

**Wiley Series in Operations Research
and Management Science**



ANALYTICS BODY OF KNOWLEDGE

Edited by **JAMES J. COCHRAN**

WILEY

INFORMS Analytics Body of Knowledge

Wiley Essentials in
OPERATIONAL RESEARCH AND MANAGEMENT SCIENCE

INFORMS Analytics Body of Knowledge

Edited by James J. Cochran

WILEY

This edition first published 2019
© 2019 John Wiley and Sons, Inc.

All rights reserved. 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 or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of James J. Cochran to be identified as the author of this work has been asserted in accordance with law.

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

ISBN: 9781119483212

Set in 10/12 pt WarnockPro-Regular by Thomson Digital, Noida, India

Contents

Preface *xv*

List of Contributors *xix*

1	Introduction to Analytics	1
	<i>Philip T. Keenan, Jonathan H. Owen, and Kathryn Schumacher</i>	
1.1	Introduction	1
1.2	Conceptual Framework	3
1.2.1	Data-Centric Analytics	3
1.2.2	Decision-Centric Analytics	4
1.2.3	Combining Data- and Decision-Centric Approaches	5
1.3	Categories of Analytics	6
1.3.1	Descriptive Analytics	7
	Data Modeling	7
	Reporting	10
	Visualization	10
	Software	10
1.3.2	Predictive Analytics	10
	Data Mining and Pattern Recognition	11
	Predictive Modeling, Simulation, and Forecasting	11
	Leveraging Expertise	12
1.3.3	Prescriptive Analytics	14
1.4	Analytics Within Organizations	16
1.4.1	Projects	17
1.4.2	Communicating Analytics	21
1.4.3	Organizational Capability	21
1.5	Ethical Implications	23
1.6	The Changing World of Analytics	25
1.7	Conclusion	28
	References	28

2	Getting Started with Analytics	31
	<i>Karl G. Kempf</i>	
2.1	Introduction	31
2.2	Five Manageable Tasks	32
2.2.1	Task 1: Selecting the Target Problem	33
2.2.2	Task 2: Assemble the Team	34
	Executive Sponsor	35
	Project Manager	35
	Domain Expert	35
	IT Expert	35
	Data Scientist	36
	Stakeholders	36
2.2.3	Task 3: Prepare the Data	36
2.2.4	Task 4: Selecting Analytics Tools	39
	Analytical Specificity or Breadth	39
	Access to Data	40
	Execution Performance	40
	Visualization Capability	40
	Data Scientist Skillset	40
	Vendor Pricing	41
	Team Budget	41
	Sharing and Collaboration	41
2.2.5	Task 5: Execute	42
2.3	Real Examples	43
	Case 1: Sensor Data and High-Velocity Analytics to Save Operating Costs	43
	Case 2: Social Media and High-Velocity Analytics for Quick Response to Customers	44
	Case 3: Sensor Data and High-Velocity Analytics to Save Maintenance Costs	44
	Case 4: Using Old Data and Analytics to Detect New Fraudulent Claims	45
	Case 5: Using Old and New Data Plus Analytics to Decrease Crime	45
	Case 6: Collecting the Data and Applying the Analytics Is the Business	45
	References	46
	Further Reading: Papers	47
	Further Reading: Books	48
3	The Analytics Team	49
	<i>Thomas H. Davenport</i>	
3.1	Introduction	49

3.2	Skills Necessary for Analytics	50
3.2.1	More Advanced or Recent Analytical and Data Science Skills	51
3.2.2	The Larger Team	53
3.3	Managing Analytical Talent	57
3.3.1	Developing Talent	58
3.3.2	Working with the HR Organization	59
3.4	Organizing Analytics	61
3.4.1	Goals of a Particular Analytics Organization	62
3.4.2	Basic Models for Organizing Analytics	63
3.4.3	Coordination Approaches	65
	Program Management Office	66
	Federation	67
	Community	67
	Matrix	67
	Rotation	67
	Assigned Customers	67
	What Model Fits Your Business?	68
3.4.4	Organizational Structures for Specific Analytics Strategies and Scenarios	70
3.4.5	Analytical Leadership and the Chief Analytics Officer	70
3.5	To Where Should Analytical Functions Report?	72
	Information Technology	72
	Strategy	72
	Shared Services	72
	Finance	73
	Marketing or Other Specific Function	73
	Product Development	73
3.5.1	Building an Analytical Ecosystem	73
3.5.2	Developing the Analytical Organization over Time	74
	References	75
4	The Data	77
	<i>Brian T. Downs</i>	
4.1	Introduction	77
4.2	Data Collection	77
4.2.1	Data Types	77
4.2.2	Data Discovery	80
4.3	Data Preparation	86
4.4	Data Modeling	93
4.4.1	Relational Databases	93
4.4.2	Nonrelational Databases	95
4.5	Data Management	97

5	Solution Methodologies	99
	<i>Mary E. Helander</i>	
5.1	Introduction	99
5.1.1	What Exactly Do We Mean by “Solution,” “Problem,” and “Methodology?”	99
5.1.2	It’s All About the Problem	101
5.1.3	Solutions versus Products	101
5.1.4	How This Chapter Is Organized	103
5.1.5	The “Descriptive–Predictive–Prescriptive” Analytics Paradigm	105
5.1.6	The Goals of This Chapter	105
5.2	Macro-Solution Methodologies for the Analytics Practitioner	106
5.2.1	The Scientific Research Methodology	106
5.2.2	The Operations Research Project Methodology	109
5.2.3	The Cross-Industry Standard Process for Data Mining (CRISP-DM) Methodology	112
5.2.4	Software Engineering-Related Solution Methodologies	114
5.2.5	Summary of Macro-Methodologies	114
5.3	Micro-Solution Methodologies for the Analytics Practitioner	116
5.3.1	Micro-Solution Methodology Preliminaries	116
5.3.2	Micro-Solution Methodology Description Framework	117
5.3.3	Group I: Micro-Solution Methodologies for Exploration and Discovery	119
	Group I: Problems of Interest	119
	Group I: Relevant Models	119
	Group I: Data Considerations	120
	Group I: Solution Techniques	120
	Group I: Relationship to Macro-Methodologies	126
	Group I: Takeaways	126
5.3.4	Group II: Micro-Solution Methodologies Using Models Where Techniques to Find Solutions Are Independent of Data	127
	Group II: Problems of Interest	127
	Group II: Relevant Models	127
	Group II: Data Considerations	128
	Group II: Solution Techniques	128
	Group II: Relationship to Macro-Methodologies	135
	Group II: Takeaways	137
5.3.5	Group III: Micro-Solution Methodologies Using Models Where Techniques to Find Solutions Are Dependent on Data	137
	Group III: Problems of Interest	137
	Group III: Relevant Models	138
	Group III: Data Considerations	138
	Group III: Solution Techniques	139

	Group III: Relationship to Macro-Methodologies	140
	Group III: Takeaways	141
5.3.6	Micro-Methodology Summary	141
5.4	General Methodology-Related Considerations	142
5.4.1	Planning an Analytics Project	142
5.4.2	Software and Tool Selection	142
5.4.3	Visualization	143
5.4.4	Fields with Related Methodologies	144
5.5	Summary and Conclusions	144
5.5.1	“Ding Dong, the Scientific Method Is Dead!”	145
5.5.2	“Methodology Cramps My Analytics Style”	145
5.5.3	“There Is Only One Way to Solve This”	146
5.5.4	Perceived Success Is More Important Than the Right Answer	148
5.6	Acknowledgments	149
	References	149
6	Modeling	155
	<i>Gerald G. Brown</i>	
6.1	Introduction	155
6.2	When Are Models Appropriate	155
6.2.1	What Is the Problem with This System?	159
6.2.2	Is This Problem Important?	159
6.2.3	How Will This Problem Be Solved Without a New Model?	159
6.2.4	What Modeling Technique Will Be Used?	159
6.2.5	How Will We Know When We Have Succeeded?	160
	Who Are the System Operator Stakeholders?	160
6.3	Types of Models	161
6.3.1	Descriptive Models	161
6.3.2	Predictive Models	161
6.3.3	Prescriptive Models	161
6.4	Models Can Also Be Characterized by Whether They Are Deterministic or Stochastic (Random)	161
6.5	Counting	162
6.6	Probability	163
6.7	Probability Perspectives and Subject Matter Experts	165
6.8	Subject Matter Experts	165
6.9	Statistics	166
6.9.1	A Random Sample	166
6.9.2	Descriptive Statistics	166
6.9.3	Parameter Estimation with a Confidence Interval	166
6.9.4	Regression	167
6.10	Inferential Statistics	169
6.11	A Stochastic Process	170

6.12	Digital Simulation	173
6.12.1	Static versus Dynamic Simulations	174
6.13	Mathematical Optimization	174
6.14	Measurement Units	175
6.15	Critical Path Method	176
6.16	Portfolio Optimization Case Study Solved By a Variety of Methods	178
6.16.1	Linear Program	178
6.16.2	Heuristic	179
6.16.3	Assessing Our Progress	179
6.16.4	Relaxations and Bounds	179
6.16.5	Are We Finished Yet?	180
6.17	Game Theory	181
6.18	Decision Theory	184
6.19	Susceptible, Exposed, Infected, Recovered (SEIR) Epidemiology	187
6.20	Search Theory	189
6.21	Lanchester Models of Warfare	189
6.22	Hughes' Salvo Model of Combat	192
6.23	Single-Use Models	193
6.24	The Principle of Optimality and Dynamic Programming	195
6.25	Stack-Based Enumeration	197
6.25.1	Data Structures	197
6.25.2	Discussion	199
6.25.3	Generating Permutations and Combinations	199
6.26	Traveling Salesman Problem: Another Case Study in Alternate Solution Methods	200
6.27	Model Documentation, Management, and Performance	206
6.27.1	Model Formulation	206
6.27.2	Choice of Implementation Language	207
6.27.3	Supervised versus Automated Models	207
6.27.4	Model Fidelity	208
6.27.5	Sensitivity Analysis	210
6.27.6	With Different Methods	211
6.27.7	With Different Variables	212
6.27.8	Stability	213
6.27.9	Reliability	213
6.27.10	Scalability	213
6.27.11	Extensibility	214
6.28	Rules for Data Use	215
6.28.1	Proprietary Data	215
6.28.2	Licensed Data	215
6.28.3	Personally Identifiable Information	216
6.28.4	Protected Critical Infrastructure Information System (PCIIMS)	216

6.28.5	Institutional Review Board (IRB)	216
6.28.6	Department of Defense and Department of Energy Classification	216
6.28.7	Law Enforcement Data	216
6.28.8	Copyright and Trademark	216
6.28.9	Paraphrased and Plagiarized	217
6.28.10	Displays of Model Outputs	217
6.28.11	Data Integrity	217
6.28.12	Multiple Data Evolutions	217
6.29	Data Interpolation and Extrapolation	217
6.30	Model Verification and Validation	218
6.30.1	Verifying	219
6.30.2	Validating	219
6.30.3	Comparing Models	219
6.30.4	Sample Data	220
6.30.5	Data Diagnostics	220
6.30.6	Data Vintage and Provenance	220
6.31	Communicate with Stakeholders	220
6.31.1	Training	221
6.31.2	Report Writers	221
6.31.3	Standard Form Model Statement	222
6.31.4	Persistence and Monotonicity: Examples of Realistic Model Restrictions	223
6.31.5	Model Solutions Require a Lot of Polish and Refinement Before They Can Directly Influence Policy	224
6.31.6	Model Obsolescence and Model-Advised Thumb Rules	226
6.32	Software	227
6.33	Where to Go from Here	228
6.34	Acknowledgments	228
	References	229

7 Machine Learning 231

Samuel H. Huddleston and Gerald G. Brown

7.1	Introduction	231
7.2	Supervised, Unsupervised, and Reinforcement Learning	232
7.3	Model Development, Selection, and Deployment for Supervised Learning	235
7.3.1	Goals and Guiding Principles in Machine Learning	235
7.3.2	Algorithmic Modeling Overview	236
7.3.3	Data Acquisition and Cleaning	236
7.3.4	Feature Engineering	237
7.3.5	Modeling Overview	238

7.3.6	Model Fitting (Training) and Feature Selection	240
7.3.7	Model (Algorithm) Selection	241
7.3.8	Model Performance Assessment	242
7.3.9	Model Implementation	242
7.4	Model Fitting, Model Error, and the Bias-Variance Trade-Off	243
7.4.1	Components of (Regression) Model Error	243
7.4.2	Model Fitting: Balancing Bias and Variance	245
7.5	Predictive Performance Evaluation	247
7.5.1	Regression Performance Evaluation	248
7.5.2	Classification Performance Evaluation	249
7.5.3	Performance Evaluation for Time-Dependent Data	253
7.6	An Overview of Supervised Learning Algorithms	254
7.6.1	k-Nearest Neighbors (KNN)	255
7.6.2	Extensions to Regression	256
7.6.3	Classification and Regression Trees	257
7.6.4	Time Series Forecasting	259
7.6.5	Support Vector Machines	261
7.6.6	Artificial Neural Networks	262
7.6.7	Ensemble Methods	265
7.7	Unsupervised Learning Algorithms	267
7.7.1	Kernel Density Estimation	267
7.7.2	Association Rule Mining	268
7.7.3	Clustering Methods	269
7.7.4	Principal Components Analysis (PCA)	270
7.7.5	Bag-of-Words and Vector Space Models	271
7.8	Conclusion	272
7.9	Acknowledgments	272
	References	273
8	Deployment and Life Cycle Management	275
	<i>Arnie Greenland</i>	
8.1	Introduction	275
8.2	The Analytics Methodology: Understanding the Critical Steps in Deployment and Life Cycle Management	276
8.2.1	CRISP-DM Phase 1: Business Understanding	278
8.2.2	JTA Domain I, Task 1: Obtain or Receive Problem Statement and Usability	278
8.2.3	JTA Domain I, Task 2: Identify Stakeholders	279
8.2.4	JTA Domain I, Task 3: Determine if the Problem Is Amenable to an Analytics Solution	281
8.2.5	JTA Domain I, Task 4: Refine the Problem Statement and Delineate Constraints	281

8.2.6	JTA Domain I, Task 5: Define an Initial Set of Business Benefits	281
8.2.7	JTA Domain I, Task 6: Obtain Stakeholder Agreement on the Business Statement	282
8.2.8	JTA Domain II, Task 1: Reformulate the Problem Statement as an Analytics Problem	283
8.2.9	JTA Domain II, Task 2: Develop a Proposed Set of Drivers and Relationships to Outputs	285
8.2.10	JTA Domain II, Task 3: State the Set of Assumptions Related to the Problem	286
8.2.11	JTA Domain II, Task 4: Define the Key Metrics of Success	287
8.2.12	JTA Domain II, Task 5: Obtain Stakeholder Agreement	287
8.2.13	CRISP-DM Phases 2 and 3: Data Understanding and Data Preparation	288
8.2.14	JTA Domain III, Task 1: Identify and Prioritize Data Needs and Sources	290
8.2.15	JTA Domain III, Task 2: Acquire Data	290
8.2.16	JTA Domain III, Task 3: Harmonize, Rescale, Clean, and Share Data	291
8.2.17	JTA Domain III, Task 4: Identify Relationships in the Data	292
8.2.18	JTA Domain III, Task 5: Document and Report Finding	293
8.2.19	JTA Domain III, Task 6: Refine the Business and Analytics Problem Statements	293
8.2.20	CRISP-DM Phase 4: Modeling	293
8.2.21	CRISP-DM Phase 5: Evaluation	294
8.2.22	CRISP-DM Phase 6: Deployment	297
8.2.23	Deployment of the Analytics Model (Up to Delivery)	298
8.2.24	Post-deployment Activities (Domain VI: Model Life Cycle Management)	301
8.3	Overarching Issues of Life Cycle Management	303
8.3.1	Documentation	303
8.3.2	Communication	305
8.3.3	Testing	307
8.3.4	Metrics	308

9 The Blossoming Analytics Talent Pool: An Overview of the Analytics Ecosystem 311

Ramesh Sharda and Pankush Kalgotra

9.1	Introduction	311
9.2	Analytics Industry Ecosystem	312
9.2.1	Data Generation Infrastructure Providers	314
9.2.2	Data Management Infrastructure Providers	315
9.2.3	Data Warehouse Providers	316

9.2.4	Middleware Providers	316
9.2.5	Data Service Providers	316
9.2.6	Analytics-Focused Software Developers	317
	Reporting/Descriptive Analytics	317
	Predictive Analytics	318
	Prescriptive Analytics	318
9.2.7	Application Developers: Industry-Specific or General	319
9.2.8	Analytics Industry Analysts and Influencers	321
9.2.9	Academic Institutions and Certification Agencies	322
9.2.10	Regulators and Policy Makers	323
9.2.11	Analytics User Organizations	323
9.3	Conclusions	325
	References	326
Appendix: Writing and Teaching Analytics with Cases		327
<i>James J. Cochran</i>		
Index		355

Preface

A body of knowledge (BOK) is a comprehensive compilation of the core concepts and skills with which a professional in a specific discipline should be familiar. BOKs are generally produced and maintained by members of an academic society or professional association, and a BOK serves as the means by which the academic society or professional association communicates its vision, both internally and externally.

The broad objective of this BOK, entitled *INFORMS Analytics Body of Knowledge (ABOK)*, is to provide those interested in the development and application of the tools of analytics with an understanding of what analytics is and how analytics can be used to solve complex problems, make better decisions, and formulate more effective strategies. *ABOK* is produced by the Institute for Operations Research and the Management Sciences INFORMS¹ and represents the perspectives of some of the organization's most respected members on a wide variety of analytics-related topics.

We use INFORMS' definition of analytics—*the scientific process of transforming data into insight for making better decisions*—as the foundation for this book. But each chapter also reflects the unique insights and experiences of the chapter's author(s). This is intentional; analytics is a nascent, diverse, and complex discipline (or perhaps a collection of disciplines) that is defined somewhat differently by various practitioners and organizations. The various perspectives within this book will provide the reader with a better understanding of this dynamic field.

This book is a valuable resource for professionals in business and industry who are looking for ways to fully and effectively integrate analytics into their organizations' problem-solving, decision-making, and strategic planning.

¹ INFORMS (www.informs.org) is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

Instructors who are developing or revising/modernizing analytics courses and programs will also find *ABOK*'s chapters illuminating.

ABOK's chapters are written by colleagues recognized for their expertise in various areas of analytics (Philip T. Keenan, Jonathan H. Owen, and Kathryn Schumacher of General Motors; Karl Kempf of Intel Corporation; Thomas H. Davenport of Babson College; Brian Downs of Accenture LLC; Mary E. Helander of the IBM T.J. Watson Research Center; Gerald G. Brown and Samuel H. Huddleston of the Naval Postgraduate School; Arnie Greenland of the University of Maryland; Ramesh Sharda of Oklahoma State University and Pankush Kalgotra of Clark University). We have solicited input from colleagues in industry, government, and academia, and each chapter has been peer-reviewed by respected colleagues in analytics in order to ensure that *ABOK* will be a useful practical resource.

An appendix on writing and teaching analytics with cases is also included in this book. This appendix is included to support the development of courses in analytics and foster the case approach in analytics courses. It is also intended to encourage colleagues in business and industry to work with academicians to develop and publish analytics cases for use in analytics courses as a means to improve student understanding and appreciation of the importance and relevance of this discipline. Although *ABOK* is not intended to be a comprehensive source for preparation for INFORMS Certified Analytics Professional (CAP®) and Associate Certified Analytics Professional (aCAP™) examinations, its contents will be very helpful to those preparing for these examinations.

Each chapter and the appendix also feature relevant portions of interviews with other well-respected practitioners and instructors of analytics. These interviews were conducted by the INFORMS' *Analytics Body of Knowledge Committee* and provided by Eric Stephens of the Vanderbilt University Medical Center; Alan Taber of Lockheed Martin Missiles and Fire Control; Jeff Camm of Wake Forest University; Katya Scheinberg of Lehigh University; Harrison Schramm of the United States Navy (retired); Greta Roberts of Talent Analytics; Susan Martonosi of Harvey Mudd College; Russell Walker of Northwestern University's Kellogg School of Management; Robert Clark of RTI International; Cole Smith of Clemson University; and Matt Drake of Duquesne University.

Major undertakings, such as a body of knowledge, can only succeed if all members of a large and talented team work toward a common objective, and *ABOK* is certainly no exception to this rule. Several colleagues from industry and academia provided detailed reviews of the chapters. Tasha Inniss of INFORMS; Cole Smith of Clemson University; Manoj Chari of SAS; J. Antonio Carbajal of Turner Broadcasting System, Inc.; Ashley Cowall of Booz Allen Hamilton; Graciela Chadwick of Chick-fil-A; Nick Wzientek of Rocky Mountain Resources; Linda Schumacher of ABB, Inc.; Alan Taber of Lockheed Martin Missiles and Fire Control; Susan Martonosi of Harvey Mudd College; Sean

MacDermant of International Paper; and Matt Drake of Duquesne University each generously reviewed chapters and provided valuable input.

INFORMS' *Analytics Body of Knowledge* Committee, which is chaired by Terry Harrison (Penn State University) and includes Michael Rappa (North Carolina State University), Jim Williams (FICO), Alan Briggs (Elder Research), Eric Stephens (Vanderbilt University Medical Center), Alan Taber (Lockheed Martin Missiles and Fire Control), Jeanne Harris (Columbia University), and Layne Morrison (IBM), has provided valuable input. Lisa Greene and Bob Clark of RTI, International were instrumental in executing the interviews and advised on several issues.

Other members of INFORMS who provided advice and feedback include Donald Baillie (Anzac Finance Solutions), James Taylor (Decision Management Solutions), Irv Lustig (Princeton Consultants), Harrison Schramm (retired Naval officer), Thomas Reid (Booz Allen Hamilton), Charley Tichenor (Marymount University), Selene Crosby (Tesoro Companies, Inc.), Jack Levis (UPS), Anne Robinson (Verizon Wireless), Mike Gorman (University of Dayton), Glenn Wegryn (independent consultant), Ira Lustig (Princeton Consultants), and Brenda Dietrich (Cornell University). Several members of INFORMS' staff, including Jeff Cohen, Bill Griffin, Tasha Inniss, Jan paul Miller, Melissa Moore, and Louise Wehrle, have made vital contributions to *ABOK*. Danielle LaCourciere, Mindy Okura-Marszycki, Lauren Olesky, Kathleen Pagliaro, and Andrew Prince of John Wiley & Sons, Inc. have also made critical contributions.

I am very excited about what *ABOK* can do for the analytics community, and I am confident you will share my enthusiasm once you have read *ABOK*. This is a living resource that will be updated and revised in the future to ensure it remains current, timely, and cutting-edge, and I encourage you to contact me with suggestions for how to improve it.

Associate Dean for Research, Professor of
Applied Statistics, and the Rogers-Spivey
Faculty Fellow
Culverhouse College of Business
The University of Alabama

James J. Cochran, PhD

List of Contributors

Gerald G. Brown

Operations Research Department
Naval Postgraduate School
Monterey, CA
USA

James J. Cochran

Culverhouse College of Business
The University of Alabama
Tuscaloosa, AL
USA

Thomas H. Davenport

Technology, Operations, and
Information Management
Babson College
Wellesley, MA
USA

Brian T. Downs

Accenture Digital
Data Science Center of Excellence
Dallas, TX
USA

Arnie Greenland

Robert H. Smith School of
Business
University of Maryland
College Park, MD
USA

Mary E. Helander

Data Science Department
IBM T. J. Watson Research Center
Yorktown Heights, NY
USA

Samuel H. Huddleston

Operations Research Department
Naval Postgraduate School
Monterey, CA
USA

Pankush Kalgotra

Graduate School of Management
Clark University
Worcester, MA
USA

Philip T. Keenan

General Motors
Global Research & Development
Warren, MI
USA

Karl G. Kempf

Decision Engineering
Intel Corporation
Chandler, AZ
USA

Jonathan H. Owen

General Motors
Global Research & Development
Warren, MI
USA

Kathryn Schumacher

General Motors
Global Research & Development
Warren, MI
USA

Ramesh Sharda

Spears School of Business
Oklahoma State University
Stillwater, OK
USA

1

Introduction to Analytics

*Philip T. Keenan, Jonathan H. Owen, and
Kathryn Schumacher*

General Motors, Global Research & Development, Warren, MI, USA

1.1 Introduction

We all want to make a difference. We all want our work to enrich the world. As analytics professionals, we are fortunate—this is our time! We live in a world of pervasive data and ubiquitous, powerful computation. This convergence has inspired new applications and accelerated the development of novel analytic techniques and tools, while breathing new life into decades-old approaches that were previously too data- or computation-intensive to be of practical value. The potential for analytics to have an impact has been a call to action for organizations of all types and sizes. Companies are creating new C-level positions and departments to grow analytic capability. A torrent of new start-ups have formed to sell analytics products and services. Even governments have created new high-profile offices to leverage analytics. These changes have driven a surge in demand for analytics professionals, and universities are creating departments, curricula, and new program offerings to fill the gap.

But what exactly do we mean when we say “analytics”? The term is widely used, but has vastly different meanings to different people and communities. A number of well-established disciplines, including statistics, operations research, economics, computer science, industrial engineering, and mathematics, have some claim to “analytics” and interpret it to have specialized meaning within their domains. The popular usage of the term is often comingled with other widely used but equally overloaded terms such as “big data,” “data science,” “machine learning,” “artificial intelligence,” and “cognitive computing.” As a result, this seemingly innocuous term has led to much confusion over the last decade as people using the same language often talk right past each other. In the authors’ own experience, frustration at all levels of an organization is inevitable when well-intentioned and intelligent people believe they have a shared

understanding—on a new project initiative, for example—only to discover weeks or months later that there was a fundamental misunderstanding of what work was to be performed or insights delivered.

In a 2016 article intended to reduce some of this confusion, Robert Rose identified three main usages of the term “analytics” [1]:

- 1) As a synonym for metrics or summary statistics
- 2) As a synonym for “data science” (another overloaded term)
- 3) As a very general term to represent a quantitative approach to organizational decision-making

Our use of the term is closest to the last of these; we consider analytics broadly as a process by which a team of people helps an organization make *better decisions* (the objective) through the *analysis of data* (the activity). This chapter gives a brief, high-level introduction to the subject. We first describe a conceptual framework for analytics, and define three primary categories of analytics (descriptive, predictive, and prescriptive). We then discuss considerations for applying analytics within an organization, and briefly discuss the ethical implications of using analytics. Subsequent chapters dive more deeply into each component of the process of applying analytics, including developing a request for a new project, building a cross-functional team, collecting data, analyzing data with a wide variety of mathematical and statistical methods, and communicating results back to the client.

INTERVIEW WITH ALAN TABER

Alan Taber, System Engineer with Lockheed Martin Missiles and Fire Control, defines analytics in the following way:

Analytics is both a mindset and a process. The mindset is that instead of simply reacting to what you perceive your environment to be that you gather data understanding the limits and bounds of that data. You feed it into a model. It can be a very detailed model or a simple model about how situations evolve over time if you do take options A or B or C, or some combination thereof, and then you test that hypothesis. You have the continual feedback loop to say if what you’re doing makes sense and also keep an eye on your surroundings

because what may have made sense a year ago or a month ago may no longer make sense. That’s the mindset, to always be paying attention rather than running on autopilot.

The process is to make sure you understand the root problem, figure out if you can frame that as a problem that’s amenable to being solved with data, figure out your data sources, and don’t limit yourself to the data you have on hand and know how to collect. If you need a different data set, go get it. Once you have your data and can run your test, do that. Over and under and around all that, you’re working with your stakeholders so that when you deploy people are

familiar enough with what you're doing that they're willing to try it out rather than saying, "I don't understand the model and therefore I'm busy, I don't have time to learn, I'm not interested." If you are overwhelming people

with information but not helping them actually solve the problems that they perceive they have, you simply will not get very far. You will have wasted all your time. So that's the mindset and that's the process.

This is an excerpt from one of a series of interviews with analytics professionals and educators commissioned by the *INFORMS Analytics Body of Knowledge* Committee.

1.2 Conceptual Framework

As shown in Figure 1.1, the generic analytics process can be viewed as a continuous cycle where the analysis of data produces insights that inform better decision-making. We use this simple figure to highlight two fundamentally different approaches to analytics: *data-centric* and *decision-centric*.

1.2.1 Data-Centric Analytics

The philosophy behind *data-centric* analysis is to "let the data speak freely." Working under this philosophy generally involves pulling together as much relevant data as possible, analyzing that data to identify patterns that lead to insight, and serving up those insights to a decision-maker who (hopefully) will make better informed decisions. As shown in Figure 1.2, this follows the natural (clockwise) flow of the analytics process.

Not surprisingly, the data-centric approach has gained popularity with the surge in "big data." Many of the analytic methodologies employed in this arena—including data mining and classification, machine learning, and artificial intelligence—increase in effectiveness with the volume of data available for analysis. Advocates believe that we are in a new "machine age" that is changing the landscape of business and the world [2–4]. Some argue that the data-centric "big data" paradigm is really about eliminating sampling error; they claim that we are

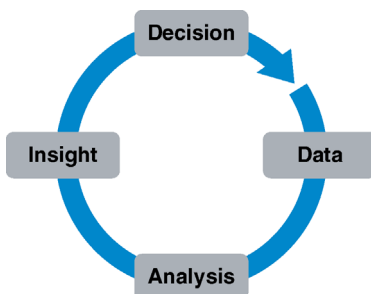


Figure 1.1 Simplified visual representation of the analytics process.

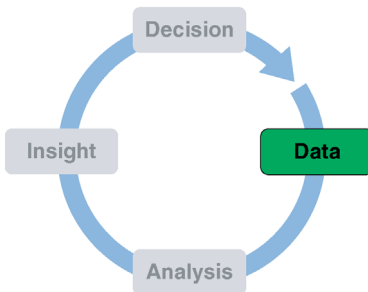
Start with the data

Figure 1.2 The *data-centric* approach starts with the data to surface insights.

no longer reliant on small samples since we have storage capacity to hold and computing power to process vast amounts of data [5]. Others have observed that the promised insights have not always materialized, and that the challenge is “to solve new problems and gain new answers—without making the same old statistical mistakes on a grander scale” [6].

1.2.2 Decision-Centric Analytics

Decision-centric analytics begins with an understanding of the *decision* that needs to be made and what *insights* would lead to better expected outcomes. Decision-centric models typically encapsulate subject matter expertise (SME) and codify domain knowledge in order to relate decision variables to the target objective. Data requirements are determined by the chosen analytical model; ideally these data already exist in a convenient form, but often they must be extracted from disparate sources or collected through new instrumentation or market research. As summarized in Figure 1.3, this approach starts with the final outcome—the decision—and works backward (counterclockwise) at each step to define and develop needed analysis and data resources.

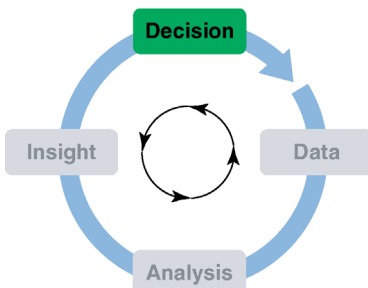
Start with the decision

Figure 1.3 The *decision-centric* approach starts with the problem and works backward.

Decisions are often defined as an “irrevocable allocation of resources” [7]. Improving decision-making requires an understanding of the desired outcome (the objective), alternative actions (decision variables), and boundary conditions (constraints), but also the richer context of possible future conditions (scenarios). It also requires that we answer several softer questions: Who is making the decision? What is her or his scope of control and influence? What information is already available to the decision-maker(s) and where are the gaps? In a decision-centric approach, many of these questions are considered as part of upfront framing activities that look ahead toward operational implementation.

1.2.3 Combining Data- and Decision-Centric Approaches

Analytic practitioners and professional communities are often predisposed to either data-centric or decision-centric approaches. In the authors’ view, this is attributable to different pedagogical perspectives and experiences. Given the centrality of computing and information technologies for handling large amounts of data, it is not surprising that many organizational IT functions are naturally aligned with a data-centric view. Business operations and the analytic teams that support them often have a natural affinity for decision-centric approaches that leverage their deep understanding of key problems and models that support improvements. Table 1.1 summarizes salient features of the two approaches.

Important opportunity arises from combining elements of the two approaches. There is undeniable potential to leverage increasingly pervasive data and computational power associated with data-centric analysis, but contextual knowledge and subject matter expertise provide needed guardrails so that the resulting insights are meaningful.

Acknowledging the natural tendencies of individuals or analytics organizations toward data- or decision-centric approaches may help practitioners to identify growth opportunities. For example, traditionally decision-centric organizations may benefit by expanding the amount of data used in their analyses, including unstructured data sources. Typically, data-centric groups may improve the fit and predictive power of their models by incorporating domain-specific expertise.

Evidence of the benefit of utilizing a combined approach is seen in recent movements to incorporate “thick data” into marketing analytics (see Refs [8,9], for example). Combining thick data, such as ethnographic studies or focus group responses (see Figure 1.5), with big data, such as transaction data, enables a more complete understanding of customers’ preferences and behaviors. Decision-centric framing, domain knowledge, and deep subject matter expertise collectively provide scaffolding that helps big data insights take shape.

Table 1.1 Comparison of data-centric and decision-centric approaches.

	Data-centric analysis (Data science, computer science)	Decision-centric analysis (Decision science, operations research)
Mantra	“Start with the data”	“Start with the decision”
Philosophy	Leverage large amounts of data. Let the data “speak freely” by identifying patterns and revealing implicit (hidden) factor relationships	Leverage domain knowledge and subject matter expertise to model explicit variable relationships
Data	More is better, especially for “big data” applications (e.g., speech or image recognition)	Custom collection of curated data sets
Computing	High-performance computing is often price of entry. Potential need for specialized processors (e.g., GPUs, TPUs) for acceptable execution speeds, especially in contexts requiring real-time analysis	Desktop or server-based computing is typical. Trade-offs between potential benefits of leveraging high-performance computing versus added overhead in development and maintenance
Pros	<ul style="list-style-type: none"> • Increasingly automatable • Potential to extract weak signals from large, unstructured data sets 	<ul style="list-style-type: none"> • Causal focus • Strategic value beyond historical observations
Cons	<ul style="list-style-type: none"> • Risk of conflating correlation with causation • Analysis inferences are limited by history • Noisy data with confounded effects 	<ul style="list-style-type: none"> • Human subject matter expertise required • Cost of data acquisition can be high
Key disciplines	<ul style="list-style-type: none"> • Computer science • Data science • Machine learning and unstructured data mining • Artificial intelligence (AI), deep learning 	<ul style="list-style-type: none"> • Management and decision sciences • Operations research • Mathematics • Classical statistics
Example applications	<ul style="list-style-type: none"> • Image classification • Speech recognition • Autonomous vehicle scene recognition 	<ul style="list-style-type: none"> • Supply chain optimization • Scenario planning • New business model development

1.3 Categories of Analytics

A well-known and useful classification scheme for analytics was proposed by Lustig et al., at IBM [10]. Based on their experience with a variety of companies across a diverse set of industries, they defined three broad categories of analytics:

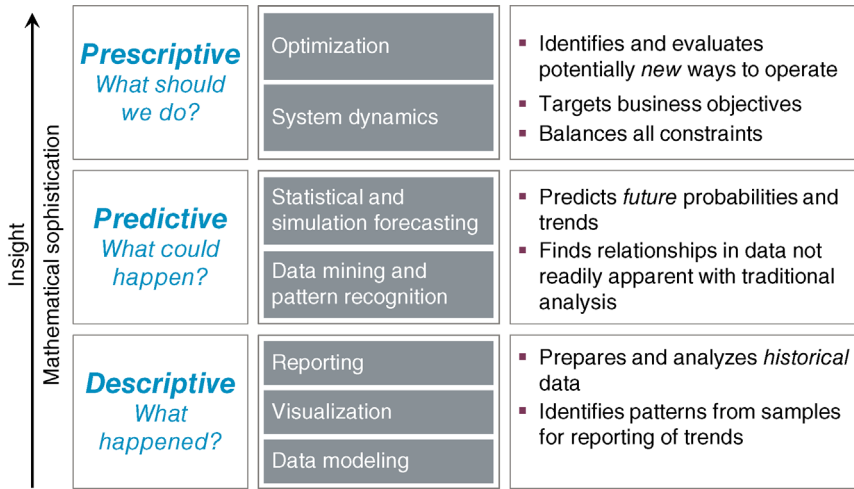


Figure 1.4 Categories of analytics.

descriptive, predictive, and prescriptive. As summarized in Figure 1.4, there is a natural progression in the level of insight provided—and potential value—as an organization moves from descriptive to predictive and ultimately to prescriptive analytics. Typically there is also a progression in the mathematical sophistication of the analysis techniques, as well as the organizational maturity required to absorb and act on resulting insights.

1.3.1 Descriptive Analytics

The purpose of descriptive analytics is to reveal and summarize facts about what has happened in the past or, in the case of real-time analysis, what is happening in the present. This is done by examining and synthesizing data collected from a variety of sources. Raw data are captured and recorded in source systems, eventually to be cleaned, retrieved, and normalized such that entities and relationships can be meaningfully understood. The audience for descriptive analytics is broad, potentially reaching all functions and levels of an organization. Descriptive analytics are at the heart of most business intelligence (BI) systems.

Data Modeling

Many organizations have access to vast quantities of data. Useful descriptive analytics generally involves processing the raw facts into higher level abstractions. Data scientists think in terms of *entities* and *relationships*. For example, a customer database might contain entities like “Household” and “Product,” linked by relationships like “Purchased,” with data elements

Table 1.2 Potential sources of data.

Source	Examples
Transaction data	Data associated with a transactional event. Example: a purchase transaction with details of the specific item purchased, where and when it was purchased, the price paid and any discounts applied, how the customer paid (e.g., cash, credit card, finance), and other contextually relevant data (e.g., inventory of other items for sale at the same time and location)
Customer data	Data associated with customers. Examples: detailed demographic or psychographic information on individuals and households, history of interactions (past purchases, Web site visits, customer service requests)
Sensor data	Data collected through electronic or mechanical instrumentation. Examples: web browser cookies tracking customer activity, electronic sensors monitoring weather conditions, airplane flight data recorder information
Public data	Open-source data from individuals, organizations, and governments. Example: aggregated census data
Unstructured	Data without known structure. Examples: text and images from social media, call center recordings, qualitative data from focus groups or ethnographic studies
Curated data	Data collected for a specific purpose with downstream analysis in mind. Examples: consumer surveys, designed market research experiments

including the demographics of the households and the price, cost and features of the products.

Sources of data can be highly varied (see Table 1.2 for examples), as can the size and information density of any given data set (see Figure 1.5). There is also high variability in the expense and effort required to collect different types of data. On one end of the spectrum, ethnographic studies require social scientists to spend many hours shopping with or interviewing individual customers, and thus the data are very carefully curated and very expensive to collect. On the other end of the spectrum, “data exhaust” is logged nearly for free, including data generated from smartphones and online activity [11]. Data exhaust is collected without a specific intended purpose and can be especially messy, so substantial cleanup effort is usually necessary before this type of data are usable.

Developing a data model that captures the structure and relationships among the different data elements is a fundamental task. Generic data models are often constructed to efficiently store ingested data, without specific analytic use cases in mind. Although such data models can be useful for general-purpose reporting and data exploration, purpose-built data models are typically needed for efficient