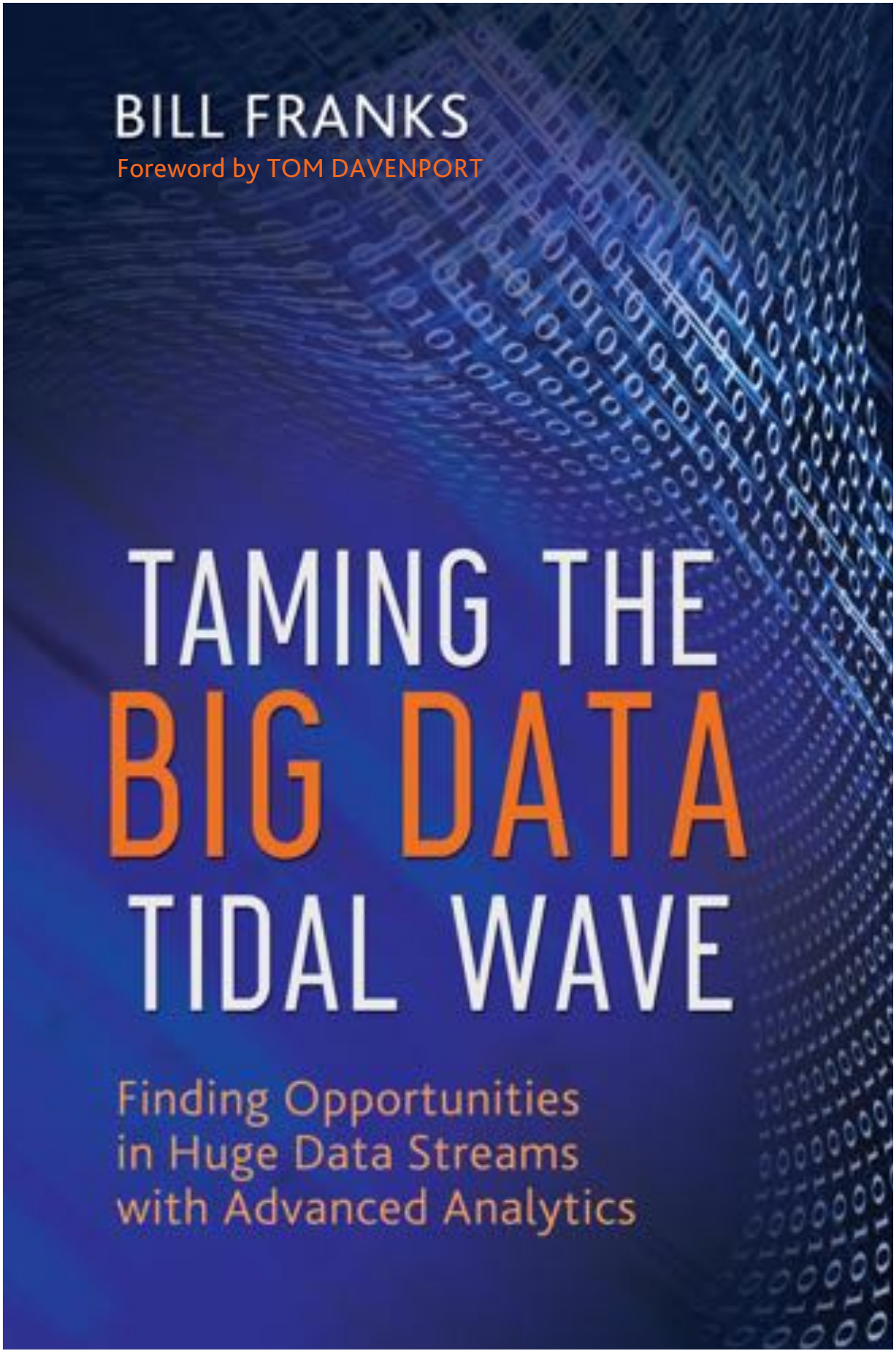
BILL FRANKS
Foreword by TOM DAVENPORT
TAMING THE
BIG DATA
TIDAL WAVE
Finding Opportunities
in Huge Data Streams
with Advanced Analytics

Praise for

TAMING THE BIG DATA TIDAL WAVE

Finding Opportunities in Huge Data Streams with Advanced Analytics

"This book . . . puts the focus squarely where it belongs . . . It's primarily about the effective analysis of big data, rather than the big data management (BDM) topic per se. It starts with data and goes all the way into such topics as how to frame decisions, how to build an analytics center of excellence, and how to build an analytical culture. You will find some mentions of BDM topics, as you should. But the bulk of the content here is about how to create, organize, staff, and execute on analytical initiatives that make use of data as the input."

—From the Foreword by Thomas H. Davenport, President's Distinguished Professor of IT and Management, Babson College, cofounder and Research Director, International Institute for Analytics

"This is a one-stop handbook for anyone who wants to understand what big data is and how to leverage it through advanced analytic processes and methods. Bill Franks intimately understands and describes how to create an entire analytics ecosystem intended to deliver competitive advantage."

—**Stuart Aitken**, CEO, dunnhumby USA

"In *Taming the Big Data Tidal Wave*, Bill Franks does a great job introducing both big data and the kind of analytics that will generate value from the waves of new data that are washing over companies. Easy to read and with helpful wrap-up sections in each chapter, the book avoids technical jargon without being lightweight. In this great introductory book, Bill makes a powerful case for analytic innovation and for getting started now."

—**James Taylor**, CEO, Decision Management Solutions and author of *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics*

"In case you ever wondered why big data is providing business value in many industries, this book gives you perspectives and answers from many angles—from the tech side, to data science, to business users and processes. In my entire career of researching and lecturing on analytics, I have never encountered a book that combines the knowledge of both information technology and business managers in such a succinct way. I would recommend it to anyone whose career intersects with big data."

—**Diego Klabjan**, Professor at Northwestern University; Director, Master of Science in Analytics

"Bill Franks provides an entertaining and consumable take on a complex and intricate topic. The mix of insights applicable to practitioners and novices alike make this a critical read for someone new to the analytics space or to anyone in the space wanting to ensure they can learn from an accomplished leader. Franks' view across multiple industries and uses of big data have positioned him well to deliver this entry into the emergence of the space."

—**Richard Maltsbarger**, Senior Vice President of Strategy, Lowe's Companies, Inc.

Also available as an e-book

sas

The Wiley and SAS Business Series presents books that help senior level managers with their critical management decisions

ISBN 978-1-118-20878-6
34995
9 781118 208786

Contents

Foreword xiii

Preface xvii

Acknowledgments xxv

PART ONE THE RISE OF BIG DATA 1

Chapter 1 What Is Big Data and Why Does It Matter? 3

What Is Big Data? 4
Is the "Big" Part or the "Data" Part More Important? 5
How Is Big Data Different? 7
How Is Big Data More of the Same? 9
Risks of Big Data 10
Why You Need to Tame Big Data 12
The Structure of Big Data 14
Exploring Big Data 16
Most Big Data Doesn't Matter 17
Filtering Big Data Effectively 20
Mixing Big Data with Traditional Data 21
The Need for Standards 22
Today's Big Data Is Not Tomorrow's Big Data 24
Wrap-Up 26
Notes 27

Chapter 2 Web Data: The Original Big Data 29

Web Data Overview 30
What Web Data Reveals 36
Web Data in Action 42
Wrap-Up 50
Note 51

Chapter 3 A Cross-Section of Big Data Sources and the Value They Hold 53

Auto Insurance: The Value of Telematics Data 54
Multiple Industries: The Value of Text Data 57
Multiple Industries: The Value of Time and Location Data 60
Retail and Manufacturing: The Value of Radio Frequency Identification Data 64
Utilities: The Value of Smart-Grid Data 68
Gaming: The Value of Casino Chip Tracking Data 71
Industrial Engines and Equipment: The Value of Sensor Data 73
Video Games: The Value of Telemetry Data 76
Telecommunications and Other Industries: The Value of Social Network Data 78
Wrap-Up 82

PART TWO TAMING BIG DATA: THE TECHNOLOGIES, PROCESSES, AND METHODS 85

Chapter 4 The Evolution of Analytic Scalability 87

A History of Scalability 88
The Convergence of the Analytic and Data Environments 90
Massively Parallel Processing Systems 93
Cloud Computing 102
Grid Computing 109

MapReduce 110
It Isn't an Either/Or Choice! 117
Wrap-Up 118
Notes 119

Chapter 5 The Evolution of Analytic Processes 121

The Analytic Sandbox 122
What Is an Analytic Data Set? 133
Enterprise Analytic Data Sets 137
Embedded Scoring 145
Wrap-Up 151

Chapter 6 The Evolution of Analytic Tools and Methods 153

The Evolution of Analytic Methods 154
The Evolution of Analytic Tools 163
Wrap-Up 175
Notes 176

PART THREE TAMING BIG DATA: THE PEOPLE AND APPROACHES 177

Chapter 7 What Makes a Great Analysis? 179

Analysis versus Reporting 179
Analysis: Make It G.R.E.A.T.! 184
Core Analytics versus Advanced Analytics 186
Listen to Your Analysis 188
Framing the Problem Correctly 189
Statistical Significance versus Business Importance 191
Samples versus Populations 195
Making Inferences versus Computing Statistics 198
Wrap-Up 200

Chapter 8 What Makes a Great Analytic Professional? 201

Who Is the Analytic Professional? 202
The Common Misconceptions about Analytic Professionals 203
Every Great Analytic Professional Is an Exception 204
The Often Underrated Traits of a Great Analytic Professional 208
Is Analytics Certification Needed, or Is It Noise? 222
Wrap-Up 224

Chapter 9 What Makes a Great Analytics Team? 227

All Industries Are Not Created Equal 228
Just Get Started! 230
There's a Talent Crunch out There 231
Team Structures 232
Keeping a Great Team's Skills Up 237
Who Should Be Doing Advanced Analytics? 241
Why Can't IT and Analytic Professionals Get Along? 245
Wrap-Up 247
Notes 248

PART FOUR BRINGING IT TOGETHER: THE ANALYTICS CULTURE 249

Chapter 10 Enabling Analytic Innovation 251

Businesses Need More Innovation 252
Traditional Approaches Hamper Innovation 253
Defining Analytic Innovation 255
Iterative Approaches to Analytic Innovation 256
Consider a Change in Perspective 257
Are You Ready for an Analytic Innovation Center? 259
Wrap-Up 269
Note 270

Chapter 11 Creating a Culture of Innovation and Discovery 271

Setting the Stage 272
Overview of the Key Principles 274
Wrap-Up 290
Notes 291

Conclusion: Think Bigger! 293

About the Author 295

Index 297

Foreword

Like it or not, a massive amount of data will be coming your way soon. Perhaps it has reached you already. Perhaps you've been wrestling with it for a while—trying to figure out how to store it for later access, address its mistakes and imperfections, or classify it into structured categories. Now you are ready to actually extract some value out of this huge dataset by analyzing it and learning something about your customers, your business, or some aspect of the environment for your organization. Or maybe you're not quite there, but you see light at the end of the data management tunnel.

In either case, you've come to the right place. As Bill Franks suggests, there may soon be not only a flood of data, but also a flood of books about big data. I'll predict (with no analytics) that this book will be different from the rest. First, it's an early entry in the category. But most importantly, it has a different content focus.

Most of these big-data books will be about the management of big data: how to wrestle it into a database or data warehouse, or how to structure and categorize unstructured data. If you find yourself reading a lot about Hadoop or MapReduce or various approaches to data warehousing, you've stumbled upon—or were perhaps seeking—a "big data management" (BDM) book.

This is, of course, important work. No matter how much data you have of whatever quality, it won't be much good unless you get it into an environment and format in which it can be accessed and analyzed.

But the topic of BDM alone won't get you very far. You also have to analyze and act on it for data of any size to be of value. Just as traditional database management tools didn't automatically analyze transaction data from traditional systems, Hadoop and MapReduce won't automatically interpret the meaning of data from web sites,

gene mapping, image analysis, or other sources of big data. Even before the recent big data era, many organizations have gotten caught up in data management for years (and sometimes decades) without ever getting any real value from their data in the form of better analysis and decision-making.

This book, then, puts the focus squarely where it belongs, in my opinion. It's primarily about the effective analysis of big data, rather than the BDM topic, per se. It starts with data and goes all the way into such topics as how to frame decisions, how to build an analytics center of excellence, and how to build an analytical culture. You will find some mentions of BDM topics, as you should. But the bulk of the content here is about how to create, organize, staff, and execute on analytical initiatives that make use of data as the input.

In case you have missed it, analytics are a very hot topic in business today. My work has primarily been around how companies compete on analytics, and my books and articles in these areas have been among the most popular of any I've written. Conferences on analytics are popping up all over the place. Large consulting firms such as Accenture, Deloitte, and IBM have formed major practices in the area. And many companies, public sector organizations, and even nonprofits have made analytics a strategic priority. Now people are also very excited about big data, but the focus should still remain on how to get such data into a form in which it can be analyzed and thus influence decisions and actions.

Bill Franks is uniquely positioned to discuss the intersection of big data and analytics. His company, Teradata, compared to other data warehouse/data appliance vendors, has always had the greatest degree of focus within that industry segment on actually analyzing data and extracting business value from it. And although the company is best known for enterprise data warehouse tools, Teradata has also provided a set of analytical applications for many years.

Over the past several years Teradata has forged a close partnership with SAS, the leading analytics software vendor, to develop highly scalable tools for analytics on large databases. These tools, which often involve embedding analysis within the data warehouse environment itself, are for large-volume analytical applications such as real-time fraud detection and large-scale scoring of customer buying propensi-

ties. Bill Franks is the chief analytics officer for the partnership and therefore has had access to a large volume of ideas and expertise on production-scale analytics and "in-database processing." There is perhaps no better source on this topic.

So what else is particularly interesting and important between these covers? There are a variety of high points:

- Chapter 1 provides an overview of the big data concept, and explains that "size doesn't always matter" in this context. In fact, throughout the book, Franks points out that much of the volume of big data isn't useful anyway, and that it's important to focus on filtering out the dross data.
- The overview of big data sources in Chapter 3 is a creative, useful catalog, and unusually thorough. And the book's treatment of web data and web analytics in Chapter 2 is very useful for anyone or any organization wishing to understand online customer behavior. It goes well beyond the usual reporting-oriented focus of web analytics.
- Chapter 4, devoted to "The Evolution of Analytical Scalability," will provide you with a perspective on the technology platforms for big data and analytics that I am pretty sure you won't find anywhere else on this earth. It also puts recent technologies like MapReduce in perspective, and sensibly argues that most big data analytics efforts will require a combination of environments.
- This book has some up-to-the-minute content about how to create and manage analytical data environments that you also won't find anywhere else. If you want the best and latest thinking about "analytic sandboxes" and "enterprise analytic data sets" (that was a new topic for me, but I now know what they are and why they're important), you'll find it in Chapter 5. This chapter also has some important messages about the need for model and scoring management systems and processes.
- Chapter 6 has a very useful discussion of the types of analytical software tools that are available today, including the open source package R. It's very difficult to find commonsense advice

about the strengths and weaknesses of different analytical environments, but it is present in this chapter. Finally, the discussion of ensemble and commodity analytical methods in this chapter is refreshingly easy to understand for nontechnical types like me.

- Part Three of the book leaves the technical realm for advice on how to manage the human and organizational sides of analytics. Again, the perspective is heavily endowed with good sense. I particularly liked, for example, the emphasis on the framing of decisions and problems in Chapter 7. Too many analysts jump into analysis without thinking about the larger questions of how the problem is being framed.
- Someone recently asked me if there was any description of analytical culture outside of my own writings. I said I didn't know of any, but that was before I read Part Four of Franks's book. It ties analytical culture to innovation culture in a way that I like and have never seen before.

Although the book doesn't shrink from technical topics, it treats them all with a straightforward, explanatory approach. This keeps the book accessible to a wide audience, including those with limited technical backgrounds. Franks's advice about data visualization tools summarizes the tone and perspective of the entire book: "Simple is best. Only get fancy or complex when there is a specific need."

If your organization is going to do analytical work—and it definitely should—you will need to address many of the issues raised in this book. Even if you're not a technical person, you will need to be familiar with some of the topics involved in building an enterprise analytical capability. And if you are a technical person, you will learn much about the human side of analytics. If you're browsing this foreword in a bookstore or through "search inside this book," go ahead and buy it. If you've already bought it, get busy and read!

THOMAS H. DAVENPORT
President's Distinguished Professor of IT and
Management, Babson College
Co-Founder and Research Director, International
Institute for Analytics

Preface

You receive an e-mail. It contains an offer for a complete personal computer system. It seems like the retailer read your mind since you were exploring computers on their web site just a few hours prior. . . .

As you drive to the store to buy the computer bundle, you get an offer for a discounted coffee from the coffee shop you are getting ready to drive past. It says that since you're in the area, you can get 10% off if you stop by in the next 20 minutes. . . .

As you drink your coffee, you receive an apology from the manufacturer of a product that you complained about yesterday on your Facebook page, as well as on the company's web site. . . .

Finally, once you get back home, you receive notice of a special armor upgrade available for purchase in your favorite online video game. It is just what is needed to get past some spots you've been struggling with. . . .

Sound crazy? Are these things that can only happen in the distant future? No. All of these scenarios are possible today! Big data. Advanced analytics. Big data analytics. It seems you can't escape such terms today. Everywhere you turn people are discussing, writing about, and promoting big data and advanced analytics. Well, you can now add this book to the discussion.

What is real and what is hype? Such attention can lead one to the suspicion that perhaps the analysis of big data is something that is more hype than substance. While there has been a lot of hype over the past few years, the reality is that we are in a transformative era in terms of analytic capabilities and the leveraging of massive amounts of data. If you take the time to cut through the sometimes overzealous hype present in the media, you'll find something very real and very powerful underneath it. With big data, the hype is driven by

genuine excitement and anticipation of the business and consumer benefits that analyzing it will yield over time.

Big data is the next wave of new data sources that will drive the next wave of analytic innovation in business, government, and academia. These innovations have the potential to radically change how organizations view their business. The analysis that big data enables will lead to decisions that are more informed and, in some cases, different from what they are today. It will yield insights that many can only dream about today. As you'll see, there are many consistencies with the requirements to tame big data and what has always been needed to tame new data sources. However, the additional scale of big data necessitates utilizing the newest tools, technologies, methods, and processes. The old way of approaching analysis just won't work. It is time to evolve the world of advanced analytics to the next level. That's what this book is about.

Taming the Big Data Tidal Wave isn't just the title of this book, but rather an activity that will determine which businesses win and which lose in the next decade. By preparing and taking the initiative, organizations can ride the big data tidal wave to success rather than being pummeled underneath the crushing surf. What do you need to know and how do you prepare in order to start taming big data and generating exciting new analytics from it? Sit back, get comfortable, and prepare to find out!

INTENDED AUDIENCE

There have been myriad books on advanced analytics over the years. There have also been a number of books on big data more recently. This book attempts to come from a different angle than the others. The primary focus is educating the reader on what big data is all about and how it can be utilized through analytics, and providing guidance on how to approach the creation and evolution of a world-class advanced analytics ecosystem in today's big data environment. A wide range of readers will find this book to be of value and interest. Whether you are an analytics professional, a businessperson who uses the results that analysts produce, or just someone with an interest in big data and advanced analytics, this book has something for you.

The book will not provide deeply detailed technical reviews of the topics covered. Rather, the book aims to be just technical enough to provide a high-level understanding of the concepts discussed. The goal is to enable readers to understand and begin to apply the concepts while also helping identify where more research is desired. This book is more of a handbook than a textbook, and it is accessible to non-technical readers. At the same time, those who already have a deeper understanding of the topics will be able to read between the lines to see the more technical implications of the discussions.

OVERVIEW OF THE CONTENTS

This book is comprised of four parts, each of which covers one aspect of taming the big data tidal wave. Part One focuses on what big data is, why it is important, and how it can be applied. Part Two focuses on the tools, technologies, and methods required to analyze and act on big data successfully. Part Three focuses on the people, teams, and analysis principles that are required to be effective. Part Four brings everything together and focuses on how to enable innovative analytics through an analytic innovation center and a change in culture. Below is a brief outline with more detail on what each part and chapter are about.

PART ONE: THE RISE OF BIG DATA

Part One is focused on what big data is, why it is important, and the benefits of analyzing it. It covers a total of 10 big data sources and how those sources can be applied to help organizations improve their business. If readers are unclear when picking up the book about what big data is or how broadly big data applies, Part One will provide clarity.

Chapter 1: What Is Big Data and Why Does It Matter? This chapter begins with some background on big data and what it is all about. It then covers a number of considerations related to how organizations can make use of big data. Readers will need to understand what is in this chapter as much as anything else in the book if they are to help their organizations tame the big data tidal wave successfully.

Chapter 2: Web Data: The Original Big Data. Probably the most widely used and best-known source of big data today is the detailed data collected from web sites. The logs generated by users navigating the web hold a treasure trove of information just waiting to be analyzed. Organizations across a number of industries have integrated detailed, customer-level data sourced from their web sites into their enterprise analytics environments. This chapter explores how that data is enhancing and changing a variety of business decisions.

Chapter 3: A Cross-Section of Big Data Sources and the Value They Hold. In this chapter, we look at nine more sources of big data at a high level. The purpose is to introduce what each data source is and then review some of the applications and implications that each data source has for businesses. One trend that becomes clear is how the same underlying technologies can lead to multiple big data sources in different industries. In addition, different industries can leverage some of the same sources of big data. Big data is not a one-trick pony with narrow application.

PART TWO: TAMING BIG DATA: THE TECHNOLOGIES, PROCESSES, AND METHODS

Part Two focuses on the technologies, processes, and methods required to tame big data. Major advances have increased the scalability of all three of those areas over the years. Organizations can't continue to rely on outdated approaches and expect to stay competitive in the world of big data. This part of the book is by far the most technical, but should still be accessible to almost all readers. After reading these chapters, readers will be familiar with a number of concepts that they will come across as they enter the world of analyzing big data.

Chapter 4: The Evolution of Analytic Scalability. The growth of data has always been at a pace that strains the most scalable options available at any point in time. The traditional ways of performing advanced analytics were already reaching their limits before big data. Now, traditional approaches just won't do. This chapter discusses the convergence of the analytic and data environments, massively parallel processing (MPP) architectures, the cloud, grid computing, and

MapReduce. Each of these paradigms enables greater scalability and will play a role in the analysis of big data.

Chapter 5: The Evolution of Analytic Processes. With a vastly increased level of scalability comes the need to update analytic processes to take advantage of it. This chapter starts by outlining the use of analytical sandboxes to provide analytic professionals with a scalable environment to build advanced analytics processes. Then, it covers how enterprise analytic data sets can help infuse more consistency and less risk in the creation of analytic data while increasing analyst productivity. The chapter ends with a discussion of how embedded scoring processes allow results from advanced analytics processes to be deployed and widely consumed by users and applications.

Chapter 6: The Evolution of Analytic Tools and Methods. This chapter covers several ways in which the advanced analytic tool space has evolved and how such advances will continue to change the way analytic professionals do their jobs and handle big data. Topics include the evolution of visual point and click interfaces, analytic point solutions, open source tools, and data visualization tools. The chapter also covers how analytic professionals have changed their approaches to building models to better leverage the advances available to them. Topics include ensemble modeling, commodity models, and text analysis.

PART THREE: TAMING BIG DATA: THE PEOPLE AND APPROACHES

Part Three is focused on the people that drive analytic results, the teams they belong to, and the approaches they use to ensure that they provide great analysis. The most important factor in any analytics endeavor, including the analysis of big data, is having the right people in the driver's seat who are following the right analysis principles. After reading Part Three, readers will better understand what sets great analysis, great analytic professionals, and great analytics teams apart from the rest.

Chapter 7: What Makes a Great Analysis? Computing statistics, writing a report, and applying a modeling algorithm are each only

one step of many required for generating a great analysis. This chapter starts by clarifying a few definitions, and then discusses a variety of themes that relate to creating great analysis. With big data adding even more complexity to the mix than organizations are used to dealing with, it's more crucial than ever to keep the principles discussed in this chapter in mind.

Chapter 8: What Makes a Great Analytic Professional? Skill in math, statistics, and programming are necessary, but not sufficient, traits of a great analytic professional. Great analytic professionals also have traits that are often not the first things that come to most people's minds. These traits include commitment, creativity, business savvy, presentation skills, and intuition. This chapter explores why each of these traits are so important in defining a great analytic professional and why they can't be overlooked.

Chapter 9: What Makes a Great Analytics Team? How should an organization structure and maintain advanced analytics teams for optimal impact? Where do the teams fit in the organization? How should they operate? Who should be creating advanced analytics? This chapter talks about some common challenges and principles that must be considered to build a great analytics team.

PART FOUR: BRINGING IT TOGETHER: THE ANALYTICS CULTURE

Part Four focuses on some well-known underlying principles that must be applied for an organization to successfully innovate with advanced analytics and big data. While these principles apply broadly to other disciplines as well, the focus will be on providing a perspective on how the principles relate to advanced analytics within today's enterprise environments. The concepts covered will be familiar to readers, but perhaps not the way that the concepts are applied to the world of advanced analytics and big data.

Chapter 10: Enabling Analytic Innovation. This chapter starts by reviewing some of the basic principles behind successful innovation. Then, it applies them to the world of big data and advanced analytics through the concept of an analytic innovation center. The goal is to provide readers with some tangible ideas of

how to better enable analytic innovation and the taming of big data within their organizations.

Chapter 11: Creating a Culture of Innovation and Discovery. This chapter wraps things up with some perspectives on how to create a culture of innovation and discovery. It is meant to be fun and lighthearted, and to provide food for thought in terms of what it takes to create a culture that is able to produce innovative analytics. The principles covered are commonly discussed and well-known. However, it is worth reviewing them and then considering how an organization can apply the well-established principles to big data and advanced analytics.

For more information on Bill Franks and his book, *Taming the Big Data Tidal Wave*, visit **www.tamingthebigdatatidalwave.com**