Wiley Series on Parallel and Distributed Computing

Albert Y. Zomaya, Series Editor

The DATA Bonanza

Improving Knowledge Discovery in Science, Engineering, and Business

EDITED BY

Malcolm Atkinson

Rob Baxter

Peter Brezany

Oscar Corcho

Michelle Galea

Mark Parsons

David Snelling

Jano van Hemert

IEEE

IEEE computer society

WILEY

The DATA Bonanza

WILEY SERIES ON PARALLEL AND DISTRIBUTED COMPUTING

Editor: Albert Y. Zomaya

A complete list of titles in this series appears at the end of this volume.

The DATA Bonanza

Improving Knowledge Discovery in Science, Engineering, and Business

Edited by

Malcolm Atkinson
Rob Baxter
Michelle Galea
Mark Parsons
University of Edinburgh, Edinburgh, UK

Peter Brezany
University of Vienna, Vienna, Austria

Oscar Corcho
Universidad Politécnica de Madrid, Madrid, Spain

Jano van Hemert
Optos PLC, Dunfermline, UK

David Snelling
Fujitsu Laboratories Europe Limited, Hayes, UK

WILEY | IEEE computer society

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created ore extended by sales representatives or written sales materials. The advice and strategies contained herin may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department with the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Atkinson, Malcolm.
The data bonanza : improving knowledge discovery for science, engineering and business / Malcolm Atkinson, Rob Baxter, Michelle Galea, Mark Parsons, Peter Brezany, Oscar Corcho, Jano van Hemert, David Snelling.
pages cm
ISBN 978-1-118-39864-7 (pbk.)
1. Information technology. 2. Information retrieval. 3. Databases. I. Title.
T58.5.A93 2013
001–dc23
2012035310

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To data-to-knowledge highway engineers, everywhere.

Contents

Contributors

M. Atkinson, School of Informatics, University of Edinburgh, Edinburgh, UK

A. Aschenbrenner, State and University Library Göttingen, Göttingen, Germany

J. Austin, Department of Computer Science, University of York, York, UK

R. Baldock, Medical Research Council, Human Genetics Unit, Edinburgh, UK

R. Baxter, EPCC, University of Edinburgh, Edinburgh, UK

P. Besana, School of Informatics, University of Edinburgh, Edinburgh, UK

T. Blanke, Digital Research Infrastructure in the Arts and Humanities, King's College London, London, UK

J. Blower, Reading e-Science Centre, University of Reading, Reading, UK

R. Cook, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

O. Corcho, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain

T. Damoulas, Department of Computer Science, Cornell University, Ithaca, New York, USA

D. Fink, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

C. Fritze, State and University Library Göttingen, Göttingen, Germany

M. Galea, School of Informatics, University of Edinburgh, Edinburgh, UK

A. Gemmell, Reading e-Science Centre, University of Reading, Reading, UK

O. Habala, Oddelenie paralelného a distribuovaného spracovania informácií̀', Ústav informatiky SAV, Bratislava, Slovakia

K. Haines, Reading e-Science Centre, University of Reading, Reading, UK

L. Han, School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University, Manchester, UK

J. van Hemert, Optos plc, Dunfermline, UK

L. Hluchý, Oddelenie paralelnèho a distribuovanèho spracovania informàciì', Ústav informatiky SAV, Bratislava, Slovakia

W. Hochachka, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

M. Holliman, Institute of Astronomy, University of Edinburgh, Edinburgh, UK

A. Hume, EPCC, University of Edinburgh, Edinburgh, UK

M. Jarka, Comarch SA, Warsaw, Poland

S. Kelling, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

T. Kitching, Institute of Astronomy, University of Edinburgh, Edinburgh, UK and Mullard Space Science Laboratory, University College London, Dorking, UK

A. Krause, EPCC, University of Edinburgh, Edinburgh, UK

R. Mann, Institute of Astronomy, University of Edinburgh, Edinburgh, UK

P. Martin, School of Informatics, University of Edinburgh, Edinburgh, UK

W. Michener, DataONE, University of New Mexico, Albuquerque, New Mexico, USA

A. Mouat, EPCC, University of Edinburgh, Edinburgh, UK

K. Noddle, Institute of Astronomy, University of Edinburgh, Edinburgh, UK

I. Overton, Medical Research Council, Human Genetics Unit, Edinburgh, UK

M. Parsons, EPCC, University of Edinburgh, Edinburgh, UK

W. Pempe, State and University Library Göttingen, Göttingen, Germany

A. Rietbrock, School of Environmental Sciences, University of Liverpool, Liverpool, UK

K. Rosenberg, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

C. Šilva, Department of Computer Science, Polytechnic Institute of New York, Brooklyn, New York, USA

A. Spinuso, School of Informatics, University of Edinburgh, Edinburgh, UK; Royal Netherlands Meteorological Institute, Information and Observation Services and Technology–R&D, Utrecht, The Netherlands

D. Snelling, Research Transformation and Innovation, Fujitsu Laboratories of Europe Limited, Hayes, UK

C. Sun Liew, School of Informatics, University of Edinburgh, Edinburgh, UK; Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

B. Simo, Oddelenie paralelnèho a distribuovanèho spracovania informàciì', Ústav informatiky SAV, Bratislava, Slovakia

V. Tran, Oddelenie paralelnèho a distribuovanèho spracovania informàciì', Ústav informatiky SAV, Bratislava, Slovakia

L. Trani, School of Informatics, University of Edinburgh, Edinburgh, UK; Royal Netherlands Meteorological Institute, Information and Observation Services and Technology–R&D, Utrecht, The Netherlands

L. Valkonen, School of Informatics, University of Edinburgh, Edinburgh, UK

G. Yaikhom, School of Informatics, University of Edinburgh, Edinburgh, UK; School of Computer Science and Informatics, Cardiff University, Cardiff, UK

Foreword

This book is a systematic approach to help face the challenge of data-intensive computing and a response to the articulation of that challenge set out in *The Fourth Paradigm*, a collection of essays by scientists and computer scientists that I co-edited. It also recognizes the fact that we face that challenge in the context of a digital revolution that is transforming communities worldwide at Internet speed.

This book proposes a strategy for partitioning the challenge, both in the ways in which we organize and in which we build systems. This partitioning builds on natural foci of interest and the examples show that this approach works well in the context of groups and organizations. The technological strategy reflects the evolving pattern of provision driven by business models and the flourishing diversity of tools and applications that enable human innovation.

We all face the need to separate concerns every time we face a data-intensive problem. This is key to making data-intensive methods routinely available and to their easy application. This leads to the recognition of effective working practices that need supporting with better 'datascopes' that are easily steered and focused to extract the relevant information from immense and diverse collections of data. The book calls for the introduction of ”intellectual on-ramps” that match the new tools to well-understood interfaces, so that practitioners can incrementally master the new data-intensive methods.

This book calls for recognition that this notion of intellectual on-ramps is worthy of study. Data-intensive computing warrants an appropriate engineering discipline that identifies effective ways of building appropriate highways from data to required knowledge. It is a call to arms for a serious attempt at initiating the professionalization of this discipline.

We are very much at the start of the digital revolution. The growth in digital data will not abate, and in certain areas it will probably accelerate. There is much to be gained by exploiting the opportunities this bonanza of data brings but the extraction of insights and knowledge from ‘Big Data’ will also certainly transform our organizations and society. Responding effectively to these changes requires the availability of ready-made tools and reusable processes together with practitioners with the skills to deploy them precisely and safely. This book frames an approach as to how these tools, processes, and skills may be developed. Such a systematic approach is now urgently needed as the opportunities are rapidly outgrowing our capabilities to assemble and run data-intensive campaigns.

This book provides a vocabulary to facilitate data-intensive engineering by introducing key concepts and notations. It presents nine in-depth case studies that show how practitioners have tackled data-intensive challenges in a wide range of disciplines. It also provides an up-to-date analysis of this rapidly changing field and a survey of many of the current research hot-spots that are driving it. For all these reasons, I believe that this book is a welcome addition to the literature on data-intensive computing.

Tony Hey

Redmond, Washington
July 2012

Preface

The world is undergoing a digital-data revolution. More and more data are born digital. Almost every business, governmental, and organizational activity is driven by data and produces data. Science, engineering, medicine, design and innovation are powered by data. This prevalence of data in all that we do is changing society, organizations, and individual behavior. To thrive in this new environment requires new strategies, new skills, and new technology. This book is the first to expound the strategies that will make you adept at exploiting the expanding opportunities of this new world.

This book identifies the driving forces that are provoking change and proposes a strategy for building the skills, methods, technologies, and businesses that will be well adapted in the emerging data-wealthy world. This strategy will change the way in which you spend your organization's (country's, company's, institution's, and profession's) resources. You will invest more in exploiting data, even if that means spending less on creating, capturing or archiving it, as today most data are underused or never used, even though they frequently *contain the latent evidence that should be leading to innovation, knowledge and wisdom.* After reading this book, you will expect an understandable path from data, via analysis, evidence and visualization, to influential outputs that change behavior. When this is not happening, your organization is underperforming and is at risk. You will come to expect that all of those with whom you deal should be competent at getting good value from data.

This book will change your skills by developing your ingenuity in discovering, understanding, exploiting, and presenting data. You will acquire a compendium of tools for addressing every stage in the data life cycle. It will change education—everyone needs survival skills for the data-wealthy world. All professionals and experts need experience and judgement for their part of the path from data to discovery, innovation, and outcomes.

This book will initiate the development of professionals who will engineer tomorrow's data highways. These highways will be designed to meet carefully analyzed and anticipated needs, to interwork with existing data infrastructure, to accelerate the journeys of millions from data to knowledge. These knowledge discovery highways will make it easy for people to get data from wherever they are stored and promptly deliver understandable information to wherever it is needed.

Who Should Read this Book

This book will be valuable to a wide range of information strategists, decision makers, researchers, students, and practitioners, from domains such as computer science (data mining, machine learning, statistics, databases, knowledge-based systems, large-scale computing), e-Science, and to workers in any discipline or industry where large-scale data handling and analysis is important.

The ideas presented are relevant to and draw on: data mining, knowledge discovery in databases, machine learning, artificial intelligence, databases and data management, data warehousing, information systems, distributed computing, grid computing, cloud computing, ubiquitous computing, e-Science (including a wide range of scientific and engineering fields dealing with large data), modeling and simulation.

This Book's Structure

This book consists of 24 chapters grouped into six parts; they are introduced here.

Part I: Strategies for success in the digital-data revolution Part I provides an executive summary of the whole book to convince strategists, politicians, managers, and educators that our future data-intensive society requires new thinking, new behavior, new culture, and new distribution of investment and effort. This part will introduce the major concepts so that readers are equipped to discuss and steer their organization's response to the opportunities and obligations brought by the growing wealth of data. It will help readers understand the changing context brought about by advances in digital devices, digital communication, and ubiquitous computing.

Chapter 1: The digital-data challenge This chapter will help readers to understand the challenges ahead in making good use of the data and introduce ideas that will lead to helpful strategies. A global digital-data revolution is catalyzing change in the ways in which we live, work, relax, govern, and organize. This is a significant change in society, as important as the invention of printing or the industrial revolution, but more challenging because it is happening globally at lnternet speed. Becoming agile in adapting to this new world is essential.

Chapter 2: The digital-data revolution This chapter reviews the relationships between data, information, knowledge, and wisdom. It analyses and quantifies the changes in technology and society that are delivering the data bonanza, and then reviews the consequential changes via representative examples in biology, Earth sciences, social sciences, leisure activity, and business. It exposes quantitative details and shows the complexity and diversity of the growing wealth of data, introducing some of its potential benefits and examples of the impediments to successfully realizing those benefits.

Chapter 3: The data-intensive survival guide This chapter presents an overview of all of the elements of the proposed data-intensive strategy. Sufficient detail is presented for readers to understand the principles and practice that we recommend. It should also provide a good preparation for readers who choose to sample later chapters. It introduces three professional viewpoints: domain experts, data-analysis experts, and data-intensive engineers. Success depends on a balanced approach that develops the capacity of all three groups. A data-intensive architecture provides a flexible framework for that balanced approach. This enables the three groups to build and exploit data-intensive processes that incrementally step from data to results. A language is introduced to describe these incremental data processes from all three points of view. The chapter introduces 'datascopes' as the productized data handling environments and 'intellectual ramps' as the 'on ramps' for the highways from data to knowledge.

Chapter 4: Data-intensive thinking with DISPEL This chapter engages the reader with technical issues and solutions, by working through a sequence of examples, building up from a sketch of a solution to a large-scale data challenge. It uses the DISPEL language extensively, introducing its concepts and constructs. It shows how DISPEL may help designers, data-analysts, and engineers develop solutions to the requirements emerging in any data-intensive application domain. The reader is taken through simple steps initially, this then builds to conceptually complex steps that are necessary to cope with the realities of real data providers, real data, real distributed systems, and long-running processes.

Part II: Data-intensive knowledge discovery Part II focuses on the needs of data-analysis experts. It illustrates the problem-solving strategies appropriate for a data-rich world, without delving into the details of underlying technologies. It should engage and inform data-analysis specialists, such as statisticians, data miners, image analysts, bio-informaticians or chemo-informaticians, and generate ideas pertinent to their application areas.

Chapter 5: Data-intensive analysis This chapter introduces a set of common problems that data-analysis experts often encounter, by means of a set of scenarios of increasing levels of complexity. The scenarios typify knowledge discovery challenges and the presented solutions provide practical methods; a starting point for readers addressing their own data challenges.

Chapter 6: Problem solving in data-intensive knowledge discovery On the basis of the previous scenarios, this chapter provides an overview of effective strategies in knowledge discovery, highlighting common problem-solving methods that apply in conventional contexts, and focusing on the similarities and differences of these methods.

Chapter 7: Data-intensive components and usage patterns This chapter provides a systematic review of the components that are commonly used in knowledge discovery tasks as well as common patterns of component composition. That is, it introduces the processing elements from which knowledge discovery solutions are built and common composition patterns for delivering trustworthy information. It reflects on how these components and patterns are evolving in a data-intensive context.

Chapter 8: Sharing and re-use in knowledge discovery This chapter introduces more advanced knowledge discovery problems, and shows how improved component and pattern descriptions facilitate re-use. This supports the assembly of libraries of high level components well-adapted to classes of knowledge discovery methods or application domains. The descriptions are made more powerful by introducing notations from the semantic Web.

Part III: Data-intensive engineering Part III is targeted at technical experts who will develop complex applications, new components, or data-intensive platforms. The techniques introduced may be applied very widely; for example, to any data-intensive distributed application, such as index generation, image processing, sequence comparison, text analysis, and sensor-stream monitoring. The challenges, methods, and implementation requirements are illustrated by making extensive use of DISPEL.

Chapter 9: Platforms for data-intensive analysis This chapter gives a reprise of data-intensive architectures, examines the business case for investing in them, and introduces the stages of data-intensive workflow enactment.

Chapter 10: Definition of the DISPEL language This chapter describes the novel aspects of the DISPEL language: its constructs, capabilities, and anticipated programming style.

Chapter 11: DISPEL development This chapter describes the tools and libraries that a DISPEL developer might expect to use. The tools include those needed during process definition, those required to organize enactment, and diagnostic aids for developers of applications and platforms.

Chapter 12: DISPEL enactment This chapter describes the four stages of DISPEL enactment. It is targeted at the data-intensive engineers who implement enactment services.

Part IV: Data-intensive application experience This part of the book is about applications that shaped the ideas behind the data-intensive architecture and methods. It provides a wealth of examples drawn from experience, describing in each case the aspects of data-intensive systems tested by the application, the DISPEL-based methods developed to meet the challenge, and the conclusions drawn from the prototype experiments.

Chapter 13: The application foundations of DISPEL The early development of DISPEL was influenced and assisted by research challenges from four very different data-intensive application domains. This chapter reviews these four domains in terms of their particular needs and requirements and how, as a suite, they provide an effective test of all key dimensions of a data-intensive system. It reviews the data-intensive strategy in terms of these applications and finds support for the approach.

Chapter 14: Analytical platform for customer relationship management This chapter demonstrates that the data-intensive methods and technology work well for traditional commercial knowledge discovery applications. Readers are introduced to the application domain through a scene-setting discussion, which assumes no prior knowledge, and are then taken through the process of analyzing customer data to predict behavior or preferences.

Chapter 15: Environmental risk management This chapter presents applications in the context of environmental risk management. The scenarios involve significant data-integration challenges as they take an increasing number of factors into account when managing the outflow from a reservoir to limit the effects downstream.

Chapter 16: Analyzing gene expression imaging data in developmental biology This chapter describes the application of data-intensive methods to the automatic identification and annotation of gene expression patterns in the mouse embryo. It shows how image processing and machine learning can be combined to annotate images and identify networks of gene functions.

Chapter 17: Data-intensive seismology: research horizons Seismology has moved from focusing on events to analyzing continuous streams of data obtained from thousands of seismometers. This is fundamental to understanding the inner structure and processes of the Earth and this chapter investigates the data-intensive architecture necessary to enable the analysis of large-scale distributed seismic waveforms.

Part V: Data-intensive beacons of success This part introduces a group of challenging, sophisticated data-intensive applications, which are starting to shape and promote a new generation of knowledge discovery technology. The chapters show that science, engineering, and society are fertile lands for data-intensive research. This part is targeted at novel application developers who like to include visionary aspects in their research.

Chapter 18: Data-intensive methods in astronomy Astronomy has been at the forefront of the digital revolution as it pioneered faster and more sensitive digital cameras, and established a new *modus operandi* for sharing and integrating data globally. These are yielding floods of data, and opening up new approaches to exploring the cosmos and testing the physical models that underpin it. This chapter describes two examples that exemplify the data-intensive science now underway.

Chapter 19: Interactive interpretation of environmental data A crucial step in any science is to explore the data available; this will often stimulate new insights and hypotheses. As the volumes of data grow and diverse formats are encountered, the effort of handling data inhibits exploration. This chapter shows how these inhibitory difficulties can be overcome, so that oceanographers and atmospheric scientists can easily select, vizualise and explore compositions of their data.

Chapter 20: Data-driven research in the humanities Researchers in the arts and humanities are using digitization to see new aspects of the many artifacts and phenomena they study. Digital resources allow statistical methods and computational matching to be employed, as well as the full panoply of text processing and collaborative annotation. In this chapter, researchers show their plans for a Europe-wide data infrastructure to facilitate this new research.

Chapter 21: Analysis of engineering and transport data Analysis of vibration data from aero-engines, turbines, and locomotives, ‘listening to the engine’, can reveal incipient problems and trigger appropriate remedial responses. To do this safely for large numbers of operational engines is a major data-intensive challenge. This chapter reports on ten years’ progress and its spin-offs into analyzing medical time series.

Chapter 22: Determining the patterns of bird species occurrence In this chapter, the ornithologists describe the challenge of estimating the populations of birds as they migrate and of inferring the factors affecting species numbers. It takes a great deal of sophisticated data analysis to extract and visualize the relevant information and much ingenuity to discover and use the data required from other disciplines.

Part VI: The data-intensive future This part presents a summary of the state of the industry and research, the observed trends and the current ‘hot-spots’ of data-intensive innovation. It provides a framework for reviewing the current activity and anticipated changes it will bring about. It offers a rich set of pointers to the literature and Web sites, built over the 15 months of the data-intensive research theme at the e-Science Institute. This should help readers select and find highly relevant further reading.

Chapter 23: Data-intensive trends This chapter summarizes the learning about data-intensive methods and their potential power. It then analyzes some of the categories of data-intensive challenge and assesses how they will develop.

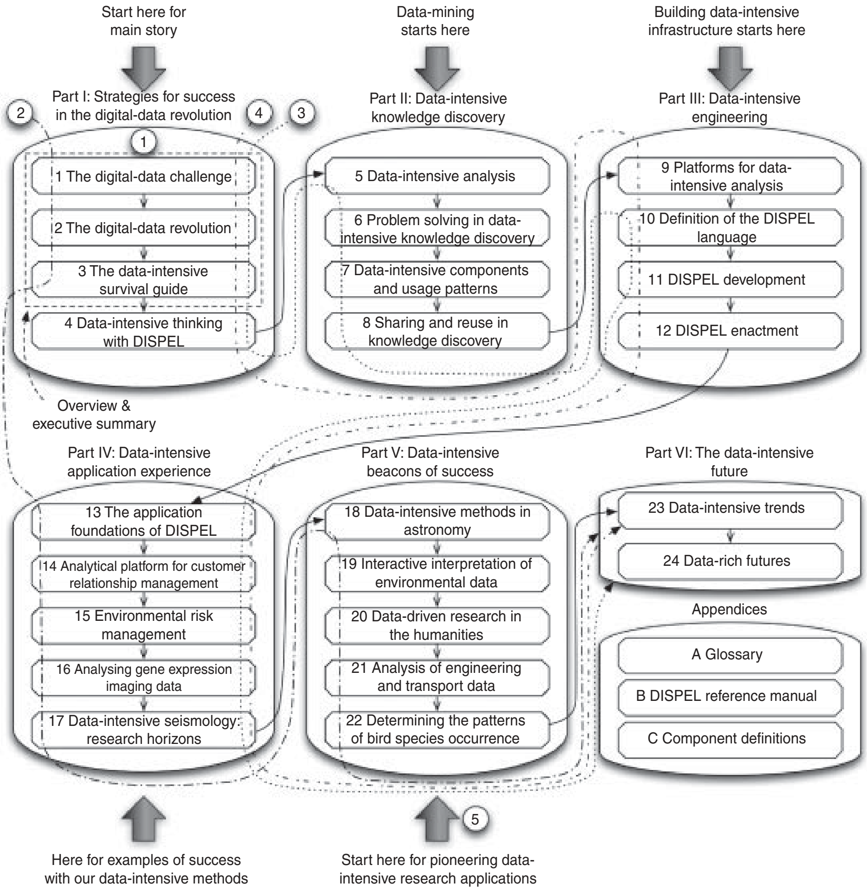
Chapter 24: Data-rich futures This chapter dangerously attempts some 'crystal gazing'. It looks first at technological factors and current research that should be observed by those who wish to further develop data-intensive strategies. It explores some of the economic factors that will shape a data-rich future and concludes with a view on the social issues that will emerge. We call on those with influence to strive for professional standards in handling data, from their collection to the actions based on their evidence.

How to Read this Book

There are two pieces of information that we wish to offer readers in order to help them get the most benefit from this book: the choice of routes through the book and conventions used in the book. We consider each of these in turn.

1. The primary story line of the book develops in the conventional way by reading the parts and the chapters in order, up to the end of Part IV. It begins with scene setting to establish a conceptual framework. It then addresses the methods and engineering needed to put the ideas into practice before successful applications in a wide range of domains. Many readers will want a selected-reading route that matches their needs; the following map and suggested routes will encourage them to plan their reading.
2. Research strategists, scientists and managers, who may be less interested in technical detail should follow Chapters 1 to 3→Part IV→Part V→Part VI.
3. Data-analysis experts should try Part I→Part II→Chapter 10→Chapter 11→Part IV→Part VI.
4. Those building data-intensive systems should read Part I→Part III→Part IV→Part VI.
5. Domain experts, who are looking for ideas that may pay off with their specialist data could read the applications first, Part V→Part IV→Chapters 1 to 3. We hope there are enough signposts that you can adjust your route easily to follow new interests.

The conventions concern the structure and the representation of certain items. As the map shows, there are six parts; each one begins with a preamble intended to orient readers who start in that part, and to relate that part to the other parts of the book. Appendix A provides a glossary, where we hope you will find useful definitions of terms used frequently in the book. Terms used infrequently may be traced via the index. Each chapter concludes with the references cited in that chapter. References to Web sites are shown as URLs in line with the text if they are short and in footnotes if they are long. The `http://` prefix is omitted.


Start here for main story
Data-mining starts here
Building data-intensive infrastructure starts here
Part I: Strategies for success in the digital-data revolution
Part II: Data-intensive knowledge discovery
Part III: Data-intensive engineering
2
4
3
1
1 The digital-data challenge
2 The digital-data revolution
3 The data-intensive survival guide
4 Data-intensive thinking with DISPEL
5 Data-intensive analysis
6 Problem solving in data-intensive knowledge discovery
7 Data-intensive components and usage patterns
8 Sharing and reuse in knowledge discovery
9 Platforms for data-intensive analysis
10 Definition of the DISPEL language
11 DISPEL development
12 DISPEL enactment
Overview & executive summary
Part IV: Data-intensive application experience
Part V: Data-intensive beacons of success
Part VI: The data-intensive future
13 The application foundations of DISPEL
14 Analytical platform for customer relationship management
15 Environmental risk management
16 Analysing gene expression imaging data
17 Data-intensive seismology: research horizons
18 Data-intensive methods in astronomy
19 Interactive interpretation of environmental data
20 Data-driven research in the humanities
21 Analysis of engineering and transport data
22 Determining the patterns of bird species occurrence
23 Data-intensive trends
24 Data-rich futures
Appendices
A Glossary
B DISPEL reference manual
C Component definitions
5
Here for examples of success with our data-intensive methods
Start here for pioneering data-intensive research applications

Readers' map of the book.

There are many programming examples, mostly represented in DISPEL. A consistent highlighting convention has been used for these, which we believe helps legibility. These are also available on the Web site (see following text), so that you can view them in your favorite editor. There are often corresponding diagrams, showing the data-flow graph that the program would generate. These pick up the same color conventions. The language DISPEL is introduced in Chapters 3, 4 and 10; the language definition is in Appendix B. The components that are used in the examples are provided from standard libraries of components; these are described in Appendix C.

Web Site with Additional Material

A Web site at www.dispel.lang.org holds material intended to help readers of this book; these include pages covering the following topics.

- An overview and table of contents of the book.
- A collection of success stories.
- Teaching material including presentations to be used in conjunction with chapters in the book.
- Copies of the program examples from the book, so that they may be used by readers.
- The libraries of components used in the book, with corresponding descriptions in a local registry.
- The DISPEL reference manual.
- Links to other sites including many of those referenced from the book, the associated open-source project and new developments.

This Web site will be updated and contributions from others will be welcome.

Acknowledgments

Many people have contributed to this book through discussions and visits over the past 6 years; we thank them all and limit the explicit list to those who worked directly on the book.

Funders The primary funder of the editorial process was the European Commission's Framework Program 7's support for the ADMIRE project (www.admire-project.eu), grant number 215024. The tour of the USA by Atkinson and De Roure, to study their use of data (bit.ly/c0G2rn), and Atkinson's work on the book was funded by the UK's Engineering and Physical Sciences Research Council (EPSRC) Senior Research Fellowship to undertake the UK e-Science Envoy role, grant EP/D079829/1. The initial workshop on data-intensive research (bit.ly/bQpu5h), where many of the commitments for contributions to the book were made, and the one-year data-intensive research theme (bit.ly/cygimA), were funded by the e-Science Institute, EPSRC grant EP/D056314/1. The work to further develop this book, the technology and the methodology described, is partially funded in the University of Edinburgh by the EPSRC NeSC Research Platform (research.nesc.ac.uk) grant EP/F057695/1, and in Universidad Politècnica de Madrid by the Spanish Ministry of Science and Technology under grant TIN2010-17060. Baxter would also like to acknowledge the direct support of EPCC at the University of Edinburgh, and support from the UK's Software Sustainability Institute under EPSRC grant number EP/H043160/1.

Helpers A big thank you to Jo Newman, who was a continuous support to the editors of the book, impeccably arranging many meetings and tele-communications, as well as thorough proof reading, checking copyright information, and communicating with all of the contributing authors. Kathy Humphry read nearly every chapter and gave good advice on each one of them.

In addition to being contributing authors, we must thank Ivan Janciak, in the University of Vienna, who made life much easier for all of us by setting up convenient macros for building the book; Paul Martin and Gagarine Yaikhom, of the Universities of Edinburgh and Cardiff respectively, for setting up the system for typesetting DISPEL highlighting its structure; and Amrey Krause and Chee Sun Liew of the University of Edinburgh, who set up the system for validating DISPEL text used in the book. Chee Sun Liew also did a great deal of LaTeX wrangling to shape the book into its final form. Ivan Janciak, Alexander Wöhrer and Marek Lenart of the University of Vienna reviewed several book chapters and helped improve the quality of figures and the formatting of the text.

We would also like to thank our colleagues Martin Šeleng and Peter Krammer from the Institute of Informatics of the Slovak Academy of Sciences for their excellent work on the data mining scenarios of the environmental risk management application, and our dear friends at the Slovak Hydro-Meteorological Institute and the Slovak Water Enterprise for their invaluable help with designing the pilot scenarios and providing real input data for them. A big thank you must go to the EPCC software engineering team for making large swathes of the prototype data-intensive platform work: Ally Hume, Malcolm Illingworth, Amrey Krause, Adrian Mouat and David Scott, with a special mention for our integration, test and build-meister Radek Ostrowski.

We heartily thank all of the open-source developers on whose work we built; all those who helped and confirm that all the remaining errors are the responsibility of the editors, led by myself.

Malcolm Atkinson

Edinburgh, UK
April 2012