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

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.

Library of Congress Cataloging-in-Publication Data

ISBN: 9781119483212

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

10 9 8 7 6 5 4 3 2 1

Contents

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 INFORMS[1] 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.

James J. Cochran, PhD

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

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

INFORMS Analytics Body of Knowledge, First Edition. Edited by James J. Cochran.

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

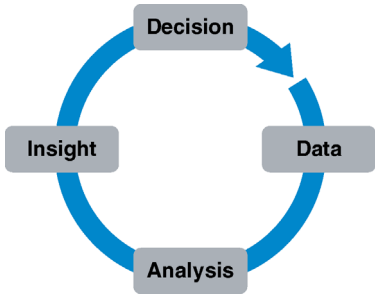

Figure 1.1 Simplified visual representation of the analytics process.

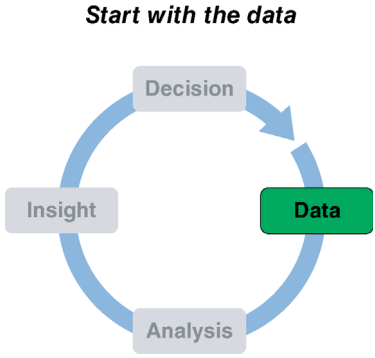

Figure 1.2 The *data-centric* approach starts with the data to surface insights.

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.

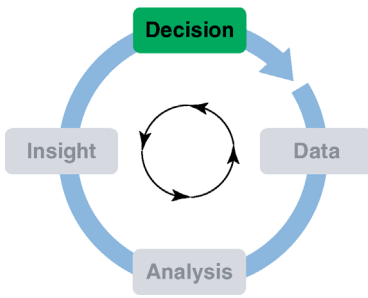

Figure 1.3 The *decision-centric* approach starts with the problem and works backward.

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	• Increasingly automatable • Potential to extract weak signals from large, unstructured data sets	• Causal focus • Strategic value beyond historical observations
Cons	• Risk of conflating correlation with causation • Analysis inferences are limited by history • Noisy data with confounded effects	• Human subject matter expertise required • Cost of data acquisition can be high
Key disciplines	• Computer science • Data science • Machine learning and unstructured data mining • Artificial intelligence (AI), deep learning	• Management and decision sciences • Operations research • Mathematics • Classical statistics
Example applications	• Image classification • Speech recognition • Autonomous vehicle scene recognition	• 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:

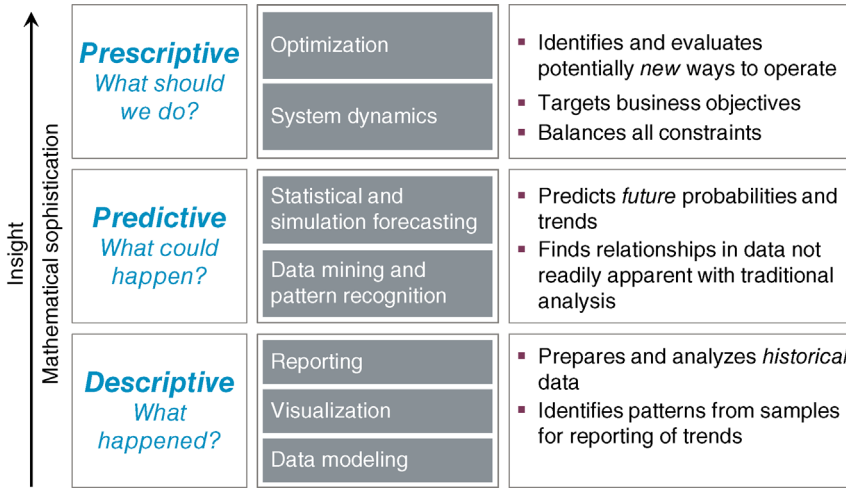

Figure 1.4 Categories of analytics.

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