school instead of a job were reading the book *Numerical Recipes* [6] from which I had learned many useful things and attending the 1988 *Interface* conference, whose speakers on the intersection of computer science and statistics were electrifying to me.

[3] Polynomial networks are something of a cross between regression and neural networks, but their structure adapts to the complexity of the problem rather than being preestablished [2].

[4] For Tolkien fans, imagine building a *Palantir*, which can reveal what might be the future, though at great cost!

There is a saying that "To a little boy with a hammer, all the world's a nail." My hammer was a polynomial network, so I set out to learn about the rest of the statistical modeling toolbox. After the intensity of my job, I found Ph.D. studies liberating. I felt as if the sun had come out and birds had begun to sing. But, with a mortgage to pay and Elizabeth staying home now to raise our children, finances would be a challenge, even with a welcome research fellowship from the Systems Engineering department.

Then, an amazing development: a former client I had brought to the consulting firm, Mark Finn, and his company Delta Financial in Virginia Beach, Virginia, offered me a serious retainer for my help with quantitative analysis for their stock market work. Mark had always treated me with respect and affection, and I did my best to make his bet on me pay off for them. Many of his employees have stayed with him for more than two decades, which is a great sign of his qualities as a boss. I share his passion for learning and in being interested in too many things at once, and we have been friends through many changes and years. Mark taught me valuable lessons about the market, including the huge influence of noise, degrees of market efficiency, and dangers of over-optimization. Though unschooled in data mining, his intuitive grasp of the key issues is complete. Mark loves to give crazy investor/inventors their first million dollars to work with. Some of them became very successful managers; most did not. Often, I was the one to decide whether someone had something real enough to bet on.

I experienced some heartbreaking stories in that role. One bright engineer, with the backing of a legendary computer company founder, had put everything into building a 10,000-line program using genetic algorithms to trade commodities. It appeared very promising, but was not yet in use, so had not generated income. To keep going, he had mortgaged his house and bet everything, even to the point where his wife left him, to pursue his inventive obsession. But when I evaluated his work, I came to discover I could exactly match his results with a two-line program. That made it clear that he was not accounting for some costs correctly (and that the complexity of the search was masking the simplicity of the resulting algorithm). My efforts shattered the apparent value of his years of effort. It was terrible news to have to deliver, and I could only pray that the diagnosis would help him regain balance.

In another case, I evaluated the results of a company that had people in suits flying around the country giving presentations to potential investors. Their back-tests claimed that they could predict the next day's stock prices with an astonishing 70% accuracy. But after a week, I found that I could exactly match their results with a moving average of 3 days' prices. Understanding that their algorithm was that simple is discouraging enough, but fatally, one of the 3 days in their model was apparently *tomorrow*! (If they had dropped yesterday's prices, then their estimate would have never looked wrong!) They had made a simple error early on, but if I had not caught it, they would likely have been eventually charged with fraud. As it was, it was just the end of their company.

Alternating, at UVA, between high finance and low homework was a great blessing. Like many who return to school after working, I was highly motivated;

I knew the value of what I wanted to learn and master. With a mortgage and a growing family (three of our five children were running around then), I needed not to linger. Still, it was such a pleasant life that I might yet be there, had not another setback/blessing struck. Delta Financial hit a snag and had to drop me. That final spring, I raced against our bank account to finish. I pulled off writing my dissertation[5] in about 12 weeks, with Elizabeth's support; she would often bring me dinner and the kids for a picnic outside my office, then take them home as I climbed back in the office window to push late into the evening.

5 Bat Ensembles

No longer slated to join Delta, I was free to accept a 2-year postdoc back at Rice. There, I did a lot of writing and conference speaking and became active with the emerging field of KDD—Knowledge Discovery in Databases (the second "D" would soon morph to stand for data mining). Also, Elizabeth's extended family was all in Houston those same 2 years—a providential development.

Freedom to study and write anything, with no clients or boss, turned out to be surprisingly disorienting! But I eventually took full advantage. A great friend and roommate from Rice, Doug Jones, was now a tenured EE professor at UIUC and a world expert in time-varying frequency analysis. We collaborated on a project to predict bat species from their echolocation signals. (Bats can do such amazing things with their chirps that the Navy and Air Force are envious!) I wrote new a *k*-nearest neighbor algorithm, which tests all possible sets of dimensions to use, and a new decision tree algorithm, which examines splits two steps ahead instead of just one. I tried those and other modeling methods out, focusing much longer on a single problem than I would ever had the luxury to do before. And then, I had an exhilarating discovery.

Neural networks and 5-nearest neighbors (after careful input creation and selection) had almost identical overall accuracy, yet they disagreed a third of the time on individual cases. Studying those disagreements, I noticed that the estimate that was more *confident* was usually correct. That is, say 0.5 was the threshold for being in or out of the class (one of the six bat species), and one estimate was 0.9 (1)

[5] I had long wanted to use inductive modeling to improve global optimization. By analogy, imagine you seek the location of the deepest part of a large lake. You can probe the lake anywhere (by lowering an anchor, say), but each such experiment is costly, so you want to get to the overall bottom (not some local minimum) as efficiently as possible. You would also like some idea of the probability that a better result is still out there. The key idea is to model the response surface from the known probe results to guide the location of the next probe and improve the model with each new experiment's result. The required properties of the surface model are very different than those for normal prediction. My search algorithm—for Global R^d Optimization when Probes are Expensive (GROPE)—generalized Kushner's two-dimensional search [4] and was the world champ for many years by the metric of requiring the fewest probes [1].

and the other 0.4 (0); then the true class was likely a 1. So averaging the probabilities would be a better answer than either alone. However, it would be looking ahead to use evaluation data results to pick and choose which algorithms to employ, so I used estimates of *all* the algorithms which training revealed might be competitive, assuming I would get a result somewhat above their mean. Instead, the result was better than all of the individual models! I called this new approach "bundling."[6] At almost the same time, others were creating useful methods like "bagging," "boosting," "Bayesian model averaging," and "committee of experts" with very similar ideas (though mostly using variations in the data to get variety rather than different algorithms). The term that won out to describe the area is "Ensemble Modeling," and years later, I had the pleasure of coauthoring a book with Giovanni Seni to clarify and summarize the powerful developments in the field [7].

While I was at Rice, Daryl Pregibon headed statistics at Bell Labs—essentially like a top university in those days—and invited me to join him in writing a book chapter for the first KDD book. It seems statisticians were always telling data miners (who were predominately computer scientists by training) that everything they were so excited about had already been done years ago in Statistics (which is basically true!). They asked, "Could we explain, in 20 pages or so, everything about statistics data miners need to be aware of?" A daunting task, but Daryl (now at Google) had been known to converse warmly with AI folks and the like, and he was up for the challenge. Our collaborative effort [3] was well received and helped me reach an entry level of attention in the field.

Oddly, though I work with, and speak about, statistics, I have had little formal training in it. I did enjoy a brilliant and inventive statistics professor at UVA—John Aitchison, who was generous with his time. But I had already written a full data mining algorithm before then—used to good effect by the Air Force and others. I was mostly self-taught. This is likely a good thing, as statistics has been measured to be the "worst caught" discipline in universities. That is, its lessons are retained by students the least! I think it is because it is taught as a branch of mathematics, rather than as an experimental discipline. The central question in statistics is whether a relationship you have just observed is real. More precisely, how likely is it to have happened by chance? Starting a century ago, geniuses figured out ways to infer such useful probabilities from scanty evidence on well-defined problems using fancy mathematical equations, but only because they did not have computers! Today, resampling simulations are the cleanest way to learn and do statistical inference. I think it is better to master that single tool that essentially always works, than dozens of specialized formulas whose assumptions have to be tightly matched to your challenge. Statistical knowledge is so important that teaching it more optimally will have great benefits for mankind.

[6] From the Chinese fable of a wise father urging his quarreling sons to stand together, while each could easily break a single stick, none could break them when bundled together.

Anyway, while I was at Rice in Texas, a successful entrepreneur/investor moved to Charlottesville, Virginia and started asking around for the right person to help him build a focused investment model for a certain niche of the stock market. When he went to the Systems Engineering department at UVA, my former professors and peers told him "Sounds like you want John Elder, but he is no longer around." Coincidentally, he next turned to the firm for which I had worked before UVA and got the exact same response. Determined (fortunately for me), the investor tracked me down in Houston. I jumped at the opportunity and ended up making such frequent trips back to Charlottesville that my old church's softball team kept me on the roster for the 2 years I was out of state.

6 Striking Out

The stock market project turned out, against all predictions of investment theory, to be very successful. We had stumbled across a persistent pricing inefficiency in a corner of the market. A slight pattern emerged from the overwhelming noise which, when followed fearlessly,[7] led to roughly a decade of positive returns that were better than the market and had only two-thirds of its standard deviation—a home run as measured by risk-adjusted return. My slender share of the profits provided enough income to let me launch Elder Research, Inc. (ERI) in 1995 when my Rice fellowship ended, and I returned to Charlottesville for good. ERI was one of the first data mining consulting firms and now, with 30 of us, is possibly the largest in the USA.

Hedge funds were our primary clients in our early years, where years of steady results gave us a strong reputation in the field. We proved that repeatable patterns were possible to find, even for something as brutally competitive as the stock market. Other types of work started to come in, arising from referrals and my public speaking and short courses. Soon, commercial clients overtook investment work as our main business. We often did not have contracts with our entrepreneurial clients, just a handshake agreement. (That almost always worked out, though now, we are more responsible, given ERI has about 100 mouths to feed.) And things moved fast. I once took a call in Virginia while driving into work in the morning and started a new project over dinner in upstate New York that night.

Then the 9/11 attacks occurred and the world changed. Business was already reeling from the recent dot-com market crash, when suddenly all assumptions had to be revisited. We lost our largest contract, and a few smaller ones, overnight as businesses dug in defensively, and it was a very rough patch for a while.

I had taught several courses on analytics for the Department of Defense and even had a senior DoD executive work with us; she had been assigned to a sabbatical to learn from ERI in lieu of graduate school. I volunteered my services to help in the

[7] The quantitative system was fearless. We were not!

aftermath of the attack and, perhaps because of that or due to my speaking exposure, was appointed by President Bush to be on a new panel to advise the DoD on technology related to national security. The panel met for 2 days every month, and in my 5 years of serving, I had the chance to work alongside many dedicated and brilliant people. I was almost certainly the most junior and least influential of the crowd. Early meetings were somewhat chaotic—due to the stakes involved, but also to the fact that everyone in the room was used to others becoming silent when they began to speak! We had to relearn the kindergarten skill of taking turns.

While I was on the panel, ERI had to essentially recuse itself from actually doing any of the work the panel might review. So, after 5 years, it was good to rotate off that role and return to the front lines of working with the data. There, our team, led by my friend and colleague, Dustin Hux (see chapter "An Unusual Journey to Exciting Data Mining Applications"), has the great fortune of working with a terrific client and his team, and the data mining results generated have made their group a government center of excellence.

7 The Power of Data Mining

Being able to address a wide variety of stimulating challenges, and having good reason to expect strong results, make data mining a satisfying field. I see predictive analytics as able to help in three ways, to (1) discover the good, (2) eliminate the bad, or (3) magnify speed through semi-automation.

Our market sectors models, mentioned above, are an example of the first, where new patterns of market behavior were discovered and harnessed to trade profitably almost every month for a decade. Another example is the new drug we helped Pharmacia & Upjohn discover (before it became part of Pfizer). They were faced with a $1B decision on whether to continue development of a potential drug whose test results, using standard methods, did not appear good enough. But we were able to transform the data, adapt the performance metric to better fit the application, and create a new visualization technique that made it clear they had a winner on their hands. P&U changed course and brought the drug to market, and many years later, it became one of only three blockbuster drugs they developed in that entire decade. Literally, a billion dollars had been bet on our conclusion, and thankfully, it turned out very well!

A prime example of the second way data mining helps—eliminating the bad—is in detecting and preventing fraud. An early and influential ERI project led by Cheryl Howard (see chapter "A Field by Any Other Name") improved models at the IRS to such an extent that analysts were finding *25 times more* fraud, given the same number of investigations, due to improved ranking of suspected returns. Similarly, an ERI team led by Dustin Hux (see chapter "An Unusual Journey to Exciting Data Mining Applications") to discover and prevent fraud and abuse at a world-leading consumer electronics firm saved a documented $67M in the first 5 years.

The third way data mining helps—semi-automation—is exemplified by a project we did for Anheiser-Busch, identifying beer products on store shelves. AB is very interested in how the placement and refreshing of products affects sales, to the degree that they pay staff to spend up to 4 h creating a "planogram" reflecting the exact product positioning of selected stores. By focusing on recognizing the image patterns of the most popular products and developing a new way to iteratively build a classification matrix for the picture, we were able to very accurately identify 90% of the products instantaneously, speeding the process up by an order of magnitude. Likewise, a small project for the Richmond Police Department helped reallocate personnel by predicting the locations of expected criminal activity for a given time period, leading to increased interventions and arrests, even while using fewer police than before.[8]

8 The Business of Data Mining

One would think that with such a powerful technology to wield and with so many provable results, it ought to be easy to run a business in this field! But wild business cycles are a real challenge. Early in the company's history when we ran into a rough patch, I could mortgage the car to make payroll. Later, with more people to cover, I had to take out a second loan on the house. (For some reason, my wife did not like this! Despite the fact that collateralized loans have the best rates…) One of my favorite cartoons depicts a data miner slumping with discouragement as the executive with his resume, says "With all your predictive modeling experience, I would have thought you would have realized that we would not hire you!"

Planning ahead seems like a luxury when keeping the business going feels like a white-water rafting adventure down an unknown river. Our small tribe is living off the land, just a couple of bad hunting trips away from nonexistence. To be prepared for whatever the future throws at us, it is essential to attract and develop great people.

A successful data miner combines business knowledge, common sense, statistical expertise, and the ability to work closely with clients to learn from them and their domain expertise. Almost no one has a degree in the field, so we seek candidates with great character and technical backgrounds who are worth investing in with expensive apprenticeship training. We seek, as Winston Churchill might have said, "a humble person with little to be humble for."

Humility is essential as nobody likes an arrogant consultant. Humble people will listen and learn from others. We are always the most clueless in the room about a

[8] Our tasks continued the groundbreaking and successful work by Colleen McCue in using data mining for public safety (see Chap. "Operational Security Analytics: My Path of Discovery").

particular application when we are first brought in to help. But we have to learn fast to translate the problem from the squishy real-world domain to the crisp technical domain where our expertise can apply. Further, humble people will admit mistakes; a proud person might work double hard to fix it before having to reveal an error, which is a recipe for disaster.

The importance of mistakes—recognizing, preventing, assessing, and correcting—cannot be overemphasized. But they must be allowed to happen. Creativity can only flourish when the penalty for error is light. The best attitude would be like that recommended by both Aristotle and Gandhi: "hate the sin, not the sinner." If people are to feel free to try their ideas (which is very motivating), they need to be allowed to fail. For the project to succeed, of course, such failures eventually have to come to an end! But winning ideas are often seeded by lessons from the early failures. It is often said that good judgment comes from experience. But one could argue that experience comes from bad judgment!

It is perhaps dangerous, as a consultant, to be known as an expert on failure, but my most popular talk and writing is on the top ten mistakes in data mining[9] (see Chap. 20, of [5][10]). When you know what to look for, you can recognize a deadly error early and help save the ship. For instance, many of our clients use up too much of their data in learning phase; they need to separate out evaluation data ("out of sample") first of all, so that they do not "look ahead." Done right, the evaluation data will have the capacity to stun them—to be a realistic test—when run through the model. (Because reality will certainly do so when an overprotected model is released!) Or researchers may incorrectly optimize overall accuracy, rather than accounting for the different costs of different types of errors. In "needle in a haystack" problems, the rare needle is often far more important to pull out than the common "hay" is to avoid. For instance, it may take seven clients with good credit to make up for a single one who defaults; those relative costs—which can be a tough business problem to derive—need to drive the solution.

Some failures are not due to mistakes, but to attempting problems which may be impossible. Often when searching for a solution for our toughest challenges, nine of ten paths will not succeed. It can be difficult for some high-achieving people to attempt such work as they have never encountered a problem they could not solve! Some of our challenges are so tough that pushing through them is like a destructive evaluation, where a bridge is loaded until it collapses, revealing its true limit. Again, it is essential to seek out people with the core character to handle such setbacks well and not tire of trying.

[9] Video excerpts on four of the top mistakes can be found on YouTube, starting with http://www.youtube.com/watch?v=Rd60vmoMMRY

[10] The book (http://www.tinyurl.com/bookERI) won the 2009 *PROSE* award for Mathematics (no doubt helped by full color figures and charts). An interview about it is at http://www.youtube.com/watch?v=4B3SritCxSk

9 Take or Give?

The business world is harsh—seemingly a constant struggle to maintain a positive margin of income to outflow. It is natural to gravitate toward an aggressive "survival of the fittest" posture to seize and prolong advantages in order to stay in the game. Under the Darwinian model, even technological edges fade though, without unrelenting effort to stay competitive.

10 But Is that All There Is? Is It Just that "He Who Dies with the Most Toys, Wins?"

I grew up in a Christian home and became a serious believer just before college. Faith in God gives me something bigger than myself to live for, as well as comfort that trials and joys are worked together, by a benevolent Creator, for good. Yet, it has taken me a long time to realize how thoroughly faith should affect my *work*—in fact, that it can be the opposite of Darwinian struggle.

The biblical definition of *love* is to put others before oneself. It is easy to see that this principle is useful for business when applied to *clients*; after all, a great majority of top companies have a strong reputation for client service. It is not that hard to humble oneself to serve those who pay you. And loyal, hardworking employees are not hard to honor either. Strong companies have learned to treat people well, to enhance long-term productivity, if for no other reason. At ERI, we want those who leave to leave happy—having felt listened to, appreciated, and treated fairly while there. A bitter departure can work a lot of harm in this small world of data mining, but a friendly one can be great for both parties. One valued ERI employee left for a dream job on Wall Street and years later became a client; I found myself reporting to my former reportee—an ideal situation. Two other senior colleagues returned after 4 and 8 years away, respectively, which was an encouraging sign that our culture is healthy.

So love as a business principle makes some sense for clients and employees. But Jesus' command went much further, even to "Love your enemies."[11] Meaning you are to put, before yourself, parties who take more than they give, such as clients who are not paying, subcontractors who do not deliver, demanding employees who underappreciate the contributions of others, or competitors who violate contracts

[11] In the Bible, the book by Matthew, Chap. 5, Verses 43–48, Jesus says: *"You have heard that it was said, 'Love your neighbor and hate your enemy.' But I tell you, love your enemies and pray for those who persecute you, that you may be children of your Father in heaven. He causes his sun to rise on the evil and the good, and sends rain on the righteous and the unrighteous. If you love those who love you, what reward will you get? Are not even the tax collectors doing that? And if you greet only your own people, what are you doing more than others? Do not even pagans do that? Be perfect, therefore, as your heavenly Father is perfect."*

and steal business. Clearly, such a standard is impossible to attain! But God promises to supply the ability to do so, if I rely on Him, and not on my own limited ability and will.

Over the years, I have transitioned in my primary role from doing the work, to leading the work, to bringing in the work, to developing the workers. I miss spending deep time devising algorithms, but find it a fascinating challenge to assemble, encourage, and learn alongside a team designed to solve tough challenges. Slowly, I have gotten to where I can rejoice when a great opportunity (elsewhere) comes along for a colleague, rather than feel only the loss to our team. But I have a long way to go in being the kind of leader who leads primarily by serving.

11 Full Circle

Now my thoughts have come full circle, and I know what I will emphasize in the morning. The technical breakthroughs—the "Eureka!" moments—are always my favorite to hear from others, and it will be worthwhile to recount a few. Especially if I can link them to the soft issues having to do with strange "carbon-based life forms" (people). We geeks often want to take our project off to work on it in isolation, and present it only after it is all polished and shiny, to receive our blue ribbon. Well, that default strategy of ours does not work well. The technical risk of a project, that we focus so much effort on taming, appears to be far outweighed by its business risk! In studying ERI's successes and failures over 15 years, it appears that 95% of our projects met their technical goals, but only 70% were implemented and thereby brought value to the client. If we want to see our work used, we have to learn and attend to the people part—the social and political dimension—of how an idea makes it all the way to fruition. And one of the key components is communicating much more frequently with all the stakeholders than would naturally occur to us.[12]

Being a business owner is often too much of an adrenaline ride, but it has enabled me to grow a team of people who work together to serve others through the powerful technology of predictive analytics. It has allowed us to make contributions in a wide variety of fields by understanding a problem in its own domain well enough to abstract it to a common space where we can find faint clues that lead to insight. From investment enhancement to medical discoveries, and product recommendation to national security, the clues in the data can be followed and harnessed for good. I have not become a Tom Swift, but I have experienced hints of Sherlock Holmes, detecting the story underlying subtle clues, or "Q" from the Bond stories—creating cool gizmos with which the good guys can fight lawlessness. This is definitely the calling for me!

[12] The Las Vegas talk went over well; the post-talk interview by SAS is at http://www.youtube.com/watch?v=mVzbEtobb2E

References

1. J.F. Elder, Efficient optimization through response surface modeling: a GROPE algorithm. Dissertation, School of Engineering and Applied Sciences, University of Virginia, 1993
2. J.F. Elder, D. Brown, in *Network Models for Control and Processing* (Chapter 6), ed. by M.D.Fraser. Induction and Polynomial Networks, Intellect, 2000
3. J.F. Elder, D. Pregibon, in *Advances in Knowledge Discovery and Data Mining*, eds. by U.Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. A Statistical Perspective on Knowledge Discovery in Databases, Chapter 4 (American Association for Artificial Intelligence, Palo Alto, California, 1996)
4. H.J. Kushner, A new method of locating the maximum of an arbitrary multipeak curve in the presence of noise. J. Basic Engineer. **86**, 97–106 (1964) (March)
5. R. Nisbet, J. Elder, G. Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Elsevier's Academic, San Diego, CA, 2009)
6. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1988)
7. G. Seni, J.F. Elder, in *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions* (Morgan and Claypool, FL, 2010)

Editor's note. Dr. Elder heads Elder Research, one of the earliest and largest consultancies in data mining and predictive analytics. He is an Adjunct Professor in the Systems Engineering Department at the University of Virginia, where he periodically teaches graduate courses in optimization or data mining. With ERI colleague Andrew Fast and others, he has written a book on *Practical Text Mining*, published by Elsevier in January 2012. He is grateful to be a follower of Christ and the father of five children.

Voyages of Discovery

David J. Hand

1 Data Mining and Me

To set the context, I would like to begin by describing what data mining is—or, at least, my view of what data mining is. A broad high-level definition, given in the opening sentence of the preface of [1], is that it is 'the science of extracting useful information from large data sets or databases'. The opening chapter of the same book has a slightly refined definition: 'Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner.'

Other authors may have slightly different definitions, but they will overlap substantially with the above. The fact that the data sets may be (and nowadays usually are) large is one of the things which has driven the development of data mining as a discipline distinct from (but with considerable overlap with) other data analytic disciplines, such as statistics, machine learning, and pattern recognition. Statistics, for example, developed over the twentieth century, initially in an age when data had to be collected manually, so limiting the sizes of data sets. A data set of a few thousand data points would have been regarded as large in the first half of that century, whereas nowadays data sets consisting of billions of data points are commonplace. Of course, this is not to say that statistics has not evolved to meet the requirements of modern problems, but merely that it had origins different from those of data mining. One particular consequence of the size of many modern data sets is that computational issues such as search, ordering, and the implications of the combinatorial explosion have become pressing: efficient ways to cope with such problems matter when one has billions of data points (and power sets of order 2^(billions)) to consider. This means that, whereas statistics has placed emphasis on modelling and inference, data mining has placed substantially more emphasis on

D.J. Hand (✉)
Department of Mathematics, Imperial College, London SW7 2AZ, UK
e-mail: d.j.hand@imperial.ac.uk

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_7,

algorithms and search. I have explored the relationship between statistics and data mining elsewhere [2].

The second definition above describes data mining as focusing on 'observational' data sets. These are data sets in which the process being studied has not been subject to any experimental manipulation but are simply collected as they come. Statistics, of course, has classically dealt with both observational and experimental data sets, so that has been one difference between the two disciplines. It is of interest to note, however, that nowadays sometimes very large data sets subject to data mining exercises are collected after experimental manipulation. This means that the restriction to 'observational data', in the definition above, is less true now than it was when it was written. Examples of such designed data sets are those collected by large retail organisations, experimenting with loyalty schemes, and data collected directly from the World Wide Web by organisations such as Amazon and Google.

The final part of the second definition above says that the aim is to 'find unsuspected relationships' or to summarise the data in novel ways. It is that which really encapsulates the notion of data mining: one is sifting through large bodies of data seeking something new.

New discoveries can come in various shapes and forms. A useful distinction is between *models* and *patterns*, though other terms are sometimes used for these. A model is a 'global, high-level' summary of a relatively large mass of data: a time-series model describing how an economy changes over time, a multivariate mixture model of a distribution of data points, a regression model relating covariates to a response, for example. A pattern is a 'local' configuration of relatively few (even single) data points, showing some kind of anomaly or departure from the background or expectation: a sudden change in a time series, an outlying data point, a few patients showing an unusual pattern of symptoms, etc. One can think of data mining for models as being analogous to coal mining (large-scale summaries) and of data mining for patterns as being analogous to diamond mining (small-scale anomalies; the term 'nugget' is often applied, in another analogy).

There is another perspective on data mining which can shed some light on why it is becoming so important. Broadly speaking, data analysts construct two kinds of models, which have gone under various names [3, 4]. *Phenomenological* models are models which are based on some kind of theory. A simple such model would be the inverse square law of gravity, for example. Such a model has parameters which need to be estimated, and these will be estimated from data using statistical methods. In contrast, *empirical* models are simply summaries of the data, not based on any underlying theory or postulated mechanism or understanding. An example would be a credit score, which is a statistical summary of data relating characteristics of an individual to their probability of defaulting on a loan. Although models for loan default could be based on psychological theories, in fact, they are actually simply based on descriptive summaries of large amounts of data describing individuals and their behaviour. Data mining might justly be described as a triumph of empirical model building. In general, almost by definition, since one is seeking the unexpected, one will not want to constrain oneself to a pre-existing theory but

will be analysing the data looking for the unexpected: the analysis process will be purely data driven. It will be empirical.

Data mining has been driven from several distinct directions. One is the investigation of large scientific databases—genomic and proteomic databases, astronomical catalogues, particle physics experiments, etc. Another is the commercial driver of large business organisations seeking to extract 'business intelligence' from the wealth of data they have describing customer behaviour. This second application domain led to something of a hiatus in the early development of the discipline. In particular, it meant there was often a failure to appreciate that the aim of an analysis was normally one of *inference* rather than description. For example, one is typically not really interested in describing how previous customers have behaved but in using the data describing that behaviour to predict (infer) how future customers are likely to behave. Recognition of this has broadened data mining from an early overemphasis on algorithms to a more sophisticated understanding of aims.

Given the nature of its subject matter—large data sets—one should expect that data mining will be particularly sensitive to two issues: one is data quality and the other is chance. This expectation turns out to be met, and these are key issues which must be taken into account in any data mining exercise [2]. Large data sets mean large scope for errors—and errors will often manifest as anomalous structures, though seldom ones of great substantive interest. Likewise, large data sets mean great scope for 'interesting' data configurations to occur by chance, but accidental configurations are not of substantive interest. Some sort of allowance or control has to be made for this. These sorts of issues make data mining a discipline of substantial intellectual depth, as well as practical challenge. And the interplay between the various strands of the discipline lends it a particular power, for example, sometimes data analytic methodologies are essentially discovered empirically ('this intuitive idea seems to work') and, only later, put on a solid theoretical base (boosting is an example of this).

Having set the context by outlining modern data mining philosophy, perhaps a slight qualification is appropriate: this is that the above comments are generalisations. No doubt particular exceptions can be found for each and every one of them. Economists, for example, have a rather narrower view of what data mining is (see [5] and the discussion comment in [6]). The overall thing to bear in mind, however, is that data mining captures the excitement of discovery. The days of physical expeditions to far-flung lands may be long gone, but the opportunities for intellectual voyages exploring data are greater than ever.

So how did I discover the scope for making such intellectual voyages?

I was lucky enough to be around in the early days of computers, and it was easy to see that these had the scope to make real some of the possibilities hinted at in the science fiction novels I was fond of reading (artificial intelligence, for example). But I was also interested in fundamental scientific questions—the mysteries of quantum physics and astrophysics, for example—as well as technological questions of how to make things happen. Mathematics seemed to be a requisite background for all of these interests, as well as meaning I could avoid too narrow a specialisation early on, so I took a degree in mathematics at Oxford University.

It became apparent during my studies that chance and probability were fundamental to many of my interests, so I followed my mathematics degree with a master's in statistics at Southampton University. While studying that, I continued my reading in parallel areas of computer science. In particular, I became intrigued by the early work on pattern classification algorithms which showed that systems with *random* choices of parameter values could learn to assign objects to correct classes (typically with some misclassification rate, of course, for real-life problems). This seemed magical: it appeared to emulate the brain, which seemed to have huge numbers of random neuronal connections, but, through repeated exposure to different stimuli and reinforcement of decisions which by chance were correct, it could learn to make decisions which were correct most of the time. Donald Michie's MENACE, an implementation of a system using coloured beads distributed across 300 matchboxes, and which learnt to play a perfect game of noughts and crosses (tic-tac-toe), illustrated how the one could achieve remarkable results using these ideas.

These interests led naturally to me taking a PhD in statistical pattern recognition—in the Electronics Department at Southampton University. This work was characterised by being at the interface of statistics and machine learning, and this has been typical of a great deal of my subsequent research, with data mining fitting this pattern. My PhD research was on supervised classification with incomplete data—methods for assigning objects to classes when the descriptive feature vectors of the objects had missing values.

On completing my PhD, I worked for Logica, a computer consultancy, for a while but found it constraining to work on problems specified by others, when there were so many more interesting problems to work on! So I left Logica to join the London University Institute of Psychiatry, working as the statistician member of a team investigating the effect of aircraft noise on mental health—a topic of particular importance in parts of west London, which is overflown by aircraft from Heathrow.

From there, I moved to a lectureship in statistics in the then Biometrics Unit at the Institute of Psychiatry. This Unit was led by Brian Everitt, a leading authority on cluster analysis. Cluster analysis and descriptive multivariate statistical methods more generally have played an important role in teasing out understanding of psychiatric illness (see, for example, [7]). I was promoted to senior lecturer in the Unit and left in 1988 to take up the chair of statistics at the Open University.

The Open University has been described as the greatest innovation in tertiary education in the UK during the second half of the twentieth century. It is a distance learning institution, which has used multiple media, including the printed word, radio and television broadcasts, and computer media and the Internet. I had the interesting experience of making several television programmes while I was there. I argued for the principle that we should use these programmes to promote the discipline of statistics: apart from students actually studying our courses, there was a casual viewer audience of some hundreds of thousands per programme, so this was a great opportunity to try to communicate the importance and excitement of statistics. I avoided the hackneyed image of a man in front of a blackboard and,

instead, went out on location—to pharmaceutical companies to illustrate how statistics was used in drug development, to blast furnaces to illustrate how statistics was used in steel manufacture, and so on.

In general, the Open University is characterised by extraordinarily high-quality teaching material, and its degree courses are of a standard comparable to leading conventional universities. Its student satisfaction scores are always excellent—probably, partly a reflection of the high degree of motivation of the students: they made an active choice to study with the University, rather than it being what someone automatically did after school.

In 1999, I moved to Imperial College, London. Imperial College is one of the UK's top universities, focusing on science, technology, and medicine. My predecessors in the chair of statistics there were George Barnard, David Cox, and Adrian Smith, all great statisticians (prompting the obvious anxiety that things had gone downhill with my appointment). In the next section, I look at my research interests and the specific kinds of problems which led to me working in data mining and to the book *Principles of Data Mining*, coauthored with Padhraic Smyth and Heikki Mannila [1].

2 Challenges

The Institute of Psychiatry turned out to be a fortuitous place for me to work since statistical pattern recognition tools had an obvious application in psychiatric diagnosis. The aim of diagnosis is to assign patients to appropriate disease classes, which was very much within the classification paradigm of my PhD work. This enabled me to apply the tools I had been studying to a wide variety of real data sets. This was a turning point in my development since it drove home the importance of real problems and real data, both to motivate methodological research and to define the practical issues. As regards the motivation, most of my subsequent work has arisen from people needing solutions to real practical problems, rather than from interesting but merely theoretical issues. Thus, for example, it is all very well developing new algorithms or methods for tackling a problem simply because one recognises that a new approach is possible, but there is little point in doing that unless the existing methods have been found wanting in some way. Too much of the data mining literature is characterised by such new development, followed by a rather unconvincing and limited comparison of the proposed new method with existing methods to try to establish circumstances under which the new method performs well.

When I refer to application to real data sets defining practical problems, I have in mind things such as data quality. Elsewhere, and this is especially true of large data sets, I have suggested that if your data seems to have no distortions, no missing values, etc., then one should be suspicious and ask what data cleaning exercises it has been subjected to—since data cleaning exercises lead to risks such as selection bias. The classic example here is the difficulty of analysing data which have

informatively missing items. This problem, in particular, is something which has been the focus of insufficient attention in the data mining literature, though statisticians have given it considerable thought (even culminating in a Nobel Prize for economics in one case).

In addition to teaching statistics, my role at the Institute of Psychiatry involved providing statistical advice to and collaborating with senior researchers from a variety of disciplines including psychiatry, psychology, sociology, pathology, endocrinology, etc. This was a superb training for a young statistician, as it required me to develop breadth of expertise: I could not simply focus down into an increasingly narrow realm built on my PhD work. It also forced me to develop skills not taught on my MSc or PhD, such as how to interact with clients, how to choose solutions which matched the problem and the client's expertise, as well as how to cope with editors in different disciplines ('XYZ analysis may be the most appropriate method, but this journal has never before published examples of XYZ analysis and our readership will not understand it'). The expertise I was forced to acquire in this way has proven to be extremely useful since (as a consequence?) I have continued to undertake a large amount of consultancy work throughout my career, with a tremendously wide range of clients, from pharmaceutical companies, to banks, to trades unions, charities, and governments. This has included legal expert witness work, which has its own challenges.

It was my experience at the Institute of Psychiatry which motivated my interest in statistical expert systems: software tools which acted as an interface between someone needing to analyse data and the statistical packages or languages which did the calculations.

My advisory role at the Institute of Psychiatry also led to my interest in two areas which have continued throughout my career: the issue of formulating the right question and the area of measurement theory. Scientific research involves many stages, but key amongst these are the formulation of the scientific question, the formulation of the statistical question, and the mapping between the two. To a large extent, that mapping is what a statistical expert system seeks to facilitate. But it is all too easy to construct a statistical question which does not properly reflect the scientific one. Indeed, often the attempt to formulate the statistical question leads one to recognise that the scientific one is poorly formulated. A simple example of this lies in the choice of mean or median when summarising the 'average value' of a set of data. These two statistics correspond to different questions asked of the data: it is not the case (for example) that the shape of the distribution should determine the choice—and an inappropriate choice could mean an incorrect answer. Illustrations of such issues are given in [8–10].

Although I have here referred to *statistics* when outlining the issue of question formulation, it applies just as strongly to data mining—indeed, perhaps even more so, given that data mining is the search for the unexpected.

In some sense, measurement theory is a key aspect of formulating the right question. When I mine a set of data to try to identify the key influences on crop harvest, should I use the weight of corn or the log weight (or, indeed, some other transformation)? This is a question of measurement scale, and different choices can

lead to different answers (for a particularly graphic illustration, see Example 3 in [8]). But measurement theory has far greater depth: after all, it is describing the mapping from the external reality (if you believe such a thing exists) to the numerical representation, and these two things are not the same. Wrestling with such matters, noting that the mathematical physics I studied in my first degree and the social, behavioural, and medical science I worked in at the Institute of Psychiatry all used the word 'measurement', but used it to mean very different things, drove me to a detailed study of such matters, leading to my Royal Statistical Society read paper [4] and my book [11]. The book, in particular, describes an attempt to unify the different conceptions of measurement in terms of position on a scale spanning the two attributes of representational and pragmatic measurement. This is not the place to go into details, but I believe that this represents a substantial clarification of these deep issues.

I have described supervised classification as a 'paradigmatic statistical problem' (p. xi of [12]), and one might replace 'statistical' by 'data mining' (indeed, in the book I went on to say that one might equally regard supervised classification as a part of computer science). By this, I meant that it 'includes in its own microcosm . . . all of the major issues of [data mining]: problem formulation, model building, estimation, uncertainty, prediction, interpretation, etc.'. And it has an extraordinary breadth of potential applications. I have already referred to how it influenced my work at the Institute of Psychiatry. But in the early 1980s, I also recognised that the ideas could be applied in the relatively new area of credit scoring, to guide decisions in the retail banking sector.

The retail banking sector is concerned with personal mortgages, bank loans, credit cards, auto finance, etc. It is of particular interest to data miners since it collects vast amounts of data relating to the behaviour of individuals. Indeed, it has been described as not so much a banking industry but really as a data industry. A generic question in the industry is 'is this applicant likely to default on a loan?' This is a natural classification problem. In the early 1980s, I began to collaborate with banks, developing models to answer this question, as well as (what was later to become a particular interest) methods to evaluate such models. This work, expanded to answer a tremendously wide variety of related questions used to guide the operations of credit and banking systems, has been an important focus of my research over the past 20 odd years. It is, of course, very much driven by real practical problems, rather than arid theory: the organisations with which I collaborate have real questions, which require real answers, and the real risks if those answers are not accurate are obvious.

An early piece of work in this area, conducted with my then PhD student William Henley, was to formalise the notion of 'reject inference'. This is the term used in the industry to describe attempts to cope with the selection bias arising from the fact that data are only collected on customers who were previously given the financial product in question. Analysis of only these customers means that one's model is inevitably biased as a model of the entire population of potential applicants for the product. This can lead to poor decisions. The key paper on this topic which

William and I wrote [13] attracted a great deal of interest—unfortunately, most of this interest came from the industry, so that it is not reflected in high citation counts!

Although my interest in credit scoring was originally motivated by the potential to apply the tools of supervised classification, I soon recognised that there was the potential for the application of a much wider range of data mining tools, as well as the need to develop novel tools to answer specific problems. I cannot go into all these areas here, but one, in particular, is worth mentioning, as it is a nice illustration of the importance of data mining. This is the area of fraud detection.

Fraud detection in retail banking is characterised by the need to examine a huge number of transactions (possibly billions) from a large number of distinct customers (possibly millions) looking for anomalous behaviour, *in real time*. A system which perfectly identified all fraudulent transactions and misclassified no legitimate transactions but could only do this 3 months after the event would be useless. Fraud detection in this area is an illustration of mining streaming data, a topic which is of increasing importance as automatic electronic data capture means that streaming data is becoming more and more prevalent.

One novel technique for fraud detection which my group developed was what we called *peer group analysis*. This involves identifying customers whose past behaviour has been similar to a target individual and detecting if and when the target's behaviour deviated from the peer group's. This has the property, in particular, that sudden changes in behaviour are, of themselves, not necessarily indicators which should arouse suspicion: a sudden change of purchasing pattern just before Christmas may well be the norm.

In fraud detection (and in all other applications of supervised classification involving two classes), there are two kinds of errors: a class 0 point can be misclassified as a class 1 point or vice versa. In order to choose between classification rules, as well as to optimise parameters, to choose predictive variables, and so on, these two sources of error must be reduced to a single numerical performance value which can then be optimised. Unfortunately, there is no single best way of combining these two types of errors. Simply counting up the total number of misclassifications and expressing it as a proportion of the total number of objects classified leads to the familiar error rate or misclassification rate, but this is equivalent to regarding the two kinds of misclassifications as equally serious. While this is sometimes true, it is more often not—typically, one type of misclassification is more serious than the reverse (e.g. misclassifying, as healthy, someone who has a disease which is potentially fatal but easily treatable is more serious than the reverse; misclassifying a fraudulent credit card transaction as legitimate is more serious than the reverse; etc.). Thus, either one must choose relative weights/severities for the two kinds of misclassification, and then combine them, weighted by the proportion of each kind, or one must fix the number of one type of error and use the number of the other type as the performance criterion, or one must use some other way of reducing the two values to a single value. (I am deliberately glossing over issues such as the possible costs of correct classification, as well as other issues, for simplicity of exposition.)

It is obvious that poor choice of performance measure can lead to poor performance: choosing a classifier on the grounds that it led to minimum misclassification rate would probably be useless in a credit card fraud detection system since, with around 0.1% of transactions fraudulent, an apparently excellent system would be achieved simply by classifying all transactions as legitimate. Thus, the choice of performance measure is critical. Because it is so important, this has been a continued focus of my research over my career. Most recently, in work which I believe could have important consequences, I showed that the very widely used 'area under the ROC curve' (AUC; or, equivalently, the Gini coefficient) is fundamentally flawed if one believes that the two kinds of misclassification can be balanced against each other (albeit with unknown relative severities) [14, 15]. The AUC is equivalent to asserting that one believes that *different* relative misclassification costs apply for different classifiers. Thus, one might say that if I use a neural network, then misclassifying a cancer sufferer as healthy is ten times as serious as the reverse, but if I use a tree classifier, it is only half as serious as the reverse. This is clearly nonsensical: the relative cost of misclassification cannot depend on how I arrive at those misclassifications. This result is surprising because the AUC is so widely used—in data mining, machine learning, pattern recognition, medicine, banking, and other areas.

As will be apparent, aspects of supervised classification have interested me ever since my PhD. The area is a large one, not least because of the multiple intellectual disciplines which make use of the tools. One consequence is that there is a considerable amount of research seeking to develop improved methods. In an article which attracted considerable attention, I explored the suggestion that much of this research made misleading claims and that, in fact, the progress was less than met the eye (I called my paper *Classifier Technology and the Illusion of Progress*, and a number of discussants published their responses). In some sense, the points I made are analogous to the issue of publication bias, though I addressed more specific technical issues of classification.

3 Limitations

A great deal of my learning has come from applying statistical and data mining tools to real problems. It is inevitable that one begins with some understanding of the technical aspects of the tools, learnt from lectures, books, and papers, and only, as one acquires more experience in applying those tools to real problems does one begin properly to appreciate their strengths and properties—and their limitations. This is one reason why people working in data analytic disciplines tend to gain strength as they get older (up to a point, presumably, though I am hoping it is far off!). This makes such disciplines very different from the popular caricature of mathematics, which would lead us to believe that mathematicians' best work is done when they are young—to be an effective data miner or statistician, one has to have some understanding of the domain of application and the peculiarities of its

data (data quality and distortion, missing values and how to cope with them, and so on), and this can come only with experience. (Incidentally, I believe that the caricature of mathematics and mathematicians is also wrong—and there are plenty of examples of mathematicians doing their best work in their sixties.)

The limitations of what can be achieved were brought home to me during a study to develop a screening questionnaire to identify elderly women who were particularly likely to develop the disease osteoporosis, in which the bones lose their strength and are susceptible to fractures. I believe that, up to that point, I had naively held the view that it was simply a question of choosing the right variables and transforming and combining them in a clever enough way to push the misclassification rate (if that was your choice of performance measure) down to whatever level you deemed acceptable. But, of course, for any given population of variables, there is a limit to how well separated the classes can be, and this limit may not be at a level for the classification rule or diagnostic instrument to be useful.

In a similar vein, early in my career, I held the belief that classification performance could be improved essentially without limit (up to the Bayes error, anyway) by developing and using cleverer and cleverer tools. This seemed to be illustrated by the general thrust of the discipline, developing linear discriminant analysis and the perceptron, quadratic discriminant analysis, logistic regression, classification trees, neural networks, and support vector machines, and a thousand variants of these and other methods. And, strangely enough, almost all of the papers presenting a new method showed that that method was superior to the earlier methods with which it was compared. Surely, a sign of the march of science, ever upwards, better and better. Or a sign of publication bias, data set selection, test problem selection, and many other things. It took me a while to appreciate these effects, which I described in my paper, mentioned above, on progress in classifier technology [16]. It also took me a while to recognise, and this is something which only came gradually as I applied the tools in more and more distinct application areas, that the non-statistical aspects of the problems might be far more important than minor issues of narrowly defined performance. In many areas, for example, interpretability of the classification rule is important (perhaps for legal reasons, perhaps for other reasons) and can outweigh quite substantial performance differences.

Of course, this is not to say that technical advances do not lead to practical advances. The advent of random forests, a radically new and highly effective tool, took me by surprise.

One important lesson researchers must learn is to cope with setbacks. Around 1990, a pharmaceutical company sought my advice on synergy in drug interactions. ‘Synergy’ is an interaction in which two or more drugs ‘work together’ to be more effective than either alone or than one might expect from the simple combination. The question was as follows: how does one define ‘what one might expect from the simple combination?’ A major review had been published in 1989, in which the author, Berenbaum, claimed to demonstrate that only the so-called isobole method was generally valid and free of assumptions. However, I proved that Berenbaum was mistaken and that the method he described made an implicit assumption which destroyed its general validity. I then went on to explore the issue in depth and to

develop a general definition which made no assumptions. Unfortunately, when I contacted the editor of the journal in which Berenbaum's review had appeared, pointing out the mistake and submitting a paper describing it, I was told that the journal only published reviews (and not corrections of errors!). I therefore tried to contact Berenbaum directly, only to discover that he had died after writing the review. Eventually, I published the paper, which appeared as [17], but unfortunately, it seems to have attracted no attention whatsoever—likewise, a subsequent paper describing the valid method [18].

This is not an isolated story. My more recent work pointing out that the AUC has a fundamental flaw has likewise met a hostile reception from referees (the idea is subtle and also points out problems with long-established and popular methods, so it is perhaps understandable). In the case of this work, I am indebted to editors who recognised the importance of the work, despite the objections of the referees, and overrode their recommendations (though I did have to revise the papers substantially!).

4 Current Research Issues and Challenges

What I see as important research issues and challenges are, naturally, very much influenced by my own research interests. Thus, I tend to assume that the 'housekeeping' issues of handling large data sets—search, sorting, ranking, etc.—will be solved by powerful computers—and by very clever people with interests in those areas. For me, more interest lies in issues such concerned with extracting reliable knowledge from observed data configurations. Thus, I am interested in inference, where it is clear that multiplicity, which is currently attracting considerable interest—with tools such as empirical Bayes methods and false discovery rate—will be an important research focus for the near future. It is interesting that most of the work on those particular classes of tools appears to be being conducted by the statistical community rather than the data mining community. Perhaps, data miners have not yet fully appreciated that inference, rather than mere description, is at the heart of many data mining problems.

More generally, new types of data will attract research interest. We have seen this in recent years with new measurement technologies in biomedicine, leading to data mining challenges in genomics, proteomics, and indeed bioinformatics in general. Biomedicine is likely to remain a focus of interest—and of research funds—and continued progress in automatic data capture means that new data mining problems will certainly arise.

One such, which also manifests itself in other areas, arises with streaming data. These are (typically multivariate) data which just keep on coming (like a stream of water from a hose), requiring online, one-pass, fast analysis, unlike classic problems where data is fixed and can be re-examined at will.

In general, data quality will become more and more important. I have a particular interest in the pattern detection aspects of data mining—detecting unusual

features or anomalies in large data sets—and have applied the ideas to problems in fraud detection, cheating in examinations, astronomy, patient safety issues, adverse drug reactions, earthquake detection, national security, and many other areas. But such problems are particularly vulnerable to poor data quality, which often manifests itself as an anomaly in the data. I hope that the attention being focused on data mining will lead to an emphasis on improved data quality in the future.

5 Research Tools and Techniques

As I have described, I have particular methodological interests in supervised classification and in anomaly detection. Supervised classification has been the focus of an enormous amount of research, stretching back to the early part of the twentieth century, but there is no reason to expect this to stop. Indeed, new application areas, new kinds of data, and new kinds of questions mean that we might expect new methods to (need to) be developed, to cope with these challenges. An example of this is the increasing focus on adaptive supervised classification methods to cope with non-stationarity and population drift in streaming data.

Rather similar comments apply to anomaly detection—with the difference that the large data sets which are becoming commonplace open up new opportunities for interesting discoveries which would have been very difficult with smaller data sets—and so many scientific discoveries have arisen by spotting something which appears at odds with the current thinking of the time [19]. There are challenging methodological questions surrounding anomaly detection, and I think (or at least hope) that these will be the focus of concerted research effort in coming years.

There is, however, a general point which might be made about methodological research, at least that which often appears in the major conference outlets. This is that much of the work describes new methodology and, perhaps, compares its performance with existing alternatives on one or a handful of data sets. If the data sets are familiar ones which are regularly used for such comparisons, this has the merit of allowing broader conclusions about the relative performance of different methods. On the other hand, it also has the potential demerit of overanalysis and overfitting to the particular collection of data sets. In the supervised classification context, for example, the huge number of methods which have been tested on the UCI data repository collection might lead one to fear overfitting. On the other hand, if the data sets are new ones, just used in a single study, one has to ask to what extent any conclusions about relative merits will generalise to other studies and data sets. Such issues are discussed in [16]. There is no ideal solution to this: it is all part of science's gradual accretion of evidence.

6 Making an Impact

The way to make an impact in science is to identify important problems and solve them. (Nothing to it, really!) Important problems are ones that others care about—there is no point in solving problems no one cares about. And there is likewise no point in finding important problems you cannot solve. However, perhaps it goes without saying that if others care about problems, then one might expect many researchers to be focussing on them. The amount of effort being expended on bioinformatics problems is an illustration of this. Likewise, apart from the trivial, it is difficult to determine beforehand whether one will be able to solve a problem.

Data mining has the advantage over other areas, that new large data sets are accumulating all the time and that new domains are being subjected to rigorous quantitative investigation. This means that the applied data miner has a vast wealth of potential application areas in which to work, involving problems that the 'owners' of those problems care about. This is true both of the commercial sector (e.g. retail enterprises, the Internet, etc.) and the scientific sector (e.g. detecting scientific fraud, particle physics and astronomy data, bioinformatics, etc.). The opportunities here tie in with the scope for consultancy work—and, as I have already mentioned, many of the interesting theoretical problems I have studied arose as spin-offs from consultancy work. The clients had particular data sets and particular questions which needed addressing, and one made an impact for them by solving those problems, but then they often led on to deeper theoretical or methodological issues which would, in turn, have a larger impact.

7 The Future

Some years ago, there was a concern that interest in data mining would peak and decline, as it became apparent that the inflated claims for what it could achieve would fail to be met. In fact, this has not happened.

I think a primary reason for this is the continued increase in the rate of growth of large data sets. Whether or not data mining lived up to exaggerated claims, the fact is that such a technology is becoming more and more important as we are faced with more and more data. A commercial operation simply cannot do without the potential business intelligence which can be extracted from the large databases describing customer behaviour—at least, not if it is to remain competitive. Likewise, data mining techniques are a necessary requirement for sifting through the large data sets which science and medicine are increasingly generating. And this applies to the social sciences and government, just as much as the natural sciences.

Given that progress in computer technology is continuing and given the growth of new media (images, audio and video recording, social networking and

other telecoms phenomena, etc.), one can confidently predict that the need for data mining, and continued progress in data mining, will continue for the foreseeable future.

There are issues which have to be addressed. One important one in the social, governmental, and commercial contexts is that of privacy and confidentiality [20]. The debate surrounding the US 'Total Information Awareness Project' is an example.

While current debate, arising from Internet phenomena such as Facebook, has suggested that notions of 'privacy' may be changing, I am not convinced. I think that such debates have only just begun—and it is clear that they are closely tied in with the technology of data mining, and what that technology enables, when applied to data of that sort and in that sort of context. I also think that legislation is not keeping up: the Internet and Internet data mining technology develop incredibly rapidly, far more rapidly than our ponderous judicial processes can update themselves. This could be a serious and uncomfortable issue in the future.

I think that the social applications of data mining technology will be amongst the most exciting in the longer term future, especially in the context of the Internet. Applications in the natural sciences may lead to major discoveries and advances, but it is applications in the social and behavioural spheres which will change the way we live. Such applications are growing up around us, often unnoticed, from contactless swipe cards monitoring our whereabouts, through mobile phone access to the Internet and banking services, and countless others. But it is important to recognise that data mining technology, like any other advanced technology (think nuclear technology and biotechnology), is amoral: it can be used for good or bad. Data mining technology could lead to Orwell's Big Brother monitoring, but it could lead to a society in which needs are better catered for. It is up to us to see that it is used in moral proper ways.

I am sure that some technical challenges will continue to be important for the long term. Examples are those relating to data quality, issues of selection bias, increasingly the challenges of streaming data, and also those of network analysis, both social and otherwise.

8 Summary

My journey through data mining has been, and indeed continues to be, the most exciting of journeys. As I said above, the days of physical expeditions to far-flung lands may be long gone, but the opportunities for intellectual voyages exploring data are greater than ever. And data mining and data mining software tools are the vehicles through which we can make those intellectual voyages. Data mining might justifiably be considered a most enviable profession, since it is in the nature of the beast that data miners are in at the kill—they are there when the discoveries are actually made. Indeed, they are likely to be the ones actually making the discoveries. What could be more exciting than that?

References

1. D.J. Hand, H. Mannila, P. Smyth, *Principles of Data Mining* (MIT, Cambridge, MA, 2001)
2. D.J. Hand, G. Blunt, M.G. Kelly, N.M. Adams, Data mining for fun and profit. Stat. Sci. **15**, 111–131 (2000)
3. D.R. Cox, Role of models in statistical analysis. Stat. Sci. **5**, 169–174 (1990)
4. D.J. Hand, Statistics and the theory of measurement (with discussion). J. Roy. Stat. Soc. A **159**, 445–492 (1996)
5. K.D. Hoover, S.J. Perez, Data mining reconsidered: encompassing and the general-to-specific approach to specification search. Econometrics J. **2**, 167–191 (1999)
6. D.J. Hand, Discussion contribution on data mining reconsidered: encompassing and the general-to-specific approach to specification search. Econometrics J. **2**, 241–243 (1999)
7. G. Dunn, P.C. Sham, D.J. Hand, Statistics and the nature of depression. J. Roy. Stat. Soc. A **156**, 63–87 (1993)
8. D.J. Hand, Deconstructing statistical questions (with discussion). J. Roy. Stat. Soc. A **157**, 317–356 (1994)
9. J. HandD, Scientific and Statistical Hypotheses: Bridging the Gap, in *Understanding Social Research: Perspectives on Methodology and Practice*, ed. by G. McKenzie, J. Powell, R. Usher (Falmer, London, 1997), pp. 124–136
10. M.G. Kelly, D.J. Hand, N.M. Adams, Defining the Goals to Optimise Data Mining Performance, in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, ed. by R. Agrawal, P. Stolorz, G. Piatetsky-Shapiro (AAAI, Menlo Park, 1998), pp. 234–238
11. D.J. Hand, *Measurement Theory and Practice: The World Through Quantification* (Edward Arnold, London, 2004)
12. D.J. Hand, *Construction and Assessment of Classification Rules* (Wiley, Chichester, 1997)
13. D.J. Hand, W.E. Henley, Can reject inference ever work? IMA J. Maths Appl. Bus. Indust. **5**, 45–55 (1993)
14. D.J. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**, 103–123 (2009)
15. D.J. Hand, Evaluating diagnostic tests: the area under the roc curve and the balance of errors. Stat. Med. **29**, 1502–1510 (2010)
16. D.J. Hand, Classifier technology and the illusion of progress (with discussion). Stat. Sci. **21**, 1–34 (2006)
17. D.J. Hand, What is synergy? Revisited J. Pharma. Med. **3**, 97–100 (1993)
18. D.J. Hand, Synergy in drug combinations, in *Data Analysis*, ed. by W. Gaul, O. Opitz, M. Schader (Springer, Berlin, 2000), pp. 471–475
19. D.J. Hand, R.J. Bolton, Pattern discovery and detection: a unified statistical methodology. J. Appl. Stat. **31**, 885–924 (2004)
20. D.J. Hand, The Information Economy: Personal Privacy and Public Protection, in *Statistics, Science, and Public Policy XI: Government, Science, and Politics*, ed. by A.M. Herzberg (Queen's University, Waterloo, 2007), pp. 37–42

A Field by Any Other Name

Cheryl G. Howard

1 What Makes a Data Miner?

I have often been asked, "What do you have to study to be a data miner?" or "What skills should we look for in hiring data miners?" I think there are three success factors that have little to do with one's specific academic training: an analytic mindset, attention to detail, and a rigorous appreciation of and adherence to the scientific method. It is also important that data mining practitioners know their limitations and understand that the success of any project depends on working closely with the people who will be using the solution. Successful data miners do not build solutions in isolation; they take the time to understand the problem and the data before even beginning to consider a technological approach.

These leads to another often-asked question: "What makes a successful data mining project?" A well-defined problem along with accessible and well-understood data is more important than the particular tools, methodologies, or algorithms used in developing a solution.

I have spent nearly 30 years in this growing and changing field that has been called many things: simulation and heuristics, artificial intelligence, machine learning, data mining, and predictive analytics. Whatever it is called, it has been fascinating to observe the academic and technological advances as well as to see commercial and public sector institutions embrace data mining as a way to solve real-world challenges.

C.G. Howard (✉)
IBM Corporation, Washington, DC, USA
e-mail: cghoward@us.ibm.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_8,

2 The Long and Winding Road Through Academia

My academic endeavors began with a rather interdisciplinary program at the University of Rochester. I went to Rochester with a strong interest in the sciences, but like most undergraduates I did not really have a clue where that would take me. Within the Psychology Department at Rochester was the Center for Visual Science. I had long been fascinated by the human visual system. When I visited the department at the end of my freshman year to ask for permission to take one of their graduate courses, the center's director, Walt Makous, asked me what I was doing that summer. I told him that I was going to Lake George, NY, to waitress in a pizzeria—the same thing I had done the previous summer. He suggested that perhaps I change my plans and work as an assistant in his laboratories. It took me all of 15 minutes to decide that was a pretty good idea—my parents enthusiastically agreed. I ended up taking many classes in the Center for Visual Science and worked there throughout my undergraduate years. I focused on physiological optics, involving the study of human vision and brain processes, but also incorporated optical engineering and the small amount of computer science that was required in engineering courses in the early 1980s. From Walt Makous, I learned the importance of the scientific method and the applicability of signal detectability theory. His bible was a classic experimental psychology text that I think all scientists should read; it gave beautifully clear explanations on how to design rigorous experiments and measure their outcomes [1]. I did not realize at the time that optimizing systems to balance the cost of two kinds of errors would become a cornerstone of my life's work. Detecting minute effects in human perceptual systems was excellent training in the need to baseline systems and to carefully control every possible variable so that observed effects could be reliably quantified.

I considered going directly to graduate school in this highly specialized field, but instead found my first job at the US Army Night Vision Laboratory at Fort Belvoir, Virginia. Although hired to work for the optics team, I found myself reassigned on the first day to a team vaguely named systems integration. It turned out to be rewarding and fascinating work; we designed and tested electro-optical devices that allowed soldiers to accomplish their mission in complete darkness.

Graduate school was still in my near-term plan, so I started looking at programs in the Washington, DC, area. I knew that I wanted to stay in the sciences, but did not really know exactly what I wanted to study, nor did I have a good sense of what the local schools had to offer. After some cursory research—unaided as we were in those days by the internet—I settled on the Computer Science program at the School of Engineering and Applied Science at the George Washington University. In my first course in that program—Introduction to Computer Systems—the instructor, Prof. Peter Bock, told the class of about 40 students, "Statistically, three of you will complete a master's degree and one of you will get a Ph.D." I figured that he was trying to scare us, and indeed he was, but his projection was also fairly accurate.

My association with Prof. Bock turned out to be long and fruitful. He directed a concentration within the Computer Science that at the time was called "simulation and heuristics." I confess to having very little idea of what that meant when I dove headlong into the program. One of the group projects in that very first course involved building a system that *learned* to play blackjack. It learned by tracking the decision it made in any given state and then changing the weighting of future decisions based on whether it won or lost. There I was, doing machine learning without even knowing that such a field existed. The program, like the field, has gone through several name changes and the concentration is now called "Machine Intelligence and Cognition."

I continued working at the Night Vision Lab where we were developing the second generation of forward looking infrared (FLIR) devices. Although the optical, electronic, and mechanical challenges we faced were fascinating, the thing that interested me most were the fledgling automatic target recognition (ATR) systems that attempted to recognize objects in the digital images generated by the FLIR systems. Target recognition was required to discriminate between friendly and enemy vehicles; this was no small challenge when combined with the myriad technical issues involved in seeing in the dark. Working on these real-world image recognition problems gave context to the work I was doing in graduate school.

Upon completing my master's degree in 1989, I continued directly into the doctoral program. By this time, the concentration had been renamed from simulation and heuristics to artificial intelligence and machine learning. I continued to work closely with Peter Bock; he had developed a theory of machine learning known as collective learning automata (CLA) theory [2]. It was a special case of histogram learning or memory-based reasoning. This class of methods were exquisitely simple, but posed some special challenges due to the expense of storage at that time. We spent a great deal of effort optimizing dynamic range and designing sophisticated hashing systems; had we access to the inexpensive memory available today, we could have skipped some of that effort, although it was certainly good training from a theoretical computer science standpoint.

In 1989, I left the Night Vision Lab to work for a defense contractor; I did so in the interest of doing more hands-on development and less project management; the work was not nearly as interesting or relevant as what I had been doing at Night Vision. Around the same time, Prof. Bock came upon the opportunity to apply CLA theory to real-world systems in the form of the adaptive learning for image and signal analysis (ALISA) system [3]. The work would be sponsored and performed at the newly established Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW), which translates loosely to Research Institute for Applied Knowledge Processing, in Ulm, Germany. An initial group of senior-level graduate students went over to set up the project. What was initially a 6-month sponsored effort was so successful that the project continued on for several more years. In 1991, the question, "What are you doing this summer?" once again changed the course of my life. Peter Bock asked if I would like to go over and join the team in Germany for a few months. I replied that if I could go for a year and if my electrical engineer husband could find a job over there, we would go.

Everything came together and I ended up working at the FAW for over 4 years. It was a fantastic experience working with top-notch researchers from all over Germany and around the world. Our project was sponsored by the German industrial firm Robert Bosch GmbH and involved applying CLA theory to a wide range of interesting industrial problems. These included discriminating between scratches on ball bearings that would cause performance issues and those that could be ignored. Another task converted acoustical signals to images using fast-fourier transforms (FFTs) and then allowed the machine to determine whether the motors producing the acoustic signals were abnormal. Both of these tasks were previously performed by human experts; the goal, as in many data mining problems, was to automatically dispatch the simple cases so that the experts could focus their knowledge on the more challenging and interesting cases.

At this time, the machine learning and AI communities were very excited about decision trees and even more so about neural networks. We often found ourselves having to defend our simple and elegant learning paradigm; it just did not have the coolness factor that artificial neural networks did. I remember talking with fellow attendees at conferences who seemed genuinely interested in the broad applicability and success of our work, but then apologized that they would not be able to attend my talk because it was running concurrently with a neural network session. We worked hard to convince the doubters that feature selection was a key success factor in learning systems; pretty much any learning paradigm will give good results if the problem is well defined and the input features well chosen.

Another key element that was often lacking in those headstrong days of rapidly emerging new paradigms and algorithms was the need to rigorously baseline the existing performance of a system before claiming that a given technology could produce great results. I recall attending a talk at the aforementioned conference where the speaker outlined two key assumptions for his currency trading application: One must start with an unlimited number of Swiss Francs and be exempt from all exchange fees. I daresay that had I an unlimited number of Swiss Francs I would have been somewhere other than at a machine learning workshop.

My doctoral work focused on the ability of a computer to learn geometric concepts in images using the CLA paradigm [4]. The fascinating thing about this work for me was the ability to build up a hierarchical learning system. Simple units detected features; slightly more cognitively complex units assembled these features and detected higher-level geometric constructs which could then be assembled into shape concepts. This approach gave our system a robust and flexible character that allowed us to successfully apply the paradigm to a broad range of problems.

3 Into the Real World

After completing my doctorate [5], I joined the research labs of the Thomson Corporation, one of the world's largest information companies. I joined Thomson expecting to apply my image analysis experience to problems of content

management in the publishing domain, but soon found myself working with highly structured data in the financial domain. I quickly learned that structured data is not necessarily any cleaner than the unstructured image or acoustic data that I had worked with previously. At Thomson, I learned a lot about data cleansing and also about the need to truly understand the underlying business objectives in order to effectively apply analytics to real-world, real-time problems. This was in the late-1990s when day trading was everybody's favorite pastime and there was an abundance of financial activity and data. We had access to a range of rapidly evolving analytic tools including S-Plus and Matlab, and there was great interest building in quickly deployable Web-based applications using java and other emerging Web-friendly approaches. Fully functioning data mining workbenches were just appearing on the horizon; we endeavored to learn as much as we could about those tools.

Although Thomson restructured and decided they no longer needed a centralized corporate research organization, I knew that applied data mining was where I wanted to focus my efforts. I initially struck out on my own as a data mining consultant, but soon joined forces with Elder Research, a small but highly regarded firm that focused exclusively on data mining. I recently moved from Elder Research to IBM, where I continue to work directly with clients to help them solve challenging data mining problems using the best available tools and methodologies.

3.1 Tax Fraud Detection

Prior to and after joining Elder Research, I had the opportunity to work on a near-ideal data mining project—detecting refund fraud for the Internal Revenue Service. The project was ideal for many reasons: We had a knowledgeable and supportive client, the previously used approach provided a performance baseline against which we could measure our progress, we were given latitude in terms of the tools and methodologies that we could use to solve the problem, and we had access to a large and surprisingly clean data set of both fraudulent and general population data.

In our client, we had an asset that I still consider the holy grail of any data mining effort; in a single person, we had someone who both knew and understood the contents and structure of the data warehouse and also had a deep knowledge of the mission that we were trying to accomplish with that data. This is not something that is easy to find, and we tried never to take that for granted. Solving a data mining challenge very often involves process-driven constraints that can limit the utility of certain data or approaches; having a client who could help us understand those constraints early in the data mining process was immensely helpful. The senior management of this project also had an incredible grasp of analytics and clearly understood the performance metrics that we presented as we moved through the program. Again, this is not something that is present in every data mining project.

Having a firm grasp on the performance of an existing system before attempting to improve it might seem like a given, but my experience has proven otherwise. However, in this project, the task of determining which tax returns should be sent to experts for further processing was part of a well-established workflow management system. The problem constraints dictated that we detect as much fraud as possible while not slowing down the processing of nonfraudulent returns. The existing linear regression-based approach was causing experts to have to review a large number of returns in order to find a single fraudulent one. Because these experts represented a limited resource, the client knew that fraudulent cases were being missed due both to not being detected by the model and by having experts not be able to focus appropriately on the truly fraudulent cases. Although the false alarm rate was clearly too high, at least we had a very good handle on how high it was.

As other authors in this volume have noted, the specific tool or machine learning paradigm used is not the most important factor in the success of a data mining effort. Having access to and understanding of reliable, relatively clean (or cleanable) data is far more important. For this project, we evaluated the available data mining workbenches, which had by now become commonly available; we chose SPSS Clementine (now known as IBM SPSS Modeler.) This tool allowed us to easily explore and compare supervised learning algorithms and to quickly test which were most effective in improving detection performance and reducing false alarm rates. We found, not surprisingly, that an ensemble model combining neural networks and boosted decision trees was most effective. It is now commonly accepted that ensemble models are often more accurate than any single algorithm [6], but ensemble approaches were less widely employed at the time that this project was beginning. Having the flexibility to recommend a toolset and to evaluate various algorithms and their ensembles contributed to the technical success of the project.

And then there is the data. The availability of good data has been known to delay or even derail many well-intentioned data mining efforts. Supervised learning is an extremely powerful technique, but it requires labeled training data—in our case, fraudulent and nonfraudulent. In reality, the "nonfraudulent" data set may contain some undetected fraud cases, but supervised learning techniques tend to be robust to even a moderate amount of noise in the training data. In the case of our tax fraud project, we had significant numbers of cases of expertly verified fraud of the type we were trying to detect. We also had a vast amount of nonfraudulent data.

All of this is not to say that there were no technical challenges in this effort. One of these was to determine how to effectively balance the training data so that the system would not be biased against identifying any cases as fraudulent. The overall rate of occurrence of tax fraud is estimated to be quite low, although another ever-present challenge in the world of fraud detection is that we can never know for certain what we might be missing. Thus, assuming that the rate of fraud is $<1\%$, we could build a system that was over 99% correct by simply declaring all tax returns to be nonfraudulent. This clearly would not meet our objectives for the model since we would be detecting 0% of the actual fraud. By training the model

with a larger ratio of fraudulent data than existed in the general population, we biased the model back in favor of effective detection.

The end result was a deployed system that reduced the false alarm rate by two orders of magnitude while significantly increasing the detection of fraudulent returns.

3.2 Insider Threat Detection

As an offshoot of my work in traditional fraud detection, I have had the opportunity to work with several enterprises on the problem of detecting improper behavior among trusted employees. Whether in commercial or public sector environments, the greatest threat to the well-being of an enterprise comes from within.

From a data mining perspective, such tasks have their own set of challenges and advantages. An advantage is that an enterprise often has a vast amount of data on the characteristics and behavior of trusted employees. Accepting a position of trust, especially in the public sector, generally gives the employer license to monitor that employee's behavior. This means that data that could not be used in customer-based data mining due to privacy concerns becomes part of the mine-able data.

A challenge of this kind of data mining is that false alarms can be extremely costly, both in terms of resources required to investigate a possible breach of integrity and the loss of good faith among employer and employee if one is falsely accused.

My experience in this area has also shown me that carefully mining the employee-based transactions of an enterprise can lead to valuable insights into the integrity of the data and can also identify management and training issues that can be addressed with the positive outcome of improving the overall functioning of the organization. What looks like anomalous behavior of one or a few employees might actually indicate that an aspect of the business process is not working as intended, even where no malicious intent exists.

4 Looking Ahead

The field of data mining, now often referred to as predictive analytics, has grown and changed significantly over the past 25 years. There is now a broad range of data mining tools that can be used to implement, test, and deploy effectively the machine learning algorithms that were still being researched when I entered the field. Of course, our academic and corporate research institutions continue to refine and expand on these algorithms and on technologies for their effective deployment; one of the things that make data mining such an interesting field is the close cooperation between researchers and practitioners.

From a practical point of view, computers and database systems have become so much faster and higher capacity that there are virtually no limits on the amount of data that can be accessed to solve problems. The challenge lays now more than ever in using the available data intelligently and in truly understanding what the data represents and how it can be used in a way that will lead to meaningful, reliable results.

Text mining and natural language understanding have also seen incredible advances in recent years, and we now have commercially available tools that can exploit the vast amounts of free text data that virtually all enterprises are generating.

The field of data mining is wide open to all who want to apply their analytic talents to solving real-world problems.

References

1. J.W. Kling, L.A. Riggs, *Woodworth & Schlossberg's Experimental Psychology* (Holt, Rinehart, and Winston, New York, 1971)
2. P. Bock, *The Emergence of Artificial Cognition: An Introduction to Collective Learning* (World Scientific Publishing Company, New Jersey, 1993)
3. P. Bock, R. Klinnert, R. Kober, R. Rovner, H. Schmidt, Gray-scale ALIAS. IEEE Trans. Knowl. Data Eng. **4**(2), 109–122 (1992)
4. C. Howard, P. Bock, Multi-class classification and symbolic cognitive processing, in *Proceedings of the Conference on Computer Analysis of Images and Patterns*, Budapest, Hungary, 1993
5. C. Howard, An adaptive learning approach to acquiring geometry concepts in images, Doctoral Dissertation, The George Washington University School of Engineering and Applied Science, Washington, DC (UMI, Ann Arbor, Michigan, 1997)
6. G. Seni, J. Elder, *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions* (Morgan & Claypool, San Rafael, California, 2010)

An Unusual Journey to Exciting Data Mining Applications

J. Dustin Hux

I found my way into this field haphazardly. There are academics with the training and graduate work that put them on the path toward being data scientists. But as an undergraduate, and even in graduate school, I did not know what data mining was, partially because the field was too new to have made it to the schools where I studied. However, my passion for statistics and for solving problems, as well as a fortunate series of jobs, led me into this field. Rather than approach it academically, I have used data mining for many, many projects that have solved real-life problems.

I grew up in a rural East Tennessee town. In high school, I was interested in science and math, but the town was pretty much made up of farmers and factory workers (from an academic perspective, the smartest people in town were the country doctors and lawyers), so I did not see these subjects in terms of a career path. But several teachers—Mrs. Hammonds, Mr. Anders, and Coach Williams, especially—made the most of the available resources. I did not really know what I wanted to do, but knew that I wanted to go to college, even though my high school did not offer a lot of guidance in potential career possibilities. My grandparents and parents really emphasized the need for a great education and were definitely very supportive.

My mother and father were entrepreneurs. My dad had a small truck driving company, and my mother ran a preschool, and from an early age, I spent a lot of time working with them. They instilled in me a strong work ethic and a passion for learning. In addition, I have done lots of manual labor, which is common for kids with my background—things like lawn care, farm labor, even maintaining a cemetery. The job that I loved most in high school was working for a livestock auction company. It was a wonderful combination of manual labor, working with

J.D. Hux (✉)
VP Analytics, Elder Research, Inc., 300 W. Main St., Suite 300, Charlottesville, VA 29901, USA
e-mail: dustin@dataminininglab.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_9,

animals (understanding their behavior and biology), and a grassroots introduction to how markets work.

After high school, I ended up at Emory & Henry College in Southwest Virginia, where I studied economics and biology and played football. Emory & Henry was the perfect school for me. It is a small liberal arts college, with only about 800 students when I was there, that emphasizes community service and leadership and, given its geographic location, provides an excellent education for many first-generation college students. Both the business and the biology departments required students to study statistics, but typical of a liberal arts college there were not a lot of options for advanced statistics classes. I definitely liked statistical modeling—really enjoyed it—which made me sort of an oddball. My girlfriend, who ultimately became my wife, was in the psychology department, which had more statistics classes available; I did not take all of those courses, but I devoured their textbooks. When I graduated, I knew that I wanted to focus on the nexus between science and business, and I had a strong interest in environmental issues, but I did not know how to make a career out of that. I did not even have a plan for how to make it happen. Luckily, Emory & Henry gave me the general tools I needed to advance my education and ultimately my career. I really cannot emphasize enough how much this college means to me and how fortunate I am to have had the opportunity to attend.

While I was in college, I had a couple of internships that helped to give me some direction. The first was with the Department of Mines, Minerals, and Energy for the State of Virginia. This was a great position because it allowed me to see some ways that science and business come together. During this internship, I spent a lot of time in Southwest Virginia helping inspect oil and gas drilling operations. I spoke with engineers from many different gas companies and came to understand how they work, specifically, what they had to notice to know where to drill and how deep to drill. It also provided me with one of my first experiences with engineering software and geospatial data. The office that I worked in was moving from paper to digital maps, and part of my job was to help with this transition. It was a job that really strengthened my organizational and planning skills and, as a bonus, allowed me to see a lot of the really rural countryside of Virginia.

My second internship was for a company called Morrison Molded Fiberglass, a company that made industrial structures from fiberglass. This was where I did my first real-world statistical analysis. I worked in their corporate marketing department on a project to try to help the company figure out which of the many relevant trade shows around the world were worth its investment of time, money, and energy and which it should not attend. That was the first place where I actually got to apply some of the statistics I had learned in my college coursework to a business environment. In some ways, I still knew nothing of data mining, but I liked this kind of project.

At that point, I was trying to decide whether I should get an MBA and go straight in the business world, whether I wanted to focus on biology and environmental science, or whether as third choice (influenced by Richard Preston's book

The Hot Zone), I should explore the virus-hunting world of the Center for Disease Control, which I knew would also require a lot of analytics.

After Emory & Henry, I moved to Charlotte, North Carolina, and went to work for Pacific Hemostasis, a company that made diagnostic controls out of human blood and plasma. These controls help the doctors of someone having surgery evaluate a sample of the patient's blood for clotting or other issues, and to examine how different drugs affect the patient's clotting during surgery. The company took human blood, put it through various chemical processes, and built the individual controls against which the blood would be tested. This allowed doctors to determine how, say, a hemophiliac's reactions might compare to the reactions of normal blood. I worked in the formulation chemistry area, which was the production environment for the company. We were the ones who actually made the product and saw it through; from the lab, we would put the samples into vials, followed by freeze-drying and storage until they were packaged up for shipping.

Getting this job was an example of unintended good fortune; it was not something that I had been looking for in particular and I did not know where it would lead me. I needed to be in Charlotte because my wife was starting graduate school there, but the economy was in a recession and jobs were scarce. I looked for any kind of work (e.g., Christmas light hanger, medical laboratory technician, photo processing center employee, and ink manufacturing worker) out of college that I could get. I was actually at the point where I had set a deadline, after which, if I did not have a job, I would go back to work for my dad's trucking company. The job offer from Pacific Hemostasis came in the day before my deadline.

I am definitely a scientist by nature. I love learning new things and would have devoured any scientific work that I could get my hands on. I got lucky because the human clotting process is one of the most intricate biological processes that there is; even though it was not my job to know about the process, I learned everything I could. At the same time, I kept learning about other scientific fields, taking postbaccalaureate courses at the University of Charlotte in environmental science, with a specific focus on climatology—the study of climate. I was studying watershed science, atmosphere and weather, while working full time.

At Pacific Hemostasis, I had my second experience with applying statistics to a business challenge. The problem was that, all of a sudden, we started having complete batches of a product line fail; the process that we were doing in the laboratory was the same, we were following the established recipe, but batch after batch of the product was not right and no one understood why. At many points in the production cycle, product samples were collected and sent to the quality control lab, which tests them and reports on their performance. We were collecting lots of data—we knew the room temperature when we were making the product, who was involved in the formulation of the product, who was on the line when it was put into the vials, all of the pressure and temperature data from the freeze-drying machines and the outcome variables—but no one knew what was driving the failure.

Although we were taking all of these samples, no one was following the resulting data, and no one was putting the results into a database. In those days, people used computers in the workplace mostly for word processing. Big companies, the

government, and universities had access to massive computer power, but midsized companies did not; most did not even have computerized billing systems. There simply was not a notion of collecting all of the data from throughout the process, creating a database and using that to try to determine where in the process something had gone wrong. So I did that. I was getting ready to head back to graduate school, but before I left, I put all that data in the database, analyzed it, and figured out the source of the problem. When we corrected the issue, the batches started working again. This was critical to the company's success. It was a relatively new company with a good product, but its clients were beginning to worry about its quality and reliability. With that correction, the company stabilized and I got an employee excellence award. This was my second taste of the value of statistics.

In 1995, I moved to Charlottesville and started graduate school in the environmental sciences department at the University of Virginia. Students in this field were required to take courses in the four tracks of geology, hydrology, ecology, and atmospheric science. So I studied all four, but I focused on atmospheric science, specifically synoptic climatology. In this field, one does not study microclimates or short-term weather; instead, the scale of synoptic climatology is larger, ranging from regional to global; we understand weather and short-term atmospheric phenomena, but the focus is more long-term (seasonal to decadal). While other areas of environmental science focus on field studies, a synoptic climatologist pulls together a multitude of data from many sources (e.g., all of the data from weather gauges in the United States and around the globe). In effect, I saw myself as a combination of physical scientist/computer scientist/statistician who focused on climate.

This is really where I honed the tools that I use as a data miner, because weather and climate science is one of the original huge data problems. The issue of weather forecasting—whether one is thinking about global warming or weather changes on a large scale—really took off in the 1940s, around World War II; it was one of many fields that really started to develop then. The first mainframe computers were built to deal with issues of weather forecasting, because with advanced data measurement systems, more information was coming in than people could handle. For instance, there are first-order stations that collect temperature data, and several other variables, every hour. There are second-order stations that measured temperature and rainfall daily. In addition, scientists are sending up weather balloons four times a day around the world. These along with data from ships, buoys, satellites, ice cores, etc., are combined to generate some massive geospatial data sets. Analyzing these allows forecasters to ask things like, "Is the climate really changing?" "How can I anticipate whether the precipitation will be rain or freezing snow?" "Where exactly will this hurricane land?" and "How is the jet stream changing over time?" These are the kinds of questions at which I was looking in graduate school. I was also fortunate to work with three great climatologists—Robert Davis (my major professor), Patrick Michaels, and Bruce Hayden—who really emphasized statistics and the need to listen to the data.

My thesis was on the January thaw. Back when we were an agrarian society, people were tied to the land; they had to pay close attention to the weather and to

other natural phenomenon because their lives depended on it. Naturally, lots of folklore developed around weather and weather patterns. The January thaw was said to be an uncharacteristic or anomalous warming that occurs in New England during the 3rd week of January. I wanted to determine whether the said warm-up is statistically real. I also hoped to find out whether there is a physical mechanism that can describe what goes on when it happens. I discovered that the January thaw was, in fact, statistically real, but I did not fully come to understand the mechanism that caused it. This was largely because by the time of my defense I had switched to data mining, so found myself pulled in different research directions.

The entire time that I was in graduate school, I also worked in the Virginia State Climatology Office. The mission of that office was to answer any questions having to do with the weather and climate in Virginia. Any questions at all. For instance, a researcher might want to know whether it had rained in Roanoke on February 2, 1969. Or an engineer working with a convenience store company to build stores with gas pumps might ask about the maximum snow load that he could expect on the top of a canopy.

We also worked with the National Weather Service, using data mining algorithms to create a model that would anticipate the precipitation type for winter storms in the mid-Atlantic region. The mid-Atlantic turns out to be one of the most difficult places in the country to determine this, because so many different geographic and atmospheric features come together there to create a rapidly changing transition zone. We succeeded in building a really good model for the National Weather Service, one that they used successfully for years; at the time, it was one of the best models for this prediction. I was also involved in a project for the Virginia Department of Environmental Quality to build ozone forecast models for the Commonwealth's metropolitan areas.

At the Climatology Office, I learned what it takes to be a consultant, in terms of using statistics and applying data mining skills, because pretty much every question boiled down to a statistical problem, "what's the probability of this?" I honed my people skills as well, in the course of trying to understand the question that someone was really asking or the problem that he was really trying to solve. I learned to seek out both the individual and the business goal. As a scientist, I was more accustomed to being in the laboratory than in the public side of things. This job allowed me to get used to talking to people. For example, we had a movie company that wanted to film a movie in Virginia and asked us to determine the optimal week for fall scenery, which turned out to mean, "when and where can we expect the peak leaf color?" It was a process of understanding what people were looking for, of trying to focus on exactly what questions to ask and which ones it was actually possible to answer, of making sure that everyone was on the same page and that we were solving the right problem.

It was a really great experience because everything that I was learning in the classroom, I was turning around and applying in real time to data. Working at the Climatology Office, I discovered that, while I am fascinated by weather and climate (these are still passions and hobbies of mine), I really just loved solving problems and working with data. It did not matter whether you call it statistics or data mining

or predictive analytics, it did not matter what domain it was in, I just wanted to be doing it.

Even when I was still in graduate school, still working on my Ph.D., and still fully funded, I realized that I was finding the day-to-day questions at the State Climatology Office less challenging. The wife of one of my colleagues was working for John Elder. She knew that the company was looking for people who were quantitative and knew statistics, so she suggested that John talk to me. I liked him and liked the work, but since I was still finishing the Ph.D., I intended to come to Elder Research only part time, at 20 h a week. I quickly discovered that you cannot really solve a problem for a client in 20 h; in fact, clients do not care about your hours, they want you to be available to them, they want their problem solved. So 20 h a week was not realistic, but the work was exciting.

This work continues to be the ultimate in good fortune. In the decade since joining ERI, I have had the chance to work on and lead teams solving problems in the commercial, financial, and scientific domains including fraud detection, cross-selling, customer profiling, product bundling, direct marketing, biometric identification, stock selection, market timing, and text mining. And, more importantly, I have found my calling in terms of a career, had the opportunity to work and talk with many of the leading practitioners in the field (many of them without having to leave the office), been a part of growing a company, and had the opportunity to have John Elder, the leading applied educator in the field, as a friend and mentor.

When I first started at Elder Research, I worked on some quantitative investing problems for a hedge fund, looking at questions like, “If this is the goal of the fund, how should I trade it, based on statistics?” “Should I be in or out, or should I short it?” I enjoyed looking at those financial questions because the data in many ways were similar to weather data; time based and continuous mostly. It was a great introduction into data mining. I was building something that would help hedge funds vet strategies to make sure they had the systems to help people trade, another example of applying the theory that I had learned to a real-world problem.

Also early in my time at Elder Research, I learned that data mining could be used in “customer relationship management (CRM).” For example, imagine I am in a catalog company, with mailing lists, subscribers, and products. I know who has purchased in the past and who has not, and I know what and how much they have purchased. I also have some demographic information about my customers (their home states, ages, and genders). Now that I have that information, and with a limited budget for mailing my catalog, to whom should I mail it? Which customers should I keep and which should I drop? We built models to help answer these questions. We also built a model for the catalog company’s Web page to show people that “customers who have bought this have also purchased this,” that is, helping the company to determine the next most likely product it should offer a customer. We did the same kind of work for one of the largest banks in the world, helping it to decide which products would best appeal to different clients; I built a system to help it make interesting offers to people.

One thing that I like about data mining is that it allows me to develop strategies that are good for the company and it is good for the consumer too. If you think back to the days of snail mail, people used to receive tons of irrelevant junk mail. These days, computers have the ability to target individuals, which saves the company money and only sends the customers information or offers in which they are most likely to be interested.

At Elder Research, I also ended up doing some biometric analysis. We had a company that was developing a medical diagnostic for testing people's blood without pricking them. The process involved shining a light on a person's skin. The skin absorbs some of the light and reflects some back out. The company could measure the spectrum of reflectivity from the skin and, from that, determine things like the glucose level in the blood. As it refined the process, the company discovered that, although the study was supposed to be blind, the spectrum from the skin has biometric qualities, which made it possible to identify individual subjects. We helped them determine whether each person's spectrum was specific enough to make this truly a unique identifying technique. Furthermore, we helped to determine possible applications for this. The company thought that the light technology might work via a sensor to lock and unlock a car or, from a law enforcement angle, hoped the technology could be used to lock a gun when an unauthorized user touched it. In the end, although the process was really good statistically at uniquely identifying people, it was not quite good enough for the handgun scenario, where you do not want even the one percent risk that the gun might not work for the law officer. It was accurate enough for use at Disney, however, where it became the biometric test that allows people to get in and out of the park without having to show a pass.

One passion of mine in the field of data mining is fraud detection. This developed from a series of projects that we did with one of the largest electronics manufacturers in the world. That company had hired someone to start up an internal global analytics group. This person realized that he and his staff needed external assistance, so he called Elder Research to help him set up an "Analytic Center of Excellence" within the company. A colleague, Karl Rexer, and I went down and initiated what would become a very successful (both for ERI and the client) multiple year, multiple project engagement.

Over the course of that first year, we did four pilot projects for the electronics manufacturer to help to prove the value of data mining to them. The first was a market basket study, trying to preconfigure large servers based on their client companies' demographics. We looked at where the client companies were located, their size, etc., and used that information to model what kinds of servers and storage ability would be best for them. The results of that project were very successful.

The second study was along the lines of target marketing, trying to find the best of the many possible service agreements, warranty options, and end-user agreements to fit each client. In the third project, we analyzed why people sometimes returned parts when there was no functional reason for the return and no problem with the product. This led to an investigation of who was returning products and why, and whether the returns were fraudulent.

The fourth project was also in the area of fraud detection, and it was my favorite. Occasionally, one of the manufacturer's products develops a problem or needs servicing. The company does not have its own service people, but instead relies on authorized service providers with whom it contracts to help purchasers. The focus of the project was on how to identify which of these service providers might be committing or might be likely to commit fraud against the company.

When we started, the electronics manufacturer had two employees engaged in this area. They did not use well the few metrics they had, and they lacked an efficient research process. We needed to make the hunt more systematic. The company had a tip line, which disgruntled workers would call, but the employees had to do the follow-up research manually, looking at the providers' transactional history, and it was a huge job. We came in and created many more metrics and statistical models, which allowed us to rank order the service providers based on a fraud score. Some of it was low-hanging fruit (a technician who managed to be in three states at the same time). Some looked at unusual patterns or the transactional numbers that the service providers were submitting. The company implemented our method, and in the first two years, it recovered 20 million dollars in fraud. Over the first five years, it recovered 67 million.

That one analysis changed what I have become passionate about within data mining; it changed my career trajectory within the field. I now have the perfect path for myself: I get to use statistics, solve hard problems, and help catch the bad guys.

I have done more and more antifraud work since then. It is unbelievable to me how smart and creative some of the fraudsters are. The talented, professional ones could use their skills and do something worthwhile, but sadly, they choose to do this instead. These days, when analyzing a fraud detection problem, we not only rely on traditional structured data, but we might also have, say, the texts of e-mail exchanges with customer service agents or someone's notes on the case. Now we have the ability to take this raw text data and extract the relevant information and incorporate it into the models alongside the structured variables.

There are always innovations in data mining and new algorithms. I get to read about the newest possibilities, create the relevant algorithms, and try them on real-world applications for a client. For me, this is not just a theoretical study, I am taking science out of academia and putting it to work for businesses—it is awesome.

9/11 was another defining point in my career as a data miner. I was in a meeting with a financier, discussing how we could help him optimize trading strategies for his hedge fund when the news came in. I knew people who worked in the World Trade Center. Those towers were falling, a plane had crashed into the Pentagon, and the world was changing drastically for our country that day. But my client was just focused on his model, he did not even want to take a break to follow the news or ponder its enormity.

This made me want to do less financial work! While I have done some of that since then, I have spent most of my time taking the tools that I have developed for fraud detection and data mining, and using them to try to help with the new challenges that we face as a country. These tools can help the government in

a number of ways, particularly in the areas of fraud, waste, and abuse. Since 9/11, there have been more opportunities for people to cheat the government. If they send in fraudulent tax forms, Elder Research helps discover that. Many more contracting companies now work for various agencies in Washington, DC; the vast majority are completely above board, but some have tried to steal taxpayer dollars. Elder Research's analysis can help to ferret out these companies as well.

I have been fortunate. While I knew what I liked to do, I did not know that there was a field where I could put my passions and my strengths to such powerful and satisfying use. I have loved solving problems in my decade at Elder Research. Because the company does not have a vertical niche, we can take on issues in any field and try to solve them. I have followed an almost random career path, but the constant has been my desire to solve problems so that businesses can function better. I love the practical application of the theories and techniques that I have studied. From the tradeshow analysis, through the January thaw and random weather questions, to hedge funds and fraud (anomaly) detection, I have had the opportunity to use and hone my data mining skills. This has culminated in the important work that I do now, building data mining solutions for government clients.

Editor's note. Mr. Hux has for several years led an award-winning team to successfully employ data mining in the service of national security.

Making Data Analysis Ubiquitous: My Journey Through Academia and Industry

Hillol Kargupta

1 Motivation

It was one of those late fall mornings in Urbana. I was working on some of the final pages of my dissertation. I got a note from Mike Welge of the National Center for Supercomputing Applications (NCSA) whom I came to know during the course of my work with my Ph.D. advisor David Goldberg. Mike was leading a data mining project for Caterpillar, the US heavy duty equipment manufacturer. Caterpillar clients bring their equipment to their worldwide service center for maintenance and repair. Their service staff types in short descriptions of the work done on the equipment and saves that information in the computer. Caterpillar wanted to link this data from different service centers, analyze, and identify which equipment and parts are failing frequently and related decision support tasks. The problem became more challenging because their employees often used different abbreviations and spelled names incorrectly to describe the work done on the equipment. Mike wanted to address this as an unstructured text data mining problem and asked me if I would like to collaborate. I joined their meetings and started thinking about the problem in a bigger context.

A Couple of months passed; I finished my dissertation, went through the final thesis defense, and moved to Los Alamos. I was working at the Los Alamos National Laboratory and my group was highly interested in data mining and high-performance computing. I kept my relationship with the NCSA as an affiliated scientist and managed to convince them to subcontract a small part of their Caterpillar project to my group at Los Alamos.

H. Kargupta (✉)
Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, MD 21045, USA

Agnik, LLC, 8840 Stanford Blvd. Suite 1300, Columbia, MD 21045, USA
e-mail: hillol@cs.umbc.edu; hillol@agnik.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_10,

As I started thinking about their unstructured text data mining problem, I wondered how our solution would be deployed in practice. I joined the project meetings and noted that the data collection step was itself challenging as the data were available at different locations of the same company. They have service centers all over the world, and they had to abide by the laws of different countries. They had different systems and that threw additional challenges. One of the main things I noted was that the data was collected at distributed locations all over the world and there were some challenges in centralizing that data. That set me thinking about other possible challenges in various applications where centralizing the data from the distributed locations may be very difficult.

It was 1996 and the field of mobile and ubiquitous computing was at its inception. I was trying to convince my wife Kakali that we should buy cell phones. However, I started thinking about what would happen if we had computers connected over wireless networks with limited bandwidth. Would we be able to centralize all the data from distributed locations and use our traditional centralized data mining algorithms to quickly analyze that data? Peer-to-peer (P2P) networks for file sharing applications and sensor networks for different monitoring applications were getting a lot of attention around that time. I started thinking about how data centralization would work in such applications. In a P2P application, can we really support downloading data to a central location from millions of nodes? In a wireless sensor network, sending a lot of data over the wireless network will increase the energy consumption and reduce the lifetime of the network. Can we really build practical sensor network applications using centralized algorithms? I was thinking about many of these questions.

Over the next few months, I convinced myself that in the near future, we were likely to face many applications where centralizing the data first may not be practical or even feasible for subsequent analysis. Wireless network bandwidth, cost of communication, privacy issues, regulatory requirements, and the large asynchronous nature of many distributed systems would require a different way to process distributed data at least as long as the results were needed in a reasonable time period.

That was the start. I started on the Caterpillar project but with a twist. I extended the scope of the problem and decided to explore how we would solve the unstructured text mining problem using fundamentally distributed algorithms. We started with relatively simple n-gram-based text representation and developed various synchronous distributed text analysis and clustering algorithms. Meanwhile, I engaged two very smart graduate student interns—Ilker and Brian—to work on this project with me. We built a prototype system PADMA, which stood for Parallel Data Mining Agents. We submitted a paper to the KDD conference [1]. Back then, it was just KDD, not today's ACM SIGKDD conference. Some of the reviewers doubted the need for distributed data analysis algorithms. However, the paper was accepted. Los Alamos liked the application and gave us the Technical Achievement award. We were happy. That is how I started work on distributed and ubiquitous data mining.

The rest of this chapter shares the stories of my journey over the last 15 years. First, in Sect. 2, I define the distributed data mining problem a little more formally, of course without getting into symbols and mathematics. The section also provides a brief overview of the milestones of my algorithmic work and commercial product development. Section 3 offers some of the lessons I learned. Section 4 describes some of my thoughts on outstanding research issues and challenges. Finally, Sect. 5 summarizes and concludes this chapter.

2 Milestones and Success Stories

I spent the early days of my career thinking about how data analysis would change in future as the nature of computing evolves over time. When I was in graduate school in Illinois, the Internet had just started booming. Having witnessed the transition of the Web browser Mosaic from the campus of the University of Illinois to the commercial product Netscape, the Internet boom of the 1990s had a long-lasting impact on me. While the data mining community was primarily focused on developing algorithms and systems for analyzing the tremendous amount of data accrued at the data warehouses, my thoughts were mostly on how data analysis would work and scale in this increasingly network distributed environment of machines and devices. For some reason, I could not accept a scenario where data analysis takes place only at a handful of places where all the data of the world are conglomerated. In my mind, data mining had to play a key role in every step of our life where data would be collected and stored without having to give it up to a big brother corporate entity for getting the intelligence out of the data.

My early effort in this area was in figuring out how this next generation of data analysis applications would look in a distributed ubiquitous environment, abstracting the problems, and developing algorithms and system design principles for solving those. The Caterpillar project with NCSA helped me better understand real-life needs and shape the abstract analytical problems.

2.1 Abstraction of DDM Problems

My quest began with the abstraction of distributed data mining applications based on how data and computing are distributed among the different participating nodes of the distributed environment. Think about your favorite data analysis algorithm and consider running that when you have the data and computing sources distributed, and not centralized. Many relatively well-understood data analysis algorithms become challenging in this distributed scenario. Unlike the traditional high-performance parallel/distributed computing literature, my focus was more on systems where the data is inherently distributed and redistribution of data is often a difficult expensive step. When we have data sources embedded in the ubiquitous processes supporting every step of our life and the data is hard to centralize because of communication cost,

scalability, privacy, and other reasons, then it must also be hard to redistribute the data for the same reasons. That made me think about how we can design data analysis algorithms where data centralization or redistribution has an inherent cost and how those should be done only with due care so that the overall performance of the algorithm is optimized.

The first step we took was to abstract DDM applications in order to pose them as algorithm-design problems. We looked at different ways to abstract distributed data mining problems. For example, we looked at the distribution of the data across different sites:

1. Heterogeneous scenario: When the data models (e.g., the schema in the relational model of data management) at different locations are different but somehow linked through some properties or features or other data items. For example, in the relational model of data, this could mean tuples in two tables at two different locations are linked through some foreign keys; in this case, the overall data set comprised of both the sites can be viewed as a join of the two tables.
2. Homogeneous scenario: When the data models at different locations are identical to each other. In the relational model, this could mean that the two tables at two different locations have the same schema; in this case, the overall data set comprised of both the sites can be viewed as the union of all tuples from both the sites.

Data distribution is not the only aspect you need to keep in mind to develop an abstract formulation of a DDM problem. Communication, computing, and role of users are some of the other aspects that we need to think about. For example, our early effort explored different aspects of distributed computing and their interaction with the notion of time in a distributed data mining [19, 20, 21, 22, 23] system:

1. Synchronous algorithms: When all the participating nodes have the same clock and the different components of algorithm work in tandem at every node, it is called a synchronous algorithm.
2. Asynchronous algorithms: When the participating nodes do not have a synchronized clock, then we need asynchronous algorithms. For example, large P2P networks over the Internet often require asynchronous algorithms since different participating nodes may have different clocks that may not be in sync with each other.

Right from the beginning, we paid a lot of attention to the communication cost, and the algorithms we came up with had a major focus on optimizing that. We explored how different modes of communication can be abstracted into the problem definition. For example, we considered the following possibilities:

1. Global communications: This is the scenario where the algorithm can communicate with any node in the network.
2. Local communications: In this case, the nodes are allowed to communicate with only a subset of nodes in the network. Usually, these nodes are restricted to the local neighborhood in the communication graph. This graph is essentially comprised of the participating nodes and the links representing the communication channels between the nodes.

These different ways to structure various aspects of distributed data mining applications and solutions allowed us to formulate the abstract problem necessary for the algorithm design.

2.2 DDM Algorithm Development

Like most centralized data mining problems, distributed data mining problems often require learning the structure or patterns from data. We call an observation a pattern when it is repeatedly observed in the data. A pattern is captured through many different types of mathematical constructs such as the probability distribution, classifiers, clusters, and many other entities. These mathematical constructs are often described with respect to the chosen representation of the data. Feature space defined by the problem and constructed features are some examples of representations. The goal of a centralized data mining algorithm is to compute such constructs from the data. On the other hand, distributed data mining deals with the problem of computing such constructs from distributed data while paying attention to the distributed computing, storage, and human resources. My initial work on distributed data mining explored how algorithms and systems for data mining can be developed for synchronous environments. The next section summarizes some of my work in that area.

2.2.1 Distributed Data Mining for Synchronous Environments

Synchronous models of distributed computing deal with environments where participating nodes with clocks are all in sync with each other and the processes and message communications occur in bounded time. Data mining in such environments can be approached in various ways. Decomposing the tasks into a set of subtasks that can be performed at different nodes, followed by an aggregation of the results of these subtasks, is a popular approach. Another possibility is partitioning the data, running the operations on each of these partitioned blocks of data, and subsequent aggregation of the results at distributed nodes.

Consider the simple example of covariance matrix computation. Once the column means are translated to zero, the covariance matrix computation problem essentially reduces to summation computation with appropriate scaling factors depending upon the number of tuples used in the summation. Computations like this can be easily decomposed into a set of subtasks when the data tuples are stored or distributed at different locations. This is because summation computation can be decomposed among a set of partial sum computations. Using a similar philosophy, we developed a series of different algorithms for data analysis in a distributed environment. For an overview of our work in that area, please see Kargupta and Sivakumar [2].

2.2.2 Distributed Data Mining in Asynchronous Environments

Asynchronous systems do not enjoy the benefits of having clocks at every node that are in sync with each other. Moreover, processes and message communications may not come with bounded time. It offers a new area of application for distributed data mining algorithms. Our work in this area started in 2003 as a growing number of P2P systems started appearing for file sharing and other related applications. Asynchronous DDM algorithms must be scalable to very large networks that are decentralized and asynchronous. Apart from P2P systems, asynchronous distributed data mining algorithms also play an important role in mobile and ubiquitous applications. The European Union KDUbiq initiative organized by Michael May and Codrina Lauth and the KD2U[1] [3] emphasized the need for this technology in the field of ubiquitous data mining. There are many ways to classify asynchronous DDM algorithms and systems. Since these are typically large systems and finding exact solutions for some problems quickly appears to be difficult, we started exploring probabilistic algorithms that often yields probabilistic and approximate results. We developed a bunch of exact and approximate asynchronous distributed data mining algorithms. Here is a brief overview of the work we did along each of these directions:

Approximate Algorithms. Approximate distributed algorithms provide an approximation of the result produced by their centralized counterparts where all the distributed data is first centralized. The approximation can be probabilistic or deterministic. Probabilistic algorithms use a wide range of techniques from random walks on graphs to sampling to analyze data in an asynchronous distributed system. Examples of these algorithms include the P2P k-means algorithm by Bandyopadhyay et al. and Datta et al. [4, 5], the newscast model by Kowalczyk et al. [6], and the ordinal statistics based distributed inner product identification in P2P networks by Das et al. [7]. We also explored some deterministic approximation techniques using techniques such as variational approximation [8]. Here are some examples of our work on asynchronous approximate DDM algorithms:

1. *Sampling Over P2P Networks*. Probabilistic approximation techniques are widely used in P2P applications. Many such approximate algorithms rely on sampling strategies from the population to draw conclusions about certain statistics. In the context of analyzing data distributed in a P2P network, summary statistics from a uniformly and independently sampled set of nodes can help in estimating aggregate information about the entire data with certain confidence bounds. This is usually communication-wise far more efficient than communication with all nodes to compute aggregate statistics. Our work in this area explored possible ways to draw uniform, independent, and identically distributed (i.i.d.) samples of nodes and data from a P2P network. Such sampling techniques can be used for various P2P applications. We demonstrated the

[1] http://www.kd2u.org.

application using k-means clustering in a distributed P2P setting. Selecting a node uniformly from all the nodes in a large, stationary P2P network in a communication-efficient manner is a non-trivial task. This is because no node has a list of the IDs of all nodes in the network. Each node is only assumed to know the IDs of its immediate neighbors. Collecting a uniform sample of data from the network is even more complex, as data distribution is highly skewed among the nodes. We posed the problem of selecting data points uniformly from a P2P network as the problem of selecting nodes from a graph by transforming the underlying network topology and data distribution to a simple, connected, undirected graph. The technique we developed relies on taking a random walk over the edges of the graph with a carefully chosen transition probability and exploiting the Markovian property of random walk to limit the length of the walk for a communication-efficient sampling procedure.

2. *Approximate P2P k-Means Clustering*. Given that we can collect an i.i.d. data sample from a P2P network, we can apply that to solve the problem of clustering data distributed over a P2P network. *K*-means clustering is a simple clustering algorithm that assumes that the number of clusters, k, is known apriori and then tries to group given data points into k clusters by ensuring that each point belongs to the cluster with the nearest average. The algorithm works by randomly initializing k cluster centroids and assigning each datapoint to its closest centroid. Then in every iteration, it updates the centroid of each cluster by taking an average of all the data points in that cluster, and reassigning datapoints to its closest updated centroid, till the centroids converge. Our proposed P2P k-means clustering algorithm runs in a fashion similar to the standard version, except that to update the centroid at every iteration, the distributed average computation problem needs to be solved in a communication-efficient manner without flooding the entire network. At every iteration, a set of nodes are uniformly sampled using the sampling technique described before, and local cluster centroids from these nodes are used to update the global cluster centroids. In this way, only a local synchronization is required among the sampled nodes as opposed to a network-wide global synchronization. The approximate cluster centroids computed certainly deviate from the actual global cluster centroids computed using the entire data on that iteration. However, exploiting the statistical properties of the i.i.d. sample, a bound can be drawn on the deviations of centroids with probabilistic guarantee. The iterative process continues till the centroids fall within an acceptable error bound or the maximum number of iterations reached. The final approximate centroids are broadcasted to every node in the network upon termination of the algorithm, and each node assigns its local data points to the closest final cluster centroid received.

Exact Algorithms. In exact distributed algorithms, the results produced are exactly the same as that of the centralized counterparts where the mining is done after centralizing all the data to a single site. We developed various types of exact distributed algorithms for P2P applications. Many of these algorithms came with a monitoring version. In other words, we also explored how we would detect changes

in the models or statistical properties over the P2P network and trigger the computation of those when the change is detected. Here are some examples of exact P2P data mining algorithms we developed:

1. *Multivariate Regression in P2P Environments*. Multivariate regression (MR) is a popular technique for building predictive models, learning classifiers, and many other applications. So we wanted to develop a P2P version of the regression technique. In this setup, each peer has a set of input features (multivariate), and the goal is to predict the value of a target. Unlike the P2P k-means algorithm, this algorithm is provably exact—each peer has the correct regression model (with respect to centralized computation). The algorithm takes a two-step approach for building and maintaining MR models in P2P networks. The first step is the monitoring phase in which, with a prior estimate (can be random) of the MR model given to all the peers, they asynchronously track any change between the model and the union of all peers' data using a provably exact local algorithm. The second step, known as the computation phase, uses the monitoring algorithm as a feedback loop for triggering a new round of MR model building if necessary. The overall algorithm takes a geometric perspective and converts the overall regression monitoring problem into a series of decision-making problems distributed over the network. Each node checks for a set of conditions and takes actions to share information based on those conditions. For an in-depth discussion, interested readers are referred to [9].
2. *P2P Decision Tree Induction*. Decision tree induction by Quinlan and Breiman et al. [10, 11] is a powerful statistical and machine learning technique widely used for data classification, predictive modeling, and more. So we wanted to develop an exact ID3-like greedy algorithm for inducing a decision tree in large P2P networks. The algorithm used a technique to decide which attribute would best divide a given set of learning examples. Based on the correctness of this technique on convergence, a decision tree can be induced in a recursive fashion by splitting the set of learning examples from the root and proceeding onward. In a P2P setup, the progression of the algorithm needs to be coordinated among all peers, or they might end up developing different trees. In smaller scale distributed systems, occasional synchronization usually addresses coordination. However, since a P2P system is too large to synchronize, we needed an asynchronous solution.

The starting point of the tree development—the root—is known to all the peers. Thus, they can all initialize to find out which attribute best splits the example set of the root. However, the peers are only guaranteed to converge to the same attribute selection eventually and may well choose different attributes in the intermediate step. The algorithm has two main functionalities. First, it manages the ad hoc solution which is a decision tree composed of active nodes. The root is always active and so is any node whose parent is active provided that the node corresponds to one of the values of the attribute which best splits its parent's examples—i.e., the ad hoc solution as computed by the parent. The rest of the nodes are inactive. A node (or a whole subtree) can become inactive because its parent (or foreparent)

has changed its preference for the splitting attribute. Inactive nodes are not discarded; a peer may well accept messages intended for inactive nodes—either because a neighbor considers them active or because the message was delayed by the network. Such a message would still update the majority voting for which it is intended. However, peers never send messages to an inactive node. Instead, they check whenever a node becomes active and whether there are pending messages, i.e., majority votes whose tests require sending messages, and if so, they send those messages. Another activity which occurs in the active nodes is the further development of the tree. Each time a leaf is generated, it is inserted into a queue. Once every few time units, the peer takes the first active leaf in the queue and develops it. Interested readers are referred to the article by Bhaduri et al. [12] for a thorough description of this algorithm.

Privacy Issues and Distributed Data Mining. Like many other fields of data mining, privacy is an important issue in many distributed data mining applications. Privacy issues and regulations often create scenario where data cannot be centralized to a single location. Our initial work on privacy preservation in data mining raised some questions about a body of work in the field of privacy-preserving data mining using additive perturbations. Back in 2003, a growing body of literature on privacy-preserving data mining was making use of additive perturbation to "protect" privacy-sensitive data while letting data mining algorithms extract patterns from that data. The problem was that additive white noise is relatively easy to filter out. We pointed out that large random matrices have a well-understood spectral distribution and that information can be used to remove the perturbation in order to extract the original privacy-sensitive data. Our work got acknowledged by winning the 2003 ICDM Best Paper Award [13]. Then, we moved on to exploring multiplicative matrices for privacy protection and analyzing various other attack techniques [14, 15]. Our most recent work in this area has been on the design of asynchronous privacy-sensitive distributed data mining algorithms that are local. Kamalika explored this area and offered a game theoretic approach to design mechanisms for privacy protection in a distributed environment. More details about this work can be found elsewhere [16].

The following section describes some of our achievements on the applications front.

2.3 Applications and Commercial Product Development

A new field is unlikely to grow fast unless the technology has direct practical applications. Successful applications and commercialization are very important for sustaining a field. This was 2003. There was hardly any company with any commercial product based on the DDM technology. I really felt that we needed to do something and put our thoughts to the task. With a background mostly in academia, this was definitely not a trivial task for me. Nevertheless, we put together a small team of smart people and got started. And yes, I also managed to close my

first sales deal during the process—convinced my wife Kakali to quit her job at Intel and join Agnik. That was the beginning of Agnik.[2] We started our journey with a small team in our university technology incubator. We focused on cars and the high-throughput onboard data streams they continuously generate and developed several products based on the car data in a distributed environment where the cars are connected over wireless networks and saw good traction in the market.

2.3.1 Vehicle-Performance Monitoring Products from Agnik

The wireless and mobile computing/communication industry is producing a growing variety of devices that process different types of data using limited computing and storage resources with varying levels of connectivity through wireless communication networks. The rich source of data from the ubiquitous components of businesses, mechanical devices, and our daily lives offers the exciting possibility of a new generation of data-intensive applications for distributed and mobile environments. Agnik's core technical focus is to mine distributed data streams in such ubiquitous environments and create analytics-driven products.

Agnik introduced the product MineFleet®, a novel mobile and distributed data mining application for monitoring vehicle data streams in real time. MineFleet® is designed for monitoring commercial vehicle fleets using onboard embedded data stream mining systems and other remote modules connected through wireless networks in a distributed environment. MineFleet® is a powerful data stream mining software for modeling, benchmarking, and monitoring of vehicle health, emissions, driver behavior, fuel consumption, and fleet characteristics (Fig. 1).

Consider a nationwide grocery delivery system which operates a large fleet of trucks. Regular maintenance of the vehicles in such fleets is an important part of the supply chain management, and normally, commercial fleet management companies are given the responsibility of maintaining the fleet. Fleet maintenance companies usually spend a good deal of time and labor in collecting vehicle-performance data, studying the data offline, and estimating the condition of the vehicle primarily through manual efforts. Fleet management companies are also usually interested in studying the driving characteristics for a variety of reasons (e.g., policy enforcement, insurance, department of transportation regulations). Monitoring fuel consumption, vehicle emissions, and identifying how vehicle parameters can be optimized to get better fuel economy are some additional factors that support ample return of investment (ROI) for systems like MineFleet®.

MineFleet® is now widely adopted in the mobile resource management and fleet management industry. The main unique characteristics of the MineFleet® system that distinguish it from traditional data mining systems are as follows:

[2] http://www.agnik.com.

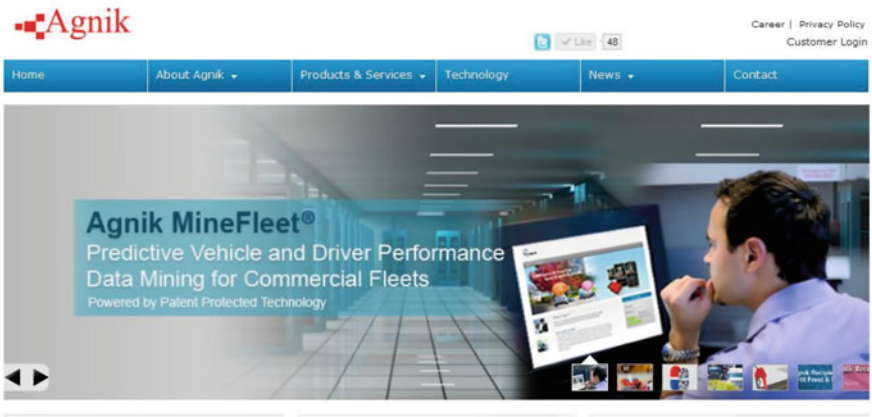


Fig. 1 MineFleet® product from Agnik

1. Distributed mining of the multiple mobile data sources with little centralization of the data.
2. Onboard data stream management and mining using embedded computing devices.
3. Designed to pay careful attention to the following important resource constraints:
   (a) Minimize data communication over the wide-area wireless network.
   (b) Minimize onboard data storage and the footprint of the data stream mining software.
4. Process high-throughput data streams using resource-constrained embedded computing environments.
5. Respect privacy constraints of the data, whenever necessary.

There are some major differences between MineFleet® and traditional telematics systems. Some of them are listed below:

1. *Advanced Data Analytics*. MineFleet® is powered by advanced distributed data mining and statistical analysis algorithms. Most telematics systems are designed for in-car infotainment and security application based on relatively simple data management operations.
2. *Onboard Data Mining*. MineFleet® offers dramatic reduction of wireless communication by performing data analysis onboard the vehicle. Unlike most conventional telematics systems, MineFleet® sends the results of the onboard analysis to the server over the wireless network, not the raw data. As mentioned earlier, if a device is monitoring hundreds of vehicle-performance parameters, it may easily collect about 10 MB of raw data in about a few hours. Sending this raw data to the server for advanced data mining at the server over the wireless network is very expensive. Most MineFleet® customers would not pay for a data plan beyond 5 MB per month. Therefore, analyzing data onboard the vehicle and

sending the resulting analytics instead of raw data are imperative. One full MineFleet® update takes about 1 K. If a vehicle runs for about 8 h a day and gets an update once an hour, then in 30 days, the vehicle would need about 240 K wireless data communication in order to send the MineFleet® analytics to the server. This dramatic reduction in communication cost is a unique feature of the MineFleet® technology which enabled more powerful data analysis and mining at a low cost.

3. *Not a Global Positioning System (GPS)-Based Tracking/Navigation System*. Unlike most conventional telematic devices, MineFleet® is primarily focused on vehicle-performance data analysis, not tracking and navigation.

These unique aspects of MineFleet® distinguish it from the conventional tracking/navigation and telematic services. The following section offers an overview of the MineFleet® architecture.

MineFleet® Overview

MineFleet® is a mobile and distributed data stream mining environment where the resource-constrained "small" computing devices need to perform various non-trivial data management and mining tasks onboard a vehicle in real time. MineFleet® analyzes the data produced by the various sensors present in most modern vehicles. It continuously monitors data streams generated by a moving vehicle using an onboard computing device, identifies the emerging patterns, and, if necessary, reports these patterns to a remote control center over low-bandwidth wireless network connection.

MineFleet® also offers different distributed data mining capabilities for detecting fleet-level patterns across the different vehicles in the fleet. This section presents a brief overview of the architecture of the system and the functionalities of its different modules. The current implementation of MineFleet® analyzes and monitors only the data generated by the vehicle's onboard diagnostic system and sometimes the GPS. The MineFleet® Onboard is designed for embedded in-vehicle computing devices, tablet PCs, and cell phones.

The overall conceptual process diagram of the system is shown in Fig. 2. The MineFleet® system is comprised of several important components that are briefly described in the following sections.

Onboard Hardware

The MineFleet® Onboard module is comprised of the computing device that hosts the software to analyze the vehicle-performance data and the interface that connects the computing device with the vehicle data bus. Figure 3 (left) shows the MineFleet® Onboard Data Mining platform (MF-DMP101) device that hosts the MineFleet® Onboard software. MineFleet® also runs on many different types of embedded devices, in-vehicle tablet PCs, laptops, cell phones, and other types of handheld devices. Several other hardware platforms (e.g., DMP-201 from Agnik

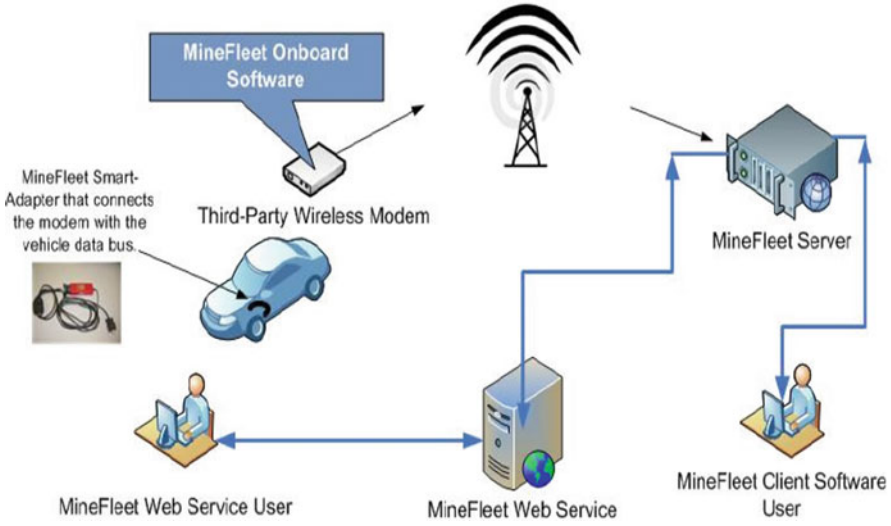


Fig. 2 MineFleet® architecture. © Copyright, Agnik, LLC

Fig. 3 MineFleet® Onboard software currently supports many different hardware. (*Left*) One such example—MineFleet® Data Mining Platform (MF-DMP101) that hosts the MineFleet® Onboard software. Mainly used for the commercial fleet market. (*Right*) Another example device MV-200 that is designed for the consumer market. © Copyright, Agnik, LLC

and other third-party vendors) are also currently available for running MineFleet® Onboard. Figure 3 (right) shows a plug-n-play dongle that is mostly used for the consumer market.

Onboard Data Stream Mining Module

This module manages the incoming data streams from the vehicle, analyzes the data using various statistical and data stream mining algorithms, and manages the

transmission of the resulting analytics to the remote server. This module also triggers actions whenever unusual activities are observed. It connects to the MineFleet® Server located at a data center through a wireless network. The system allows the fleet managers to monitor and analyze vehicle performance, driver behavior, emissions quality, and fuel consumption characteristics remotely without necessarily downloading all the data to the remote central monitoring station over the expensive wireless connection.

MineFleet® Server

The MineFleet® Server is in charge of receiving all the analytics from different vehicles, managing those analytics, and processing them further as appropriate. The MineFleet® Server supports the following main operations: (1) interacting with the onboard module for remote management, monitoring, and mining of vehicle data streams and (2) managing interaction with the MineFleet® Web Services. It also offers a whole range of fleet-management-related services that are not directly related to the main focus of this paper. The server is connected with a relational database management system where it stores the analytics received from the vehicles in the fleet. All the onboard diagnostic provisioning and updates are performed over-the-air. Using an easy-to-use Web-based interface, members of the support team from Agnik and its resellers perform these over-the-air operations.

MineFleet® Web Services

This module offers a Web-browser-based interface for the MineFleet® analytics. It also offers a rich class of API functions for accessing the MineFleet® analytics which in turn can be integrated with third-party applications. Figure 4 shows one of the interfaces of the MineFleet® Web Services. MineFleet® is currently offered by many vendors that have already integrated their Web-based mobile resource management product with the MineFleet® Web services.

Privacy Management Module

This module plays an important role in the implementation of the privacy policies. This module manages the specific policies regarding what can be monitored and what cannot be. It also allows the fleet manager to create an environment where the MineFleet® technology can be used for saving money, sharing benefits without violating the privacy of the drivers.

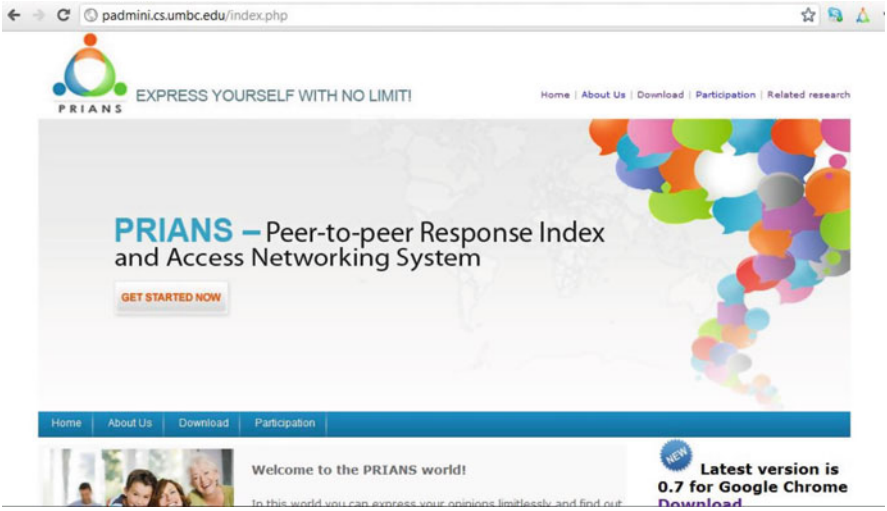


Fig. 4 PRIANS Web site

Mobile Interfaces

MineFleet® technology is now an integrated Smartphone Apps. It supports both Bluetooth and wide-area wireless modem (GSM/GPRS and CDMA)-enabled onboard devices that can be interfaced with Smartphones. As you walk into your car, your phones now start monitoring your car and offer various vehicle health updates, parental control tools, and varieties of location-based services.

2.3.2 PRIANS for P2P Response Indexing and Access Networks

We have been working with P2P systems for close to 10 years now. We have been developing asynchronous algorithms for P2P data mining applications. We have also been developing P2P data mining systems for several years. One of our early applications was the PADMINI system [17] that allowed highlighting a particular portion of the text on any Web site and tagging it with a set of labels. The data was stored in a distributed manner in a P2P network. The distributed labeled text was used for learning classifiers for future labeling of texts on the Web. My ex-graduate students Haimonti and Tushar [18] explored some of these applications (Fig. 5).

Over several years, the concept evolved. We decided to further simplify the application. We built a simple browser plug-in that allows you to type in comments on any Web site that you visit. The data is again stored in a P2P network. The distributed data is analyzed using asynchronous P2P data mining algorithms, and the results drove many of the features such as hot topics discussed over the Internet. Figure 6 shows a screenshot of the PRIANS plug-in. Once you visit a Webpage and

Fig. 5 Agnik's Smartphone Apps for vehicle, driver, and location analytics. © Copyright, Agnik, LLC

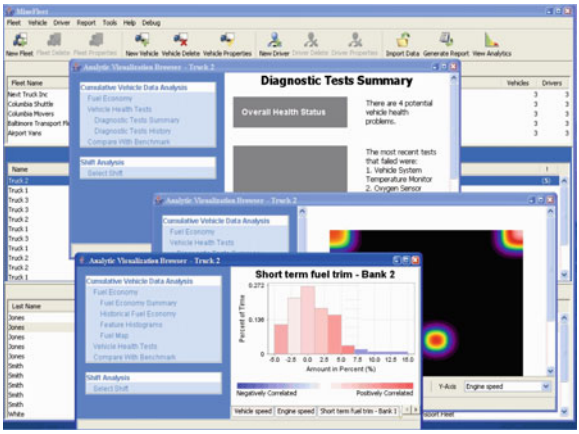


Fig. 6 MineFleet® Server. © Copyright, Agnik, LLC

you feel like commenting on the content or posting a link on that Web site, you simply click on the icon for the PRIANS plug-in posted on the right-hand corner of the browser, pull down the interface, and comment. If you want to share that with your friends through Facebook or Twitter, you can do that. PRIANS offers many interesting features. However, the main story is something bigger than these individual features (Fig. 7).

PRIANS allows one to create content on the Web and own it using the distributed resources offered by a P2P network. It does not rely upon the traditional client-server architecture where the content you generate is the property of the business that owns the server. Currently popular social networking sites are examples of that. The content that you generate is the property of those businesses and that is not the case for PRIANS. The goal is to create a system that will allow creating content, owning it, and making it visible wherever you want on the Web.

Fig. 7 PRIANS plug-in for the Chrome browser

3 Lessons in Learning from Failures

My journey in the field of distributed data mining had many challenges and failures. They span from tough reviews in a conference to the difficulty of taking a product to the market. I recall that in the late 1990s we used to get reviews that would fundamentally doubt the need for developing data mining algorithms in a distributed environment. We would see questions like why we cannot collect all the data in a central location and then analyze it. Back in those days, sensor networks, P2P computing, cloud computing were not that fashionable particularly for the data mining community.

Here are a few important things that I learned during this journey:

1. Convince yourself first and then believe in yourself.
2. Continue to challenge yourself as you explore new science, develop new technology, and offer new solutions.

Often in academia, we see researchers do research and leave it up to the practitioners to take it to practice. I am a strong believer of doing otherwise—explore, develop, and take it to practice. I felt that at least in my area of research, unless we check out these different facets, none of the individual steps may be complete. For example, when I got involved with Agnik and started building the MineFleet® system, I learnt many new things which exposed the holes in my

understanding of the distributed data mining field and applications. I understood better why certain assumptions are okay to make and some are not, why a feature of a product was popular in the market, and why some other features would never fly. This really helped me better pose the research problems and create solutions that make better sense.

4 Current Research Issues and Challenges

Despite the continued effort of the community for more than 15 years, the field of distributed data mining offers many challenges today on both the algorithmic and applications front. Some of these issues are discussed below:

1. Algorithm development: Most of the distributed data mining algorithms rely upon distributed techniques for computing various types of primitive aggregates such as sum, average, max, and L2 norm. Usually, such distributed primitive aggregates are used as a building block to design more complex distributed data mining algorithms. For example, if you are interested in computing a covariance matrix from distributed data, then you may reduce the covariance computation to a set of sum computations and then use the distributed algorithms for sum computations to solve the problem. We have several communication-efficient deterministic and approximate algorithms for computing some of these primitives. However, for many others, we need more efficient algorithms. For example, consider the problem of computing the similarity matrix from distributed data. Each site has a set of data tuples, and we need to compute the similarity matrix where the (i, j)th entry of that matrix represents the similarity between the ith and the jth tuples. This is an important problem that shows up in many applications. We need efficient algorithms for computing such quantities.
2. Applications development: The field of distributed and ubiquitous data mining requires more real-life successful applications. Although, the technology has been used in different research applications in industry, there are only a handful of commercial companies and products that are fundamentally based on distributed data mining technology. With the growing market for Apps for Smartphones, location-based services, and cloud computing, there should not be any shortage of ideas that entrepreneurs can come up with.

5 Summary

Distributed data mining algorithms fundamentally face the abstract problem of computing a collection of functions from distributed data using decentralized resources. Sometimes, the target functions to be computed are defined ahead of time. Sometimes, they evolve out of the local interactions among the participating

entities. This is a common process in many natural and artificial complex systems. Complex systems such as ant colonies, school of fishes, human society, and many others sense and process data in a distributed manner. Most natural and artificial systems do not count on a completely centralized system to process the data observed in a distributed manner.

The field of distributed data mining is fundamentally motivated by such observations. Although there will be plenty of centralized applications that will work very well in many domains, there will probably be many other applications where distributed solutions will work a lot better. A growing number of commercial applications of distributed and ubiquitous data mining technology is starting to corroborate this observation.

Fifteen years ago, when I started this journey, I did not expect that the world would be so connected and computing would be so ubiquitous as it is today. But I definitely thought that distributed computing will touch every major part of our life. That drove my interest on data analysis in such distributed and ubiquitous environments. Over the years, we faced many questions (including many tough reviews). However, our belief in our work, the desire to learn from real-life problems, and a continuous stream of smart graduate students, postdocs, and visitors in our lab made this journey a wonderful experience so far.

As distributed and ubiquitous data mining hits the commercial realm in a big way, more algorithmic challenges and the need for innovative applications will grow. We will need the next generation of researchers and practitioners who will continue to innovate and make the field richer.

Acknowledgments The author would like to thank all the organizations that funded his research over the last 15 years including the National Science Foundation, NASA, Department of Defense, Department of Energy, Caterpillar, IBM, and many other organizations. The current work is funded by the NASA grant and AF MURI grant.

References

1. H. Kargupta, I. Hamzaoglu, B. Stafford, Scalable, distributed data mining using an agent-based architecture, in *Proceedings of Knowledge Discovery and Data Mining*, ed. by D. Heckerman, H. Mannila, D. Pregibon, R. Uthurusamy (AAAI, Palo Alto, CA, 1997), pp. 211–214
2. H. Kargupta, K. Sivakumar, Existential pleasures of distributed data mining, in *Data Mining: Next Generation Challenges and Future Directions*, ed. by H. Kargupta, A. Joshi, K. Sivakumar, Y. Yesha (AAAI, Palo Alto, CA, 2004)
3. M. May, L. Saitta, Ubiquitous knowledge discovery: challenges, techniques, applications. *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence Series* **6202** (2010)
4. S. Bandyopadhyay, C. Giannella, U. Maulik, H. Kargupta, K. Liu, S. Datta, Clustering distributed data streams in P2P environments. Inf. Sci. **176**(14), 1952–1985 (2006)
5. S. Datta, C. Giannella, H. Kargupta, Approximate distributed k-means clustering over a peer-to-peer network. IEEE Trans. Knowl. Data Eng. **21**, 1372–1388 (2009)
6. W. Kowalczyk, M. Jelasity, A.E. Eiben, Towards data mining in large and fully distributed peer-to-peer overlay networks, in *Proceedings of BNAIC*, 2003, pp. 203–210

7. K. Das, K. Bhaduri, K. Liu, H. Kargupta, Distributed identification of top-*l* inner product elements and its application in a P2P network. TKDE **20**(4), 475–488 (2008)
8. S. Mukherjee, H. Kargupta, Distributed probabilistic inferencing in sensor networks using variational approximation. JPDC **68**(1), 78–92 (2008)
9. K. Bhaduri, H. Kargupta, A scalable local algorithm for distributed multivariate regression. Stat. Anal. Data Min. **1**(3), 177–194 (2008)
10. J.R. Quinlan, Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
11. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and regression trees* (Wadsworth, Belmont, CA, 1984)
12. K. Bhaduri, R. Wolff, C. Giannella, H. Kargupta, Distributed decision tree induction in P2P systems. Stat. Anal. Data Min. **1**(2), 85–103 (2008)
13. H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, On the privacy preserving properties of random data perturbation techniques, in *Proceedings of the IEEE International Conference on Data Mining*, Melbourne, FL, 2003, pp. 99–106
14. K. Liu, C. Giannella, H. Kargupta, A survey of attack techniques on privacy-preserving data perturbation methods, in *Privacy-Preserving Data Mining: Models and Algorithms*, ed. by C. Aggarwal, P.S. Yu (Springer, Berlin, 2008), pp. 357–380. Chapter 15
15. K. Liu, H. Kargupta, J. Ryan, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowl. Data Eng. **18**(1), 92–106 (2006)
16. H. Kargupta, K. Das, K. Liu, A game theoretic approach toward multi-party privacy-preserving distributed data mining, in *11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Polland, September 2007
17. T. Mahule, K. Borne, S. Dey, S. Arora, H. Kargupta, PADMINI: a peer-to-peer distributed astronomy data mining system and a case study, in *Proceedings of the Conference on Intelligent Data Understanding*, 2010
18. H. Dutta, X. Zhu, T. Mahule, H. Kargupta, K. Borne, C. Lauth, F. Holz, G. Heyer, TagLearner: a P2P classifier learning system from collaboratively tagged text documents, in *Proceedings of the International Conference on Data Mining (ICDM), Workshop on Mining Multiple Information Sources*, December, 2009
19. H. Kargupta, A. Joshi, K. Sivakumar, Y. Yesha (eds.), *Data Mining: Next Generation Challenges and Future Directions* (AAAI, Palo Alto, CA, 2004)
20. H. Kargupta, P. Chan, *Advances in Distributed and Parallel Knowledge Discovery* (AAAI, Palo Alto, CA, 2000)
21. H. Kargupta, K. Sivakumar, W. Huang, R. Ayyagari, R. Chen, B. Park, E. Johnson, Towards ubiquitous mining of distributed data, in *Data Mining for Scientific and Engineering Applications*, ed. by R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, R. Namburu (Kluwer, Dordrecht, 2001), pp. 281–306
22. H. Kargupta, B. Park, D. Hershberger, E. Johnson, Collective data mining: a new perspective toward distributed data mining, in *Advances in Distributed and Parallel Knowledge Discovery*, ed. by H. Kargupta, P. Chan (AAAI, Palo Alto, CA, 2000), pp. 133–184
23. B. Park, H. Kargupta, Distributed data mining: algorithms, systems, and applications, In *Data Mining Handbook*, ed. by N. Ye (Lawrence Earlbaum Associates, 2002).

Operational Security Analytics: My Path of Discovery

Colleen McLaughlin McCue

1 Early Years

My life is an example of the saying, "an apple doesn't fall far from the tree." My father worked in the early days of satellite imagery and the space program. My mother retired from a very successful career at theWashington State Department of Corrections. I work in the field of geospatial predictive analytics, creating models of violent crime and terrorism. While this seems like a logical career choice for a woman raised by a rocket scientist and a parole officer, my journey was neither planned nor straightforward. I am very fortunate, though, in that I absolutely love what I do and am "living the dream" as they say.

My "formative years" were spent in Downers Grove, Illinois, which is a suburb of Chicago. Although I did not appreciate it at the time, I was a real geek. My first love was science, and my high school coursework included advanced classes in most of the physical and natural sciences. The affinity for science did not translate to math, however, which in my mind was a necessary evil to the advancement of science. In my senior year of high school, I was selected to participate in the Fermi National Accelerator Lab's inaugural "Saturday Morning Physics" program. This was my first exposure to big science, and it was fantastic to see professional scientists pursuing their dreams and making discoveries that changed how we viewed the world. They told neat "science" jokes that enabled the geeky science students to feel cool because we understood them and generally created a great feeling of belonging and "fit" to this community. Dr. Leon Lederman organized the program and was an enthusiastic and energetic leader. He went on to win the Nobel Prize for work with neutrinos, and I have thoroughly enjoyed following his career and the success of the program.

C.M. McCue (✉)
GeoEye, Herndon, VA, USA
e-mail: mccue.colleen@geoeye.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_11,

My college career began at the University of Illinois at Chicago in the fall of 1981. I started as an engineering student but realized very quickly that I had gotten into the wrong car when the career train left the station. Immediately after the close of my first term, I withdrew from the University to regroup and consider my options. In an effort to stay engaged academically (and ensure my parents that I was not completely abandoning higher education), I enrolled in a couple of courses at the local community college. One of these courses was a general psychology course taught by a practicing clinician with a great sense of humor and some very interesting "strange but true" stories from his clinical practice. After completing the academic year at the community college, I made arrangements to return to the University of Illinois at Chicago as a psychology major.

Psychology and the behavioral sciences really appealed to my interest in understanding how things work, and I very quickly embraced physiological psychology and the neurosciences. My strong background in science also proved extremely useful—reinforcing this as a great fit for both my skill set and interest. Courses in abnormal psychology and deviance held a strong appeal to me, and I was particularly intrigued by the study of serial murderers. In one course, we studied the Hillside Strangler case in which a particularly clever psychologist was able to identify and successfully unravel the killer's attempt to deceive the investigative team. In a series of interviews, Dr. Martin Orne was able to demonstrate that Kenneth Bianchi was fabricating mental illness in an effort to support his plea of "not guilty" by reason of insanity. Subsequent discovery of psychology textbooks and police procedure manuals in the suspect's home confirmed his plan and provided insight regarding the source of his "illness." This ability to deconstruct behavior and identify "normal" trends and patterns made a huge impression on me and is something that has influenced my approach to behavior going forward. Years later, during a course on death investigation at the FBI Academy, I had the opportunity to listen to audiotapes of the murders that Bianchi and his cousin Angelo Buono Jr. perpetrated. The material was horrific and extremely difficult to experience, but I also began to understand the importance of being able to transcend the specific details of violent crime in order to identify general trends, patterns, and themes that could be used to segment, categorize, and better understand these crimes in an effort to generate a "profile" of the likely killer and apprehend them. That Dr. Orne was able to move beyond the horrific aspects of these incidents and identify the flaw in the killer's plan that broke the case still amazes me.

Dr. Orne's work also highlights the importance of being able to characterize and understand "normal" [1] patterns of crime in support of identifying crimes that differ or otherwise deviate from other crimes similar in offense category but that differ significantly in detail, execution, and frequently intent. Our experience suggests that these "abnormal" crimes tend to be at increased risk for escalation and pose a serious threat to our communities [2]. The use of anomaly detection to identify "riskier" patterns of crime, including indicators of hostile surveillance on a particular individual or target suggestive of a preoperational surveillance and attack planning, is something that would become increasingly important to my work later and would cause us to reprioritize resources in an effort to get in front of

specific crime patterns and series associated with an increased risk for escalation. Overall, this particular case underscored many of the elements required for effective behavioral analysis of violent crime that I could come to truly understand only later in my career.

2 My Life as an Early Career Researcher

By the time I graduated, I already had taken several graduate level courses in physiological psychology, psychopharmacology, and neuroanatomy. I am not sure when or even if I ever actively decided to go to graduate school. It just seemed like the logical extension of the program of study that I was pursuing. I was accepted into the doctoral program at Dartmouth College and arrived on the campus for the first time in the fall of 1985 when I matriculated, having decided to attend Dartmouth "sight unseen." While my approach to selecting a program was relatively uninformed, I was incredibly fortunate to receive my graduate education from a truly exceptional program. My current field of work did not exist when I received my training, and it is a testament to the program at Dartmouth that I was able to make several, extremely radical lateral moves during the course of my career. It has only been in retrospect that I have appreciated the high-quality training that I received in scientific method and process.

While he was not my thesis advisor, Dr. Michael Fanselow had the biggest impact on my approach to science and my career going forward. I remember making a joke once that Dr. Fanselow could attend a seminar on astrophysics and by the end of the seminar would not only ask an extremely intelligent, highly insightful question, but would inevitably be able to link the material in the seminar back to Pavlovian conditioning, his area of expertise, and cite at least three experiments that Pavlov had completed earlier and published in 1927 in his seminal body of work, "conditioned reflexes [3]." Although this is an absurd example, it underscored Dr. Fanselow's ability to transcend the details of a specific program of research or scientific discipline and understand it within the context of good scientific method and process. As in the Bianchi case, I was impressed by and sought to emulate those who were able to identify and characterize common trends, patterns, and themes and use this information to create models that could be applied to novel examples in support of better understanding and insight. The ability to see universal trends, patterns, and constructs within and across nature was something that fascinated me and was an important theme that supported some of my more innovative work. I have not only been very open about, but have taken pride in the fact that almost everything that I have done in operational security analytics has been pulled directly from existing science, law enforcement domain expertise, and business analytics. I would later use concepts including optimal foraging, market segmentation, and retail supply chain management to describe human predatory behavior, motive determination, and public safety resource deployment.

During graduate school, my parents continued to encourage and support me actively—my father in ways that would become particularly meaningful only years later. In an effort to find something of common interest between an engineer with a modeling and simulation background and a budding neuroscientist, my father immersed himself into the artificial intelligence domain, which was developing rapidly at that time in parallel with advancements in cognitive neuroscience, which fell into my purview. As these two related yet disparate fields developmentally leapfrogged over and passed each other, my father remained current in the literature and continued to send me books and articles on the subject in support of this common area of interest. He even loaded an early copy of "Brain Maker" on his home computer in an effort to better understand and operationalize the concepts. Years later, I would have the opportunity to actually use a neural network model in a particularly challenging crime analysis problem, already primed after years of informal tutoring by my father.

I successfully defended my thesis and was awarded a Ph.D. from Dartmouth in 1989 and immediately moved to Richmond, Virginia, to start a postdoctoral fellowship at the Department of Pharmacology and Toxicology at the Medical College of Virginia, Virginia Commonwealth University. By that point in time, I was working on a relatively well-developed program of research on pediatric pain and perinatal exposure to drugs of abuse. The Medical College of Virginia had strong ties to and funding from the National Institute of Drug Abuse at the National Institutes of Health and was an excellent location for the next phase of my training. During my postdoctoral fellowship, I acquired additional skills in molecular biology and was involved in some very exciting basic research, including the analysis of the developmental expression of various receptor systems and DNA-binding proteins. I also extended this work to include complementary clinical research on pediatric pain management.

At this point, I was on a pretty clear path to a career in basic science research in an academic setting. I was spending most of my time in the lab designing experiments, supervising the technical staff, publishing papers, and supporting graduate training. Again, basic science research really appealed to my ongoing drive to "create science" and participate on the ongoing effort to unlock the secrets of nature and better understand how the brain worked. Albert Einstein was quoted as saying that his "sense of God [was his] sense of wonder about the universe," and I have always shared that "holy curiosity." Even today, I marvel at the unfathomable complexity of nature achieved through the unique combination of relatively simple, common elements. The human genome is captured in unique combinations of four common nucleotides: adenine, cytosine, guanine, and thymine. Similarly, pain transmission, human movement, drug action, memory, and emotion all are mediated through a unique combination of neurons, neurotransmitters, receptors, second messengers, and other cellular components. It is in the unique combinations of these common elements that the complexity is achieved. Looking beyond this complexity to identify common elements and patterns that could be used to organize, segment, and understand nature was emerging as a constant theme in my view of the world and approach to science. It also began

to manifest itself functionally in good pattern recognition skills, which would serve me well going forward.

3 Career Path Change

By the time that my fellowship began to wind down and I started considering the next opportunity, I had developed strong ties to Richmond, which required that I expand my career path somewhat in order to stay in the community. The Virginia Department of Juvenile Justice had recently received a federal grant to establish a correctional facility exclusively devoted to substance abuse treatment for drug-involved juvenile offenders. One goal of this program was to incorporate rigorous program evaluation in an effort to identify and effectively document program elements that were effective or promising. The opportunity to join this program from the beginning and support the development and execution of their evaluation effort was very exciting. It represented a significant deviation from my current career path, but the opportunity was too tempting to refuse.

Almost immediately, it became apparent that there were vast, untapped data resources just waiting to be explored. The chief psychologist, Dr. Dennis Waite, had established a comprehensive database years previously that included data on each offender's psychological, medical, educational, and criminal history, as well as the results from a battery of tests, assessments, and interviews conducted at intake. The amount of information was staggering, and I immediately began to explore it in an effort to better understand the nature and quality of the data. Consistent with the fact that substance use and abuse effectively spanned the public health and public safety domains, I also retained adjunct faculty status at the Medical College of Virginia through 2004, which was extended to include appointments in the Departments of Surgery, Emergency Medicine, and a few years later, Pediatrics. This provided unique opportunities for multidisciplinary research, data sharing, and cross training, and eventually the "Cop & Docs" program, which received the IACP Community Policing Award in 2002.

Shortly after taking the position, I encountered one of the department physicians at a luncheon. Dr. Patricia Reams casually mentioned to me that they were seeing an increased number of juvenile offenders being incarcerated with previous firearm injuries. As I recall, she said that they were encountering "a lot of kids with holes in them" and wondered whether it was something that I could look into as part of my research tasks. Dr. Reams had been instrumental in recruiting and hiring me, so I decided to pursue her question, not realizing at the time that it would totally change the direction and overall nature of my career.

Incorporating her request into the existing program of research, I began to study the relationship between drugs, guns, and violence among juvenile offenders. In some of our earliest work, we found that different people were shot for different reasons under different circumstances. The "I was in the wrong place at the wrong time" explanation quickly was replaced by the "I was in the wrong place at the

wrong time doing the wrong things with the wrong people" description of the relationship between involvement in juvenile offending and violent injuries [4]. Further extending the research, we found that that constellation of factors associated with an increased risk for violent victimization was specific to the pattern of offending [5]. In other words, juvenile drug sellers were shot for different reasons than violent juvenile offenders, than their girlfriends, and so on. While this might seem obvious, it seriously called into question generic "gunshot wound survivors" groups that were being created in hospitals responding to the epidemic of violence in their communities. By bringing everyone together for "treatment," we ran the risk of combining victims with perpetrators, as well as rival gangs and other factions. Similarly, a quick review of a "drug-involved offenders" program revealed that users and sellers were being combined. While this was expeditious from a scheduling perspective, and created unique business opportunities for the sellers, it was not optimal for the users and did not align well with the treatment goals of reducing drug selling and use. Therefore, by segmenting these various criminal populations into functionally related, relatively homogenous groups, associated patterns of risk emerged that could be effectively leveraged in support of targeted approaches to treatment and reinjury prevention.

Some of our early results did not align well with the existing literature, though. For example, we found that juvenile drug sellers were less likely to use drugs, as compared to other juvenile offenders. This stood in marked contrast to the existing literature, which suggested that sellers got involved in selling drugs to support personal use. In light of this and other findings, it became increasingly important to identify domain experts who would help me vet and validate my models, so I turned to the local police department. They immediately confirmed that drug selling is an extraordinarily predatory criminal environment. Those who use what they are selling do not last long. The findings from the other studies may have reflected biased recruitment methods and/or the incentives for participation that skewed the research samples in favor of folks who needed money most and were willing to divulge their involvement in criminal activity for cash. Over time, I developed a relationship with the Richmond Police Department and began to have regular meetings with the officer in charge of the Violent Crimes Division, the late Captain Donnie Robinson, to review, vet, validate, and refine the models that we were developing. As we continued to meet, I realized that Captain Robinson had an uncanny ability to "predict" crime trends, patterns, outcomes, motives, and suspects. I was pretty sure that he did not have a crystal ball in his office and began to understand that crime, even the most aberrant patterns of violent crime, tended to be relatively homogeneous and predictable if viewed in the proper context. Over time, my "predictions" became increasingly accurate, and I started using this "institutional knowledge" as the starting point for our research—operationalizing and extending the "gut instincts" and domain expertise of my new law enforcement colleagues.

As an extension of my research into this area, I started going out to crime scenes in an effort to better understand the data, including data quality, and also to understand the end user requirements and constraints. Again, this activity went

back to my graduate training. One of the most important lessons that I learned during my graduate training was the importance of "knowing" your subjects. It is absolutely essential to observe the subjects and actively consider alternate hypotheses for the outcome. It became increasingly important in my "new" career to get out in the field to understand where the data came from, the limitations and general ugliness, as well as the end user constraints and requirements. It also gave me credibility with the end users. Over time, they gained respect for what I was doing, and I was able to build trust, which further enhanced and enabled my "education" as my colleagues shared even more of their domain expertise, tacit knowledge, and insight. During this time, I also had the opportunity to meet with members of the FBI Child Abduction and Serial Killer Unit (CASKU), otherwise known as the "profilers." They provided enormous support to the research that I was doing on juvenile murderers, which afforded me unique insight into their process and methods. Over time, one thing that I realized was that their methods, while truly amazing in their accuracy and insight, generally went back to good case management, solid pattern recognition skills, and case-based reasoning. After one lively discussion regarding my newfound insight, I proposed that we could leverage their methods and process and use mathematical algorithms to model violent crimes. Somewhat skeptically, they encouraged me to try, and the result was that we were able to identify and differentiate drug-related homicides from all others in juvenile homicides using discriminant analysis [6, 7]. This represented the beginning of what would become a serious and long-term engagement in the use of data mining and predictive analytics to study and model violent crime.

4 Using Data Mining

Data mining has been described as a process of confirmation and discovery—confirmation of the things that we know (or think that we know!), extension of this existing knowledge base, and discovery of new knowledge and insight. I really like the "Wal-Mart emergency response plan"[1] as a real-world illustration of the concepts of confirmation and discovery and use it frequently when I teach. Wal-Mart is a true "analytic competitor." They have been tracking point of sale data

[1] Regarding the Wal-Mart emergency response plan, I had the opportunity to speak at the Wal-Mart headquarters in Bentonville, Arkansas, and was somewhat intimidated about including this example but am very glad to say that they not only confirmed it but seemed very pleased to learn how their experience has been able to provide inspiration and method to the public safety and national security analysis community. This model works extremely well for teaching and has almost become viral in the law enforcement community. I have seen others use it in their lectures and even attended a briefing in Washington, DC, where the speaker attributed the original discovery of this relationship to the LAPD Chief Charlie Beck and my coauthor, reporting that he was collecting data on Pop-Tart sales from the local Wal-Mart stores in advance of bad weather moving into Los Angeles [10].

for a very long time and using this information to optimize their supply chain. They have mastered "just-in-time" supply chain analytics and have the ability to flex rapidly in response to changing events. Over time, they have found that sales of certain items increase in advance of bad weather. These items include duct tape, bottled water, and Pop-Tarts—strawberry Pop-Tarts, to be precise. Bottled water and duct tape make sense—people need to ensure that they have a supply of potable water should municipal supplies be disrupted, and both FEMA and DHS have told us to buy duct tape. The strawberry Pop-Tarts, on the other hand, do not really fit the model. This finding is surprising. We can speculate that they are easy to store and prepare, even after a power loss, but it does not really matter. The only important consideration from the Wal-Mart perspective is that if the weather is bad, then people will want to purchase Pop-Tarts, and they need to adjust their supply chain accordingly. From a data mining perspective, the increased sales of water and duct tape is "confirming." The increased sale of Pop-Tarts is surprising and represents "discovery." Most of my work has been related to the confirmation and extension of existing knowledge in the field, although there have been more than a few surprises over the years. This type of discovery can be particularly exciting, particularly if the insight gained is actionable and can be used to improve public safety and security.

In 2000, I had the opportunity to work for the Richmond Police Department. As I recall, the "recruitment" took place at a crime scene. It was in the middle of the night, cold and rainy, and I believe that there was some suggestion that I would not be able to get back into the warm police cruiser until I had made a decision. My husband, NCIS Supervisory Agent Richard J. McCue, encouraged me to take the position. I was concerned about making another huge lateral move in my career, but he correctly pointed out that the insight and domain expertise that I would gain by working for a law enforcement organization would be invaluable. The insider knowledge that I would acquire and relationships that I would make would more than offset this minor "course correction" in my career path. My husband has been my strongest advocate and greatest critic, and I need both. His counsel always has been insightful, and I know that he has my best interests at heart, so I accepted the position based on his encouragement. While I did not know it at the time, my tenure with the Richmond Police Department would have a profound impact on my career and would include the 9/11 attacks, the DC Sniper series, the anthrax series, and a few other high-profile child abductions.

Like most people, I recall exactly where I was when I first heard about the attacks on 9/11. In my case, I was in Washington, DC, attending "advanced intelligence analysis" training that included education on pencil and paper methods of analysis. In the days and weeks that followed the attack, I became increasingly frustrated because I knew that we could do better. At that time, I had been working on creating models of violent crime in an effort to enhance investigative efficacy and pace. As part of that research, I had been studying the use of advanced analytics in other professional domains, including medicine and business, and knew that companies like Amazon and Wal-Mart had better analytic capabilities than most, if not all of the law enforcement agencies in our county. I also knew that none of the basic science research that I had been doing in this area was appropriate for direct

transfer to the operational environment and began a program of research and development on the translation of advanced analytics output to the operational public safety and security environment that would continue to this day.

One of the first projects involved the development of a model that could be used to support information-based approaches to patrol deployment. Taking a cue from the "just-in-time" supply chain analytics used by retail organizations, we began to view law enforcement patrol deployment as a resource allocation problem. Ideally, we would like to have a police officer available when and where they are likely to be needed without deploying them at times and to locations where they are not likely to be needed because this is wasteful and also can be perceived as intrusive. At the time, deployment decisions were being made based on historic precedent, gut instincts, requests from citizens that wanted to see a "cop on every corner," and a variety of other reasons that generally had no grounding in crime trends and patterns. Following this lead from the supply chain analytics community, two measureable outcomes for police deployment emerged. First, if we could identify when and where police resources would be needed, then we could proactively deploy them to those locations. Ideally, the presence of police officers will serve as a deterrent—effectively preventing crime, which could be documented in reported crime statistics. In the event that a crime did occur, however, the officers will be prepositioned closer and will be able to respond more rapidly, increasing the likelihood that they will be able to catch the bad guy. Again, this could be documented in arrest statistics. Integrating additional information regarding crime type extended our "just-in-time" policing concept into a "risk-based deployment" model that could be used to support patrol deployment decisions [8].

We had our first test of the "risk-based" deployment concept on New Year's Eve 2004 [9]. During the late 1990s and early 2000s, Richmond, Virginia, struggled with serious violent crime, and New Year's Eve historically was associated with numerous citizen calls for service for random gunfire in the community. At the police department, leave books were closed leading up to the holiday, and everyone was expected to work, from the newest line staff to the supervisors and command staff and even the chief of police. Deployment was heavy, and there was a perceived need to fill the streets with officers in an effort to create a police presence that would ideally reduce crime, but at a minimum, demonstrate to the community that we were aware of the problem and trying to address it. We had enjoyed some early wins using data mining and predictive analytics to surface actionable crime trends and patterns, so the police chief, Colonel Andre Parker, asked us to review the New Year's Eve deployment plan to see if we could find anything actionable that could be used to support his effort to make a difference in the holiday.

After reviewing the data, we identified a relatively small number of areas associated historically with increased activity on New Year's Eve and also were able to document the fact that most of the calls occurred within the hour surrounding midnight. To provide an additional layer of "protection" to our model, we also pulled in some additional areas that had been associated with recent outbreaks of violence but ended up with a relatively short list of focus areas. The major in charge of patrol services, Dave McCoy, created a deployment plan

based on our analysis. After providing heavy coverage for the areas identifying as high likelihood for complaints, he realized that we had 50 "extra" officers that would not be needed. Ultimately, it was decided to use the deployment plan developed based on the model and give 50 officers the night off. To say that I was nervous that night would be a huge understatement! I stayed awake all night completely worried about what might happen if there was anarchy on the streets, and they were unable to respond effectively because they were short handed because the psychologist/data miner told them to let 50 officers take the night off. In the end, though, everything went extremely well. After analyzing the results, we were able to document marked decreases in citizen complaints for random gunfire—a 47% reduction compared to the previous year—effectively supporting the hypothesis that if you could identify when and where bad things were likely to happen and proactively deploy resources to these locations, then you could reduce crime through a strong police presence. We also documented a 246% increase in the number of weapons seized during the initiative, which was consistent with the ability to respond more rapidly because the resources had been prepositioned where they were likely to be needed. There also was an unintended benefit to this effort, in that we were able to save $15,000 in personnel costs alone. While this makes complete sense in retrospect, it never had occurred to us that effective resource deployment essentially represented optimization, which enabled us to do more with less.

In the end, this initiative resulted in a win, win, win, win for everyone involved. First and foremost, the increase in public safety, as defined by decreased citizen complaints for random gunfire and increased recoveries of weapons, addressed directly the goals of risk-based deployment. Second, the optimization of resources resulted in a saving of $15,000 in direct personnel costs alone. This figure did not include the additional, undocumented savings associated with police time related to arrest and processing of the suspects, court costs associated with prosecution, jail and prison costs related to incarceration, and the incalculable costs associated with the fear of crime and lost opportunity associated with high levels of crime, particularly violent crime in a community. Related to this point, there was a direct win for the citizens living of the areas of Richmond normally inundated with random gunfire on New Year's Eve. Finally, there was a huge win for the Richmond Police Department. Fifty officers were able to take the night off and welcome the New Year with their friends and families. Those who were working that night were busy responding to calls and making arrests; they were making a difference. When it was all over, they knew that their presence on the streets was directly responsible for the marked reductions in crime documented. For many in law enforcement, the real benefits of the job are not money. Rather, the police officers that I know and have worked with joined the force in order to fight crime, make a difference in the communities that they serve, and change outcomes for the citizens that look to them to serve and protect. The translation of the just-in-time supply chain model to risk-based deployment enabled us to proactively place them when and where they were needed so that they could engage the community and make a difference.

Validation of the risk-based deployment model on New Year's Eve 2004 would change the way that we view police deployment going forward and would be replicated and extended in numerous other departments across the country. Ultimately, it would enable us to do more with less, which in a few short years would be turned into doing almost everything with next to nothing as the recession significantly impacted law enforcement budgets—forcing many of them to layoff sworn personnel for the first time in their history [10]. This initiative also was a big win for me personally as it operationally validated many of the concepts that we had been developing. It was the first real demonstration that data mining and predictive analytics could be used in the operational public safety environment to change outcomes and make our communities safer.

5 Data Mining Textbook

With few exceptions, my mother was continuously employed from the time when I was in grade school until her retirement a few years ago—effectively managing both family and a very successful career in community corrections. In doing so, she led by example. After 9/11, my husband volunteered for and accepted multiple overseas assignments, leaving me as a functional "single" parent of five children. He had special skills and training in counterterrorism, executive protection, and force protection, and we all were very proud to support him in his work. Given the size of our family and the budget constraints associated with two civil service salaries, though, my opportunities for entertainment outside the house were severely limited, so I used the time to write a textbook on the use of data mining and predictive analytics in the public safety and security setting [8]. I had been working on ways to translate our work using data mining and predictive analysis in the applied law enforcement setting and wanted to share these with our professional community. The working title of the book was "Data Mining for Doorkickers," and I was completely committed to writing a book that could be easily read and used by both the analyst and operator communities and would also represent a starting point for the truly collaborative work that I believed was possible between these two groups.

In my book, I mapped the analytic process that I developed, "Actionable Mining and Analysis," which extended the Cross Industry Standard Process for Data Mining (CRISP-DM) [11] to address the unique challenges and requirements associated with data mining in the operational public safety and security environment. Specifically, the emphasis on operationally relevant and actionable analysis included special provisions for data collection and preprocessing, evaluation of the models, and the generation of output that could be translated directly to and used in the operational environment for decision support. This emphasis on "operationally relevant and actionable" represented development that was necessary to the translation of a successful program of academic research, creating models of violent crime to analytic process and output that would support decision making in the

operational public safety and security environment. Finding a way to translate this work to the applied setting was not trivial.

One of the first tasks was structuring the process to accommodate the fact that data were not necessarily available when they were needed. For example, my earlier academic research generally had been conducted on closed cases, which enabled us to create some relatively accurate models using victim, suspect, and crime scene data. In the real world, however, the goal generally is to create a short list of suspects in an effort to focus the case. Therefore, any model requiring suspect information would have limited value. Similarly, structuring the data to match existing boundaries and schedules was important to facilitating direct translation to the operational environment and related deployment decisions.

A second challenge included evaluation of the models. Overall accuracy almost never represents a good method given the relatively infrequent nature of crime. While this might seem to be a good thing, the ability to effectively model infrequent events creates unique challenges that generally preclude the use of overall model accuracy as an effective metric. For example, when creating a model for the escalation of armed robberies into aggravated assaults, we found that only 3% of these robberies actually resulted in the victim being shot at or actually assaulted [12]. While this would be a good thing if you are the victim of an armed robbery, it requires different thinking regarding how best to evaluate an associated model. For example, we could create a model that would say that an armed robbery will never escalate and be correct 97% of the time. While this would be enviable accuracy for a public safety-related model, it would have very little utility to law enforcement decision makers. Therefore, we needed to consider the nature, direction, and magnitude of the specific errors and their potential consequences to public safety and related operations. It is for this reason that the unique attributes of each pattern of crime must be considered and that we pay special attention to the confusion matrix generated as we evaluate models within the context of the specific public safety requirements and constraints.

This point underscores the fact that there are no “free lunches” in security analytics, meaning that it is unlikely that we ever will develop a program, technology, or specific algorithm that will effectively address all crimes and all hazards given the unique pattern of associated requirements and constraints associated with different patterns of crime and related threats. For some patterns of offending, merely increasing the accuracy of our models above chance represents a marked improvement in decision making. Other models, including those that are used to establish investigative direction or surface possible suspects, require a greater degree of accuracy given the consequences, but also can be more opaque given the model deployment requirements. A related challenge is that we rarely receive nonevent data, which is like trying to study disease without a healthy comparison group.

Creating a reliable model is only the beginning in operational security analytics, though. In many cases, the real challenge only begins when we try to generate analytic output that can be translated directly to the applied setting and used to support decision making. This requirement for “operationally relevant and

actionable" output would represent our greatest challenge, but also the area where we truly differentiated ourselves from other groups working in this space. Our first attempts at this were well intentioned but clumsy and resulted in several early failures. Like any good research program, though, we learned as much, if not more from the losses, which then were translated into improvements in the next version.

One of our first approaches to model deployment involved the creation of an interface with a series of pull-down menus that enabled the end user to input various data points associated with relevant variables. The generated score was deployed in a slick output page that included the department's seal and an image of a card from the game "Clue." We quickly learned that this method of delivering the model results represented a poor fit for deployment decisions and that one end user in particular was running every possible permutation of the model in an effort to create a "schedule" that included the various likelihood estimates in the created cells. This method underscored the importance of time and space in deployment decisions. Therefore, using this as a starting point, we then adjusted the models slightly and depicted the results in a mapping environment, effectively creating a heat map that conveyed the likelihood of future crimes as increased intensity on the map using the department's existing patrol deployment boundaries. This slight change in the depiction of our results was game changing. The field staff immediately reached back to the Crime Analysis Unit with enthusiasm and requested more of the same. They also were able to incorporate their domain expertise and tacit knowledge in the evaluation and interpretation of these results—effectively extending from the analytic product and using it to inform their decision making, tactics, and strategy going forward. From this process, we realized the importance of time, space, and the nature of the incident or threat in the creation of analytic product for the operational public safety and security environment. These elements were actionable, while almost everything else was not.

This early experiment in "geospatial predictive analysis" was crude at best—requiring the unnatural pairing of existing data mining and mapping tools, which resulted in some truly insightful and actionable analytic output—but placed constraints on both methods. Other teams were finding similar success in this area, though, including a team at SPADAC that was able to effectively integrate robust predictive analytics within the geospatial environment. By incorporating geospatial data as actual inputs into the models, additional fidelity and refinement of the models was achieved. This also enabled the analyst to effectively convey the nonrandom nature of crime, as well as the subtle selection process that criminals and terrorists engage in when selecting potential victims and target locations. I had the good fortune to join the team at SPADAC, now GeoEye, and support some very exciting work for public safety and national security clients that fully leverages the promise of geospatial predictive analysis in support of information-based approaches to prevention, thwarting, mitigation, and response.

6 Challenges and Success Factors

One common theme among data miners seems to be that good analysts are hard to find. Moreover, an intuitive data miner, particularly one with domain expertise, is exceptionally rare. In my experience, analysis still is not valued within the operational public safety and security environment, and it is not unusual to see analyst positions created as a form of career advancement for administrative personnel. Similarly, analyst positions in operational organizations frequently are filled by operational personnel assigned to "light duty," further underscoring the limited value that they place on these roles. While there are agencies that focus almost exclusively on analysis (e.g., NSA, CIA), those with operational and enforcement missions frequently discount the value of analysis. In many cases, there are organizational and structural impediments that limit the direct interaction between analysis and operations, which further limits the value that analysis can bring to decision making. In my opinion, one of the greatest gifts to the community has been Tom Davenport's text, "Competing on Analytics" [13]. I feel so strongly about the importance of context that I now encourage people to purchase Tom's book and read it before even considering the purchase of my own. Many of the concepts reviewed in his text have been distilled to the DELTA model, which stands for data, enterprise, leadership, targets, and analysts [14].

Looking back on my time with the Richmond Police Department, I now realize that we had some key elements in place that truly enabled our success. The first was leadership. I worked with and then for Colonel Jerry Oliver at the Richmond Police Department. His commitment to innovation created the type of enabling environment that was necessary to achieve the success that we were able to enjoy. He brought out the best in people who had the privilege of working for him and made great things happen. The second element working in our favor was the enterprise. Many members of the command staff were pursuing graduate degrees, which sometimes brought an element of the absurd to the crime scenes as we would be standing next to a dead drug dealer and someone would request a quick tutorial on the difference between nominal, categorical, and interval data. From my perspective, though, it was an amazing opportunity to significantly expand our research "team" and extend our capabilities even further into the organization. We also had a "target-rich" environment in which to work. Richmond regularly found itself in the top per capita crime rate listings. While this was terrible for the citizens, it provided us with a "tremendous opportunity to succeed." Our data were not great, but our analytic abilities exceeded most federal agencies given the support that we received from SPSS. Several years later, Curtis Abel said to me that making SPSS analytic resources available to the Richmond Police Department was one of the best decisions he ever made and I agree. By providing us with tools and extensive technical assistance, we were able to create the science and novel analytic strategies that drove development of the public safety resource optimization solution that ultimately received the 2007 Gartner Business Intelligence Excellence Award.

I left the Richmond Police Department in 2004 and have had the pleasure of serving on teams at a series of nonprofit and then commercial consulting firms.

Through these engagements, I have had the opportunity to work on truly hard problems. I thoroughly enjoy finding the "word problem" and then solving it, and the projects that I was able to support have included engagements with law enforcement, the Department of Defense, the Department of State, corporate security, and other organizations. The problems are always different and generally hard, but they almost always have come back to the identification of actionable trends and patterns that can support information-based decisions regarding resource allocation in support of prevention, mitigation, thwarting and response. The discovery aspect of science and data mining is what truly fascinates me and has kept me energized and engaged over the years. To be able to formulate questions and gain insight into nature, particularly human behavior, is what I most enjoy. I have enjoyed some amazing opportunities and professional experiences and have been privileged to be involved in work that changed outcomes for people. When I worked for RTI International, I was fond of extending the organization's mission of "improving the human condition," to include, "by keeping people safe."

One unique challenge of working in the operational public safety and national security environment is the ongoing nature and the urgency of the work. Crime occurs at inconvenient times and inconvenient locations. Failure to make quick progress in a case increases the likelihood that it will not be solved and that the bad guy will go on to commit another crime in the future. After being shot, Bob Marley was quoted as saying, "The people who were trying to make this world worse are not taking the day off. Why should I?" In my experience, death does not take holidays. This point was underscored by a colleague a couple of years ago. We were working on the problem of improvised explosive devices (IED). IEDs are responsible for an inordinate number of injuries and death in Iraq and Afghanistan with absolutely horrific consequences for those unfortunate enough to encounter these instruments of death and destruction. We were discussing a specific analytic task related to identifying the emplacement of these devices in an effort to surface them proactively and reduce their impact on our deployed forces. Toward the end of the discussion, someone asked about the suspense date or deadline for the assignment. There were a couple of dates offered when one of my colleagues quietly suggested that we should try to complete the assignment before the next young person was killed. Ultimately, that is the goal of operational security analytics. To identify trends, patterns, relationships, associations, sequences, and whatever else is required to provide the insight necessary to prevent, thwart, mitigate, and respond more effectively—to use data mining and predictive analytics to change outcomes.

Confucius has been quoted as saying, "Find a job that you love and you'll never work a day in your life." By any measure, I have enjoyed an amazing career that has been filled with tremendous opportunities, as well as opportunities for tremendous growth. Whether by chance or divine intervention, my career path has been defined by a series of apparently disparate and unrelated opportunities and positions strung together by an ongoing desire to contribute to our knowledgebase, create insight, and to serve. It has been an amazing ride thus far and I cannot wait so see what comes next!

References

1. C. McCue, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis* (Butterworth-Heinemann (Elsevier), Burlington, MA, 2006)
2. C. McCue, G.L. Smith, R.L. Diehl, D.F. Dabbs, J.J. McDonough, P.B. Ferrara, Why DNA databases should include all felons. Police Chief **68**, 94–100 (2001)
3. I.P. Pavlov, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford University Press, London, 1927) (translated by G.V. Anrep)
4. C.R. McLaughlin, S.M. Reiner, B.W. Smith, D.E. Waite, P.N. Reams, T.F. Joost, A.S. Gervin, Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). Am. J. Public Health **86**, 751–752 (1996)
5. C.R. McLaughlin, S.M. Reiner, B.W. Smith, D.E. Waite, P.N. Reams, T.F. Joost, A.S. Gervin, Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. Free Inq. Creat. Sociol., Special Issue: Gangs, Drugs and Violence **24**, 157–165 (1996)
6. C.R. McLaughlin, J. Daniel, T.F. Joost, The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. J. Forensic Sci. **45**, 349–353 (2000)
7. McCue, C. Juvenile Murderers, in *Managing Death Investigation*. ed. by Arthur E. Westveer (US Department of Justice, Federal Bureau of Investigation, Washington, DC, 2000), pp. 481–489
8. C. McCue, *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis* (Butterworth-Heinemann (Elsevier), Burlington, MA, 2006)
9. C. McCue, A. Parker, P.J. McNulty, D. McCoy, Doing more with less: data mining in police deployment decisions. Violent Crime Newsletter, US Department of Justice **1**(Spring), 4–5 (2004)
10. C. Beck, C. McCue, Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? Police Chief, November 76 (2009). http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=1942&issue_id=112009. Accessed 17 May 2012
11. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: *Step-by-Step data mining guide*. CRISP-DM Consortium (1999)
12. C. McCue, P. McNulty, Police Pursuits from a Research Perspective: A Word about Data-Mining from Dr. Colleen McCue and Paul McNulty. Cutting Edge of Technology: Managing Police Pursuits. Findings from IACP's Police Pursuit Database. Executive Brief, Winter 2004, pp. 7–9
13. T.H. Davenport, J.G. Harris, *Competing on Analytics: The New Science of Winning* (Harvard Business School Press, Boston, 2007)
14. T.H. Davenport, S.L. Jarvenpaa, *Strategic Use of Analytics in Government*. U+IBM Center for The Business of Government (http://www.businessofgovernment.org) (2008)

An Enduring Interest in Classification: Supervised and Unsupervised

G.J. McLachlan

1 Motivation

1.1 Introduction

I have researched in the field of discriminant analysis for over 40 years and for nearly as long in the field of cluster analysis. Thus, I think it is fair to say that I have had an enduring interest in discriminant and cluster analyses, that is, in classification both supervised and unsupervised. The latter terminology is used outside of statistics in fields such as artificial intelligence, machine learning, and pattern recognition. However, the gap between these fields and statistics has narrowed appreciably over the years, and discriminant analysis and cluster analysis are also often referred in statistics as supervised classification and unsupervised classification, respectively.

Given that classification methods are among the main techniques applied in data mining, I have taken a keen interest in data mining almost from the time of its inception. In particular, I was taken at the time by the fact that data mining provided a source of real important examples, where there was a need for appropriate classification methods.

However, even in the early days of data mining, the data sets were sufficiently large, complex, and noisy to preclude routine application of most existing methods of statistics. For example, the linear discriminant function of Fisher [30] was not designed for data sets where the number of variables p is large relative to the number of observations n. It is robust to some extent under departures from normality as it was not derived under any distributional assumptions other than the two classes to which an entity can belong have the same covariance matrix.

G.J. McLachlan (✉)
Department of Mathematics, University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia
e-mail: gjm@maths.uq.edu.au

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_12,

But its direct implementation requires the inverse of the pooled sample covariance matrix $\boldsymbol{S}$, which poses problems if $\boldsymbol{S}$ is non-singular ($n \leq p$).

Now one obvious and straightforward way to avoid such problems is to replace $\boldsymbol{S}$ by a diagonal matrix with the same diagonal elements as those in $\boldsymbol{S}$. That is, proceed as if the variables are uncorrelated. This approach is known as naive Bayes; see, for example, Hand and Yu [43]. Since then, there have been theoretical results provided [7] to show that this rather crude approach can still be effective in high-dimensional data sets. Another approach is to use some form of regularization [32,38].

In the time before such statistical-based discriminant procedures were developed for large data sets with many variables, the classifiers developed in the context of machine learning dominated data mining. In the unsupervised classification case, clustering techniques favoured by computer scientists and engineers were also dominant. There was a challenge therefore to develop more principled methods of classification in a data mining context, based on sound theoretical and statistical foundations.

Since the late 1970s, I have been working on a model-based approach to clustering via finite mixture models. For the clustering of multivariate continuous data, attention has been focussed on mixtures of multivariate distributions. It was therefore natural for me to consider the application of normal mixture models to unsupervised classification problems in data mining. But as the normal mixture model with unconstrained component-covariance matrices is a highly parameterized one, there was a need for a reduction in the number of parameters. This led me to develop mixtures of so-called factor analyzers that adopt a factor-analytic representation of the component-covariance matrices to reduce the number of parameters; see McLachlan et al. [69] for a recent account of our work on factor models. The idea of using mixtures of factor models appears to have been first considered in a machine learning context by Geoffrey Hinton and colleagues; see, for example, Hinton et al. [46].

In the past decade, the aforementioned challenge of the development of soundly based statistical procedures for data mining has become much more difficult with the flood of data becoming available with the advancements in technology. Examples of such data sets can be found in bioinformatics. For example, with microarray experiments, there is the task of having to analyse the expression levels of thousands of genes p from a series of n microarray experiments where, say, each microarray might correspond to a separate tissue sample. Often n is no more than 50 or 100 (it can even be much smaller than 50), and so is much smaller than the number of genes p. That is, the number of observations n is much smaller than the number of variables p, which defines the so-called "big p, small n" problem, which is currently an area of much interest in statistics. Another example is the new generation of non-Sanger-based sequencing technologies (next-generation) that has delivered on its promise of sequencing DNA at unprecedented speed. These sequencing technologies have provided core facilities with the ability to produce large amounts of sequence data.

With the scientific mass production of data gaining momentum at the turn of this century, I also stepped up the intensity of my research in supervised classification, which had somewhat been reduced in the years following the completion of my monograph [68] on discriminant analysis and statistical pattern recognition. Clearly, the "big p, small n" problem in supervised classification is more straightforward than the unclassified case due to the existence of training data of known origin with respect to the underlying classes. Nevertheless, there are a number of challenging problems, including the selection of suitable variables, the design of accurate discriminant functions (classifiers) via regularization or other means, and the correction of the selection bias inherent in the traditional estimates of the accuracy of a classifier in its application to new data beyond the training data from which it has been formed.

1.2 Bridging the Gap Between Statistics and Machine Learning

A number of authors have contributed to bridging the gap between statistics and the fields of machine learning, pattern recognition, and artificial intelligence. Early attempts at this were centred on the use of neural networks which had been developed almost exclusively in the field of artificial intelligence. For example, the papers of Geman et al. [35], Cheng and Titterington [13], and Ripley [94] provided an excellent account of neural networks in a statistical framework, as did the book by Bishop [8] and the papers in the volume edited by Cherkassky et al. [14]. Also, the book by Ripley [95] played a major role in explaining the techniques and jargon from the artificial intelligence, computer science, and pattern recognition fields in terms familiar to statisticians. More recently, the books by Hand et al. [42], Hastie et al. [44], and Bishop [9] have done an excellent job in continuing to narrow the aforementioned gap.

1.3 My Background in Statistics

My research in statistics commenced with work on my Ph.D. thesis in the Statistics Section of the Department of Mathematics at the University of Queensland. My thesis was in the field of discriminant analysis (supervised classification) and was supervised by Stephen Lipton who had come to Australia from Rothamsted Experimental Station in England. I occupied a room next to Mildred Prentice (née Barnard), who had returned to teaching after a long absence from academia while raising her family. She had done her thesis some 40 years earlier at Rothamsted Experimental Station under the supervision of the famous Sir Ronald Fisher who was the inventor of discriminant analysis [30]. She was the author of the classic paper [5] on the application of Fisher's linear discriminant function to four series of Egyptian skulls.

My thesis was in the area of the estimation of error rates of discriminant functions (allocation rules or classifiers). It was a topical area at the time and in the previous decade. For example, Lachenbruch et al. [51]) had completed a Ph.D. thesis in 1965 at the University of North Carolina, focussing on the leave-one-out (LOO) method (i.e. n-fold cross-validation); see Lachenbruch and Mickey [52]. In the United Kingdom, Hills [45] at a Research Methods Meeting of the Royal Statistical Society had read a paper on the error rates of allocation rules, which was a major source of information in the area.

But I had no idea of the interest that would be directed to discriminant analysis due to the many new applications that would arise in data mining and other scientific fields in which this statistical technique plays an invaluable role. At the end of the decade in which I completed my thesis, the topic of error-rate estimation got a boost in mainstream statistics with the advent of the bootstrap by Efron [19] who used this problem as one of his main examples to illustrate his powerful new resampling approach. But even then, I did not foresee the tremendous interest that would be given to discriminant analysis in the coming years.

My postdoctoral work led subsequently to my interest in another field in statistics that is a core technique in data mining, namely, cluster analysis or unsupervised classification. My transition from the field of supervised classification to the unsupervised situation was a smooth one in that I first looked at the case of a partially classified sample, where the aim was to make use of the unclassified data in order to produce an improved classifier over the one based solely on the classified data whose class labels are known. This research led to the development of an iterative reclassification scheme [63] that can now be viewed as a hard-thresholding version of the EM algorithm for the fitting of a mixture of two normal multivariate distributions to the data; see Little [55]). I also developed a refined version [65] in which unclassified observations were given reduced weight relative to the classified ones in the formation of the final classifier on the basis of the combined data. This so-called classification maximum likelihood (ML) approach to estimation based on the basis of the partially classified data is biased except in special circumstances as identified in my study. This is because it uses hard rather than soft thresholding in the iterative assignment of the unclassified data points in the estimation process. At the time, I was unaware that this problem was also being studied by another Australian, Terry O'Neill, in a Ph.D. thesis [89] at Stanford University under the supervision of Brad Efron; see also O'Neill [90].

With the publication of the expectation–maximization (EM) algorithm by Dempster et al. [15], there was a framework for the iterative calculation of ML estimates from data that can be viewed as being incomplete. In particular, it could be applied to fit a finite mixture of distributions to data that were partially or completely unclassified. The incompleteness for this problem was with respect to the vector $\boldsymbol{z}$ containing the unknown class labels of the unclassified observations in the sample. I subsequently focussed my attention to forming classifiers and clustering procedures from partially and completely unclassified samples, respectively, by maximum likelihood via the EM algorithm. I referred to this approach to clustering as a mixture likelihood approach. I wrote my first paper on this approach

in 1978 [33] and summarized the results in McLachlan [67], and then in a monograph [70]. A fuller account that included more recent results was given in McLachlan and Peel [79]. The mixture likelihood approach is now more commonly referred to as model-based clustering [4].

There are a number of other monographs on mixture models. For example, in one of the first books on this topic, Titterington et al. [98] give an authoritative account of the properties of finite mixture distributions. In a more recent book, Lindsay [54] provides an excellent account, concentrating on the theory and geometry of mixture distributions.

As remarked by Aitkin et al. [1], in the reply to the discussion of their paper, "when clustering samples from a population, no cluster method is a priori believable without a statistical model". Concerning the use of mixture models to represent non-homogeneous populations, they noted in their paper that "Clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory". Previously, Marriott [56] had noted that the mixture likelihood-based approach "is about the only clustering technique that is entirely satisfactory from the mathematical point of view. It assumes a well-defined mathematical model, investigates it by well-established statistical techniques, and provides a test of significance for the results".

In addition to its usefulness in clustering, I also have sustained my interest in the use of finite mixtures of distributions for providing a flexible way of modelling complex data.

2 Milestones and Success Stories

2.1 Ph.D. and Postdoctoral Results

My first major milestone was the successful completion of my Ph.D. thesis [59]. A number of papers were published on the results, which provided a comparison of available methods of error-rate estimation, as well as providing an almost unbiased parametric estimator of the error rate of Fisher's linear discriminant function [60–62]. It also investigated discriminant analysis when the training samples are misclassified [58].

My postdoctoral work was focussed on the formation of classifiers based on partially classified data, as discussed above. This work [63,65], which can be viewed as a hard-thresholding version of the EM algorithm to this problem, prepared me for the adoption of the EM algorithm to these problems with its appearance late in 1977. This approach to clustering has grown in popularity over the years. It also provides an appealing semiparametric framework in which to estimate unknown distributional shapes. This is because the set of all normal mixture densities is dense in the set of all density functions under the L_1 metric.

2.2 EM: One of the Top-Ten Algorithms in Data Mining

The mixture model-based approach to clustering using the EM algorithm for its implementation has become a popular method of clustering [103]. The initial success of mixture models for clustering and inference led me to write the monograph [70] with a former Ph.D. student in the area, Kaye Basford. I continued with my interest in supervised classification and in 1992 wrote a Wiley monograph [68] on the topic titled, "Discriminant Analysis and Statistical Pattern Recognition".

I was very much taken by the EM algorithm, not just for the fitting of mixture models but also for the fitting of models in any situation where the computation of the maximum likelihood estimates is enhanced by viewing the data as being incomplete. For instance, the EM algorithm can be used in the training of neural networks [87]. In 1997, I coauthored a Wiley monograph [76] with Thriyambakam Krishnan from the Indian Statistical Institute. A second edition was published in 2008. Due to its role in the fitting of mixture models and their extensions such as the hidden Markov model (HMM) in data mining applications, the EM algorithm was identified as one of the most influential algorithms (among the top ten) as identified by a top-tier conference in the field, the 2006 IEEE Conference on International Data Mining (ICDM); see Wu et al. [103] and McLachlan and Ng [77].

In 2000, I wrote a Wiley monograph [79] on the advances in the field of mixture distributions, titled, "Finite Mixture Models". It was coauthored with my former Ph.D. student, David Peel. In 2004, with my new interest in bioinformatics, I coauthored a Wiley monograph [75] with colleagues, Kim-Anh Do, a former honours student under my supervision and now a professor at the MD Anderson Cancer Centre at the University of Texas, and Christophe Ambroise, now a professor at the University of Evry, Paris. This monograph is titled "Analyzing Microarray Gene Expression Data".

2.3 Data Mining Competitions

Data mining has received increased exposure in recent times through the increasing number of competitions being held in this field. Perhaps the best known is the Netflix competition; see Koren [49] for a description of the winning solution.

My interest in these competitions arose from my collaboration with a colleague in the statistics section of my university, Vladimir Nikulin. He has entered a number of competitions in recent times with a high degree of success. His wins include the International 2009 Pittsburgh Brain Connectivity IEEE ICDM Competition [84] and the RSCTC2010 Discovery Challenge: Mining DNA Microarray Data for Medical Diagnosis and Treatment [102].

3 Lessons in Learning from Failures

From my analyses of large data sets, I have had the opportunity to experience at first hand the fact that important lessons can be learnt from one's failures. Failures can come in many forms. I shall consider only one form to give an example in this section. It concerns a problem on error-rate estimation that I studied back in 1977.

As is well known, the apparent error rate A provides too optimistic an estimate of the true expected (unconditional) error rate of a classifier [64]. This is because the classifier is applied to the same data from which it has been assessed. One way to reduce its bias is to form a linear combination $A^{(w)}$ of A with an estimator that overestimates the error rate such as the half-sample cross-validated error rate $A^{(\mathrm{HCV})}$; that is,

$$A^{(w)} = (1 - w)A + wA^{(\mathrm{HCV})}.$$

In a response to the proposal by Toussaint and Sharpe [99] in which the weight w was taken to be 0.5, I investigated asymptotically the choice of w so that $A^{(w)}$ is an unbiased estimator in the particular case of $g = 2$ classes each having a multivariate normal distribution with a common covariance matrix. I showed that the optimal choice of w was approximately equal to 0.66, depending on the number of variables p, the Mahalanobis distance between the two classes, and the ratio of the class-sample sizes [66].

With this choice of w, $A^{(w)}$ is essentially the 0.632 estimator $A^{(0.632)}$ of Efron [20], since $A^{(\mathrm{HCW})}$ is almost the same as the LOO bootstrap error rate. Using an ingenious argument, Efron [20] was able to derive his 0.632 estimator $A^{(0.632)}$ without making any assumption about the underlying distributions of the classes.

At the time of the derivation of my estimator $A^{(w)}$, I did not think that the choice of w approximately equal to 0.66 would hold if the normality assumption of the class distributions were relaxed. However, if I had performed some simulations for non-normal class distributions, I would have at least been alerted to the possibility that the optimal choice of w was not sensitive to the assumption of normality.

4 Current Research Issues and Challenges

In the current era where new technologies are producing a flood of data, applications in data mining can involve exploring massively huge data sets. In the past, large data sets have led some to cautionary notes about an over reliance on models in making predictions; see, for example, Breiman [10] who contrasts the "models" of traditional statistics with the black box algorithms developed by other disciplines. But as Efron points out in his discussion of this paper, "The whole point of science is to open up black boxes, understand their insides, and build better boxes for the purposes of mankind".

The need for valid statistical models is therefore greater than ever. Besides using the data to provide a reasonable fit to the truth, models provide a framework in which to assess uncertainty, which is the very backbone of statistical learning [100]. Models are also needed as they allow true background knowledge to be incorporated in an effective way.

In mainstream statistics, the purpose of a statistical model is to postulate a set of realistic assumptions about the distribution generating the data. But with high-dimensional data, a plausible specification of the underlying distribution is extremely difficult without assumptions that would appear to be overly restrictive. This is because with the present day of data sets, there are usually many more variables p in the model than the available number n of observations, leading to the "big p, small n" problem. In the latter situation, the data sets are usually sparse, which means that the distances between points are not small, so that local information in the data is limited for the fitting of a model. A number of authors have defined this concept of sparseness in a more formal sense; see the theoretical results of Hall et al. [39].

5 Research Tools and Techniques

5.1 Finite Mixture Models

Finite mixture distributions provide a flexible and mathematical-based approach to the modelling and clustering of data observed on random phenomena. The seminal paper of Dempster et al. [15] on the EM algorithm greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood (ML) is a classic example of a problem that is simplified considerably by the EM's conceptual unification of ML estimation from data that can be viewed as being incomplete.

With the mixture model-based approach to clustering, the observed p-dimensional data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to have come from a mixture of an initially specified number g of component densities in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. The mixture density of $\boldsymbol{y}_j$ is expressed as

$$f\left(\boldsymbol{y}_j; \boldsymbol{\Psi}\right) = \sum_{i=1}^{g} \pi_i f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right) \quad (j = 1, \ldots, n), \tag{1}$$

where the component density $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ is specified up to a vector $\boldsymbol{\theta}_i$ of unknown parameters ($i = 1, \ldots, g$). The vector of all the unknown parameters is given by

$$\boldsymbol{\Psi} = \left(\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_g^{\mathrm{T}}\right)^{\mathrm{T}},$$

where the superscript T denotes vector transpose.

The mixture model (1) provides a probabilistic clustering of the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into g clusters in terms of their estimated posterior probabilities of component membership of the mixture. The posterior probability that the jth feature vector with observed value $\boldsymbol{y}_j$ belongs to the ith component of the mixture can be expressed by Bayes' theorem as

$$\tau_i\left(\boldsymbol{y}_j;\boldsymbol{\Psi}\right) = \pi_i f_i\left(\boldsymbol{y}_j;\boldsymbol{\theta}_i\right)/f\left(\boldsymbol{y}_j;\boldsymbol{\Psi}\right) \quad (i = 1,\ldots,g). \tag{2}$$

An outright assignment of the data is obtained by assigning each data point to the component to which it has the highest estimated posterior probability of belonging.

With this approach to clustering, each cluster imposed on the data corresponds to a component in the mixture model. For instance, with normal mixtures to be discussed below, there is the assumption that the data have a normal distribution in a cluster. The presence of outliers or skewness in the clusters can be accommodated by the fitting of a mixture model with more normal components than the number of clusters. But there is the problem of trying to identify which components correspond to a cluster. To avoid the fitting of mixtures with more components than the number of clusters to handle outliers, McLachlan and Peel [78] proposed the use of mixtures of t-distributions; see also McLachlan and Peel [91]. The t-distribution provides a longer-tailed alternative to the normal distribution. Hence, it provides a more robust approach to the fitting of normal mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of t components gives less extreme estimates of the posterior probabilities of component membership of the mixture model. Concerning situations with skewed clusters as in the clustering of flow cytometric data, Sam (Kui) Wang and I have been working with Saumyadipta Pyne and colleagues from the Broad Institute of MIT and Harvard University on the use of mixtures of skew t-distributions [92].

5.2 ML Estimation of Mixture Models via EM

The parameter vector $\boldsymbol{\Psi}$ in the mixture model (1) can be estimated by maximum likelihood. The objective is to maximize the likelihood $L(\boldsymbol{\Psi})$, or equivalently, the log likelihood $\log L(\boldsymbol{\Psi})$, as a function of $\boldsymbol{\Psi}$, over the parameter space. That is, the ML estimate of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = 0, \tag{3}$$

where

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log f\left(\boldsymbol{y}_j;\boldsymbol{\Psi}\right) \tag{4}$$

is the log likelihood function for $\boldsymbol{\Psi}$ formed under the assumption of independent data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$.

Solutions of (3) corresponding to local maximizers of (4) can be found by applying the EM algorithm. We now give a brief description of the EM algorithm before discussing its implementation for mixtures.

5.3 EM Algorithm

The EM algorithm is an iterative algorithm; in each iteration of which, there are two steps: the Expectation Step (E-step) and the Maximization Step (M-step). A brief history of the EM algorithm can be found in McLachlan and Krishnan [76]). Within the incomplete-data framework of the EM algorithm, we let $y = (\boldsymbol{y}_1^T, 1/4, \boldsymbol{y}_n^T)$ be the observed data vector, we let $\boldsymbol{x}$ denote the vector containing the augmented or so-called complete data, and we let $\boldsymbol{z}$ denote the vector containing the additional data, referred to as the unobservable or missing data.

The EM algorithm approaches the problem of solving the "incomplete-data" log likelihood (4) indirectly by proceeding iteratively in terms of the complete-data log likelihood, $\log L_c(\boldsymbol{\Psi})$. As it depends explicitly on the unobservable data, the E-step is performed on which $\log L_c(\boldsymbol{\Psi})$ is replaced by the so-called Q-function, which is its conditional expectation given the observed data $\boldsymbol{y}$, using the current fit for $\boldsymbol{\Psi}$. More specifically, on the $(k + 1)$th iteration of the EM algorithm, the E-step computes

$$Q\left(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}\right) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi})|\boldsymbol{y}\},$$

where $E_{\boldsymbol{\Psi}^{(k)}}$ denotes expectation using the parameter vector $\boldsymbol{\Psi}^{(k)}$. The M-step updates the estimate of $\boldsymbol{\Psi}$ by that value $\boldsymbol{\Psi}^{(k+1)}$ of $\boldsymbol{\Psi}$ that maximizes the Q-function, $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, with respect to $\boldsymbol{\Psi}$ over the parameter space [76].

The E- and M-steps are alternated repeatedly until the changes in the log likelihood values are less than some specified threshold. The EM algorithm is numerically stable with each EM iteration increasing the likelihood value as

$$L\left(\boldsymbol{\Psi}^{(k+1)}\right) \geq L\left(\boldsymbol{\Psi}^{(k)}\right).$$

It can be shown that both the E- and M-steps will have particularly simple forms when the complete-data probability density function is from an exponential family [76]. Often in practice, the solution to the M-step exists in closed form.

In the fitting of the mixture model (1) within the EM framework, each $\boldsymbol{y}_j$ is conceptualized to have arisen from one of the g components of the mixture model (1). We let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ denote the unobservable component-indicator vectors, where the ith element z_{ij} of $\boldsymbol{z}_j$ is taken to be one or zero accordingly as the jth observation $\boldsymbol{y}_j$ does or does not come from the ith component $(i = 1, \ldots, g)$. The complete-data vector $\boldsymbol{x}$ is declared then to be

$$\boldsymbol{x} = \left(\boldsymbol{y}^{\mathrm{T}}, \boldsymbol{z}^{\mathrm{T}}\right)^{\mathrm{T}}.$$

The complete-data log likelihood is given by

$$\log L_{\mathrm{c}}(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left\{\log \pi_i + \log f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right)\right\}. \tag{5}$$

As the complete-data log likelihood function $\log L_{\mathrm{c}}(\boldsymbol{\Psi})$ is linear in the unobservable z_{ij}, the E-step on the $(k + 1)$th iteration is effected by replacing each z_{ij} by its conditional expectation given the observed data $\boldsymbol{y}_j$, using $\boldsymbol{\Psi}^{(k)}$. This conditional expectation is equal to $\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})$, which is the posterior probability that $\boldsymbol{y}_j$ belongs to the ith component of the mixture, using the current fit $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}(i = 1, \ldots, g; j = 1, \ldots, n)$. From (2),

$$\tau_i\left(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}\right) = \pi_i^{(k)} f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i^{(k)}\right) / f\left(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}\right). \tag{6}$$

5.4 Normal Mixture Models

In the case of the mixtures of multivariate normal densities, the ith component density is given by

$$f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right) = \phi\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \quad (i = 1, \ldots, g), \tag{7}$$

where $\phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Thus, under the normal mixture model without any constraints on the component-covariance matrices $\boldsymbol{\Sigma}_i$, the density of $\boldsymbol{y}_j$ is given by

$$f\left(\boldsymbol{y}_j; \boldsymbol{\Psi}\right) = \sum_{i=1}^{g} \pi_i \phi\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right). \tag{8}$$

For normal components, the M-step exists in closed form [79].

In the case of unrestricted component-covariance matrices $\boldsymbol{\Sigma}_i$, $L(\boldsymbol{\Psi})$ is unbounded, as each data point gives rise to a singularity on the edge of the parameter space. Consideration has to be given to the problem of relatively large (spurious) local maxima that occur as a consequence of a fitted component having a very small (but non-zero) generalized variance (the determinant of the covariance matrix). Such a component corresponds to a cluster containing a few data points

either relatively close together or almost lying in a lower-dimensional subspace in the case of multivariate data.

McLachlan et al. [81] have developed the program EMMIX as a general tool to fit mixtures of multivariate normal or *t*-distributed components by ML via the EM algorithm to continuous multivariate data. It also includes many other features that were found to be of use when fitting mixture models. These include the provision of starting values for the application of the EM algorithm, the provision of standard errors for the fitted parameters in the mixture model via various methods, and the determination of the number of components.

With applications where the log likelihood equation has multiple roots corresponding to local maxima, the EM algorithm should be applied from a wide choice of starting values in any search for all local maxima. In the context of finite mixture models, an initial parameter value can be obtained using the *k*-means clustering algorithm, hierarchical clustering methods, or random partitions of the data. With the EMMIX program, there is an additional option for random starts whereby the user can first subsample the data before using a random start based on the subsample each time. This is to limit the effect of the central limit theorem, which would have the randomly selected starts being similar for each component at least in large samples.

There is some other EM-based software for mixture modelling via maximum likelihood. For example, Fraley and Raftery [31] have developed the MCLUST program for the fitting of mixtures of normal components under various parameterizations of the component-covariance matrices. The reader is referred to the appendix in McLachlan and Peel [79] for the availability of software for the fitting of mixture models.

Concerning the number of components g in the mixture model, we can make a choice as to an appropriate value of the number of components (clusters) g by consideration of the likelihood function. In the absence of any prior information as to the number of clusters present in the data, we can monitor the increase in log likelihood function as the value of g increases. At any stage, the choice of $g = g_0$ versus $g = g_0 + 1$ can be made by either performing the likelihood ratio test or by using some information-based criterion, such as the Bayesian information criterion (BIC). Unfortunately, regularity conditions do not hold for the likelihood ratio test statistic λ to have its usual null distribution of chi-squared with degrees of freedom equal to the difference d in the number of parameters for $g = g_0 + 1$ and $g = g_0$ components in the mixture model. The EMMIX program provides a bootstrap resampling approach to assess the null distribution (and hence the P value) of the statistic $(-2\log\lambda)$. Alternatively, one can apply BIC, although regularity conditions do not hold for its validity here either. The use of BIC leads to the selection of $g = g_0 + 1$ over $g = g_0$ if $-2\log\lambda$ is greater than $d\log(n)$, where d is the number of unknown parameters in the model.

In applying normal mixture models as above to cluster multivariate (continuous) data, it is assumed as in most typical cluster analyses using any other method that (a) there are no replications on any particular entity specifically identified as such; (b) all the observations on the entities are independent of one another.

A situation where (a) and (b) both do not hold concerns the clustering of gene profiles on the basis of their values over a series of tissues samples measured under different conditions. Firstly, (a) does not hold since there is a known structure on the replications of the genes and, secondly, (b) does not hold since the gene profiles are not all independently distributed. In order to handle (a) and (b) in such situations, Ng et al. [88] have proposed the use of mixtures of linear mixed models which can be fitted using the program called EMMIX-WIRE.

5.5 *Mixtures of Factor Analyzers*

The g-component normal mixture model (1) with unrestricted component-covariance matrices is a highly parameterized model with $d = (1/2)p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$. Banfield and Raftery [4] introduced a parameterization of the component-covariance matrix $\boldsymbol{\Sigma}_i$ based on a variant of the standard spectral decomposition of $\boldsymbol{\Sigma}_i (i = 1, \ldots, g)$. But if the number of variables p is large relative to the sample size n, it may not be possible to use this decomposition to infer an appropriate model for the component-covariance matrices. Even if it is possible, the results may not be reliable due to potential problems with near-singular estimates of the component-covariance matrices when p is large relative to n.

Hence, some of dimension reduction and/or regularization is required before the fitting of normal mixtures to high-dimensional data. In recent times, there have been a number of papers written on variable selection for mixture models; see Raftery and Dean [93] and Maugis et al. [57] and the references therein. However, this methodology cannot be applied directly to high-dimensional data as noted by Khalili et al. [48].

In a series of papers, we have investigated the use of mixtures of factor analyzers to enable model-based density estimation to be undertaken for high-dimensional data, where p is large relative to n; see McLachlan and Peel [80] and McLachlan et al. [72,82]). The latter paper considers the fitting of mixtures of t-factor analyzers. With the factor-analytic representation of the component-covariance matrices, we have that

$$\boldsymbol{\Sigma}_i = B_i B_i^{\mathrm{T}} + D_i \quad (i = 1, \ldots, g), \tag{9}$$

where $\boldsymbol{B}_i$ is a $p \times q$ matrix and $\boldsymbol{D}_i$ is a diagonal matrix. The number of factors q is chosen to be small.

In practice, there is often the need to reduce further the number of parameters in the specification of the component-covariance matrices. To this end, we have recently proposed the use of common component-factor loadings, which considerably reduces further the number of parameters [3]. Moreover, it allows the data to be displayed in low-dimensional spaces. With this factor model, the matrix $\boldsymbol{B}_i$ of factor loadings in (9) is specified as

$$B_i = BK_i \tag{10}$$

where $\boldsymbol{K}_i$ is a $q \times q$ matrix ($i = 1, \ldots, g$). The specification (10) considerably reduces the number of parameters for large p since the $p \times q$ matrix of factor loadings $\boldsymbol{B}$ no longer depends on the number of components g.

5.6 Variable Selection in Cluster Analysis

When it is computationally feasible to fit the factor models as defined in the previous section, dimension reduction is effectively being done as part of the primary analysis. However, for many data sets in data mining applications, the number of variables p will be too large to fit these models directly without first performing some form of dimension reduction. In the context of clustering, McLachlan et al. [73] proposed a three-step procedure (called EMMIX-GENE) in which on the first step the variables are considered individually by performing a test of a single t-component distribution versus a mixture of two t-components for each variable. Variables found not to be significant according to this test are discarded. Then on the second step, the retained variables (after appropriate normalization) are clustered into groups on the basis of Euclidean distance. Finally, on the third step, the observations can be clustered by the fitting of mixtures of normal distributions or factor analyzers to representatives of the groups of variables.

Recently, Chan and Hall [12] have considered a method for variable selection where the variables are considered individually by performing a non-parametric mode test to assess the extent of multimodality.

Another approach to dimension reduction is projection, which is very popular. For example, principal component analysis (PCA) is commonly used as a first step in the analysis of huge data sets, prior to a subsequent and more in-depth rigorous modelling of the data. A related projection method is matrix factorization, which is described briefly in the next section.

5.7 Matrix Factorization

We follow the usual practice in bioinformatics of letting

$$A = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$$

be the $p \times n$ data matrix. The usual statistical practice is to take the transpose of $\boldsymbol{A}$, $\boldsymbol{A}^{\mathrm{T}}$, as the data matrix. Without loss of generality, we assume that the overall mean of $\boldsymbol{A}$ is zero.

In recent times, much attention has been given to matrix factorizations of the form,

$$A = C_1 C_2, \tag{11}$$

where $\boldsymbol{C}_1$ is a $p \times q$ matrix and $\boldsymbol{C}_2$ is a $q \times n$ matrix and where q is chosen to be much smaller than p. For a specified value of q, the matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are chosen to minimize

$$\|A - C_1 C_2\|^2, \tag{12}$$

where $\|\cdot\|$ is the Frobenius norm (the sum of squared elements of the matrix). With this factorization, dimension reduction is effected by replacing the data matrix $\boldsymbol{A}$ by the solution $\hat{\boldsymbol{C}}_2$ for the factor matrix $\boldsymbol{C}_2$; the ith row of $\hat{\boldsymbol{C}}_2$ gives the values of the ith metavariable for the n entities. Thus, the original p variables are replaced by q metavariables. When the elements of $\boldsymbol{A}$ are non-negative, we can restrict the elements of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ to be non-negative. This approach is called non-negative matrix factorization (NMF) in the literature [17,53]. We shall call the general approach where there are no constraints on $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, general matrix factorization (GMF).

The classic method for factoring the data matrix $\boldsymbol{A}$ is singular-value decomposition (SVD); see Golub and van Loan [36]. It follows from this theorem that the value of $\boldsymbol{C}_2$ that minimizes (12) over the set of all $q \times n$ matrices of rank q is given by the matrix whose columns are the eigenvectors corresponding to the q largest eigenvalues of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.

Now SVD, effectively PCA, imposes orthogonality constraints on the rows of the matrix $\boldsymbol{C}_2$. However, this ignores the non-independence of biological processes, which is equivalent to non-orthogonality of the rows of $\boldsymbol{C}_2$; see, for example, Kossenkov and Ochs [50]. On the other hand, GMF which has no constraints on $\boldsymbol{C}_2$ provides a factorization into a lower-dimensional subspace with no orthogonality constraints on its basis vectors. Thus, GMF has the flexibility to model, for example, biological behaviour in which the gene signatures overlap. In contrast, PCA with its orthogonality constraints is overly constraining for such data and is thus not suited to isolating gene signatures that have appreciable overlap. Also, PCA is based on finding the directions of greatest variance, but the sample covariance matrix provides misleading estimates where the number of variables p is much greater than the number n of observations [47].

Nikulin and McLachlan [84] have developed a very fast approach to the GMF (11), using a gradient-based algorithm that is applicable to an arbitrary (differentiable) loss function; see also [85]. Witten et al. [101] and Nikulin and McLachlan [86] have considered a penalized approach to PCA in order to provide sparse solutions.

5.8 Variable Selection in Discriminant Analysis

The methods of variable selection discussed in the previous section for cluster analysis obviously also apply in discriminant analysis. But in the latter context, we might want to make use of the response variables observed with the feature vectors $\boldsymbol{y}_j$ in order to obtain a more effective reduction in the number of variables. To this end, we can make use of methods of predictor selection in a regression context, since two-class discriminant analysis can be put in a regression context by defining the class label to be zero/one or one/minus one.

Many methods of variable selection have been proposed in the regression literature, including those based on penalized least squares or penalized pseudo-likelihood; see, for example, the LASSO [96] and the Dantzig selector [11]. These methods have been investigated in various studies on theoretic and algorithmic issues; see, for example, Efron et al. [24].

But frequently in discriminant analysis, variable selection is undertaken in some ad hoc manner before a more formal method of feature selection is adopted in conjunction with the choice of prediction rule. For example, one such commonly used method in the case of two classes is to rank the features on the basis of the magnitude of the (pooled) two-sample t test. It follows from Fan and Lv [27] that this so-called independence screening method has asymptotically under certain regularity conditions a "sure screening property"; that is, with probability very close to 1, the independence screening technique retains all of the important features in the model; see also Hall et al. [41]. This approach is called sure independence screening (SIS), and its extension to cover cases where the regularity conditions may fail is called iterated sure independence screening (ISIS); see Fan et al. [29], who extend SIS and ISIS to much more general models, and Fan and Lv [28], who give an overview of variable selection in high-dimensional feature space. Although not directly applicable, the sure screening property of the SIS approach after some adaptation can be used to give theoretical justification to the nearest-shrunken-centroids procedure of Tibshirani et al. [97]; see Fan and Fan [26].

Even in situations where it is possible to form directly a classifier on the basis of all the available variables, it is not advisable as the presence of noise in the many variables with little or no signal can severely limit the predictive performance of the classifier. The latter phenomenon of noise accumulation in high-dimensional classification and regression has long been observed by statisticians and computer scientists [29]. Fan and Fan [26] give a simple expression on how dimensionality impacts asymptotically on the error rate, while Hall et al. [40] have studied a similar problem for distance-based classifiers.

Besides the problem of accurate class prediction for an observation subsequent to the training data, there is also interest in discriminant analysis in forming a classifier on the basis of only a small subset of the variables and the associated problem of finding those variables (so-called biomarkers in bioinformatics) that are most useful in distinguishing between the classes. For example, Geman et al. [34] consider the development of molecular diagnostic tools based on only a few genes.

5.9 Selection Bias in Discriminant Analysis

There has been ever increasing interest in the use of microarray experiments as a basis for the provision of prediction (discriminant) rules for improved diagnosis of cancer and other diseases. Typically, the microarray cancer studies provide only a limited number of tissue samples from the specified classes of tumours or patients, whereas each tissue sample may contain the expression levels of thousands of genes. Thus, researchers are faced with the problem of forming a prediction rule on the basis of a small number of classified tissue samples, which are of very high dimension. Usually, some form of feature (gene) selection is adopted in the formation of the prediction rule, not only to avoid the noise accumulation problem but also to find suitable biomarkers. As the subset of genes used in the final form of the rule have not been randomly selected but rather chosen according to some criterion designed to reflect the predictive power of the rule, there will be a selection bias inherent in estimates of the error rates of the rules if care is not taken.

In the context of applying a support vector machine (SVM) using recursive feature elimination (RFE) as proposed by Guyon et al. [37], Ambroise and McLachlan [2] pointed out how a biased estimate of the error will result if cross-validation is applied in its traditional way. For example, if the error rate of a classifier based on a selected subset s_{p_o} of variables is estimated by cross-validation using the same set s_{p_o} on each fold, then an optimistic assessment will result. Rather, the set of variables for each fold should be reselected using the same method of selection that was originally obtained from using the full sample of observations. There will also be a bias if the final classifier is selected by optimizing some criterion over subsets of various sizes and this fact is not taken into account in constructing the estimated error rate of the final selected classifier.

The appropriate schemes for implementing cross-validation to ensure that a selection bias is circumvented have been considered by McLachlan et al. [74] and Zhu et al. [104,105]. A colleague Camille Maumet has developed an R program called *R MicroArray Gene-expression-based Program In Error* rate estimation (RMAGPIE) that is available at the Bioconductor Web site (http://www.bioconductor.org/help/bioc-views/release/bioc/html/Rmagpie.html). It provides an estimate of the error rate of a (gene signature-based) classifier such as nearest-shrunken-centroids or SVM that avoids the selection bias as well as reporting the subset of genes that appear to be most useful in discriminating between the classes.

5.10 Multiple Testing

As discussed in the previous sections on variable selection for discriminant and cluster analyses in the context of high-dimensional data, consideration might be given to carrying out a test for each variable considered separately. Given that in this context there are so many variables to be tested, issues with multiple testing

become relevant. In recent times, much attention has been focussed on controlling the false discovery rate (FDR), as introduced by Benjamini and Hochberg [6]. The FDR is essentially the proportion of false rejections of the null hypothesis in a series of tests. The literature in this area has grown enormously in the past few years with major contributions from leading statisticians including Brad Efron who has written extensively on it in a series of papers which may be found in his recent monograph Efron [22]. In the latter, he proposes an empirical Bayes approach that combines Bayesian and frequentist ideas.

Drawing on concepts such as the local FDR [25] and the empirical null [21], McLachlan et al. [71] proposed the fitting of a mixture of two univariate normal distributions in proportions π_0 and $(1 - \pi_0)$ to the z-scores corresponding to the P values obtained from a series of N tests. Here the mixing proportion π_0 represents the proportion of null effects. The value of one minus the P value for the j test is converted to a z-score, z_j, by the probit transformation,

$$z_j = \Phi^{-1}(1 - P_j) \quad (j = 1, \ldots, N).$$

In the cases where one can be confident that the outcomes of the N tests can be treated as independent of one another and where the P values P_j have a uniform distribution under the null hypothesis, we can take the first component of the normal mixture model (corresponding to the null component) to be the standard normal. Otherwise, we can estimate its mean and variance (i.e. use an empirical null), although care has to be taken due to identifiability problems. An initial value for π_0 can be obtained using the results in Donoho and Jin [18] on higher criticism thresholding.

6 Making an Impact

Perhaps the most direct way to have a short-term impact on the field is to write papers that address unresolved issues associated with topical problems in the field.

Also, presenting lectures at International Conferences provides another forum for the wide dissemination of one's work. However, for new researchers, there are limited chances to give plenary or keynote talks at conference or workshops.

As noted above, papers that are written on topical issues can make a short-term impact through their citations. They will have a long-term impact if interest in the topic is one that will persist well into the future. Also, books written on topics that are in the mainstream of statistics and machine learning can have a long-term impact if they provide an authoritative account of the area; in particular, if the area is short on authoritative references.

7 Future Insights

7.1 Short Term

With the vast amount of data being collected these days, the need for data mining will grow considerably. There would appear to be no limits to applications of data mining. For example, data mining algorithms are being developed for application to real-time data that record personal activities, conversations, and movements in an attempt to improve health, guide traffic, and advance the scientific understanding of human behaviour; see, for example, Mitchell [83].

As statistical methodology is advanced to handle the increase in the dimension size of the feature observations and the amount of data collected, technology also advances resulting in data sets of much higher dimension which throw up further challenges to the development of suitable statistical procedures and to computational problems concerning their implementation. For example, just as the statistical analysis of microarray gene-expression data has become more routine after a decade or so, a dramatic surge in the size of the data sets has commenced as the cost of next-generation sequencing (NGS) becomes ever more affordable.

7.2 Long Term

In the long term, I see the development of more sophisticated algorithms to exploit the ever increasing amount of data that will be available as, for example, from smart phones and their equivalents in the future. Also, high-throughput machines will continue to provide a flood of data for analysis in various scientific fields such as biology, medicine, and economics. Moreover, in the business world, there is the increasing recognition of the need to draw insight from the current surge in data. There is no doubt that data mining applications to such data will present formidable challenges. It is clear that the capacity to analyse the available data will always in the foreseeable future lag behind the ability to record these data.

On the vast amounts of data being collected by scientists, Brad Efron [23] in a recent interview commented “In some ways I think that scientists have misled themselves into thinking that if you collect enormous amounts of data, you are bound to get the right answer. You are not bound to get the right answer unless you are enormously smart. You can narrow down your questions; but enormous sets of data often consist of enormous numbers of small sets of data, none of which by themselves are enough to solve the thing you are interested in, and they fit together in some complicated way”.

8 Summary

In this chapter, I have concentrated on my work over the years in the fields of discriminant and cluster analyses. These techniques are frequently employed in many applications in the field of data mining. Cluster analysis is a powerful tool for the exploratory analysis of high-dimensional data sets. Discriminant analysis is needed to make reliable predictions from data of known origin with respect to a number of predefined classes and to identify those variables most useful in this task.

We have discussed how the dimension and/or size of the data sets encountered in typical data mining applications challenges our traditional methods of classification. In the context of supervised classification, it is possible for at least some classifiers such as naive Bayes or rules based on projections of the data to be formed. However, it is advisable to perform some form of variable selection to avoid any problems with the phenomenon of noise accumulation. Fan and Fan [26] have demonstrated asymptotically under certain conditions that discriminant functions including naive Bayes and those based on random projections can perform as poorly as random guessing unless the signal-to-noise ratio is sufficiently large.

In presenting tools for data mining, attention has been focussed on a model-based approach to clustering using normal mixture models and extensions fitted via the EM algorithm. As a normal mixture model with unrestricted component-covariance matrices is a highly parameterized one, some form of variable selection and/or regularization is needed. The use of factor models is highlighted in this context.

In discriminant analysis, attention is given to issues associated with the need to reduce the number of variables to avoid the aforementioned phenomenon of noise accumulation. Another issue considered is the need to ensure that estimates of error rates are formed in a proper way so as to avoid overly optimistic biases due to selection, which becomes a practical issue when a relatively small subset of variables is selected from a much larger set of variables in some optimal (non-random) way.

The problem of variable selection with the analysis of high-dimensional data can raise a number of issues. One of these is large-scale multiple testing. We have given a brief account of one approach based on mixture models for the estimation and control of the FDR under dependence.

The ongoing development of statistical methods for the analysis of high-dimensional data is essential as data sets grow in dimension and size to challenge existing methods concerning their applicability in terms of their validity, implementation, accuracy, efficiency, and robustness, among other considerations. These data sets occur in data mining applications over a wide range of fields, ranging from the hard sciences such as physics, astronomy, and biology through the social sciences such as economics to applications in the medical sciences and engineering. As prophesied by Donoho [16], this century is “surely the century of data”.

Acknowledgements The work for this chapter was supported by a grant from the Australian Research Council.

I would like to thank my collaborators on various aspects of my research relevant to data mining over the years, including Peter Adams, Christophe Ambroise, Jangsun Baek, Kaye Basford, Richard Bean, Liat Ben-Tovim Jones, Karen Byth, Igor Cadez, Kim-Anh Le Cao, Soong Chang, Jonathan Chevelu, Kim-Anh Do, Lloyd Flack, S. Ganesalingam, Doug Hawkins, Tian-Hsiang Huang, Peter Jones, Murray Jorgensen, Nazim Khan, Thriyambakam Krishnan, Charles Lawoko, Andy Lee, Jess Mar, Camille Maumet, Christine McLaren, Emmanuelle Meugnier, Katrina Monico, Angus Ng, Vladimir Nikulin, David Peel, Saumyadipta Pyne, Barry Quinn, Suren Rathnayake, Mohamed Shoukri, Padhraic Smyth, Erick Suarez, Deming Wang, Kui (Sam) Wang, Bill Whiten, Leesa Wockner, Ian Wood, Kelvin Yau, and Justin Zhu.

References

1. M. Aitkin, D. Anderson, J. Hinde, Statistical modelling of data on teaching styles (with discussion). J. R. Stat. Soc. B **144**, 419–461 (1981)
2. C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on basis of microarray gene expression data. Proc. Natl. Acad. Sci. USA **99**, 6562–6566 (2002)
3. J. Baek, G.J. McLachlan, L. Flack, Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 1298–1309 (2010)
4. J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)
5. M.M. Barnard, The secular variations of skull characters in four series of Egyptian skulls. Ann. Eugen. **6**, 352–371 (1935)
6. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57**, 289–300 (2005)
7. P.J. Bickel, E. Levina, Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**, 989–1010 (2004)
8. C.M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995)
9. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2007)
10. L. Breiman, Statistical modeling: the two cultures (with discussion). Stat. Sci. **16**(2001), 199–231 (2001)
11. E. Candes, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Ann. Stat. **35**, 2313–2404 (2007)
12. Y.B. Chan, P. Hall, Using evidence of mixed populations to select variables for clustering very high-dimensional data. J. Am. Stat. Assoc. **105**, 798–809 (2010)
13. B. Cheng, D.M. Titterington, Neural networks: a review from a statistical perspective (with discussion). Stat. Sci. **9**, 2–54 (1994)
14. V. Cherkassky, J.H. Friedman, H. Wechsler (eds.), *From Statistics to Neural Networks: Theory and Pattern Recognition Applications* (Springer-Verlag, Berlin, 1994)
15. A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)
16. D. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, in *Aide-Memoire of the Lecture in AMS Conference "Math Challenges of 21st Century"*, 2000
17. D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts? in *Advances in Neural Information Processing Systems*, ed. by S. Thrun, L. Saul, B. Schölkopf, vol. 16 (MIT, Cambridge, MA, 2004)

18. D. Donoho, J. Jin, Higher criticism thresholding: optimal feature selection when useful features are rare and weak. Proc. Natl. Acad. Sci. USA **105**, 14790–14795 (2008)
19. B. Efron, Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)
20. B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. **78**, 316–331 (1983)
21. B. Efron, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Am. Stat. Assoc. **99**, 96–104 (2004)
22. B. Efron, *Large-Scale Inference* (Cambridge University Press, Cambridge, MA, 2010)
23. B. Efron, A life in statistics – Bradley Efron. Significance **7**, 178–181 (2010)
24. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression (with discussion). Ann. Stat. **32**, 409–499 (2004)
25. B. Efron, R. Tibshirani, Empirical Bayes methods and false discovery rates for microarrays. Genet. Epidemiol. **36**, 70–86 (2002)
26. J. Fan, Y. Fan, High dimensional classification using features annealed independence rules. Ann. Stat. **70**, 2605–2637 (2008)
27. J. Fan, J. Lv, Sure independence screening for ultra-high dimensional feature space (with discussion). J. R. Stat. Soc. B **70**, 849–911 (2008)
28. J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**, 101–148 (2010)
29. J. Fan, R. Samworth, Y. Wu, Ultrahigh dimensional feature selection: beyond the linear model. J. Mach. Learn. Res. **10**, 2013–2038 (2009)
30. R.A. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
31. C. Fraley, A.E. Raftery, MCLUST: software for model-based cluster analysis. J. Classif. **16**, 297–306 (1999)
32. J.H. Friedman, Regularized discriminant analysis. J. Am. Stat. Assoc. **84**, 165–175 (1989)
33. S. Ganesalingam, G.J. McLachlan, The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika **65**, 658–662 (1978)
34. D. Geman, C. d'Avignon, D.Q. Naiman, R.L. Winslow, Classifying gene expression profiles from pairwise mRNA comparison. Stat. Appl. Genet. Mol. Biol. **3**(1), Article 19 (2004)
35. S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma. Neural. Comput. **4**, 1–58 (1992)
36. G. Golub, C. van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, MD, 1983)
37. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002)
38. Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays. Biostatistics **8**, 86–100 (2007)
39. P. Hall, J.S. Marron, A. Neeman, Geometric representation of high-dimension, low-sample size data. J. R. Stat. Soc. B **67**, 427–444 (2005)
40. P. Hall, Y. Pittelkow, M. Ghosh, Theoretic measures of relative performance of classifiers for high-dimensional data with small sample sizes. J. R. Stat. Soc. B **70**, 158–173 (2008)
41. P. Hall, D.M. Titterington, J.-H. Xue, Tilting methods for assessing the influence of components in a classifier. J. R. Stat. Soc. B **71**, 783–803 (2009)
42. D.J. Hand, H. Mannila, P. Smyth, *Principles of Data Mining* (MIT, Cambridge, MA, 2001)
43. D.J. Hand, K. Yu, Idiot's Bayes – not so stupid after all? Int. Stat. Rev. **69**, 385–399 (2001)
44. T. Hastie, R. Tibshirani, J.H. Friedman (1st edn.) (2001) *Elements of Statistical Learning*, 2nd edn. (Springer, New York, 2009)
45. M. Hills, Allocation rules and their error rates (with discussion). J. R. Stat. Soc. B **28**, 1–31 (1966)
46. G.E. Hinton, P. Dayan, M. Revow, Modeling the manifolds of images of handwritten digits. IEEE Trans. Neural Netw. **8**, 65–73 (1997)

47. I.M. Johnstone, A.U. Lu, On consistency and sparsity for principal components analysis in high dimensions. J. Am. Stat. Assoc. **104**, 682–693 (2009)
48. A. Khalili, J. Chen, S. Lin, Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. Biostatistics **12**, 156–172 (2011)
49. Y. Koren, The BellKor solution to the Netflix Grand Prize, 2009. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf
50. A.V. Kossenkov, M.F. Ochs, Matrix factorization for recovery of biological processes from microarray data, in *Methods in Enzymology*, ed. by M.L. Johnson, L. Brand, vol. 467 (Academic, New York, 2009), pp. 59–77
51. P.A. Lachenbruch, Estimation of error rates in discriminant analysis, Unpublished Ph.D. thesis, University of Los Angeles, 1965
52. P.A. Lachenbruch, M.R. Mickey, Estimation of error rates in discriminant analysis. Technometrics **10**, 1–11 (1968)
53. D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
54. B.G. Lindsay, *Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 5 (Institute of Mathematical Statistics and the American Statistical Association, Alexandria, VA, 1995)
55. R.J.A. Little, Contribution to the discussion of the paper by A.P. Dempster, N.M. Laird, D.B. Rubin. J. R. Stat. Soc. B **39**, 25 (1975)
56. F.H.C. Marriott, *The Interpretation of Multiple Observations* (Academic, London, 1974)
57. C. Maugis, G. Celeux, M.-L. Martin-Magniette, Variable selection for clustering with Gaussian mixture models. Biometrics **65**, 701–709 (2009)
58. G.J. McLachlan, Asymptotic results for discriminant analysis when the initial samples are misclassified. Technometrics **14**, 415–422 (1972)
59. G.J. McLachlan, *The Errors of Allocation and their Estimators in the Two-Population Discrimination Problem*, Abstract of unpublished Ph.D. thesis, University of Queensland. Bull. Aust. Math. Soc. **9**, 149–150 (1973)
60. G.J. McLachlan, Estimation of the errors of misclassification on the criterion of asymptotic mean square error. Technometrics **16**, 255–260 (1974)
61. G.J. McLachlan, The relationship in terms of asymptotic mean square error between the separate problems of estimating each of the three types of error rate of the linear discriminant function. Technometrics **16**, 569–575 (1974)
62. G.J. McLachlan, An asymptotic unbiased technique for estimating the error rates in discriminant analysis. Biometrics **30**, 239–249 (1974)
63. G.J. McLachlan, Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. J. Am. Stat. Assoc. **70**, 365–369 (1975)
64. G.J. McLachlan, The bias of the apparent error rate in discriminant analysis. Biometrika **63**, 239–244 (1976)
65. G.J. McLachlan, Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. J. Am. Stat. Assoc. **72**, 403–406 (1977)
66. G.J. McLachlan, A note on the choice of a weighting function to give an efficient method for estimating the probability of misclassification. Pattern Recogn. **8**, 147–149 (1977)
67. G.J. McLachlan, The classification and mixture maximum likelihood approaches to cluster analysis, in *Handbook of Statistics*, ed. by P.R. Krishnaiah, L. Kanal, vol. 2 (North-Holland, Amsterdam, 1982), pp. 199–208
68. G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, New York, 1992)
69. G.J. McLachlan, J. Baek, S.I. Rathnayake, Mixtures of factor analyzers for the analysis of high-dimensional data, in *Mixture Estimation and Applications*, ed. by K. Mengersen, C. Robert, D.M. Titterington (Wiley, Hoboken, NJ, 2011)
70. G.J. McLachlan, K.E. Basford, *Mixture Models: Inference and Applications to Clustering* (Dekker, New York, 1988)

71. G.J. McLachlan, R.W. Bean, L. Ben-Tovim Jones, A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics **22**, 1608–1615 (2006)
72. G.J. McLachlan, R.W. Bean, L. Ben-Tovim Jones, Extension of the mixture of factor analyzers model to incorporate the multivariate *t* distribution. Comput. Stat. Data Anal. **51**, 5327–5338 (2007)
73. G.J. McLachlan, R.W. Bean, D. Peel, A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**, 413–422 (2002)
74. G.J. McLachlan, J. Chevelu, J. Zhu, Correcting for selection bias via cross-validation in the classification of microarray data, in *Beyond Parametrics in Interdisciplinary Research: A Festschrift to P.K. Sen*, ed. by N. Balakrishnan, E. Pena, M.J. Silvapulle (IMS Lecture Notes-Monograph Series, Hayward, CA, 2008), pp. 383–395
75. G.J. McLachlan, K.-A. Do, C. Ambroise, *Analyzing Microarray Gene Expression Data* (Wiley, Hoboken, NJ, 2004)
76. G.J. McLachlan, T. Krishnan (1997) *The EM Algorithm and Extensions*, 2nd edn. (Wiley, New York, 2008)
77. G.J. McLachlan, S.K. Ng, The EM algorithm, in *The Top-Ten Algorithms in Data Mining*, ed. by X. Wu, V. Kumar (Chapman & Hall, Boca Raton, FL, 2009), pp. 93–115
78. G.J. McLachlan, D. Peel, Robust cluster analysis via mixtures of multivariate *t*-distributions. Lect. Notes Comput. Sci. **1451**, 658–666 (1998)
79. G.J. McLachlan, D. Peel, *Finite Mixture Models* (Wiley, New York, 2000)
80. G.J. McLachlan, D. Peel, Mixtures of factor analyzers, in *Proceedings of the Seventeenth International Conference on Machine Learning*, ed. by P. Langley (Morgan Kaufmann, San Francisco, CA, 2000), pp. 599–606
81. G.J. McLachlan, D. Peel, K.E. Basford, P. Adams, The EMMIX software for the fitting of mixtures of normal and *t*-components. J. Stat. Software **4**(2), 1–14 (1999)
82. G.J. McLachlan, D. Peel, R.W. Bean, Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**, 379–388 (2003)
83. T.M. Mitchell, Mining our reality. Science **326**, 1644–1645 (2010)
84. V. Nikulin, G.J. McLachlan, On a general method for matrix factorisation applied to supervised classification, in *Proceedings of 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop*, Washington, DC, ed. by J. Chen et al. (IEEE Computer Society, Los Alamitos, CA, 2009), pp. 43–48
85. V. Nikulin, T.-H. Huang, S.K. Ng, S.I. Rathnayake, G.J. McLachlan, A very fast algorithm for matrix factorization. Stat. Probab. Lett. **81**, 773–782 (2010)
86. V. Nikulin, G.J. McLachlan, Penalized principal component analysis of microarray data, in *Lecture Notes in Bioinformatics*, ed. by F. Masulli, L. Peterson, R. Tagliaferri, vol. 6160 (Springer, Berlin, 2010), pp. 82–96
87. S.K. Ng, G.J. McLachlan, Using the EM algorithm to train neural networks: misconceptions and a new algorithm for multiclass classification. IEEE Trans. Neural Netw. **15**, 738–749 (2004)
88. S.K. Ng, G.J. McLachlan, K. Wang, L. Ben-Tovim, S.W. Ng, A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics **22**, 1745–1752 (2006)
89. T.J. O'Neill, *Efficiency Calculations in Discriminant Analysis*, Unpublished Ph.D. thesis, Stanford University, Stanford, CA, 1976
90. T.J. O'Neill, Normal discrimination with unclassified observations. J. Am. Stat. Assoc. **73**, 821–826 (1978)
91. D. Peel, G.J. McLachlan, Robust mixture modelling using the *t* distribution. Stat. Comput. **10**, 335–344 (2000)
92. S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L.M. Maier, C. Baecher-Allan, G.J. McLachlan, P. Tamayo, D.A. Hafler, P.L. De Jager, J.P. Mesirov, Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. USA **106**, 8519–8524 (2009)

93. A.E. Raftery, N. Dean, Variable selection for model-based clustering. J. Am. Stat. Assoc. **101**, 168–178 (2006)
94. B.D. Ripley, Neural networks and related methods for classification (with discussion). J. R. Stat. Soc. B **56**, 409–456 (1994)
95. B.D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, 1996)
96. R. Tibshirani, Regression shrinkage and selection via lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
97. R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat. Sci. **18**, 104–117 (2003)
98. D.M. Titterington, A.F.M. Smith, U.E. Makov, *Statistical Analysis of Finite Mixture Distributions* (Wiley, New York, 1985)
99. G.T. Toussaint, P.M. Sharpe, An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis. Comput Biol Med **4**, 269–278 (1975)
100. M.J. van der Laan, S. Rose, Statistics ready for a revolution: next generation of statisticians must build tools for massive data sets. *Amstat News*, September Issue, 2010
101. D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**, 515–534 (2009)
102. M. Wojnarski, A. Janusz, H.S. Nyugen, J. Bazan, C.J. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, H. Luo, J. Gao, Y. Shen, V. Nikulin, T.-H. Huang, G.J. McLachlan, M. Bosnjak, D. Gamberger, RSCTC 2010 discovery challenge: mining DNA microarray data for medical diagnosis and treatment, in *Lecture Notes in Artificial Intelligence* **6086** (Proceedings of RSCT 2010), ed. by M. Szczuka et al. (Springer, Berlin, 2010), pp. 4–19
103. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, S.K. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**, 1–37 (2008)
104. X. Zhu, C. Ambroise, G.J. McLachlan, Selection bias in working with the top genes in supervised classification of tissue samples. Stat. Methodol. **3**, 29–41 (2006)
105. J.X. Zhu, G.J. McLachlan, L. Ben-Tovim, I. Wood, On selection biases with prediction rules formed from gene expression data. J. Stat. Plan. Inference **38**, 374–386 (2008)

The Journey of Knowledge Discovery

Gregory Piatetsky-Shapiro

> All truths are easy to understand once they are discovered; the point is to discover them.
>
> Galileo Galilei

1 Learning Mathematics

One of my earliest memories is from the time when I was about 4 years old, walking with my father and solving problems such as "if the sum of 2 numbers is 10 and the difference is 2, what are the numbers." I was very happy when I found that adding the sum and the difference and dividing by two gives one of the numbers. My father, Ilya Piatetski-Shapiro, was one of the leading mathematicians in the Soviet Union, and he passed along to me an appreciation of mathematics and an analytical frame of mind, but not enough mathematical talent to follow in his footsteps.

Nevertheless, I was sufficiently good at math to be accepted into the seventh grade of a leading mathematical school in Moscow, school #2. The school had many brilliant students. One of my classmates (Alexander Shen) was a winner of the Soviet Union Math Olympiad, and another was a winner of the Soviet Union Physics Olympiad. I enjoyed solving puzzles and participated in Moscow Mathematical Olympiads, but without any major successes. Here is one very elegant problem which I remember from those Olympiads.

> Given a convex polyhedron with N sides, prove that there are at least two sides with the same number of edges.

I remember submitting answers to combinatorics problems to a Moscow math journal for students called "Quant" and feeling very proud upon seeing my name in

G. Piatetsky-Shapiro (✉)
KDnuggets, Brookline, MA, USA
e-mail: Gregory@kdnuggets.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_13,

print when I was among the first to submit the solution. There was a great atmosphere of learning and discovery in school, and I acquired a number of lifelong friends from that time. One of them is Alex Tuzhilin, now a professor at NYU and a leading researcher in the area of data mining and recommendation systems. He was also my classmate in Moscow. I can probably claim a small credit of making him interested in the field of knowledge discovery when both of us were in college.

I also played a lot of chess at that time. Our school had quite a few very strong chess players and was close to the Young Pioneers' Palace in Moscow, where I went to learn and play chess. One of my schoolmates, Valery Chekhov, became World Junior Chess Champion in 1975; Arthur Yusupov was a contender for the World Chess Championship 1986–1992. Anna Aksharumova became the Soviet Union's Women Chess Champion and later became a US Women Chess Champion. In many chess blitz games I played with her, I managed to win one or two. At one point, my father became concerned that I was playing too much chess and asked one of his students to talk to me to convince me to play less chess, forgetting that this student was himself a chess master. I was not the best mathematician or chess player in our school, but I probably was the best chess player among strong mathematicians and the best mathematician among strong chess players.

2 Emigrating and Discovering Computers

In the 1960s and 1970s, the Soviet Union had strong official anti-Jewish policies. The future for Jewish students like me was bleak—unlike my Russian schoolmates I had little chance of being accepted in a good college or university. My father and mother, like other Soviet Jews, faced discrimination at work. For example, my father was invited to present a paper at the International Mathematics Congress in Stockholm—a very high honor for a mathematician, but he was not allowed to attend because he was Jewish.

By 1970, my parents were divorced, but both of them wanted to leave the Soviet Union, in large part to give me a chance for a better life. My mother and I emigrated to Israel in March of 1974. After we left, my father applied for permission to emigrate but received a "refusal." He was told that his former wife and son also were "refused." Thus, my mother and I became perhaps the only people to be refused emigration after already leaving. My father was able to leave in 1976, after a strong campaign by American scientists.

I was only 15 when I finished high school in Moscow. When I started at Tel Aviv University in the fall of 1974, I was probably the youngest student. Naturally, I started in the math department.

Fortunately for me, among the obligatory calculus and algebra classes, there was also an introduction to computers (then taught in FORTRAN). I immediately loved the computer class, especially the creative process of finding an algorithm to make the computer solve a problem. No doubt, I was greatly influenced by many science fictions and robot novels I read in my youth (Stanislaw Lem, A&B Strugatsky, and Isaac Asimov come to mind). I wanted to write a fun program that would show

some artificial intelligence and decided to write a program for playing a game. I spent a lot more time writing programs on my own rather than doing class assignments. I spent the spring semester of 1975 at Technion (Israel's equivalent of MIT). At Technion, I took a class in APL, which is a very elegant language, with very powerful array and matrix operators denoted by its own unique syntax of symbols and Greek letters. IBM at that time even produced special APL keyboards. The simplest game I could program was Battleships and I spent a lot of time after classes debugging my APL code. Finally, the program was finished. With great anticipation, I played one game and was utterly defeated by my own program. I never played that game again, but I enjoyed writing the program and learned a lot about programming in the process.

In my third and final year at Tel Aviv University (college there was 3 years), I was struggling with more advanced math and really hated functional analysis. My father suggested that I should apply to a US graduate school to study computer science, which was just emerging as a separate discipline. I was accepted at several schools in the USA including Stanford and Yale, but the best scholarship offer was from New York University's Courant Institute and I went there.

3 New York University, Artificial Intelligence, and Databases

I arrived in the USA on Labor Day, September 5, 1977, with one small suitcase and $10 in my pocket. NYU is in the middle of Greenwich Village, which at that time was a little run-down, but a very lively area. I threw myself enthusiastically into studying computer science. I enjoyed the classes on hardware, where we built an actual microprocessor (and programmed it in hardware to play an optimal game of Nim). I learned a lot in the classes on logic and computability taught by Martin Davis. We learned about NP-completeness and algorithms complexity, which was a new and mostly uncharted area. I especially liked classes on natural language processing and artificial intelligence taught by Prof. Ralph Grishman. However, my scholarship ended after 4 years. With my Ph.D. thesis still far from finished, I needed to find a job.

Soon after I arrived in the USA, I went to Brookline (a suburb of Boston, MA) to visit Vlad Rutenburg, another good friend from my school in Moscow. He organized a party for his birthday, where I met Marina, a beautiful young woman who was also from Moscow. She was also studying in New York at that time—uptown at Barnard College (a part of Columbia). We fell in love, and by 1981, we were married. I moved to Boston where Marina's family lived. I started to work with her father, a computer scientist, at Strategic Information, a company that was developing reporting and database systems for financial applications.

Doing theoretical research for a Ph.D. was hard, but since I was developing advanced database systems at work, I wanted to leverage it to get some interesting results for a Ph.D. and have an application of machine learning. I came up with the idea of a self-optimizing database system.

At that time, in-memory databases were unimaginable, and query optimization depended on a good selection of secondary indices. When I started to study this problem, I quickly found that the optimal index selection can be reduced to a set cover (a known NP-complete problem) which meant that index selection was also NP-complete. This was my first publication in SIGMOD record [1].

The database systems we were developing needed good query optimization, and this frequently required estimating the number of records satisfying a condition of the type $X > constant$.

A system would be able to compute such estimates quickly if it had precomputed a histogram for field X. At that time, query optimizers used histograms with equal-width discretization, e.g., a field salary could be discretized into bins [0, 10K, 20K, . . ., 100K]. We found that there were many fields with very skewed distribution, so that equal-width discretization would put almost all values into one bin. I came up with the idea of equal-height discretization (e.g., the first bin has 10% of the records and the second bin has the next 10% of records). I also showed how to use the scheme to have a guaranteed upper bound on the errors and published the results in SIGMOD 1984 [2].

The equal-height discretization later proved very useful for clustering when our goal was to have meaningful rules in a banking application. As a first step, many numeric fields had to be discretized, but instead of strict equal-height bins, which could look ugly, e.g., 103.46–657.98, we rounded the bin boundaries to the most significant 2–3 digits, to make them humanly readable, e.g., 100–700.

For my dissertation [3], I combined these results and also studied the behavior of greedy algorithms for set cover, which had practical applications to index selection. I found that on average, a greedy set cover was finding solutions very close to optimal, and was able to find an elegant proof of a minimum bound for a greedy set cover.

$$\text{Greedy set cover} \geq 1 - \frac{1}{e} \simeq 63\% \text{ of the optimal cover.}$$

My thesis, “Self-Organizing Database System – a Different Approach to Query Optimization,” combined a practical application to database systems, machine learning, and mathematical proofs. It was well received at NYU. In 1984, I got an award for the best Ph.D. dissertation in computer science, and in 1985, I got the award for best dissertation in all natural sciences at NYU.

4 GTE Laboratories: On the Way to Discovery

In 1985, I left Strategic Information and joined GTE Laboratories. At that time, telephone service in the USA was in transition from a longtime AT&T monopoly to a more deregulated environment. AT&T was split in 1984 into one long distance telephone company and seven “Baby Bells.” GTE (originally standing for General

Telephone and Electric) was the largest non-Bell telephone company, with about 20 million local telephone customers, mainly in Texas, Florida, and LA areas.

GTE Labs was a smaller version of AT&T Bell labs. During the regulated telephone service era, the tail end of which I caught in 1985, GTE could afford a research lab which would work on longer term research problems.

I joined the Knowledge Based System department, managed by Shri Goyal, and started to work on natural language interfaces to databases. This was very interesting but far from any practical applications. Next year, our project team, lead by Gabriel Jacobson, started to work on a project named CALIDA on integrating heterogeneous databases [4]. We had Xerox Lisp machines and were writing LISP code to access other large databases via a network.

The integration and join of large databases could take many hours, if not days, if done inefficiently. In 1986, while working on CALIDA, I found a query that could be optimized by a factor of 100 if the optimizer knew a simple pattern like "file1 join with file2 will always have field1 = A."

I started to look at ways of automatically discovering such patterns. In 1987, I attended Gio Wiederhold's tutorial "Extracting Knowledge From Data" at the ICDE conference in Los Angeles. Gio and his student Blum [5] had developed Rx, the first program that analyzed historical data from about 50,000 Stanford patients, and looked for unexpected side effects of drugs. The program discovered some side effects that were unknown to its authors, and the approach looked very promising. That research was very exciting, and I wanted to do something similar at GTE.

However, I could not quite convince GTE management that discovery in data was a good idea. One senior manager told me that he thought that data mining was a solved problem—I could apply a decision tree (and building a decision tree was a solved problem, in his opinion) to the database and presto—I would have all the results I needed.

In 1988, I attended an AAAI workshop in Minneapolis on "Databases and Expert Systems" organized by Larry Kershberg. The workshop had interesting presentations and was a good way to get researchers to interact. Putting together a workshop seemed relatively easy (little did I know), and I decided to organize a workshop on discovery in data at next year's IJCAI-89. That would be a great way to stimulate more research in the field and to convince my management at GTE Laboratories that discovery in data was a good idea.

5 First Workshop on KDD: Knowledge Discovery in Databases

What should I call this workshop? The name "data mining," which was already used in the database community, seemed prosaic, and besides, statisticians used "data mining" as a pejorative term to criticize the activity of unguided search for any correlations in data that was likely to find something even in random data. Also, "mining" sounded prosaic, and "data mining" gave no indication of what we were mining for. "Knowledge mining" and "knowledge extraction" did not seem much

better. I came up with "Knowledge Discovery in Databases," which emphasized the "discovery" aspect and the focus of discovery on "knowledge."

With encouragement and help from Jaime Carbonell (CMU), William "Bud" Frawley (GTE Labs), Kamran Parsaye (IntelligenceWare), Ross Quinlan (U. of Sydney), Michael Siegel (BU), and Sam Uthurusamy (GM Research), I put together a Knowledge Discovery in Databases (KDD-89) workshop at IJCAI-89 in Detroit [6]. The KDD-89 workshop was very successful [7], receiving 69 submissions from 12 countries. It was the largest workshop at IJCAI-89, with standing-room-only attendance. KDD-89 had nine papers presented in three sessions, on Data-Driven Discovery, Knowledge-Based Approaches, and Systems and Applications. While some topics discussed at KDD-89 have faded from the research agenda (e.g., Expert Database Systems), other topics such as Using Domain Knowledge, Dealing with Text and Complex Data, and Privacy remain just as relevant today.

6 First Knowledge Discovery Project

After the success of the KDD-89 workshop, I convinced GTE management that knowledge discovery was a good research idea with many applications, and was put in charge of a new project, which I named "Knowledge Discovery in Data." I believe it was the first project with such a name. We worked on several small tasks dealing with fraud detection and GTE Yellow Pages until we came up with a really good application to health care.

6.1 KEFIR: Key Findings Reporter

Already in 1995, US health-care costs consumed 12% of the GDP—Gross Domestic Product—and were rising faster than the GDP.[1]

Some of the health-care costs are due to potentially fixable problems such as fraud or misuse. Understanding where the problems are is the first step to fixing them. Because GTE, a large telephone company, was self-insured for medical costs, it was very motivated to reduce them. GTE's health-care costs in the mid-1990s were in hundreds of millions of dollars.

The task for our project in support of GTE Health Care Management was to analyze employee health-care data and identify problem areas which could be addressed. With Chris Matheus and Dwight McNeil, we developed a system called Key Findings Reporter, or KEFIR [9]. The KEFIR approach was to analyze all possible deviations, then select the most actionable findings using interestingness [10] (see Fig. 1). KEFIR also augmented key findings with explanations of

[1] As of 2010, US health-care costs were estimated at 15% of GDP [8].

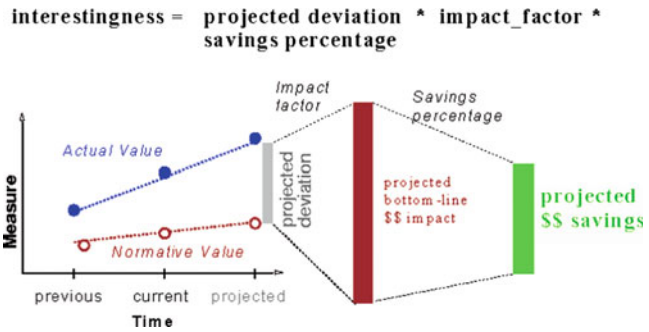


Fig. 1 KEFIR measure of interestingness

plausible causes and recommendations of appropriate actions. Finally, KEFIR converted findings to a user-friendly report with text and graphics [11].

KEFIR was a very innovative first-time project and received a top GTE technical award.

Currently, we can see some of the same ideas in Google Analytics Intelligence, which automatically finds significant deviations from the norm across multiple hierarchies.

6.2 CHAMP

After KEFIR, in 1996, I worked on predicting customer "churn" (attrition). GTE Wireless had about 3.5 million customers in 1996 and was adding new customers at a rate of 1.5 million/year. At the same time, it was losing customers at a rate of 900,000 customers a year. With average revenue of $55 per customer per month, this amounted to losing $600 million a year.

Working with Brij Masand and several other researchers, we developed a system called CHAMP (Churn Analysis, Modeling, and Prediction) [12]. CHAMP had a three-level architecture, with a data engine retrieving the data directly from the data warehouse via SQL; discovery engine combining decision trees, neural networks, and other modeling approaches; and a browser-based front-end. CHAMP scored all wireless customers and transferred the results back to the data warehouse.

One of the important ideas in CHAMP was *automatically finding data changes*. Our initial neural-net models used a field which encoded the phone type. At one point, we found that the models had become inaccurate. Examination of data showed that the definition of that field had changed and it was encoding another feature of the phone. After that, we developed a profiler program (using KEFIR technology) that automatically compared the statistics of all fields in the new month vs. those of the old month, and identified significant changes that required updating and rebuilding the models.

We also experimented with a single model for all regions and found that building customized models for each of the five major service areas improved prediction quality significantly.

CHAMP was applied to all of GTE's 4 million cellular customers, with an estimated savings of $20–30 million/year.

7 Many Names of Data Mining

The mining and knowledge discovery field has been called by many names.

In 1960s, statisticians have used terms like "data fishing" or "data dredging" to refer to what they considered a bad practice of analyzing data without a prior hypothesis.

The term "data mining" appeared around the 1990s in the database community. Some started to use "*database* mining™," but found that this phrase was trademarked by HNC (now part of Fair, Isaac) and could not be used. Other terms used include Data Archeology, Information Harvesting, Information Discovery, and Knowledge Extraction.

I coined the term "Knowledge Discovery in Databases" (KDD) for the first workshop on the same topic (1989), and this term became popular in academic and research communities. However, the term "data mining" became more popular in the business community and in the press.

In 2003, "data mining" acquired a bad image because of its association with US government program called TIA (Total Information Awareness). Headlines like "Senate Kills Data Mining Program," ComputerWorld, July 18, 2003, referring to a US Senate decision to close down TIA, helped increase the negative image of "data mining."

In 2006, the term "analytics" jumped to great popularity, driven by the introduction of Google Analytics (Dec 2005) and later by a book "Competing on Analytics" [13]. According to Google Trends, "analytics" became more popular than "data mining," as measured by Google searches, around 2006, and has continued to climb ever since (see Fig. 2).

As of Jan 2011, Google search for "knowledge discovery" finds 2.5 M pages, while search for "data mining" finds about 13M pages, and search for "analytics" finds 120M pages.

7.1 Knowledge Discovery Nuggets and Knowledge Discovery Mine

I started the Knowledge Discovery Nuggets e-mail newsletter in 1993 as a way to connect researchers who attended the KDD-93 workshop in Anaheim, CA. The first issue, in text format, is still online [14].

In 1994, our group at GTE discovered a great new thing called the World Wide Web, and in May of 1994, I launched a web site called Knowledge Discovery Mine [15] at info.gte.com/~kdd (URL no longer valid). I received many inquiries about

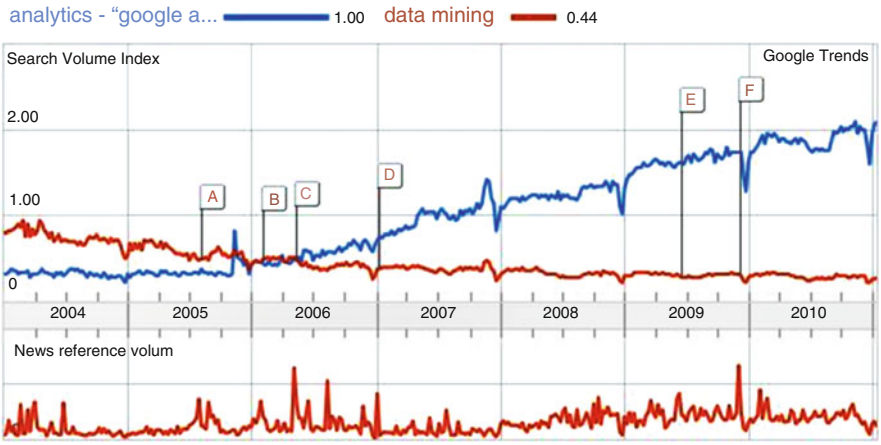


Fig. 2 Google Trends for analytics—"Google Analytics" vs. data mining

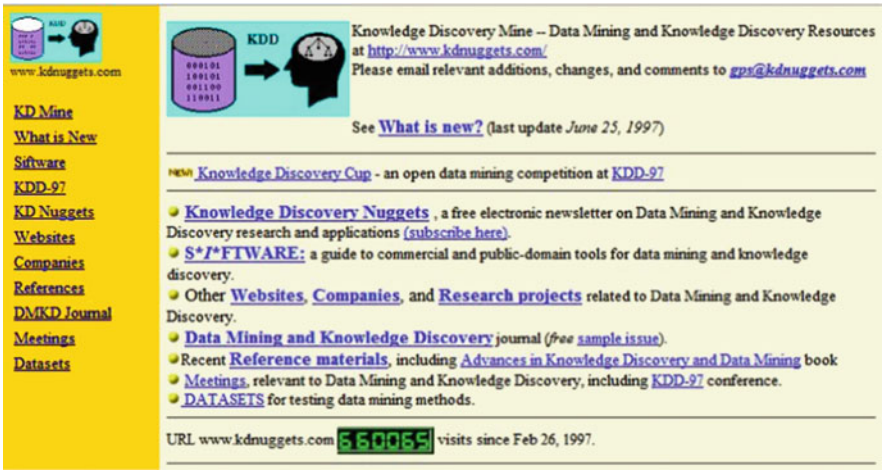


Fig. 3 A screenshot of early KDnuggets homepage in 1997, obtained from archive.com (*counter value not correct*)

software, meetings, and other activities related to KDD, so having a web site with the directory of everything related to knowledge discovery seemed like an easy way to answer these questions at once.

I am grateful to Michael Beddows and Chris Matheus who helped me maintain Knowledge Discovery Mine when I was at GTE. Shortly before I left GTE in 1997, I moved the Knowledge Discovery Mine to a new domain I called *KDnuggets.com* (see Fig. 3) for Knowledge Discovery Nuggets.

Having both an active newsletter and a web site was a virtuous combination, and the number of subscribers to KDnuggets e-mail newsletters grew from 50 who received the first issue in 1993 to 6,000 by the end of 1998 and about 11,000 by the

end of 2001, when the dot-com bubble burst. Currently, the number of e-mail subscribers oscillates around 12,000, but other KDnuggets channels like RSS and @kdnuggets on twitter have a growing number of followers.

While researchers constituted the majority of the subscribers in the first few years, now the majority is from commercial domains. About half of the subscribers are from .com and .net domains, but there are subscribers from about 95 countries, and on all continents except Antarctica.

8 Knowledge Stream Partners: The Life in a Start-Up

By 1997, it was clear to me that KDD had many applications outside of the narrow scope of GTE. In January 1997, I got a call from Robert van der Hooning, a business school professor and an entrepreneur. He created a start-up (eventually called Knowledge Stream Partners or KSP) with the motto of *delivering the right offer to the right customer at the right time*. Data mining played a key role in understanding what was the right offer, the right customer, and the right time. Robert got a contract from one of the largest Swiss banks and assembled a team of experts in data warehousing, user interfaces, and project management. My role was to lead the data mining group.

Robert was a great leader from whom I learned a lot. One of his directives was "Hire people smarter than yourself." I was always tempted to reply—Robert, it is easy for you to do—but I followed his advice and was able to hire an excellent data mining group, which included Steve Gallant, Sam Steingold, Natasha Markuzon, and Moninder Singh.

We were probably among the first people building attrition, cross-sell, and other models for actual banking customers. While these models were not perfect, they were sufficiently better than random to be useful. We built a number of cross-sell and attrition models for a Swiss client, learned to say "Gruezi" (Swiss greeting), and enjoyed visiting beautiful Zurich.

Our best work was probably building a segmentation model using AUTOCLASS (Bayesian clustering program developed by NASA). When we applied AUTOCLASS to raw numeric data, the results were not comprehensible, but when we started by discretizing the numeric variables using an intelligent equal-height method (a variation on my Ph.D. approach), it turned out that the program produced very meaningful segments. For example, our models automatically identified several groups of accounts with distinct behavior (the simplest to describe was "accounts used by landlords to hold the security deposits").

After Switzerland, we also worked with several of the largest banks and insurance companies in New York and built successful predictors of customer behavior. One large insurance company wanted to see how quickly it needed to adjust the annuity rates in order to reduce attrition. Out data analysis showed, sadly for the consumer, that attrition was not dependent much on the rates.

However, the more frequently the company mailed the customer, the higher the attrition rate was.

8.1 Estimating Campaign Benefits and Modeling Lift

In most direct marketing applications like attrition or cross-sell, the goal is to find a subset of customers most likely to respond, and models typically create a list of customers sorted by decreasing likelihood of response. A typical measure of quality of such models is called Lift. Lift compares the target frequency of a subset with the target frequency in the entire population and is defined as

$$\text{Lift(Subset)} = \frac{\text{NumTargets(Subset)}}{\text{NumTargets(All)}}.$$

By 1999, I worked on a number of projects and noticed a curious phenomenon—lift curves seemed very similar. For example, the lift at 10% of the list was usually around 3. Brij Masand [16] and I investigated about 30 different direct marketing models from several industries, which used neural nets, decision trees, and other methods. Since lift varies at different points, we looked initially at lift at $T\%$ of the list where T is the fraction of targets in the population and found this rule of thumb (Fig. 4):

For targeted marketing campaigns, a good model lift at T, where T is the target rate in the overall population, is usually sqrt(1/T) +/- 20%.

We generalized the lift formula to all points as $Lift(P) = 1/\text{sqrt}(P) = P^{-0.5}$.

Our meta-analysis of the direct marketing problems found that lift curve was typically in the range $P^{-0.4} < Lift(P) < P^{-0.6}$. This formula also allows estimation of potential campaign benefits in advance of the modeling exercise. For more details, see [16].

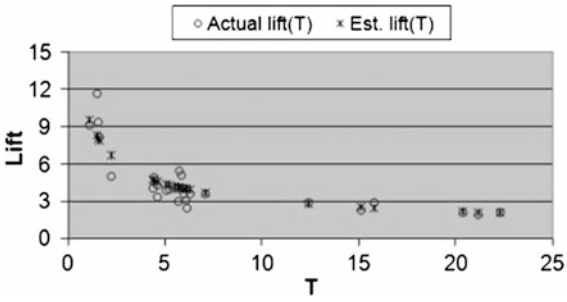


Fig. 4 Actual lift at special point T (target frequency) vs. estimated lift

8.2 The Unexpected Effect of Y2K

The Y2K bug, as it was called, was due to many programmers using only the last two digits to specify the year. (*Note*: Abbreviating year to the last two digits is not unique to twentieth century. Painters at the end of nineteenth century also used only the last two digits of the year when signing the painting, e.g., "Renoir 73" in Fig. 5.)

What seemed normal to programmers in the 1970s and 1980s, when they were motivated by saving a few bytes of memory, could have led to significant problems once the year changed from 1999 to 2000. Then the two-digit representation "00" of 2000 would lead to incorrect logic—e.g., Jan 1, 2000 (represented as 000101) would be smaller (usually interpreted as earlier) than Dec 31, 1999 (represented as 991231).

The actual problems on Jan 1, 2000 were minor, but it was no doubt due to the significant work done by almost all banks and financial companies ahead of time. As part of their preparation for Y2K, large financial companies had frozen all the new projects and developments in the last part of 1999. Since all of our clients were big banks or other financial companies, this resulted in our start-up not getting any projects at the end of 1999. Robert, our CEO, kept KSP going but had to reduce our salaries significantly and started to look for a buyer for KSP.

The extra free time gave me more time to work on KDnuggets, and reduced salary was a good motivation to introduce advertising, which I did in early 2000. My very first client for banner ads was Megaputer, a US company with a development group in Russia working on analytic and text mining software.

First ad on Kdnuggets (120x60 gif)

Fig. 5 "Renoir 73." Signature from 1873

With time, I expanded the number of ads to accommodate the demand and had ads from most leading data mining and analytics companies. I wrote my own Javascript code for ad rotation and added text links under ads, which significantly increased the click-thru rate for the first few years.

8.3 Riding the Dot-Com Bubble

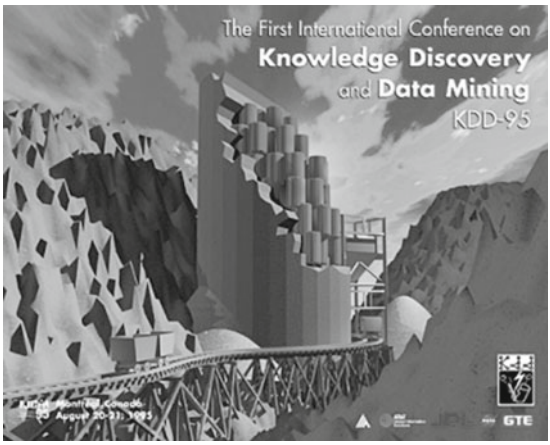
Xchange (originally Exchange Applications) was a Boston area start-up which was developing CRM and campaign management solutions. They went public early in the dot-com era, and although they were not very profitable, their sales numbers went up every quarter. In retrospect, I think this was partly because they obscured their financials by buying new companies. In any case, their stock valuation soared to $1 billion dollars in early 2000.

In April 2000, KSP was acquired by Xchange for about $50 million of Xchange stock, and on paper, my stock options were worth a few millions. Unfortunately, this was an all stock transaction, and Xchange stock started to decline dramatically shortly afterward. The acquisition condition required KSP staff to wait 6 months before cashing the stock options, and in those 6 months, Xchange stock went from about $50/share to about $1/share, putting the stock options underwater. In early 2001, it became clear to me that Xchange was going nowhere. KDnuggets, however, was doing quite well, and I left Xchange in May 2001 to try the uncertain but exciting life of independence.

9 SIGKDD and KDD Conferences

After organizing the first three workshops on KDD in 1989, 1991, and 1993, I lacked the enthusiasm for running another workshop and wanted to bring in fresh blood. One of my best ideas was to ask Usama Fayyad to chair the KDD-94 workshop in Seattle. Usama attended earlier KDD workshops, and at one point, I even offered him a summer job at GTE Labs, which he wisely declined and went to NASA instead. Usama did a great job organizing KDD-94 with Sam Uthurusamy. In 1995, Usama, being more ambitious than I, said "why don't we make a workshop into a full-fledged conference?" We did that next year in Montreal with KDD-95, First Conference on Knowledge Discovery and Data Mining.

The conference was very successful, drawing top researchers from AI, machine learning, and database fields. The poster reception which combined technical posters with an enticing buffet, wine, beer, and delicious French pastries set the standard for future KDD conferences.



KDD-95 Conference Poster

I created a steering committee to run the technical part of KDD conferences, but the conference was still managed by AAAI. This freed me from worry about logistics and financial issues, but it was constraining our growth and appeal in other fields. The KDD field was by nature very interdisciplinary and had strong connections to machine learning, statistics, and database areas. However, the statistics and database researchers and practitioners were not likely to go to an artificial intelligence conference.

We wanted to separate ourselves from AAAI conferences, but the legal and financial hurdles of creating a completely separate KDD society seemed too great. At about this time, I was contacted by Won Kim, a leading database researcher who had just completed 8 years as a chair of SIGMOD. Won was well versed in ACM organization and was interested in the data mining field. He suggested creating a special interest group—SIG inside ACM, similar to SIGMOD. Thus, SIGKDD was born in 1998 with Won Kim as a chair. I served as a director and a member of the Executive Committee till 2005, when Won retired, and I was elected as SIGKDD chair in 2005.

Among my achievements was instituting the SIGKDD Best Dissertation Award (http://www.sigkdd.org/awards_dissertation.php) which attracted many excellent submissions and helped data mining gain more respect in academic departments. I also helped to create the SIGKDD Curriculum proposal for a degree in data mining and knowledge discovery (http://www.sigkdd.org/curriculum.php). The SIGKDD balance sheet increased, and the number of KDD conference attendees rose to around 700–800.

The SIGKDD main activity was to organize the annual KDD conference. Having conferences run by volunteers frequently creates a situation of semicrises, and it is SIGKDD chair's job to solve the crises. Each conference has an organizing committee of about 30 people, all volunteers, reporting to the general chair. If all goes well, like it did at KDD-06 (chaired by Lyle Ungar) and KDD-08 (chaired by Ying Li), the SIGKDD chair does not need to be involved much after the initial selection of the conference location and key organizers.

For the KDD-07 conference in San Jose, Rakesh Agrawal agreed to be the general chair, but soon afterward, he moved from IBM to Microsoft and stopped

replying to my e-mails or preparing for the conference. By the end of KDD-06, no organization had been selected for KDD-07, no hotel had been chosen, and the next conference was about to collapse. I took the unprecedented step of asking Rakesh to resign as the general chair and was very lucky that Pavel Berkhin agreed to step in as general chair at the last moment. Pavel was already extremely busy with his job as a director at Yahoo—there were days when he did not have a free minute—but he brought his organizational skill, charm, intelligence, and energy to the task and rescued KDD-07. Under Pavel, KDD-07 chairs created a very useful wiki to hold the knowledge of all the functional chairs that continues to be used today.

KDD conferences were always held in North America, but a number of people were suggesting locations abroad. In 2005, Christophe Giraud-Carrier invited me to give a talk at Brigham Young University in Provo, Utah. BYU is a Mormon university, and it was the first time my talk (on microarray data analysis) was preceded by a benediction. The talk went well, and afterward, Christophe suggested that as a top-tier conference, KDD should also give European researchers more opportunities to attend and participate. I agreed with him and thought that Paris would be a great location for a conference, especially since SIGMOD, a top database conference always held in North America, had just had a first-ever European conference in Paris in 2004.

I convinced Francoise Fogelman, a Parisian, a scientist, and a VP at French data mining company KXEN, to be the European co-chair. John Elder was the US-based co-chair. We started planning the conference in 2007 and quickly found that everything was more expensive and more complicated to do in Europe. Managing a budget in two currencies and dealing with VAT was a nightmare (much of it shouldered by Ismail Parsa, KDD-09 treasurer). By the time the financial crisis struck in the fall of 2008, the hotel contract had already been signed, and we were worried that not enough people would show up.

But finally, the conference was a great success, with 9 tutorials, 11 workshops, about 100 excellent papers, which drew almost 900 people. Many presentations are available at videolectures.net. The highlight was probably the conference poster/reception session held at the historic Hotel de Ville.

Hotel de Ville, Paris

KDD-09 Paris poster

10 Independence

Since 2001, I have been working as an independent consultant on a large variety of problems. I worked with companies in e-commerce, web, telecom, pharma, life sciences, industrial software, state government, and other areas. There were many different types of problems, including microarray data analysis, proteomics, text mining, fraud and piracy detection, nonpayment of child support, user/feature clustering, online recommendation systems, software analysis, link analysis, and customer grouping. I also served as an expert witness in several cases.

Most of this work is confidential for the customers, but there are a few interesting projects I can describe.

10.1 Best Practices: Microarray Clementine Application Templates

At KDD-2001, I attended an inspiring keynote talk by Russ Altman on "Challenges for Knowledge Discovery in Biology." Whereas in business analytics, success is usually measured in money, in biology, there are opportunities to both advance scientific knowledge and save and improve lives of patients by finding diagnostic tests for different diseases and ultimately helping to cure cancer, Alzheimer, Parkinson, and a myriad other genetic-related ailments.

Microarray data analysis [17] is an especially interesting area, both in terms of its potential and its data analysis challenges. Microarray chips measure the expression levels of many genes simultaneously. While a large set of all relevant genes is desirable for exploration, for diagnostics or identification of therapeutic targets, the smallest reliable set of genes is needed. Unlike typical business data, where there are millions of records but only hundreds of columns, microarray data typically have many thousands of columns (genes) but typically only about a hundred samples or less.

As a result, it becomes very easy to find patterns due to randomness and extra steps of data cleaning and error elimination are needed. Also, the classification of samples by human experts may be subject to errors, and a single misclassification may skew the predictive model significantly.

In 2002, I attended a biological conference where SPSS was showing Clementine Data Mining Software, along with templates which capture the entire process for telco, financial, and other applications. However, SPSS did not have a template for microarray data analysis. I convinced SPSS that I can develop such templates and enlisted the help of Sridhar Ramaswamy, a brilliant young doctor and a researcher at MIT Whitehead Institute. Tom Khabaza of SPSS helped with Clementine, and together we developed templates for microarray data analysis which captured the best practices (Fig. 5 shows a partial example of such template).

Among the important ideas were multilevel, leave-one-out cross-validation; train many independent and randomized neural nets; and combine their predictions in a voting scheme. Thus, instead of assigning a sample to a single class, we would compute for each sample the probability of it belonging to each class. This allowed us to detect misclassified instances when an item was wrongly labeled. The key parameter was the number of best genes per class that we selected using cross-validation. Typically when the number of genes was increasing, the average classification error decreased up to a certain point, and then started to increase again. Our method was to select the number of genes in the middle of plateau. Finally, we selected the best number of genes per class also using cross-validation, and only after that applied the best model to the held-out set that was not used in training before.

For a well-known ALL/AML microarray data [18], the cross-validation curve shows the desired behavior, with ten genes being at the center of the optimum plateau. We selected this gene subset to build a new neural net model on the full training data (using 70% as the training data) and applied it to the test data (34 samples), which up to this point had not been used in any way in model development.

Evaluation of the test data gave us 33 correct predictions out of 34 (97% accuracy), which compares well with many previously reported results on the same data.

Note: the single misclassification was on sample 66 which has been consistently misclassified by other methods and is believed to be incorrectly labeled by the pathologist [19], which suggests that the classifier was actually 100% correct. These results are as good as any reported on this dataset, which was the focus of CAMDA-2000 conference/competition [20]. Equally good results were obtained on other, more complex datasets.

Our work came to be known as the SPSS Clementine Application Templates (CATs) for Microarrays and was described in [21], which won an application paper award at KDD-2003.

A nice side effect of this paper is that my Erdos number (http://en.wikipedia.org/wiki/Erdos_number) became 4 as I was a coauthor with Ramaswamy, who had many papers with a leading genetics researcher Eric Lander who has an Erdos number of 2.

10.2 Alzheimer Detection

Another very interesting project with a biological application was my work on early diagnosis of Alzheimer from spinal fluid. Alzheimer is a disease that develops over many years, and even if there is no effective treatment, an early diagnostic test can be very valuable and could lead to new treatments.

Alzheimer affects the brain, and there are reasons to believe that cerebrospinal fluid (CSF) can contain proteins that are early indicators of Alzheimer. To analyze proteins, biologists use mass spectrometry. The problem of "short" and "fat" data is just as severe for mass spec data as for microarray data analysis, since there are typically 20,000 or more protein fragments per sample.

I worked with a small biotech company in Massachusetts that collected data from clinical trials on about 100 senior citizens. For each patient, we had mass spectrometry data on their CSF and a diagnosis by a physician on whether the patient had Alzheimer or not. When I started to apply the microarray-style data analysis template described in the previous section, it turned out that a complex model was not necessary since there were about seven variables (proteins) that were perfect predictors of Alzheimer. 100% of patients with these proteins had Alzheimer, and 100% of patients without these proteins did not have it. Such perfect predictors are very rare in real-world data mining, and when I find them, my first thought is that there is a "false predictor"—a data element which is collected after the sample class is determined. For example, when predicting telephone customer attrition, we may find a variable X which encodes bill amount in the month after the attrition. The rule $X < 1$ may be an excellent predictor of attrition in data, but since X is determined after the attrition happens, it is not useful in predicting it.

We double-and triple-checked the data, but did not find any false predictors. One of the proteins was related to vitamin C, which is biologically plausible—it implied

that people with higher vitamin C levels had lower chance of getting Alzheimer. Although I thought that there must be an error somewhere in data collection process (e.g., samples from patients with Alzheimer were collected in different bottles and some proteins from the bottles leaked into the samples), I started to take extra vitamin C just in case. The only way to test the results was to conduct another study, which took many months. When new data was received, I put it into my program holding my breath.

The results were very disappointing—no previous predictors were significant in the new sample, and in fact, there was little correlation between proteins in old and new samples. Alas, it seemed that mass spec technology was not reliable enough to be used as a diagnostic. However, this case confirmed my intuition of suspecting models that predict unusually well.

10.3 Data Mining Course

In 2003, I was contacted by Prof. Gary Parker of Connecticut College who asked me if I was interested in creating teaching material for an introductory data mining course. I had taught a course only once—introduction to computer science using Basic, when I was a graduate student at NYU, but I was considering an academic career and wanted to try it. Besides, my mother-in-law would frequently say to me "Stop lecturing," and I took it as an indication that I have a professorial style of explanations and should try lecturing to actual students.

Data mining is best learned by doing. I wanted my students to use professional-level data mining tools on actual real-world data. The college could not afford commercial data mining tools like SAS or SPSS, so I selected Weka [22], which at that time was practically the only well-developed free data mining tool. The leaders of the Weka project, Prof. Ian Witten and Dr. Eibe Frank, from U. of Waikato, New Zealand, also had a book [23], which became the main textbook for the course.

However, I wanted to take a more practical approach, leverage my industrial experience, and cover applications like summarization and deviation detection, targeted marketing, consumer modeling, and microarray data analysis.

The course would build toward a final project, which would be a KDD-Cup style competition between student groups. For the competition, I was able to use microarray data of pediatric brain tumors of five different types. We had 92 samples and about 7,000 variables (genes) for each sample. Correct classification was critical for finding the best treatment and saving the children's lives, so students were motivated to strive for maximum accuracy.

Preparing all the lectures and assignments was hard and intensive work, but not all rewards were financial. I still remember my exhilaration when one of my best students had a "Eureka" moment and actually understood an important idea I was trying to explain. Such moments are probably what motivates people to become teachers.

The course assignments went over all the key steps in the knowledge discovery process, including data preparation, feature reduction, and classification. Microarray

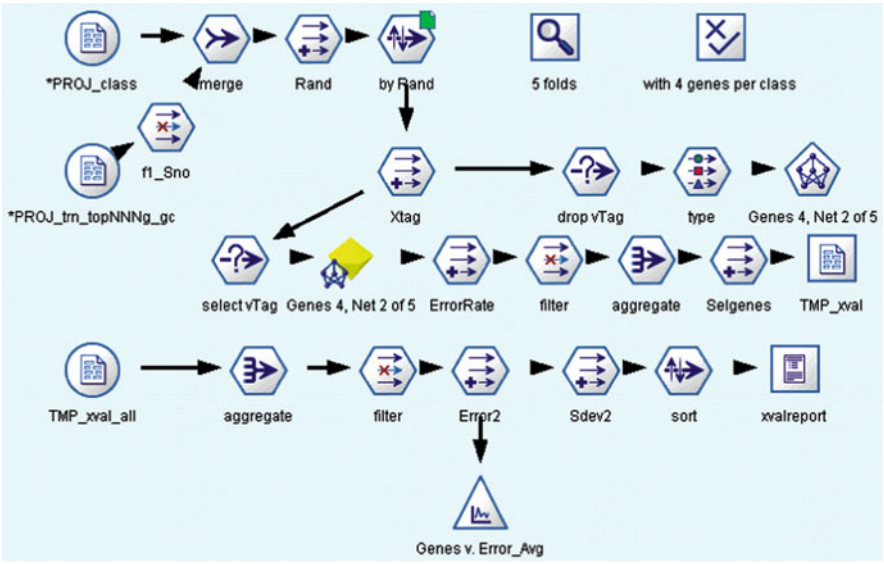


Fig. 6 Clementine template for microarray classification

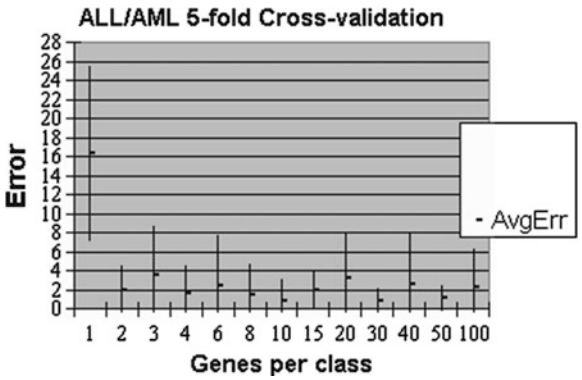


Fig. 7 Cross-validation errors for different gene subsets, for ALL/AML data, each cross-validation repeated ten times. The central point is the average error for each cross-validation. Bars indicate one st. dev. up and down

data analysis has a number of additional steps, such as randomization testing and identifying potentially mislabeled instances.

I divided the data into training and test sets, and eliminated the labels from the 23 samples in the test set. My best program was able to predict 21 or 22 out of 23 correctly. The errors were due to some of the samples having a really mixed, borderline classification.

Gregory Piatetsky-Shapiro (*left*) and his Conn. College students. Prof. Gary Parker 3rd from right.

I was fortunate in getting some of the brightest Connecticut College students to enroll in the course, and to my great satisfaction, all of the teams got at least 20 from 23 samples correctly and were presented with certificates (Fig. 6).

In 2004, thanks to the Keck and Hughes grant, I was able to make the course materials available online in 2004 for teachers of data mining [24].

The lectures and materials were viewed by tens of thousands of people and used for teaching by hundreds of professors all over the world.

10.4 Finding Counterfeit Items in Online Auctions

The Internet has greatly facilitated legitimate commerce but also provided an opportunity for many not quite legitimate ventures. For example, many buyers who were searching for luxury goods would see offers online to buy them at greatly reduced prices from suppliers in places like China. These items were almost always counterfeit. One famous jewelry company has conducted a test buying process of their "products" sold at a large online auction site and found that about 75% were fake. The jewelry company sued the auction site, but the auction site said that it did not actually see the items and could not identify if they were genuine or not. The jewelry company replied that data mining can be used to identify items which were likely to be counterfeit and hired me as an expert.

This was a very interesting case. My task was to come up with a method to show how one could identify suspicious items. The auction site was correct that it was hard to identify if a single item was suspect. I proposed a better approach, which

was to analyze the seller. Suspect sellers were much easier to identify. They tended to have many items on sale—some had over a hundred—while legitimate sellers of that brand typically had only a few items. The suspect sellers usually did not have much selling history; they were based in places like China, Hong Kong, or Taiwan which were close to counterfeiting operations. The online auctions tended to be 1 day or instant, making it easy to complete the sale before monitoring could catch it. The suspect sellers tended to use throwaway names like hvchdgvx. There were other, less obvious factors as well. Using these factors, I built a classifier that ranked sellers. The big difficulty was verifying the classifier—determining if the seller was actually suspect. In addition to visual inspection of all the seller items and history, which could use observations that are hard to computerize, I found a way to leverage the auction site to provide me with the needed confirmation. The auction site reps also had some rules (not very effective at that time) and were monitoring the sales. They were removing the sellers they determined to be suspect, but it usually took the reps a few days. Thus, for several weeks one summer, I would get up at before 6 a.m. and, accompanied by my cat "Ryzhi," would go to the computer and spend a couple of hours collecting the data on my client luxury items on sale that morning. I was also building a database of sellers and checking them on the following days. If some sellers were suspended or deleted, that confirmed my rules.

My program and rules were quite successful, and on some days, over 90% of items found on the online auction site under my client brand were strongly suspect.

The legal process went on, and I even testified in court. Although the court eventually ruled in favor of the auction house, the lawsuit—including my demonstration of identifying suspect items by focusing on sellers—forced the auction site to adopt much stronger rules against sellers of counterfeit items. As a result, currently there are very few suspect items of my client brand on the auction site.

11 Summary

The knowledge discovery research community has grown tremendously since 1989. From one workshop with about 50 attendees, we have progressed to over 25 international conferences in 2010 (www.kdnuggets.com/meetings), tens of thousands of publications, and several journals focused on the field.

Our increasingly digital world is flooded with data, and analytics and data mining play a central role in helping us make sense of it. Data mining—collecting information and finding patterns in it—can be considered one of the oldest human activities. Our ancestors, who were better in learning the behavior of prey and predators, had better chances of survival and bequeathed us genes that predispose us to look for patterns. We are so hardwired to look for patterns that we frequently mistake random patterns for true ones—just look at the success of astrology.

Our roles as data miners are to help the public, business, and government to better understand the world and to make predictions based not on random correlations but on valid theory. I feel privileged to be a part of this journey.

Acknowledgments I am very grateful to my family and especially my wife for her support and encouragement. Part of this was presented to the KDD-99 10-year anniversary panel and published in SIGKDD Explorations, 2000. I thank Mohamed Medhat Gaber for his encouragement to write this chapter and for his patience.

References

1. Gregory Piatetsky-Shapiro, "The optimal selection of secondary indices is NP-complete", SIGMOD Record, Jan 1983
2. Gregory Piatetsky-Shapiro, Charles Connell, "Accurate estimation of the number of tuples satisfying a condition", in *Proceedings of ACM-SIGMOD Annual Conference*, Boston, June 1984
3. Gregory Piatetsky-Shapiro, "A Self-Organizing Database System—A Different Approach to Query Optimization", Ph.D. Thesis, Courant Institute Report 147, New York University, April 1984
4. G. Jakobson, C. Lafond, E. Nyberg, G. Piatetsky-Shapiro, "An intelligent database assistant", IEEE Expert **1**(2) (1986)
5. Robert L. Blum, Gio Wiederhold, "Studying hypotheses on a time-oriented clinical database: an overview of the RX project", in *Proceedings of the 6th Symposium on Computer Applications in Medical Care* (IEEE 82, Washington, DC, 1981), pp. 725–735
6. Gregory Piatetsky-Shapiro, KDD-89, Knowledge Discovery in Databases Workshop, http://www.kdnuggets.com/meetings/kdd89/
7. Gregory Piatetsky-Shapiro, Knowledge discovery in real databases: a workshop report, AI Magazine **11**(5) (1991)
8. Healthcare Costs Around the World (2011), http://www.visualeconomics.com/healthcare-costs-around-the-world_2010-03-01/. Retrieved 20 Jan 2011
9. C. Matheus, G. Piatetsky-Shapiro, D. McNeill, Selecting and reporting what is interesting: the KEFIR application to healthcare data, in *Advances in Knowledge Discovery and Data Mining*, ed. by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (AAAI/MIT Press, Cambridge, CA, 1996)
10. Gregory Piatetsky-Shapiro, Data Mining Course, Connecticut College, 2003, Lecture 16: Summarization and Deviation, http://www.kdnuggets.com/data_mining_course/dm16-summarization-deviation-detection.ppt
11. KEFIR Overview Report, http://www.kdnuggets.com/data_mining_course/kefir/overview.htm
12. Brij Masand et al., CHAMP: a prototype for automated cellular churn prediction. Data Min. Knowl. Disc. J. **3**(2), 219–225 (1999)
13. Tom Davenport, Jeanne G. Harris, *Competing on Analytics Competing on Analytics: The New Science of Winning* (Harvard Business School Press, Boston MA, 2007), p. 240
14. Knowledge Discovery Nuggets 93:1 (1993), http://www.kdnuggets.com/news/93/n1.txt
15. Knowledge Discovery Mine Announcement (1994), http://www.kdnuggets.com/news/94/n8.txt
16. Gregory Piatetsky-Shapiro, Brij Masand, Estimating campaign benefits and modeling lift, in *Proceedings of KDD-99 Conference* (ACM Press, New York, NY, 1999)
17. G. Piatetsky-Shapiro, P. Tamayo, SIGKDD Explorations Special Issue on Microarray Data Mining, Guest Editors, Dec 2003
18. T.R. Golub et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
19. P. Tamayo, Personal Communication, 2002.
20. CAMDA 2000, in *Proceedings of Critical Assessment of Microarrays Conference*, Duke University, 2000

21. G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, Capturing best practices for microarray data analysis, in *Proceedings of ACM KDD-2003 Conference* (Honorary Mention as the 2nd best application paper)
22. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
23. Ian Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Morgan Kaufman, San Francisco, 1999).
24. Gregory Piatetsky-Shapiro, Data Mining Course, Presentations, Teaching Modules, and Course Notes for an Undergraduate 1-semester Course on Data Mining (2004), www.kdnuggets.com/data_mining_course/

Data Mining: From Medical Decision Support to Hospital Management

Shusaku Tsumoto

1 Introduction: Short Answer to the Editor

Q: What brought you to the data mining area?
A: I have two objectives. One is updating the knowledge base for a medical expert system which I was involved in developing when I was a student of medical school [1]. The other one is to establish a methodology to reuse data when all the hospital information is electronized. I graduated from Osaka University, School of Medicine, in 1989, which incidentally is the same year that the knowledge discovery and data mining (KDD) workshop was held in IJCAI89. I was lucky that I began my career as a researcher just when data mining was born.

Q: What are the key milestones in your data mining research?
A: 1. Progress in electronization of clinical data and hospital information
2. Rule induction resulted in interesting and unexpected discoveries from data
3. Jan Zytkow's discovery challenge
4. IEEE ICDM (especially submission data mining)
5. Motoda's active mining project
6. ACM health care IT SIG and SRII

Q: Describe your success stories and how you learned from failures.
A: My research achievements are described in the subsequent sections: they include development of new methods for rule induction, temporal data mining, and application of those methods, along with other conventional methods, to medical data.

S. Tsumoto (✉)
Faculty of Medicine, Department of Medical Informatics, Shimane University,
89-1 Enya-cho Izumo, Shimane 693-8501, Japan
e-mail: tsumoto@computer.org

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_14,

Q: What are your predictions for the future in the area?
A: In medical environments such as hospitals and clinics, electronization is being rapidly introduced. All the medical equipment, such as devices of laboratory and radiological examinations, can send information through LAN; a hospital can be viewed as a "local" cyberspace. Connecting such hospitals or clinics will provide a larger cyberspace in the regional health-care network. Temporal data mining and network mining may play important roles in dealing with such data.
Thus, all the data mining techniques will give powerful tools to decision making in medicine and health care, such as analysis of hospital data, discovery of useful knowledge, and detection of risk factors. From this point of view, medical data mining is now at a new starting point.

2 Starting Point: Analysis of Meningoencephalitis

2.1 Rule Induction Results

A rule induction method proposed by Tsumoto and Ziarko [2] generated 67 results for viral meningitis and 95 for bacterial meningitis, which included the following rules, not anticipated by domain experts, as shown below:[1]

```
1. [WBC < 12000]^[Sex=Female]^[CSF_CELL < 1000] -> Viral
                              (Accuracy:0.97, Coverage:0.55)
2. [Age > 40]^[WBC > 8000] -> Bacterial
                              (Accuracy:0.80, Coverage:0.58)
3. [WBC > 8000]^[Sex=Male] -> Bacterial (Accuracy:0.78,
   Coverage:0.58)
4. [Sex=Male]^[CSF_CELL>1000]-> Bacterial
                              (Accuracy:0.77, Coverage:0.73)
5. [Risk_Factor=n]->Viral
                              (Accuracy:0.78, Coverage:0.96)
6. [Risk_Factor=n]^[Age <40] -> Viral
                              (Accuracy:0.84, Coverage:0.65)
7. [Risk_Factor=n]^[Sex=Female] -> Viral
                              (Accuracy:0.94, Coverage:0.60)
```

These results show that sex, age, and risk factor are very important for diagnosis, but have not been examined fully in the literature [3].

[1] Rules from 5 to 7 are newly induced by introducing attributes on risk factors.

From these results, the author examined relations among sex, age, risk factor, and diagnosis and discovered the interesting relations among them:

1. The number of examples satisfying [Sex = Male] is equal to 63, and 16 of 63 cases have a risk factor: 3 cases of DM, 3 cases of LC, and 7 cases of sinusitis.
2. The number of examples satisfying [Age $\geq$ 40] is equal to 41, and 12 of 41 cases have a risk factor: 4 cases of DM, 2 cases of LC, and 4 cases of sinusitis.

Diabetes Mellitus (DM) and Liver Cirrhosis (LC) are well-known diseases in which the immune function of patients becomes very low. Also, sinusitis has been pointed out to be a risk factor for bacterial meningitis [3]. It is also notable that men suffer from DM and LC more than women.

These results gave me a strong motivation to take up data mining research and establishing a discussion forum for domain experts and data miners, which led to the Japanese workshop on Discovery Challenge [4] held from 1999 to 2004 and extended into Motoda's Active Mining Project [5]. Also, these projects are connected with Jan Zytkow, who started the PKDD conference series and its Discovery Challenge in 1999.

In this context, I introduced visualization techniques to understand the nature of the rules not anticipated by experts more easily: a combination of rule induction and multidimensional scaling (MDS) [6]. Three-dimensional assignment of rules by MDS was very useful for the detection of unanticipated rules.

3 Induction of Structured Rules

As shown in the above section, rules obtained from data are too simple and always need heavy interpretation by domain experts. So the next step is to give structured assumptions in medical rule-based reasoning to guide the rule-induction process. This is also connected with the induction of rules for RHINOS [1, 7], which makes a differential diagnosis of headache.

After I graduated from medical school, in order to learn more about headaches, I became a resident of neurology in Chiba University Hospital and then worked at the emergency department in Matsudo Municipal. During my training, I realized that the reasoning style of RHINOS captures the diagnostic process, of neurology, which I call "focusing mechanism." Figure 1 illustrates this process, which consists of the following reasoning processes: exclusive reasoning and inclusive reasoning. If a patient does not have a symptom that is necessary to diagnosis a disease, exclusive reasoning removes the disease from diagnostic candidates. Secondly, inclusive reasoning suspects a disease in the output of the exclusive process when a patient has symptoms specific to a disease. Then, finally, since symptoms which cannot be explained by diagnosis suggest the possibility of complications, we look for patients with complicated diseases. These two steps are modeled as results of two kinds of rules: negative rules (or exclusive rules), which correspond to exclusive reasoning, and positive rules, which correspond to inclusive reasoning. In the

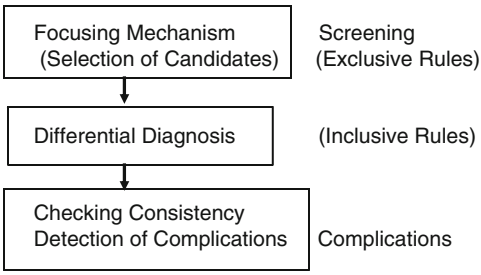


Fig. 1 Illustration of focusing mechanism

next two subsections, these two rules are represented as special kinds of probabilistic rules.

I have found that these diagnostic model can be formally described in the framework of rough sets [8] and developed a system PRIMEROSE-REX which induced RHINOS-type diagnostic rules. Experimental evaluation showed the proposed system outperforms conventional rule induction methods [7]. The results were satisfactory with respect to performance.[2] However, the rules are still simple, compared with the rules acquired from domain experts. Although I tried to extend the method [9, 10], the results were not satisfactory; this may be viewed as the limitation of expiricism. Rules should be induced by combination of deduction and induction. But I think that recent progress in graph or network mining may give insights into structured rule deduction.

4 Risk Aversion in Emergency Department

Section 3 dealt with rather fundamental research. Another dimension is to extend application areas.

I focused on how data mining can contribute to risk management, especially the prevention of accidents or incidents in medical care. For this purpose, we started analysing incident reports. During the next 6 months, from October 2001 to March 2002, we collected incident reports in an emergency department [11] to detect risk factors for medication errors. We applied C4.5 [12] to this dataset. Simple rules need a heavy interpretation process. A very heavy process exhausts domain experts and should be avoided when necessary. However, the interpretation process may evoke intensive discussion among medical staff, which may lead to prevention or risk avoidance through deduced rules. This case study is a good example.

[2] I received the Ph.D. degree from Tokyo Institute of Technology in 1997 for working on this subject.

4.1 Rule-Induction Results

The following rules were obtained:

(medication error):If the number of disturbing patients is more than one,
then medi-cation errors occur: probability (90%, 18/20).

(medication error):If nurses' work interrupted,then medication errors occur: probability (80%, 4/5).

By the addition of "the environmental factors," these high-probability rules of medication errors were extracted.

4.2 Rule Interpretation

With these results, the nurses discussed their system for medication checking.

In the emergency room, the nurses in charge of the shift prepared the medication (identification, quantity of medicines, etc.). The time of preparation before the beginning of the shift was occasionally less than 30 min when the liaison conference between shifts took place. In such cases, the sorting of medicines could not be made in advance and had to be done during the shift.

If the nurses' concentration was disturbed by restless patients in such situations, double check of the preparation of medicine could not be made, which leds to medication errors.

4.3 Utilization of Knowledge

Therefore, it was decided that two nurses who had finished their shifts would prepare medicines for the next shift and one nurse in charge of medication would check the dose and identification of medicines alone (triple check by a total of three nurses). The check system was improved as a result of their discussion. Incident reports were collected during the last 6 months (April–October 2002).

After the introduction of the triple-check system, the total number of medication errors during the last 6 months decreased from about 200 to less than 10 cases. It was considered that the nurses' medication work was improved by the triple-check system during the last 6 months.

This gives me an opportunity to introduce the concept of "risk-mining process" [13].

5 Active Mining and Trajectory Mining

5.1 Active Mining Process

After introducing the Japanese Discovery Challenge in 1999, many data miners joined my SIG workshop, and we launched a new national project called "Active Mining" [5] (2001–2004). Active mining was proposed as a new direction in the knowledge discovery process for real-world applications handling various kinds of data with actual user need. Our ability to collect data, be it in business, government, science, or even personal, has been increasing at a dramatic rate, which we call the "information flood." However, our ability to analyze and understand massive data lags far behind our ability to collect them. The value of data is no longer in "how much of it we have." Rather, the value is in how quickly and effectively the data can be reduced, explored, manipulated, and managed. For this purpose, KDD emerges as a technique that extracts implicit, previously unknown, and potentially useful information (or patterns) from data. However, recent extensive studies and real-world applications show that the following requirements are indispensable to overcome the information flood: (1) identifying and collecting the relevant data from a huge information search space (active information collection), (2) mining useful knowledge from different forms of massive data efficiently and effectively (user-centered active data mining), and (3) promptly reacting to situation changes and giving necessary feedback to both data collection and mining steps (active user reaction).

Prof. Hiroshi Motoda was the chief of this project and we had ten research groups, which used the following two common datasets: hepatitis data and chemistry data. The former gave a test bed to temporal data mining and the other gave a test bed to graph mining. This project provided an opportunity to create the core group of Japanese researchers in data mining.

5.2 Trajectory Mining

Problems in temporal data mining gave us ideas on trajectory mining [14]. Conventional temporal mining deals with the temporal evolution of one parameter, but if we select two parameters and depict temporal changes on a two-dimensional plane, we obtain a two-dimensional trajectory. If we can calculate similarities between trajectories, we can classify them and extract information on the chronology of two variables.

The main contributions of the paper are twofold. First, it proposes a new approach to comparing trajectories of medical data by shape comparison technique, not by standard time-series analysis technique. Second, it introduces a two-step method for deriving the dissimilarity between trajectories; multiscale matching is first applied to find structurally similar parts, and after that value-based dissimilarity

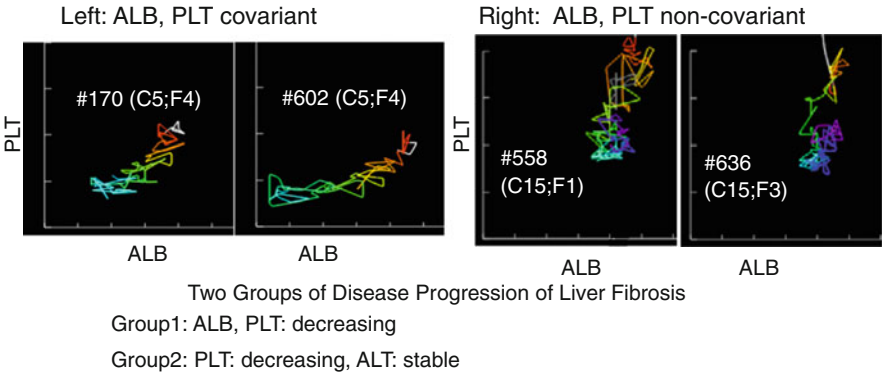


Fig. 2 Discovery from two-dimensional trajectory mining

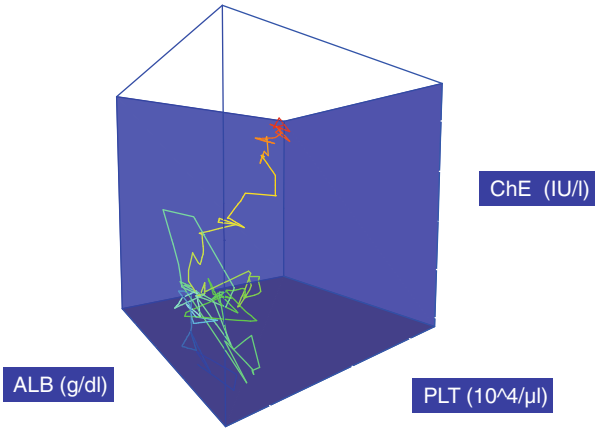


Fig. 3 Discovery from three-dimensional trajectory mining

is derived for each of the matched pairs and accumulated as the final dissimilarity between trajectories. This scheme makes the dissimilarity more informative as it takes not only value-based features but also structural features into account.

One discovery of the application of this method to hepatitis data is that we have two interesting groups in hepatitis C as shown in Fig. 2.[3] One has a pattern in which albumin (ALB) and platelet (PLT) decrease in a covariant way, and the other has one in which PLT decreases but ALB does not.

Recently, this method has been extended to three-dimensional trajectories [15]. Figure 3 gives an illustrative example from a cluster whose ALB, PLT, and Chorin Esterase (ChE) are decreasing, but ALB decreases first, then PLT and ChE covariantly decrease. In this way, three-dimensional trajectory may give more

[3] Temporal evolution is shown as red to blue.

interesting discoveries. It will be our future work to extend the method into multidimensional cases.

6 Toward Data-Mining-Oriented Hospital Services

On the other hand, clinical information has been stored electronically as a hospital information system (HIS). The database stores all the data related to medical actions, including accounting information, laboratory examinations, and patient records described by medical staff. Incident or accident reports are not exceptions: they are also stored in HIS as clinical databases. For example, Fig. 4 shows the structure of the HIS in Shimane University Hospital. As shown in the figure, all the clinical inputs are shared through the network service in which medical staff can retrieve their information from their terminals [13, 16].

As all the clinical data are distributed, stored, and connected as a large-scale network, HIS can be viewed as a cyberspace in a hospital: all the results of clinical actions are stored as "histories." It is expected that similar techniques in data mining, Web mining, or network analysis can be applied to the data. Dealing with a cyberspace in a hospital will create a new challenging problem in hospital management in which spatiotemporal data mining, social network analysis, and other new data mining methods may play central roles. I am focusing on temporal data mining to analyse histories stored in HIS [17, 18].

6.1 Basic Unit in HIS: Order

The basic unit in HIS is an "order," which is a kind of document or message which conveys an order from a medical practitioner to others. For example, a prescription can be viewed as an order from a doctor to a pharmacist, and a prescription order is executed as follows:

1. Outpatient clinic.
2. A prescription given from a doctor to a patient.
3. The patient brings it to medical payment department.
4. The patient brings it to pharmaceutical department.
5. Execution of order in pharmacist office.
6. Delivery of prescribed medication.
7. Payment.

The second to fourth steps can be viewed as information propagation: thus, if we transmit the prescription through the network, all the departments involved in this order can easily share the ordered information and execute the order immediately. This also means that all the results of the prescription process are stored in HIS.

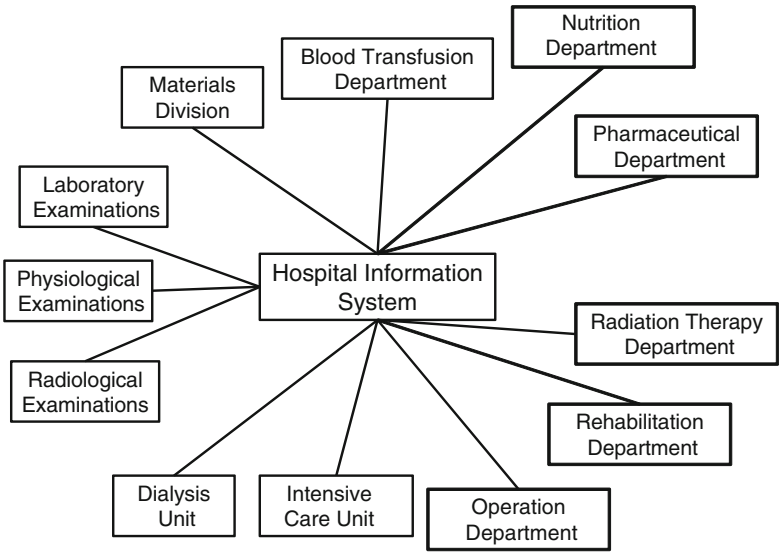


Fig. 4 Hospital information system in Shimane University

These sharing and storing processes, including histories of orders and their results, are automatically collected as a database: HIS can also be viewed as the cyberspace of medical orders.

6.2 Temporal Trends of #Orders

To get an overview of the total activities of the hospital, we can also check the temporal trend of each order as shown in Figs. 5 and 6. The former figure depicts the chronological overview of the number of each order from June 1 to 7, 2008, and the latter shows that of June 2, 2008.

Vertical axes denote the averaged number of each order, classified by the type of orders. The horizontal axis gives each time zone. The plots show the characteristics of each order. For example, the number of records by doctors has its peak at 11 a.m., which corresponds to the peak at the outpatient clinic, whose trend is very similar to reservation at the outpatient clinic. The difference between these two orders is shown in the column representing 1–5 p.m., which corresponds to the activities in the wards.

6.3 Analysis of Trajectory of #Orders

By applying trajectory mining methods, two interesting groups were found: the first one gives a pattern where orders are given both in wards and outpatient clinics.

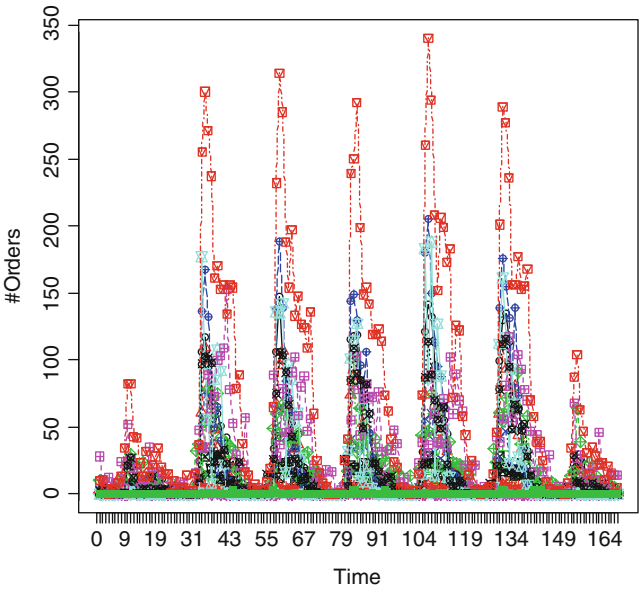


Fig. 5 Trends of number of orders (June 1–6, 2008)

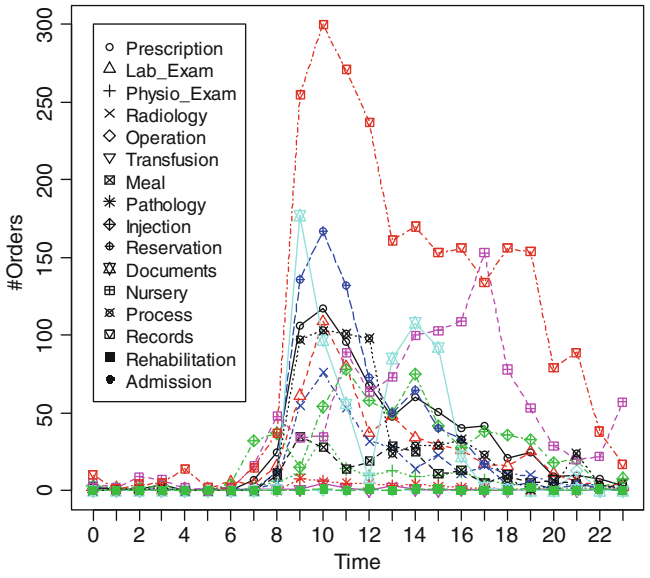


Fig. 6 Trends of number of orders (June 2, 2008)

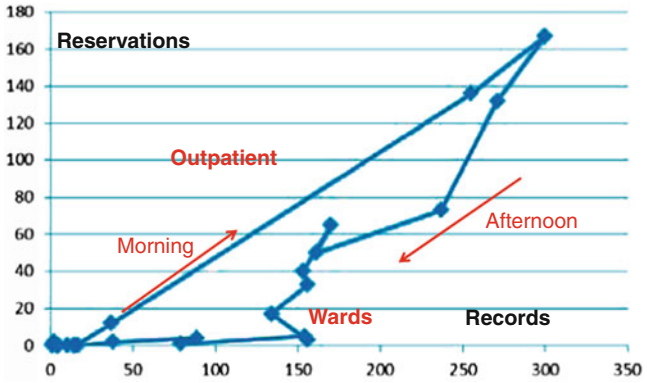


Fig. 7 Trajectory between #Reservations and #Records (June 2, 2008)

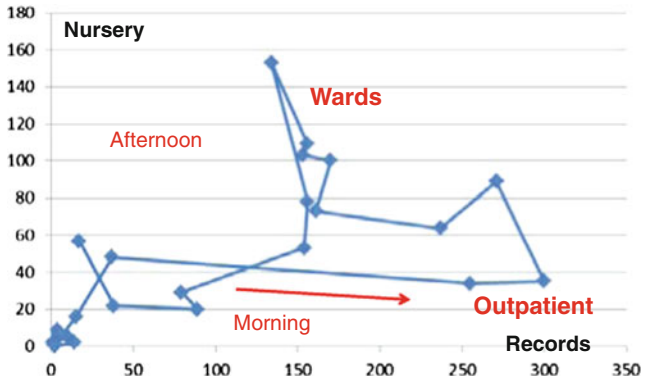


Fig. 8 Trajectory between #Nursery Orders and #Records (June 2, 2008)

The other one gives a pattern where orders are provided mainly in the wards. A typical example in the first cluster is shown in Fig. 7, while one in the second cluster is in Fig. 8.

7 Future Insights

The above developments are my major achievements, but I think that many fundamental questions have been left unsolved and that real application of data mining to hospital data has just started.

7.1 Key Research Areas

Extraction of structure from data is still important. Graph or network mining, including Bayesian network mining, has been introduced, and assumes a "specific" and core structure. From my experience, discovery does not emerge only from a simple structured object. When discovery is due to such a simple pattern, other types of efforts, such as dense discussions among experts, may be left behind. Thus, a suitable size of patterns is important to enhance the discovery process, though it is not clear whether this process will be domain dependent or independent. We, as data miners, assume that postprocessing will occur after patterns are extracted, but really, patterns may be extracted under the guidance of postprocessing. Thus, the important fundamental issues are listed below. I have classified them into two major categories: technical and application issues, but actually, they are not independent, being rather closely connected with each other.

1. Techinical issues
   (a) Pattern generation
      - Extraction of a suitable size of complex patterns in rule mining and network and temporal data mining
      - Extraction of multidimensional patterns
      - Clustering of multidimensional structured and temporal data
   (b) Postprocessing
      - Interactive interpretation of patterns
      - Objective evaluation of structured patterns
2. Application issues
   (a) Preprocessing
      - Flexible data warehousing
      - Application of domain ontologies
   (b) Pattern generation
      - Patterns guided by domain knowledge
      - Ontology mining or ontology validation
   (c) Efficient KDD process in a large-scale system

   I hope that these issues will be solved in the near future.

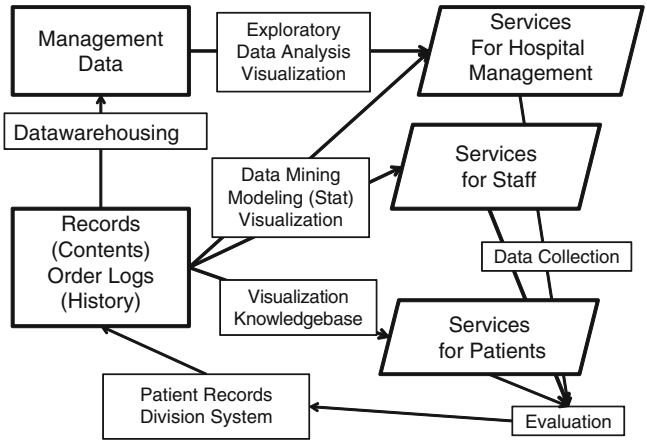


Fig. 9 Data-oriented hospital services and management

7.2 My Final Goal: Framework on Data-Mining-Based Hospital Services

Figure 9 shows our goal for hospital services, which consists of the following three layers of hospital management: services for hospital management, services for medical staff, and services for patients. Data mining in HIS plays a central role in achieving these layers.

The lowest layer is called services for patients which supports the improvement of health-care service delivery for patients. This is a fundamental level of health-care services in which medical staff directly provide medical services to the patients. Patient records and other results of clinical examinations support the quality of this service. The next layer is called services for medical staff which supports decision making of the medical practitioner. Patient histories and clinical data are applied to data mining techniques, which give useful patterns for medical practice. Especially detection of risk for patients, such as drug-adverse effects or temporal status of chronic diseases, will improve the quality of medical services. The top layer is called services for hospital management. This level is achieved by capturing the total activity global behavior of a hospital: bridging the gap between the microscopic behavior of the medical staff and the macroscopic behavior of the hospital is very important to deploy medical staff in an optimal way for improving the performance of the hospital.

Data mining will be at the core of these processes.

References

1. Y. Matsumura, T. Matsunaga, Y. Maeda, S. Tsumoto, H. Matsumura, M. Kimura, Consultation system for diagnosis of headache and facial pain: "rhinos", in *LP. Lecture Notes in Computer Science*, ed. by E. Wada, vol. 221 (Springer, Berlin, 1985), pp. 287–298
2. S. Tsumoto, W. Ziarko, N. Shan, H. Tanaka, Knowledge discovery in clinical databases based on variable precision rough set model, in *The Eighteenth Annual Symposium on Computer Applications in Medical Care* (1995), pp. 270–274
3. R. Adams, M. Victor, *Principles of Neurology*, 5th edn. (McGraw-Hill, New York, 1993)
4. S. Tsumoto, K. Takabayashi, Data mining in meningoencephalitis: the starting point of discovery challenge, in *ISMIS. Lecture Notes in Computer Science*, ed. by M. Kryszkiewicz, H. Rybinski, A. Skowron, Z.W. Ras, vol. 6804 (Springer, Berlin, 2011), pp. 133–139
5. S. Tsumoto, T. Yamaguchi, M. Numao, H. Motoda (eds.) *Active Mining, Second International Workshop, AM 2003*, Maebashi, Japan, October 28, 2003, Revised Selected Papers, Lecture Notes in Computer Science, vol. 3430 (Springer, Berlin, 2005)
6. S. Tsumoto, S. Hirano, Visualization of differences between rules' syntactic and semantic similarities using multidimensional scaling. Fundam. Inform. **78**(4), 561–573 (2007)
7. S. Tsumoto, Automated extraction of medical expert system rules from clinical databases on rough set theory. Inf. Sci. **112**(1–4), 67–84 (1998)
8. Z. Pawlak, *Rough Sets* (Kluwer, Dordrecht, 1991)
9. S. Tsumoto, Extraction of experts' decision rules from clinical databases using rough set model. Intell. Data Anal. **2**(1–4), 215–227 (1998)
10. S. Tsumoto, Extraction of structure of medical diagnosis from clinical data. Fundam. Inform. **59**(2–3), 271–285 (2004)
11. S. Tsumoto, K. Matsuoka, S. Yokoyama, Risk mining: mining nurses' incident factors and application of mining results to prevention of incidents, in *RSCTC. Lecture Notes in Computer Science*, ed. by S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H.S. Nguyen, R. Slowinski, vol. 4259 (Springer, Berlin, 2006), pp. 706–715
12. J. Quinlan, *C4.5 – Programs for Machine Learning* (Morgan Kaufmann, Palo Alto, CA, 1993)
13. S. Tsumoto, S. Hirano, Risk mining in medicine: application of data mining to medical risk management. Fundam. Inform. **98**(1), 107–121 (2010)
14. S. Tsumoto, S. Hirano, Detection of risk factors using trajectory mining. Journal of Intelligent Information System **36**(3), 403–425 (2011)
15. S. Hirano, S. Tsumoto, Multiscale comparison and clustering of three-dimensional trajectories based on curvature maxima. International Journal of Information Technology and Decision Making **9**(6), 889–904 (2010)
16. E. Hanada, S. Tsumoto, S. Kobayashi, A "ubiquitous environment" through wireless voice/data communication and a fully computerized hospital information system in a university hospital, in *E-Health, IFIP Advances in Information and Communication Technology*, ed. by H. Takeda, vol. 335 (Springer, Boston, 2010), pp. 160–168
17. S. Tsumoto, S. Hirano, Y. Tsumoto, Information reuse in hospital information systems: a data mining approach, in *Proceeding of IEEE IRI 2011* (in press)
18. S. Tsumoto, S. Hirano, Y. Tsumoto, Temporal data mining in history data of hospital information systems, in *Proceeding of IEEE SMC 2011* (in press)

Rattle and Other Data Mining Tales

Graham J. Williams

1 A Voyage to Data Mining

My own voyage to data mining started long before data mining had a name. It started as a curiosity that a young scientist had in searching for interesting patterns in data. In fact, the journey began in 1983 as an artificial intelligence Ph.D. student at the Australian National University, under Professor Robin Stanton.

Data mining emerged in 1989 from the database research community. They had by then established the foundations for the relational database theory that underpinned the storing of masses of data. The research question had now become how to add value to the ever growing amount of data being collected. Data mining was the answer. The question remains with us still today, and many advances have been made in our ability to extract knowledge from data.

It was not long before the machine learners and the statisticians started to become involved in data mining. For those researchers with an interest in the application of research results to real tasks, data mining provided a useful focus. Data mining was born as a discipline where new research in machine learning had application in industry and government.

My early efforts in artificial intelligence focused on knowledge representation. I used the concept of frames to implement the FrameUp language [17]. This lead to a simple automated reasoning system called HEFFE—the household electric fault finding expert [17]. Rather amusingly, the system always thought to first ask "is it turned on at the power point?" Fortunately, it then progressed to a more sophisticated dialogue, exposing some complex reasoning to diagnose faults. The research hinted at my interest in helping people and solving real problems.

I soon noticed that the main problem for such a reasoning system was in obtaining the knowledge to reason with. An initial interest in knowledge acquisition

G.J. Williams (✉)
Togaware Pty Ltd., Canberra, ACT, Australia
e-mail: Graham.Williams@togaware.com

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_15,

grew into a research pathway into machine learning. It were early, yet exciting days for research in machine learning in the 1980s. At the time, decision trees, and in particular the Iterative Dichotomiser 3 (ID3) algorithm [12], were emerging as a popular and successful approach. Every paper published in machine learning seemed to reference Ross Quinlan's seminal work.

The first decision tree induction algorithm was developed by Quinlan whilst visiting Stanford University in 1978. Interestingly, at about the same time and university, a number of statisticians were developing and deploying similar ideas. Their version of the concept became known as classification and regression trees [5]. Like many great ideas, it was simple yet powerful.

For my Ph.D. research, I implemented the algorithm in the C programming language. This provided a platform for experiments in machine learning. My observations of the mathematics behind the algorithm led me to develop the idea of building several decision trees from the same data set. The Ph.D. developed the idea into an algorithm for multiple inductive learning, which was a simple idea that we might today call an ensemble approach to machine learning.

With the growing interest in data mining in the 1990s, it was not long before decision trees (and classification and regression trees) became a foundational technology for the new research area. Decision trees were an obvious technology for turning data into knowledge.

My research and keen interest in the application of technology navigated my journey towards data mining. Decision tree induction (or the top-down induction of decision trees, as we used to call it), and predictive modelling in general, provided the foundation for today's data mining.

In practice, we data mine using all the tools we have available. We work on real problems in extracting knowledge from massive collections of data. This is the excitement of data mining in practice. We have the opportunity to apply an arsenal of tools and techniques across multiple disciplines. It opens the door to a lifetime's learning, as we learn to learn.

2 Signposts Along the Way

2.1 Expert Systems

My early research in artificial intelligence involved the development of expert systems. The knowledge embodied in such systems was usually either hand crafted, as rules, or "discovered" through the use of knowledge acquisition tools like ripple down rules [6] or through decision tree induction. It was decision tree induction that, as a machine learning researcher, took my interest.

Expert systems were a very practical technology. Whilst undertaking my Ph.D. research, I took up an opportunity that sidetracked me from my Ph.D. for a short while, to work on practical applications. I gained many insights and much

understanding. Some of these examples below provide a flavour of how practical problems can help form real research questions.

2.2 Bush Fires at Kakadu

The first opportunity for some practical expert systems came with a project from the Commonwealth Scientific and Industrial Research Organisation (CSIRO, which was then Australia's premier government-funded research organisation). The Division of Land and Water Research, under Dr. Richard Davis, was developing a bush fire prediction expert system for Kakadu National Park.

Kakadu is located in the Northern Territory of Australia and experienced regular devastating bush fires often started by reckless travellers. Each year, the vegetation/fuel grew unchecked, and when a fire took hold, large areas of the park would be destroyed. Over centuries, the indigenous population, the Aborigines, had developed sophisticated knowledge of how to manage their environment through fire and to avoid the devastation of an uncontrolled wildfire.

Through fires being lit at particular times under particular conditions, the Aborigines could control the type and density of vegetation into the future. Through this, they then also encouraged animals to graze certain vegetation, providing the Aborigines with their daily food source. The careful management of the vegetation also ensured it maintained a suitable density for habitation and to facilitate hunting and travel.

Our task, whilst developing a new framework for expert systems, was to capture some of this indigenous expert knowledge within a spatially oriented expert system. The knowledge base for the expert system was developed in collaboration with Aboriginal elders. The system was designed to predict the extent of a bush fire at any time, through spatial reasoning, and thereby allow a plan to be developed to allow for controlled burning and the appropriate management of these controlled burns.

Working as part of the team on this project was very satisfying. The project was successful, developing new technology in the form of a Prolog-based spatial expert system [7]. As a young researcher, I had the opportunity to publish my first paper and to present it at an international conference in France. It was awarded as the best student paper [21] at the conference.

2.3 Decision Trees and 4GLs

Continuing with this interest in the practical application of the research, I had an opportunity to lead a team of developers in building a knowledge acquisition tool into a fourth generation environment. A Melbourne-based company, BBJ

Computers, had a successful fourth generation language (4GL) called Today. The Today product allowed database systems to be rapidly developed.

My task, leading a small team, was to implement the decision tree induction algorithm, building on the ideas I was developing for my Ph.D. It was to be implemented as an integrated component of Today, so that customers were able to not only build their database systems but also have a tool that allowed them to discover knowledge from the data thus collected in their databases. The concept had attracted the interest of the managing director, after seeing the Australian press coverage of my award for the bush fire expert system.

After a year (1987–1988), we had a system implemented and being sold to customers in Australia and Europe. This development was an early example, perhaps one of the earliest, of incorporating knowledge discovery or data mining (though called machine learning at that time) into a database context.

2.4 Car Loans

In 1989, I had another opportunity to use my growing expertise in decision tree induction to build an expert system for industry. Through a consultancy with Esanda Finance (a subsidiary of the ANZ Banking Group), lead by Vish Vishwanathan, our task was to automate much of the decision-making for granting loans for the purchase of motor vehicles.

This was a time when desktop computers were just starting to be used by the financial controllers of car yards. Previously, when a customer was asked to take out a loan to purchase a vehicle, a lending clerk would seek information on the phone. Using further information offline, a decision was made. That decision was then fed back to the sales yard, perhaps that same day or within a few days.

Over a number of years, Esanda had collected data about the profitability of each of the car loans. In particular, there was quite a good record of those clients who had defaulted on their payments and those who were very reliable in their repayments. Such data provided an excellent source for building a decision tree model, which we did.

The resulting decision tree model was implemented into their loans system. This then allowed the car yard's financial controller to simply enter the client's details and obtain a decision. That decision might be yes, no, or offline. An offline decision was one where further consideration was required, and perhaps, a decision by a human expert was required. The more complex cases remained the domain of the experts, and they no longer had to deal with the often less interesting, simple decisions.

We deployed the system successfully. Over many years, and into the early years of this century, I had the occasional report that the rules developed for the decision-making system back in 1989 were still in operation. No doubt, some of the characteristics will have changed over time, but the foundations were solid. Like many expert systems, it had become part of the furniture.

2.5 A Framework for Learning Models

Moving on from expert systems, the bigger challenge was to allow machines to automatically learn as they interact with their world. From an artificial intelligence perspective, this meant machine learning.

Machine learning can provide a framework which describes how we build models of the world. Much of what we do as computer programmers can be seen in the light of building models. Our stylised models help us to understand the world and to facilitate our automated reasoning. Whether we write a program that implements a spreadsheet or a program to predict medical outcomes, we are modelling what happens in the world. The models often help us identify the key elements as the foundation of our understanding. Machine learning is about automatically building models of the world—automatic programming.

A formal framework that helps me picture each new machine learning algorithm (in the context that they are all about building models) has three elements: a language for representing a model, a measure that is used to assist in building a good model, and a search algorithm that seeks for a good model in a respectable amount of time. All data mining algorithms can be characterised within such a framework, and doing so allows us to compare and contrast them.

We have then a search-oriented framework for understanding data mining algorithms. The language describes how we express a model. This might be as a decision tree or as a neural network or as a linear formula. Once we have a language, we can often allow for an infinite number of sentences to be written in that language (i.e. each possible sentence is a candidate model expressed in that language). How then do we identify the best (or perhaps just a good) sentence/model from all of these possibilities? That is where a heuristic search algorithm comes into the picture. Using a measure of goodness that can be applied to a model, we can heuristically search through this infinite search space for the "best" model.

This framework, and explaining a number of data mining algorithms within the framework, is covered in more detail in a recent book [20]. The framework has served well over many years as a way to understand data mining (and originally machine learning) as the task of searching for the right model.

2.6 Ensembles of Decision Trees

In parallel to having a number of excursions into the practical development of expert systems, my Ph.D. research [18] was developing the then new idea of building multiple decision trees. At the time, we were familiar with an issue that the knowledge acquisition community was exploring—the issue of multiple experts and how to capture knowledge from them and unify it into a single system. Reasoning systems and formal systems of logic were being developed to handle

conflicting assertions and deductions. However, the idea that we might build multiple decision trees was yet to have its day.

Back in the 1980s, our "large" data sets were quite small compared to today's data-rich world. Indeed, today, research based on such small data sets would hardly be credible. For my Ph.D. research, I had access to a data set of 106 observations. The observations were from a collection of remote sensing data combined with local observational data. It was this small collection of data that allowed me to explore the mathematics of the algorithm in more detail and to watch the calculations as they were being performed.

The data was provided by colleagues at CSIRO. The Australian continent was divided into 106 regions. For each of these regions, a number of observations of different characteristics were made, including soil types, vegetation type, and vegetation coverage (with some data from remote sensing satellite imagery). Also recorded for each region was a measure of the feasibility for grazing cattle. It was this feasibility that we were attempting to model as our target variable.

Having coded the ID3 algorithm in C, many models were built with this tiny data set (as well as with the other commercial data sets I had access to through my expert systems developments). A fundamental operation of the decision tree algorithm is to choose from amongst the available variables one that "best" partitions the data set, with respect to the target variable. It had become quite clear that often, when the algorithm had to make this choice, the mathematics for making that choice gave ambiguous answers. That is, two or more variables were almost (or sometimes precisely) equally good according to the measure being used.

Without a clear answer, I wondered about the rationale for choosing one variable over another. The choice seemed rather arbitrary. So instead, I started to build all the equally (or nearly equally) good decision trees. The task then was to understand how best to combine them into a single model. Some structural integration was developed, as well as the idea of having the models vote for their outcome. The approach ended up being quite similar to, though rather primitive compared to, what emerged later as ensemble approaches to building models.

2.7 Next Port of Call: CSIRO

After completing my Ph.D., I had the opportunity to join the CSIRO. In 1995, under the insightful guidance of Peter Milne, we started the first data mining research group, as such, in Australia. At the same time, another colleague, Warwick Graco, had been delivering data mining projects for one of the Australian government departments (the Health Insurance Commission). I believe he had set up the first applied data mining team in Australia. It was not long though before data mining was sweeping through industry in Australia. The finance industry, the insurance industry, and government were all starting to investigate this new technology that promised to deliver knowledge from their rapidly bulging warehouses of data.

CSIRO, working collaboratively with colleagues from other institutions and industry, provided the opportunity to explore a variety of approaches for discovering knowledge from data. We explored many quite different approaches, including evolutionary systems [14], a B-spline-based approach to multivariate adaptive regression splines [1], and finite mixtures [24], to name some.

We also developed a successful unsupervised technique that we still use today, called hot spots [22]. In this approach to data mining of extremely large data sets, we cluster our data and then use decision tree induction to symbolically describe the clusters. This was then combined with the idea from evolutionary computing of measuring the fitness of a sub-population. Our sub-populations were described by symbolic rules generated from the decision trees. We referred to each sub-population or cluster that scored particularly high, according to the measure, a hot spot.

CSIRO provided an excellent environment for applied research. The opportunity to work with industry drove much of the research. It also provided much satisfaction in solving real problems using data mining. We worked in finance, in insurance, and with government.

One quite strong and pioneering body of work was in the area of health. Working with the Australian federal government and the state governments, we brought together, for the first time, data from the two levels of government. The federal government maintained data about a patient's visit to a doctor and their prescriptions. The state governments maintained data about episodes in hospital. Bringing such data together for over 20 million people proved to be a powerful tool for discovering, for example, hitherto unknown adverse reactions to pharmaceuticals that lead to patients ending up in hospital.

2.8 Data Treasures: Australian Taxation Office

In 2004, the Australian Taxation Office was in the early stages of a major project to substantially replace its ageing processing systems. A forward-looking component of the project was the formal introduction of analytics into the system. The organisation set up a data mining team which became the largest data mining team in Australia, with 18 new staff as a corporate resource. Its aims include to improve fraud identification and to assist taxpayers in meeting their obligations.

I had the opportunity to join early on to lead the deployment of the technology across the organisation. This included setting up the state-of-the-art infrastructure for data mining, based around a collection of GNU/Linux servers running mostly free and open source software (including the R statistical software).

My basic principle of analytics and data mining, in any industry, is that the focus needs to be on the data miners, not on the tools. Any relevant set of tools will do, and I have found that the open source offerings are more extensive than the commercial offerings.

Part of my role was then to introduce data mining technology to a broader community of up to 150 data analysts in the organisation. This was delivered through the shared data mining infrastructure and a regular weekly meeting of the Analytics Community of Practice, where topics covered new technology, infrastructure training, project developments, and much more.

The outcomes and benefits of just a small number of projects have been made publicly available through press releases of the commissioner of taxation. These have included projects that identify non-lodged tax returns (more difficult to accurately determine than one might imagine). This project, in 1 year alone, identified over $100 million of additional liabilities. Another publicly announced project described the benefits of an analytical system that reviews each tax return lodged (e.g. through electronic lodgements). In the context of the population of lodgements, the models developed are able to identify the risk of the lodgement being fraudulent.

Most projects in such an organisation have significant impact. By ensuring taxpayers are all meeting their obligations, the overall system is fairer to everyone. Any taxpayer illegally reducing their tax liability only results in an increased burden for the rest of the population.

A range of technology is used within the analytics capability of the Australian Taxation Office. Traditional linear regression modelling can go a long way. This is augmented by decision trees and ensembles including random forests and boosting. Neural networks and support vector machines also make a contribution in descriptive and predictive modelling. Self-organising maps (SOMs), a variety of approaches to clustering, and hot spots analysis all contribute to the insights gained from the models built.

One of the more reassuring observations in working in the area of fraud detection and non-compliance is that most people most of the time are doing the right thing. Society relies on people working together, and organisations like the tax office work to assist the taxpayer in doing the right thing. Whenever patterns extracted from the data begin to indicate the emergence of systemic misunderstandings, then appropriate action can be undertaken to rectify the issue. This can be through advertising campaigns or by providing letters to taxpayers that can better explain the taxpayer's obligations. Such transparency provides positive support to the community.

The smooth running of governments tasked with the delivery of services to its citizens depends on having the finances to deliver those services. Budgets can be seriously affected by non-compliance and fraudulent activity. When data mining identifies fraud, then more serious action needs to be undertaken. This can (and does) result in the perpetrators facing the legal system.

The tax office actively works towards bringing together all of its knowledge into a shared knowledge-based framework to improve its delivery of service to people and the government. Like any large data-oriented organisation, and most organisations are data-oriented, there remain many opportunities for the deployment of data mining. This will continue to require the skilled knowledge of the data

miners, who remain a scarce resource, together with the new technology emerging from current research in data mining and machine learning.

2.9 Reflections

Over the years, the application of research to real problems and real data has facilitated the delivery of research that has real impact. It does have a cost though, in an academic context focused on counting papers. There do remain, for the practicing data miner, opportunities to write research papers and to contribute to the research community. I continue to contribute as co-chair or member of a number of data mining and artificial intelligence conference steering committees, for example.

The impact of one's research and development can be measured in terms of how one changes an organisation (in addition to how our work changes the research directions). To see a single data mining project prevent significant fraud, to make business operate more efficiently, to save lives by identifying adverse outcomes from drug combinations, or to reduce the danger and consequences of out of control wildfires are all important outcomes from the application of our research. Data mining has contributed strongly to each of these.

Another impact one can have is on sharing the technology we use to deliver outcomes from data mining. We have the opportunity to share with many colleagues around the world our advances through open source software. The trend of doing so is growing, but perhaps too slowly. More researchers need to see the benefits of sharing the results of their research in this way. The positive feedback and recognition from this should not be underestimated.

Also, of course, the opportunity to share experience and guide research arises with the supervision of research, often through our Ph.D. students. More are needed to fill the growing need for data miners out there in real world projects. Data mining research, through a Ph.D., for example, is a great passage through the straights to a fulfilling career in data mining.

In the end though, the most satisfying is to know that we are contributing positively to society, to improve and advance society, and to better understand our world. The technology of data mining is becoming accessible to a larger user base. We will increasingly see many more people able to discover new knowledge through the assistance of data mining tools.

3 Rough Waters

As a rather young and less experienced student and researcher, exploring and developing new ideas, I gradually learnt a couple of useful lessons. With a passion, one should explore his new ideas until they either deliver or else gain the insights

that tell us why they will not. Do not let others discourage the journey, though be aware of their wisdom. Also, new ideas need to be communicated with clarity and conviction. Sometimes ideas may be rejected because they have not been communicated well.

I learnt these early on.

I have described above my Ph.D. research where I had the idea to build multiple decision trees, rather than relying on a single tree. I found that combining the decision from multiple trees gave better results, surprisingly. After quite a bit of experimentation and developing the idea, I was quite sure this was a significant new finding for the world of machine learning: combining multiple models, just like getting a team of experts to make a combined decision.

I wrote a journal paper describing some experiments in combining decision trees [15] and then a conference paper on what I called multiple inductive learning or the MIL algorithm [16]. The paper was accepted for presentation at the very first Australian Joint Artificial Intelligence Conference, held in Sydney, in 1987.

I remember the session well. It was Tuesday, November 3, 1987. Australians will know that the first Tuesday of November is when "the nation stops" for a horse race in Melbourne—the Melbourne Cup. .The artificial intelligence community must not have been so enamoured with horse racing. My paper was scheduled to coincide with the running of the Melbourne Cup.

Being scheduled to clash with the most famous horse race in Australia was just the beginnings of some trepidations. Professor J. Ross Quinlan, the pioneer of decision tree induction, was to be the session chair. As a Ph.D. student, I was looking forward to the opportunity to present my research on decision tree induction to him. It is not often we have such an opportunity.

I began my presentation. It started with a review of how decision tree induction works. (I know the session chair already knew that bit, and probably most of the audience knew that bit too, but the slides had already been prepared, so I pushed on.) I then presented the idea of building a number of decision trees and combining them into a single model. That was the MIL algorithm. The results presented in the slides clearly demonstrated the improved performance one obtained when multiple trees were combined to work together. Phew—that seemed to go okay.

The session chair took the option to ask the first question, but not before announcing the results of the Melbourne Cup. (I guess some might have been following the horse race on their little transistor radios during the presentation, back in those days.) I forget the actual wording of the question, and it might actually have been a comment rather than a question. The point of building more than a single decision tree to obtain a model had not been well communicated. It seemed like a rather odd thing to do. Surely, we should aim to get the simplest, best, single model.

That took something of the wind out of my sails. It was an awkward moment whilst I came to realise that either it was a poor idea or I was not particularly convincing in presenting the evidence. Maybe the interest in the Melbourne Cup was too much of a distraction.

Though a little demoralised, I stuck with the concept of building multiple models, though not with the vigour I could have. I wrote a Ph.D. thesis around the topic [18], obtained my doctorate, and moved on. Nonetheless, as I developed my career as both a researcher and consultant, over and over again I found that the concept of ensembles was always with me and delivering results.

It was interesting to watch similar ideas emerge from other directions. Today, of course, we have random forests [4] and boosting algorithms [8] that deliver excellent results for many situations in data mining. Ensembles continue to make a lot of sense—ask Netflix [2].

It is obvious now, but not then—new ideas need to be worked. Others may take time to come along the journey with you.

4 Steaming Ahead

A key point that we come to understand in data mining and machine learning is that the different algorithms perform their tasks admirably, but often almost equally well, or at least similarly well. If there are nuggets of knowledge to be discovered in our data, then various algorithms will give similar results. Often, from a machine learning practitioner's point of view, the difference may sometimes come down to what knowledge is discovered, rather than the accuracy of a model. We know full well from computer science and artificial intelligence that a change in how we represent a problem or represent our knowledge may be the difference between solving a problem, or not, and gaining insights from the knowledge, or not.

The key trick, in all of data mining—whether it be text mining, predictive analytics, or social network analysis—the key to successful data mining is to live and breathe the data. Grasp the data and turn it into a form that can be data mined (generally flat tables). Then apply the depth and breadth of our tools to the data to build new models of the world. Once we have the right set of features, irrespective of the structure of the original data (database, text, links, audio, video), building models is simple, and any tool might do.

How we represent our discovered knowledge, and how we combine discovered knowledge into universal knowledge bases that can reason about the world, continues to be a goal for a lot of research. It is a worthy goal.

On a more concrete footing, researchers will have their pet ideas about the most interesting areas for research in the next few years. I would not labour the near future too much, because the interesting research around data mining is, I think, for the longer term. Over the next few years, we will see much the same from data mining research as we have for the past few years. New tweaks on old algorithms will continue to be implemented. And new areas of application will lead to some variety in how we think about the algorithms we are using. We might make some steps also towards better representations of our learned knowledge.

Specific areas that I see continuing to grow strongly include social network analysis and text mining, along with mining of other media such as video and audio.

But data mining should become more personal. The personal mining of podcasts, for example, to find those that might be of interest to us, may be a key example of the kind of challenge. It will continue to replicate how we mine spatial, temporal, relational, and textual data, where we extract a textual representation from the original representation and turn that into a flat structure which we then data mine. This continues to work quite well, and the sophistication is in how to actually extract the "data" from these different representations. I look forward to further research advances around how we do this most effectively.

Another area of growing interest is in the sharing and deployment of models. A very early meeting in Newport Beach, California (1997), at the Third International Conference on Knowledge Discovery and Data Mining (KDD97), introduced me to the concept of a new standard for exchanging predictive models amongst different tools. The Predictive Modelling Markup Language (PMML) [9, 10] has developed over the years to become, now, a mature standard for the interchange of models. Active research continues to ensure the PMML standard captures not only the models but also the transforms required on our data to be used by the model.

PMML is important because it allows us to build models and have them deployed on other platforms. These other platforms may be carefully tuned deployment platforms that can score large volumes efficiently, as demonstrated by the ADAPA real-time PMML-based scoring tool. Open standards that allow the interchange of models between closed source and open source software will be increasingly important.

There is also a growing appreciation of the analyst first movement. The movement recognises that the first most priority is with the analyst or the data miner, not the tools being used. A corollary of the movement might be that the freely available open source tools, in the hands of the most skilled analysts, will deliver more than the most expensive data mining tools, in the hands of the less skilled analyst. Thus, a focus on open source data mining has brought me on quite a journey to the development of Rattle [20].

5 Charting the Route: Open Source Tools

Data mining is becoming a technology that is freely available to anyone who wishes to make use of it. Over the years, the technology has only been generally available to researchers and through large statistical software vendors. The vendors have been able to command significant prices for what is essentially a collection of easily implemented algorithms. But large organisations (the customers of the software vendors) have not understood what data mining is. These customers have been sold on the idea that software, rather than the skills of an analyst, is what makes a difference. It would be nice if that was true (see Sect. 6), but we have a long way to go yet.

The algorithms which are commonly implemented in a data mining suite have included k-means clustering, association rules, decision trees, regression, neural

networks, and support vector machines. Each of these basic data mining algorithms is actually quite simple to implement (for a software engineer), and I (and many others) have implemented, over the years, versions of each of them.

All of the data mining algorithms are available freely as open source software. Indeed, newly developed algorithms, emerging from the research laboratories, are now often available freely as open source software for many years before the vendors are able to catch up. The problem, though, is the lack of skilled people to be able to make use of the tools. Many organisations are instead quite happy to spend millions on the commercial products to deliver the mythical silver bullet, rather than investing in the people who can more effectively deliver with the technology.

With the growing popularity of the free and open source R statistical software [13], widely tested and used, implementations of all of these data mining algorithms have been available for many years. Other free and open source offerings include Weka [23], written in Java, KNIME [3], and RapidMiner [11]. However, the problem remains that we need to have a high level of skill to use many of these tools.

On joining the Australian Taxation Office to lead the roll out of data mining technology across a very large data-rich organisation, I quickly realised the issue facing many such organisations. There was a large population of quite skilled data analysts, quite happy with extracting data from large data warehouses, but not familiar with the advances in data mining that could quite quickly add considerable value to that data. Expensive data mining software could not be most effectively used because of a lack of understanding of how data mining worked. The first project or two delivered quite good results, but as the technology began to be employed seriously, the limited depth of knowledge about data mining began to inhibit progress.

I also began questioning the use of the commercial and closed source software when the same functionality was freely available. Using R, I was able to replicate all of the modelling performed by the expensive tools. But there is no question that R is a programming language, for what I call programming the data analyses, and requires a high level of skill. Data mining is about living and breathing our data, and not about pushing buttons to get results.

With the goal of providing more data miners with access to the most powerful suite of data mining algorithms and providing an approach to facilitate an understanding of what is being done, I began work on Rattle [19, 20].

5.1 A Beacon to Light the Way: Rattle

Rattle provides a very simple and easy-to-use graphical interface for data mining. Whilst others have delivered quite sophisticated, attractive interfaces for data mining, including the very common process diagram interfaces, Rattle continues to provide a basic but readily usable interface. The real focus is on migrating the

user from the basics of data mining to the full power of programming with data. The goal is to turn a data analyst into a data miner, augmenting an SQL programming skill with the power of a fully fledged statistical programming language like R.

Over 6 years of development, Rattle has become a mature product, freely available. (It is also available as a plug-in for a commercial business intelligence product from Information Builders.) It is used for teaching data mining and widely used by consultants for delivering data mining projects, without the overhead of expensive software.

Rattle can be used by anyone! It is a simple installation. With a few clicks (after we have the right data), we can have our first models. But of course, that is simply the beginning of a long journey to becoming a skilled data miner programming our analyses in R. As we gradually allow these skills to become more readily accessible, the algorithms will become "second nature" and part of the common toolbox.

5.2 Are We There Yet: Freely Receive and Freely Give

Data mining is essentially a practical application of research to real world analysis of data. We will continue to see a growing number of research papers published, incrementally increasing our global knowledge. But the real impacts are when we apply the technology. This can be where we can make significant impacts and change how organisations do things. It can be very satisfying. But the real key to this, I believe, is in freely making available the fruits of your research to all.

Scientific research has always held high the principle of repeatability and sharing of results. Peer review is important for any scientific endeavour, and advances are made by freely discussing and critiquing our work with others. Today's focus on commercialising everything leads us to hide and protect everything we do, just in case we have the big winner that will ensure our personal financial security.

Financial security is important, but so is the need for our society as a whole to benefit from all the good that we can offer. Too often I have seen students, perhaps encouraged by their advisers or their institutions, to not make their implementations of their research available, just in case they can make some significant money from it. The majority of this code then simply disappears forever. What a waste. Even simply hiding the implementations for a year or two can waste other people's time and effort that could be better spent on pushing forward the science rather than unnecessarily replicating the implementations.

For the good of all of us, do consider making your software implementations of new algorithms available. We can then try out your algorithms on our data and repeat your experiments to see how well your results generalise to other problems. We can more efficiently compare your approach to other approaches, in unbiased experiments, and share the results of these experiments to help improve results all round.

My recommendation is to package up your algorithm to make it available, for example, in R or in Weka, or simply out there.

6 Rowing in Unison: A Bright Future

Perhaps the most interesting thing to do every now and again is to sit back and wonder where the world is heading. Many writers and researchers do this, and it is instructive, though not necessarily accurate, to do so. As Alan Kay (inventor of the computer desktop-window-mouse paradigm we are still using today) said in 1971, "The best way to predict the future is to invent it". Science fiction provides a most fruitful avenue for exploring possible futures and developing an understanding of the consequences. Some predictions come true, some do not, but the fact of exploring the ideas and possibilities influences the very future we are heading towards.

For me, I see the longer-term future of data mining, leading towards delivering on some of the goals of machine learning and artificial intelligence—goals that have been around since the 1950s: for our computers to behave intelligently by learning from their interactions with the world. But the world is going to be quite different. We need to put data mining out there as a technology for accessible by all who wish to do so. And where we obtain the data to be mined will be radically different to where it is today (centralised versus distributed).

Data mining technology must become commonplace and even disappear into the fabric of our daily life. The story was the same for expert systems of the 1980s. Expert systems are no longer specifically talked about, but are ever present. They underpin many decision-making systems in use today, from supporting doctors in the interpretation of medical images to deciding whether you are a good customer for the bank.

The research for this is not necessarily about better algorithms for machine learning. We will, nonetheless, see further incremental improvements in our algorithms over time. We may see some significant conceptual shifts in our paradigms that suddenly make significant advances in collecting and using knowledge.

The direction I see that is needed for data mining is more about how we freely make the technology available to anyone. Rattle is a small, but practical, attempt to move in this direction, providing a free and open source tool. Others include KNIME [3] and RapidMiner [11]. The aim is to make it simple for the less statistically and computationally sophisticated to do the right thing in analysing data.

A longer term goal in many areas of research relating to modelling and computer science has been in intelligent tool selectors (or intelligent model selection). In the data mining context, given a data set for analysis, the user needs to be guided accurately and with statistical validity in the right direction. The intelligent guidance provided by the data mining platform of the future will consult, collaborate, and co-design[1] the analysis with the user. The ensemble of the data mining tool and the expert user will work to deliver new knowledge from the ever growing supplies of data.

[1] A motto borrowed from the Australian Taxation Office: http://www.ato.gov.au/corporate/content.asp?doc=/content/78950.htm.

7 New Horizons: Private Distributed Data

Reflecting on how we are moving forward with the management of data into the future, we might start to see some interesting trends and postulate some interesting directions. Today, we have moved into a phase of the deployment of technology which is very much oriented around putting commercial interests well and truly before individual or societal interests. The pendulum is about to start swinging the other way, we hope, for the good of society.

The concepts of cloud computing and our willingness to (perhaps unwittingly) hand over so much personal data to be controlled by social network hubs have considerable currency. Facebook, as an example, probably makes a claim to the ownership of all of the data that hundreds of millions of users are freely providing to it. The owners of these massive data stores may, at their own discretion or under duress from well-intentioned (or not) authorities, decide to do with that data whatever they like, without recourse or reference to those who supplied the data. The WikiLeaks episode of 2010–2011 made this very clear, with US authorities requiring social networking sites to hand over related data.

We now see, though, a growing interest and concern for privacy amongst the general users of the Internet. Consequently, we are also seeing the emergence of alternative, privacy preserving, and cloud and social networking applications. There is starting to emerge new technology that allows a user, or a household, to retain all of their data locally within a tiny plug device or perhaps within a personal smartphone. The technology allows these same users to continue to participate actively in the networked social communities as they do now. Plug devices (or smartphones), and the Debian-based Freedom Box,[2] are technology pointing the way to a possible future.

Such a device will be low powered with many days of battery charge in case of outages and will connect through wireless or Ethernet to the Internet. The network connection will normally be via an ISP over the phone line, as now, but with backup through the mobile network (3G and 4G) and further backup through local wireless mesh networks. The mesh network operates by connecting to your neighbours' device, which connects to their neighbours, and so on.

Under this scenario, data will be distributed. All of my e-mails, all of my social networking comments and photos, all of my documents, all of my music and videos, all of my medical records, and more will live on the small, high-capacity, but very energy-efficient device. All of this will be securely available to myself, wherever I am anywhere on the network, whether through a desktop computer or my smartphone. The data will also be available to those to whom I give access—perhaps to share social-network-type interactions or for my doctor to review my complete medical history. I retain ownership of my own data and provide it to others explicitly at my discretion.

[2] https://freedomboxfoundation.org/.

There will also be intelligent agents developed that will look after the security and backup of the data stored on the device. The intelligent agents will monitor and report to me who is accessing the data, to keep me aware of such things, when I want to be.

A new economy will develop. My intelligent agent, looking after my interests, will negotiate with centralised services, such as with Google, perhaps, to share some of my data. A service like Google may then use this data for large-scale analysis, perhaps for targeted advertising, or improving searching and spam filtering, etc. Those negotiations may involve financial rewards (or enticements) when I make some of my data available. But I have the ultimate control over what data I make available and when I make it available.

As this scenario begins to develop over the coming decade, how we do data mining will fundamentally have to change. Some advocates of the distributed data model rally against the prospect of data mining, seeing the distributed data as a mechanism to defeat data mining. However, we should not allow data mining to become the enemy, but instead, as I describe in the previous section, allow all to understand what data mining is about and even facilitate many more to easily access the technology. We can then choose to opt in to have our data analysed as part of the worldwide population for benefits that we clearly understand and agree to, but to be in control of when we make it so available. We may receive rewards or enticements to do so or other benefits, but we emphasise that it is under our personal control.

The data mining of distributed data in the new distributed personal server world will present many interesting challenges. Can we, for example, perform distributed data mining at this fine-grained level whilst not removing the raw data from a user's personal server? Or can we guarantee anonymity in the analysis of the data as it is scooped from a million personal servers? There may need to be developments around how our intelligent agents will collaborate with several million other agents to allow patterns of interest to be discovered, using distributed and privacy preserving approaches to data mining.

In summary, we will see a movement of individuals, slowly but surely, moving from centralised services to distributed, personal services. Individuals will again own their personal data and can make that data available on a per request basis, accepting payment or other benefits for making the data available. The data remains on their own personal server, but it is queried by the data mining tools across the network, with the network acting as a large distributed database. This early phase of the movement provides an opportunity for some pioneering work in distributed, fine grained, privacy preserving data mining, linked also to the concept of intelligent agents looking after the interests of the individual personal servers.

8 Summary

We continue to see an increasing demand for data miners and data mining tools. There is today a lot of active research in data mining, though it is an increasingly crowded space. Advances in data mining research deliver very small steps, one step at a time.

Pondering over where computers and technology are heading, after watching the incredible growth of social networking, we begin to see the amazing uptake of smart mobile devices (smartphones). We are now beginning to see an increasing concern for privacy and questions being raised about the ownership of data. New technology will arise over the coming years which will deliver personal servers (plug devices and smartphones) to individuals and households, where data remains with the individual. The data of the world will be very fragmented and finely distributed. Significant new challenges for data mining arise in this scenario, and little research has focused on data mining of such fine-grained sources of data.

To finish this journey, let me recount something I once heard Marvin Minsky (one of the pioneers of artificial intelligence) say: it was to the effect that when a research area gets crowded, it is time to move on. Look for new pastures where you can make a major splash, rather than staying within the crowded stadium with much competition and only little progress for each step forward. Invent the future and develop the technology that is needed to work there.

Finally, ensure that our future is free—make your research and algorithms and implementations freely available for all to benefit. Allow all the choice to participate.

References

1. S. Bakin, M. Hegland, G.J. Williams, Mining taxation data with parallel bmars. Parallel Algorithm. Appl. **15**, 37–55 (2000)
2. R.M. Bell, J. Bennett, Y. Koren, C. Volinsky, The million dollar programming prize. IEEE Spectr. **46**, 28–33 (2009)
3. M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, *KNIME: The Konstanz Information Miner*. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007) (Springer, Heidelberg, 2007)
4. L. Breiman, Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. L. Breiman, J. Friedman, R. Olshen, C. Stone, *Classification and Regression Trees* (Wadsworth and Brooks, Monterey, CA, 1984)
6. P. Compton, R. Jansen, Knowledge in context: a strategy for expert system maintenance, in *Proceedings of the 2nd Australian Joint Conference on Artificial Intelligence* (1988), pp. 292–306
7. J.R. Davis, P.M. Nanninga, G.J. Williams, Geographic expert systems for resource management, in *Proceedings of the First Australian Conference on Applications of Expert Systems* (Sydney, Australia, 1985)
8. Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in *Proceedings of the Second European Conference on Computational Learning Theory* (Springer, London, 1995), pp. 23–37
9. A. Guazzelli, W.-C. Lin, T. Jena, *PMML in Action*, CreateSpace (2010)
10. A. Guazzelli, M. Zeller, W.-C. Lin, G. Williams, Pmml: an open standard for sharing models. R J. **1**(1), 60–65 (2009). http://journal.r-project.org/2009-1/RJournalfi2009-1fiGuazzelli+et+al.pdf
11. I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, Yale: rapid prototyping for complex data mining tasks, in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. by L. Ungar, M. Craven, D. Gunopulos, T. Eliassi-Rad (ACM, Philadelphia, PA, 2006), pp. 935–940

12. J.R. Quinlan, Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
13. R (1993) *A Language and Environment for Statistical Computing*, Open Source, http://www.R-project.org
14. G.A. Riessen, G.J. Williams, X. Yao, Pepnet: parallel evolutionary programming for constructing artificial neural networks, in *Evolutionary Programming VI*, ed. by P.J. Angeline, R.G. Reynolds, J.R. McDonnell, R. Eberhart. Lecture Notes in Computer Science, vol. 1213 (Springer, Indianapolis, IN, 1997), pp. 35–46
15. G.J. Williams, Some experiments in decision tree induction. Aust. Comput. J. **19**(2), 84–91 (1987). http://togaware.com/papers/acj87fidtrees.pdf
16. G.J. Williams, Combining decision trees: initial results from the MIL algorithm, in: *Artificial Intelligence Developments and Applications: Selected papers from the first Australian Joint Artificial Intelligence Conference, Sydney, Australia, 2–4 November, 1987*, ed. by J.S. Gero, R.B. Stanton (Elsevier Science Publishers B.V., North-Holland, 1988), pp. 273–289
17. G.J. Williams, Frameup: a frames formalism for expert systems. Aust. Comput. J. **21**(1), 33–40 (1989). http://togaware.com/papers/acj89fiheffe.pdf
18. G.J. Williams, Inducing and combining decision structures for expert systems, Ph.D. thesis, Australian National University, 1991, http://togaware.com/papers/gjwthesis.pdf
19. G.J. Williams, Rattle: a data mining GUI for R. R J. **1**(2), 45–55 (2009). http://journal.r-project.org/archive/2009-2/RJournalfi2009-2fiWilliams.pdf
20. G.J. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Use R! (Springer, New York, 2011)
21. G.J. Williams, J.R. Davis, P.M. Nanninga, Gem: a microcomputer based expert system for geographic domains, in *Proceedings of the Sixth International Workshop and Conference on Expert Systems and Their Applications* (Avignon, France, 1986), Winner of the best student paper award
22. G.J. Williams, Z. Huang, Mining the knowledge mine: the hot spots methodology for mining large real world databases, in *Advanced Topics in Artificial Intelligence*, ed. by A. Sattar (Springer, London, 1997), pp. 340–348
23. I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. (Morgan Kaufmann, San Francisco, CA, 2005). http://www.cs.waikato.ac.nz/~ml/weka/book.html
24. K. Yamanishi, J-i Takeuchi, G.J. Williams, P. Milne, Online unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min. Knowl. Discov. **8**, 275–300 (2004)

A Journey in Pattern Mining

Mohammed J. Zaki

1 Motivation

The traditional research paradigm in the sciences was hypothesis-driven. Over the last decade or so, this hypothesis-driven view has been replaced with a data-driven view of scientific research. In almost all fields of scientific endeavor, large research teams are systematically collecting data on questions of great import. Knowledge and insights are gained through data analysis and mining, feeding this inversion of science, i.e., rather than going from hypothesis to data, we use data to generate and validate hypotheses and to generate knowledge and understanding. The same can be said for applications in the commercial realm.

The pace of data gathering is expected to continue unabated into the distant future, with increasing complexity. New methods are required to handle the increasingly interlinked, yet distributed and heterogeneous data, spanning various scales and modalities. Data mining offers the possibility of making fundamental discoveries at every turn, whether in the fundamentals of data mining itself or in the target application domains. These are exciting times for data mining and for data miners, given that we have only started to scratch the surface in terms of what is possible.

2 Milestones and Success Stories

My research has heavily focused on frequent pattern mining. Starting with efficient sequential and parallel algorithms for itemset mining, my work evolved to consider sequences, trees, and currently graph mining.

M.J. Zaki (✉)
Rensselaer Polytechnic Institute, Troy, NY, USA
e-mail: zaki@cs.rpi.edu

M.M. Gaber (ed.), *Journeys to Data Mining*,
DOI 10.1007/978-3-642-28047-4_16,

Interestingly, my path to data mining was somewhat tortuous. I started my Ph.D. career in compilers, working under the guidance of my first advisor, Wei Li, who is currently at Intel. He had the great foresight (this was in 1995) of pointing me toward data mining, albeit from the perspective of compiling for large-scale data-intensive applications. My first exposure to data mining was the work of Rakesh Agrawal and Ramakrishnan Srikant on the Apriori method for frequent itemset mining [1]. From that point, I was basically enthralled with data mining. When my advisor left academia, I finished my Ph.D. under the guidance of Mitsunori Ogihara, focusing on scalable sequential and parallel methods for mining itemsets, sequences, and decision trees [2]. During my Ph.D., I also had the good fortune of interning, first with the MineSet team at Silicon Graphics, Inc., and next with the Quest team at IBM Almaden Research Center, under Rakesh Agrawal and Howard Ho.

My research has focused on the issues of scalability and efficiency for the various data mining tasks, especially pattern mining. My first foray into data mining was to develop a parallel version of the Apriori itemset mining algorithm for shared memory machines [3]. The main issues were how to balance the work across the processors for the candidate generation and support counting steps. I also developed parallel methods for sequence mining [4] and decision tree classification [5, 6].

2.1 Itemset Mining

One of my most cited works is the Eclat algorithm for frequent itemset mining [7, 8]. Eclat was designed to exploit the itemset lattice via a hybrid depth-first search, using the simple yet efficient vertical tidset representation of the database (akin to an "inverted index"). The main idea was to separate the pattern enumeration step from the frequency counting step. Efficient and independent generation of the candidates was obtained by the use of self-contained equivalence class-based candidate extension, and efficient counting was obtained via tidset intersections. This approach allowed flexibility in designing effective search space traversal strategies, independently from counting. The different variants of the Eclat approach were also parallelized [9]. Follow-up work on efficient methods for itemset mining led to CHARM for mining closed frequent itemsets [10] and GenMax for mining maximal frequent itemsets [11].

2.2 Closed Itemsets

As I explored more about lattices, I stumbled upon the seminal and elegant lattice-theoretic framework of Formal Concept Analysis (FCA) developed by Rudolf Wille [12]. FCA provides much of the theoretical underpinnings for itemset mining and association rules, using the notion of formal concepts (also known as closed itemsets) and rules between concepts. In particular, association rules can be

considered to be partial implications [13], with the key difference being that in association rules, only frequent and strong rules are considered. My initial work [14] helped establish this close link between frequent closed itemsets and FCA. This connection was also independently established in [15].

The FCA connection also provided the framework to establish that there are lots of redundancies in the association rules. This led me to develop approaches for mining non-redundant association rules [16, 17]. Since frequent closed itemsets provide a loss-less summary of all possible frequent itemsets, I endeavored to develop an efficient method for mining all the frequent closed itemsets, resulting in CHARM [18]. I later extended CHARM to also generate the lattice and the set of non-redundant and conditional rules [10].

2.3 Sequence Mining

The focus on vertical data representation also led to the development of the SPADE [19, 20] algorithm for frequent sequence mining. SPADE relies on two novel concepts. For sequence enumeration, I proposed the lattice-based self-contained equivalence classes for both sequence and itemset extensions, and for fast support computation, I proposed temporal and set joins over id-lists, which store for each frequent sequence the sequence ids and time stamps where the last element of the sequence occurs. Like Eclat, SPADE allows flexible candidate space searching, with efficient frequency counting. A constrained version called cSPADE [21] was also developed that allowed the mining of sequences with minimum and maximum gaps between elements, maximum window size, and also generating discriminative sequences for use as features in a classification setting [22].

2.4 Tree Mining

With the advent of semistructured and XML data, I was attracted to the problem of mining more complex patterns like trees and graphs. I proposed TreeMiner, the first method for mining all the embedded frequent trees in a database of trees [23, 24]. The main contribution here was the use of a novel equivalence class and right-most path-based tree extension that avoids duplicate candidates, and a novel scope-list vertical index that supports constant time child and sibling tests to find tree occurrences. This approach was extended to mine embedded, and induced, as well as ordered and unordered frequent subtrees in the SLEUTH method [25]. TreeMiner was also used to mine rules for XML classification [26].

2.5 Generic Pattern Mining

A novel vertical approach was also used for a graph mining algorithm implemented in the Data Mining Template Library (DMTL) [27]. DMTL[1] is an open-source, generic library (written in C++) for mining increasingly complex patterns ranging from itemsets to sequences, trees, and graphs. DMTL has been downloaded over 6,100 times by researchers from all over the world. The DMTL effort also highlights one of the long-term goals of my research, namely, to develop a common "grand unified theory" of the various data mining tasks; DMTL does this for pattern mining. DMTL utilizes a generic data mining approach, where all aspects of mining are controlled via a set of properties. DMTL uses a novel property-based pattern mining approach based on a property hierarchy for the different pattern types. The hierarchy captures the relationships between different pattern types (e.g., whether a pattern is connected, directed, acyclic, rooted, or ordered) and associates these with specific pattern types (set, sequence, chain, (free) tree, directed acyclic graph (DAG), or (un)directed graph). For example, an itemset pattern is defined as a disconnected set of nodes having unique labels. A sequence on the other hand is defined as directed, acyclic, rooted structure that has in-degree and out-degree at most one for each node. Likewise, other types of patterns are defined as a list of formal properties. In the generic paradigm, algorithms (e.g., for pattern isomorphism and frequency checking) work for any pattern type, since they accept only a list of pattern properties. The user can mine custom pattern types by simply defining the new pattern types in terms of their properties, without the need to implement a new algorithm. Another novel feature of DMTL is that it provides transparent persistency and indexing support for both data and mined patterns for effective computation over massive datasets.

2.6 Graph Mining and Sampling

More recently, the focus of my research has moved away from complete combinatorial search to sampling-based approaches for graph mining [28–30] and graph indexing [31]. Typical graph mining methods follow the combinatorial pattern enumeration paradigm and aim to extract all frequent subgraphs, perhaps subject to some constraints. In many real-world cases, enumerating all frequent patterns is not necessarily the primary objective and may not even be feasible if the graphs are large. Rather, mined patterns are likely to be used as inputs for a subsequent analysis/modeling step, and as such, a relatively small representative set of patterns may suffice. For example, frequent patterns obtained from graphs and networks can be used to build classification models. Further, the lack of interpretability and the

[1] http://sourceforge.net/projects/dmtl/files/.

curse of dimensionality due to a large set of redundant patterns can cause problems for subsequent steps like clustering and classification. Many successful applications of pattern mining thus require the result set to be a summary, rather than a complete set of the frequent pattern space. We have formulated a novel paradigm for mining interesting graph patterns, based on the concept of output space sampling (OSS) [30]. In this paradigm, the objective is to sample frequent patterns instead of complete enumeration. The sampling process automatically performs interestingness-based selection by embedding the interestingness score of the patterns in the desired target distribution. Another important feature is that OSS is a generic method that applies to any kind of pattern, such as a set, a sequence, a tree, and of course a graph. OSS is based on Markov Chain Monte Carlo (MCMC) sampling. It performs a random walk on the candidate subgraph partial order and returns subgraph samples when the walk converges to a desired stationary distribution. The transition probability matrix of the random walk is computed locally to avoid a complete enumeration of the candidate frequent patterns, which makes the sampling paradigm scalable to large real-life graph datasets. Output space sampling is an entire paradigm shift in frequent pattern mining that holds enormous promise. While traditional pattern mining strives for completeness, OSS targets to obtain a few interesting samples. The definition of interestingness can be very generic, so user can sample patterns from different target distributions by choosing different interestingness functions. This is very beneficial as mined patterns are subject to subsequent use in various knowledge discovery tasks, like classification and clustering, and the interestingness of a pattern varies for various tasks. OSS can adapt to this requirement just by changing the interestingness function. OSS also solves pattern redundancy problem by finding samples that are very different from each other. We have used OSS to mine different types of graph summaries, such as uniform samples, support-biased samples, maximal pattern samples, discriminative subgraph samples, and so on.

2.7 Applications in Bioinformatics

I have restricted the discussion above to my research work in data mining. However, my other passion is the application of data mining methods in bioinformatics applications, as well as indexing complex data. Very briefly, I have worked on several interesting problems in bioinformatics, such as protein contact map mining [32], protein folding pathways [33], protein structure indexing and disk-based genome scale indexing via suffix trees [34, 35], structured sequence motif search and extraction [36, 37], protein shape matching for docking [38], non-sequential and flexible protein structure alignment [39, 40], microarray gene expression mining in 2D and 3D matrices [41, 42], and applications of boolean expression mining for redescribing genesets [43, 44].

2.8 Textbook Project

Another project that is keeping me busy currently is the forthcoming text book on "Fundamentals of Data Mining Algorithms," coauthored with Wagner Meira, Jr. This has been a tremendous learning experience for me, especially since I am implementing almost all of the major data mining algorithms. My current advice is that *if you want to learn a subject, you should teach a course on it, and if you want to truly understand a subject, you should write a text book on it*.

3 Lessons in Learning from Failures

Research surely has its elating times and brooding times. Success invariably hollows on the heels of (many) failures. For instance, newer and faster algorithms follow from the "failure" of the previous methods in some crucial aspect. In that sense, science is a search for failures or, to state it more mildly, a search for weaknesses and shortcomings.

Over the years, one fact I have come to realize is that one must always try the simplest solutions first. While this should be obvious, it is surprising how many times one opts for a rather complex approach, without doing basic sanity checks. This is especially true for large-scale data. Simple solutions may work remarkably well, where more complex methods may be too costly to run.

4 Current Research Issues and Challenges

Increasingly, today's massive data is in the form of complex graphs or networks. Examples include the World Wide Web (with its Web pages and hyperlinks), social networks (wikis, blogs, tweets, and other social media data), semantic networks (ontologies), criminal/terrorist networks, biological networks (protein interactions, gene regulation networks, metabolic pathways), corporate networks (knowledge networks in the form of documents, policies, etc.; Intranet and other IT network), and so on. With the explosion of such networked data, there is a pressing need for data mining, analysis, and querying tools to rapidly make sense of and extract knowledge from these massive and complex graphs, with millions to billions of nodes and edges.

Scalable methods for mining such large graphs for frequent subgraph patterns remain an open challenge. Scalable graph indexing methods to answer graph reachability queries, as well as more complex graph algorithms over such massive, enriched graphs, are also lacking. The mining and querying problem is made more challenging since one typically has to integrate information from multiple data sources, each providing only a slice of the information. However, an integrated

analysis is crucial to discovering novel patterns and knowledge from these complex and massive datasets. Another challenge is that often, the complete graph may be unbounded and not fully known, i.e., we may need to infer or mine the relationships, and we may have to develop methods to account for uncertainty (e.g., via confidences on the edges). We also need dynamic approaches for both indexing and mining, since the underlying data can change in time.

Developing novel methods for the integrated mining, analysis, and indexing of very large (possibly unbounded), complex, and dynamic graphs and networks is clearly one of the major research challenges in data mining. Due to the ubiquitous nature of graph data, such techniques will be highly relevant and useful for the exploratory analysis of complex graph datasets in a variety of application domains ranging from social network analysis to network biology.

5 Research Tools and Techniques

The fundamental tools and techniques of research in data mining include the ability to do a thorough literature search, to be able to quickly prototype algorithms, and to write good papers.

The first task has become a lot easier with search engines and online bibliographic databases. It would be inexcusable not to cite the most relevant work, regardless of when and where it was published. On the other hand, one has to be careful not to get overwhelmed by the sheer information. One approach that I have usually found to be effective is to first think through the possible solutions before reading the latest literature. The goal is to retain independent thought on the issue and not to blindly adhere to the accepted thinking.

I am somewhat of a dabbler in different programming languages. I currently recommend a scripting language like Python for rapid prototyping of algorithms. I am even using R in my data mining courses; it is surprisingly versatile. The productivity enhancement can be quite substantial, especially since data parsing, pre-processing, and post-processing do comprise a significant fraction of the time for doing data mining. Of course, when one desires pure performance, one can always roll out a C++ implementation.

While I cannot claim to be an expert in writing good papers, I do think that some tools can help render great graphs and figures for papers. I have some across many papers with very poor quality figures. I now use PSTricks[2] extensively with LaTeX. There is a steep learning curve, but the end product is unmatched by any other tool. PSTricks has all but replaced tools I used in the past, like gnuplot for plotting graphs and xfig for making diagrams.

[2] http://tug.org/PSTricks.

6 Making an Impact

A young researcher should always be on the lookout for new frontiers, since that is where she will make the most impact, at least in the short term. Advisors can help point the students in the right direction. I got my break when my advisor pointed me to data mining, which was in its infancy in the mid-1990s.

Regardless of the research area, one should be ready to challenge accepted paradigms; one should not be afraid to try simple solutions that others may have missed. Finally, the importance of conveying the research results clearly cannot be overemphasized. The artifacts of research like software, or data, should be made public where possible. If citations are the currency to measure impact, then availability of such artifacts can help a lot.

7 Future Insights

Data mining is a key component of the new data science revolution underway in virtually all fields of science, including but not limited to bioinformatics, astro-informatics, ecological-informatics, geo-informatics, chem-informatics, materials-informatics, etc.

The science of data seeks to make data a first-class object to be studied in its own right, looking for fundamental theories and models of data that span the disciplinary boundaries. While the fundamental scientific questions will be different across fields, the data management, analysis, and mining aspects will share many commonalities. Data science aims to study and characterize precisely these universal aspects of data. It is also worth emphasizing that the data encompasses the knowledge that is derived from it, as well as the process (or workflow) leading to those knowledge nuggets, since post-discovery such knowledge and its context becomes data for subsequent steps. The key elements of the new science of data will involve at least the following aspects:

- New data models for streaming and dynamic data.
- Ability to handle massive and distributed datasets in the tera- and petascale.
- Ability to handle various data modalities, such as flat files, tables, matrices, images, video, audio, text, hypertext, “semantic” text, as well as trees, graphs, and networks.
- Novel mining, learning, and statistical algorithms that offer timely and reliable inference. Online and approximate methods will be essential.
- Self-aware, intelligent continuous data monitoring, management, and analysis.
- Data and information integration.
- Data and model compression for massive data.
- Service-oriented architecture and middleware for supporting the complete data science workflow.

- Data and knowledge provenance to provide the rationale as well as repeatability, a key element of any scientific endeavor.
- Data security and privacy. Safeguarding privacy or security is crucial at the same time allowing for data analysis and mining.
- Novel data sensation methods spanning visual, aural, and tactile interaction, for better understanding of the data and knowledge.

It is exciting to be witnessing the data revolution, with data mining smack at its core. Data miners have a bright future ahead!

References

1. R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in *20th VLDB Conference*, Sept 1994
2. M.J. Zaki, Scalable data mining for rules. Technical Report URCSTR-702 (Ph.D. Thesis), University of Rochester, July 1998
3. M.J. Zaki, M. Ogihara, S. Parthasarathy, W. Li, Parallel data mining for association rules on shared-memory multi-processors, in *Supercomputing'96*, Nov 1996
4. M.J. Zaki, Parallel sequence mining on shared-memory machines. J. Parallel Distrib. Comput. **61**(3), 401–426 (2001). Special issue on High Performance Data Mining
5. M.J. Zaki, C.-T. Ho, R. Agrawal, Parallel classification for data mining on shared-memory multiprocessors, in *15th IEEE International Conference on Data Engineering*, Mar 1999. See IBM Technical Report RJ10104 [6] for a more detailed version of this paper
6. M.J. Zaki, C.-T. Ho, R. Agrawal, Parallel classification for data mining on shared-memory systems. Technical Report RJ10104, IBM, 1999
7. M.J. Zaki, Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. **12**(3), 372–390 (2000)
8. M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, New algorithms for fast discovery of association rules, in *3rd International Conference on Knowledge Discovery and Data Mining* (*KDD*), Aug 1997
9. M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, Parallel algorithms for discovery of association rules. Data Min. Knowl. Discov. Int. J. **1**(4), 343–373 (1997). Special issue on Scalable High-Performance Computing for KDD
10. M.J. Zaki, C.-J. Hsiao, Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. Knowl. Data Eng. **17**(4), 462–478 (2005)
11. K. Gouda, M.J. Zaki, Genmax: an efficient algorithm for mining maximal frequent itemsets. Data Min. Knowl. Discov. Int. J. **11**(3), 223–242 (2005)
12. B. Ganter, R. Wille, *Formal Concept Analysis: Mathematical Foundations* (Springer, Berlin, 1999)
13. M. Luxenburger, Implications partielles dans un contexte. Math. Inf. Sci. Hum. **29**(113), 35–55 (1991)
14. M.J. Zaki, M. Ogihara, Theoretical foundations of association rules, in *3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, June 1998
15. N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Pruning closed itemset lattices for associations rules, in *14ème Journèes Bases de Donnèes Avancèes (BDA)*, 1998
16. M.J. Zaki, Generating non-redundant association rules, in *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2000
17. M.J. Zaki, Mining non-redundant association rules. Data Min. Knowl. Discov. Int. J. **9**(3), 223–248 (2004)

18. M.J. Zaki, C.-J. Hsiao, CHARM: an efficient algorithm for closed itemset mining, in *2nd SIAM International Conference on Data Mining*, Apr 2002
19. M.J. Zaki, Efficient enumeration of frequent sequences, in *7th ACM International Conference on Information and Knowledge Management*, Nov 1998
20. M.J. Zaki, SPADE: an efficient algorithm for mining frequent sequences. Mach. Learn. J. **42** (1/2), 31–60 (2001). Special issue on Unsupervised Learning
21. M.J. Zaki, Sequences mining in categorical domains: incorporating constraints, in *9th ACM International Conference on Information and Knowledge Management*, Nov 2000
22. N. Lesh, M.J. Zaki, M. Ogihara, Scalable feature mining for sequential data. IEEE Intell. Syst. Appl. **15**(2), 48–56 (2000). Special issue on Data Mining
23. M.J. Zaki, Efficiently mining frequent trees in a forest, in *8th ACM SIGKDD International Conference Knowledge Discovery and Data Mining*, July 2002
24. M.J. Zaki, Efficiently mining frequent trees in a forest: algorithms and applications. IEEE Trans. Knowl. Data Eng. **17**(8), 1021–1035 (2005). Special issue on Mining Biological Data
25. M.J. Zaki, Efficiently mining frequent embedded unordered trees. Fundamenta Informaticae **66**(1–2), 33–52 (2005). Special issue on Advances in Mining Graphs, Trees and Sequences
26. M.J. Zaki, C.C. Aggarwal, Xrules: an effective structural classifier for xml data. Mach. Learn. J. **62**(1–2), 137–170 (2006). Special issue on Statistical Relational Learning and Multi-Relational Data Mining
27. V. Chaoji, M.A. Hasan, S. Salem, M.J. Zaki, An integrated, generic approach to pattern mining: data mining template library. Data Min. Knowl. Discov. **17**(3), 457–495 (2008)
28. V. Chaoji, M.A. Hasan, S. Salem, J. Besson, M.J. Zaki, ORIGAMI: a novel and effective approach for mining representative orthogonal graph patterns. Stat. Anal. Data Min. **1**(2), 67–84 (2008)
29. M.A. Hasan, M.J. Zaki, Musk: uniform sampling of k maximal patterns, in *9th SIAM International Conference on Data Mining*, Apr 2009
30. M.A. Hasan, M.J. Zaki, Output space sampling for graph patterns, in *Proceedings of the VLDB Endowment* (*35th International Conference on Very Large Data Bases*) **2**(1), 730–741 (2009)
31. H. Yildirim, V. Chaoji, M.J. Zaki, Grail: scalable reachability index for large graphs. *Proceedings of the VLDB Endowment* (*36th International Conference on Very Large Data Bases*) **3**(1), 276–284 (2010)
32. M.J. Zaki, S. Jin, C. Bystroff, Mining residue contacts in proteins using local structure predictions. IEEE Trans. Syst. Man Cybern. B **33**(5), 789–801 (2003). Special issue on Bioengineering and Bioinformatics
33. M.J. Zaki, V. Nadimpally, D. Bardhan, C. Bystroff, Predicting protein folding pathways. Bioinformatics **20**(1), i386–i393 (Aug 2004). *Supplement on the Proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology*
34. F. Gao, M.J. Zaki, PSIST: indexing protein structures using suffix trees, in *IEEE Computational Systems Bioinformatics Conference*, Aug 2005
35. B. Phoophakdee, M.J. Zaki, Genome-scale disk-based suffix tree indexing, in *ACM SIGMOD International Conference on Management of Data*, June 2007
36. Y. Zhang, M.J. Zaki, Exmotif: efficient structured motif extraction. Algorithms Mol. Biol. **1**(21), (2006)
37. Y. Zhang, M.J. Zaki, Smotif: efficient structured pattern and profile motif search. Algorithms Mol. Biol. **1**(22), (2006)
38. Z. Shentu, M.A. Hasan, C. Bystroff, M.J. Zaki, Context shapes: efficient complementary shape matching for protein-protein docking. Prot. Struct. Funct. Bioinformatics **70**(3), 1056–1073 (2008)
39. S. Salem, M.J. Zaki, C. Bystroff, Iterative non-sequential protein structural alignment. J. Bioinformatics Comput. Biol. **7**(3), 571–596 (2009). Special issue on the best of CSB'08
40. S. Salem, M.J. Zaki, C. Bystroff, FlexSnap: flexible nonsequential protein structure alignment. Algorithms Mol. Biol. **5**(12), (2010). Special issue on best papers from WABI'09

41. L. Zhao, M.J. Zaki, Microcluster: an efficient deterministic biclustering algorithm for microarray data. IEEE Intell. Syst. **20**(6), 40–49 (2005). Special issue on Data Mining for Bioinformatics
42. L. Zhao, M.J. Zaki, TriCluster: an effective algorithm for mining coherent clusters in 3d microarray data, in *ACM SIGMOD Conference on Management of Data*, June 2005
43. M.J. Zaki, N. Ramakrishnan, L. Zhao, Mining frequent boolean expressions: application to gene expression and regulatory modeling. Int. J. Knowl. Discov. Bioinformatics **1**(3), 68–96 (2010). Special issue on Mining Complex Structures in Biology
44. L. Zhao, M.J. Zaki, N. Ramakrishnan, Blosom: a framework for mining arbitrary boolean expressions, in *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2006