

# 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](mailto:mohamed.gaber@port.ac.uk)

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 BLOKDD 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.



<http://www.springer.com/978-3-642-28046-7>

Journeys to Data Mining  
Experiences from 15 Renowned Researchers  
Gaber, M.M. (Ed.)  
2012, VIII, 244 p., Hardcover  
ISBN: 978-3-642-28046-7