
THE DATA BONANZA

Improving Knowledge Discovery for Science, Engineering and Business

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

Peter Brezany
Research Group for Scientific Computing, University of Vienna, Austria

Oscar Corcho
Facultad de Informática, Universidad Politécnica de Madrid, Spain

Jano van Hemert
Optos PLC, Dunfermline, UK

David Snelling
Fujitsu Laboratories Europe Limited, UK

 **WILEY-
INTERSCIENCE**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright ©2012 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

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 or extended by sales representatives or written sales materials. The advice and strategies contained herein 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:

The DATA Bonanza: Improving Knowledge Discovery for Science, Engineering and Business

Malcolm Atkinson *et al.*

“Wiley-Interscience.”

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

Preface	xvii
PART I STRATEGIES FOR SUCCESS IN THE DIGITAL-DATA REVOLUTION	
1 The digital-data challenge	7
1.1 The digital revolution	7
1.2 Changing how we think and behave	8
1.3 Moving adroitly in this fast changing field	9
1.4 Digital-data challenges exist everywhere	10
1.5 Changing how we work	11
1.6 Divide and conquer offers the solution	12
1.7 Engineering data-to-knowledge highways	14
2 The digital-data revolution	17
2.1 Data, information and knowledge	18
2.2 Increasing volumes and diversity of data	21
2.3 Changing the ways we work with data	30
3 The data-intensive survival guide	39
3.1 Introduction: challenges and strategy	40
3.2 Three categories of expert	41
	vii

3.3	The data-intensive architecture	43
3.4	An operational data-intensive system	44
3.5	Introducing DISPEL	46
3.6	A simple DISPEL example	47
3.7	Supporting data-intensive experts	48
3.8	DISPEL in the context of contemporary systems	50
3.9	Datascope	52
3.10	Ramps for incremental engagement	55
3.11	Readers' guide to the rest of this book	57
4	Data-intensive thinking with DISPEL	63
4.1	Processing elements	64
4.2	Connections	66
4.3	Data streams and structure	67
4.4	Functions	68
4.5	The three-level type system	73
4.6	Registry, libraries and descriptions	82
4.7	Achieving data-intensive performance	86
4.8	Reliability and control	107
4.9	The data-to-knowledge highway	114
	PART II DATA-INTENSIVE KNOWLEDGE DISCOVERY	
5	Data-intensive analysis	127
5.1	Knowledge discovery in Telco Inc.	128
5.2	Understanding customers to prevent churn	129
5.3	Preventing churn across multiple companies	134
5.4	Understanding customers by combining heterogeneous public and private data	138
5.5	Conclusions	144
6	Problem solving in data-intensive knowledge discovery	147
6.1	The conventional lifecycle of knowledge discovery	148
6.2	Knowledge discovery over heterogeneous data sources	155
6.3	Knowledge discovery from private and public, structured and non-structured data	157
6.4	Conclusions	161
7	Data-intensive components and usage patterns	163
7.1	Data source access and transformation components	164
7.2	Data integration components	170
7.3	Data preparation and processing components	171

7.4	Data mining components	171
7.5	Visualisation and Knowledge Delivery components	173
8	Sharing and reuse in knowledge discovery	177
8.1	Strategies for sharing and re-use	178
8.2	Data-analysis ontologies for data-analysis experts	181
8.3	Generic ontologies for metadata generation	184
8.4	Domain ontologies for domain experts	185
8.5	Conclusions	186
PART III DATA-INTENSIVE ENGINEERING		
9	Platforms for data-intensive analysis	193
9.1	The hourglass reprise	194
9.2	The motivation for a platform	195
9.3	Realisation	196
10	Definition of the DISPEL language	199
10.1	A simple example	200
10.2	Processing elements	201
10.3	Data streams	208
10.4	Type system	211
10.5	Registration	217
10.6	Packaging	219
10.7	Workflow submission	219
10.8	Examples of DISPEL	221
10.9	Summary	229
11	DISPEL development	231
11.1	The development landscape	231
11.2	Data-intensive workbenches	233
11.3	Data-intensive component libraries	240
11.4	Summary	241
12	DISPEL enactment	243
12.1	Overview of DISPEL enactment	243
12.2	DISPEL language processing	245
12.3	DISPEL optimisation	247
12.4	DISPEL deployment	257
12.5	DISPEL execution and control	259

PART IV DATA-INTENSIVE APPLICATION EXPERIENCE

13	The application foundations of DISPEL	267
13.1	Characteristics of data-intensive applications	267
13.2	Evaluating application performance	270
13.3	Reviewing the data-intensive strategy	272
14	Analytical platform for customer relationship management	275
14.1	Data analysis in the telecoms business	276
14.2	Analytical customer relationship management	276
14.3	Scenario 1: churn prediction	278
14.4	Scenario 2: cross-selling	281
14.5	Exploiting the models and rules	283
14.6	Summary: lessons learned	286
15	Environmental risk management	289
15.1	Environmental modelling	290
15.2	Cascading simulation models	291
15.3	Environmental data sources and their management	293
15.4	Scenario 1: ORAVA	297
15.5	Scenario 2: RADAR	301
15.6	Scenario 3: SVP	306
15.7	New technologies for environmental data mining	310
15.8	Summary: lessons learned	311
16	Analysing gene expression imaging data	315
16.1	Understanding biological function	316
16.2	Gene image annotation	317
16.3	Automated annotation of gene expression images	319
16.4	Exploitation and future work	327
16.5	Summary	331
17	Data-intensive seismology: research horizons	337
17.1	Introduction	338
17.2	Seismic ambient noise processing	339
17.3	Solution implementation	341
17.4	Evaluation	352
17.5	Further work	355
17.6	Conclusions	356
PART V DATA-INTENSIVE BEACONS OF SUCCESS		
18	Data-intensive methods in astronomy	365

18.1	Introduction	366
18.2	The virtual observatory	366
18.3	Data-intensive photometric classification of quasars	367
18.4	Probing the dark universe with weak gravitational lensing	371
18.5	Future research issues	374
18.6	Conclusions	375
19	Interactive interpretation of environmental data	379
19.1	Introduction	380
19.2	The current state of the art	381
19.3	The technical landscape	384
19.4	Interactive visualisation	386
19.5	From visualisation to inter-comparison	389
19.6	Future development: the environmental cloud	392
19.7	Conclusions	394
20	Data-driven research in the humanities	401
20.1	Introduction	401
20.2	The tradition of digital humanities	404
20.3	Humanities research data	406
20.4	Use case	409
20.5	Conclusion and future development	412
21	Analysis of engineering and transport data	415
21.1	Introduction	415
21.2	Applications and challenges	416
21.3	The methods used	418
21.4	Future developments	422
21.5	Conclusions	423
22	Determining the patterns of bird species occurrence	425
22.1	Introduction	425
22.2	Data discovery, access and synthesis	427
22.3	Model development	431
22.4	Managing computational requirements	432
22.5	Exploring and visualizing model results	433
22.6	Analysis results	436
22.7	Conclusion	436
PART VI THE DATA-INTENSIVE FUTURE		
23	Review of data-intensive trends	443

23.1	Reprise	443
23.2	Data-intensive applications	450
23.3	Data infrastructure, economy, society and professionalism	457

Appendices **464**

Appendix A:	Glossary	465
-------------	----------	-----

Appendix B:	DISPEL reference manual	473
-------------	-------------------------	-----

B.1	Workflow model	474
-----	----------------	-----

B.2	Script composition	477
-----	--------------------	-----

B.3	Type System	481
-----	-------------	-----

B.4	Statements	489
-----	------------	-----

Appendix C:	Component definitions	497
-------------	-----------------------	-----

Index		527
-------	--	-----