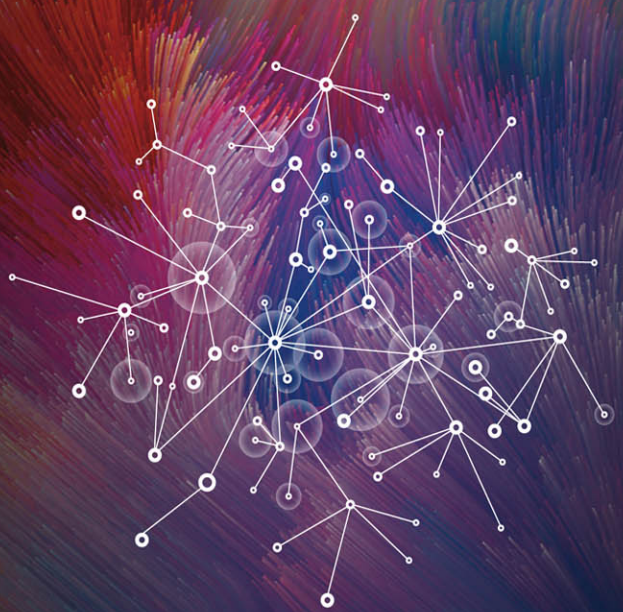
CENGAGE Learning®
Professional • Technical • Reference

DATA DIVINATION

BIG DATA STRATEGIES



PAM BAKER
WITH SPECIAL CONTRIBUTION BY BOB GOURLEY

Data Divination: Big Data Strategies

Pam Baker

Cengage Learning PTR

CENGAGE Learning®
Professional • Technical • Reference

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

CENGAGE Learning®
Professional • Technical • Reference

Data Divination: Big Data Strategies
Pam Baker

Publisher and General Manager, Cengage Learning PTR: Stacy L. Hiquet
Associate Director of Marketing: Sarah Panella
Manager of Editorial Services: Heather Talbot
Product Manager: Heather Hurley
Project/Copy Editor: Kezia Endsley
Technical Editor: Rich Santalesa
Interior Layout: MPS Limited
Cover Designer: Luke Fletcher
Proofreader: Sam Garvey
Indexer: Larry Sweazy

Library of Congress Control Number: 2014937092

ISBN-13: 978-1-305-11508-8

ISBN-10: 1-305-11508-2
eISBN-10: 1-305-11509-0

Cengage Learning PTR
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **international.cengage.com/region**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

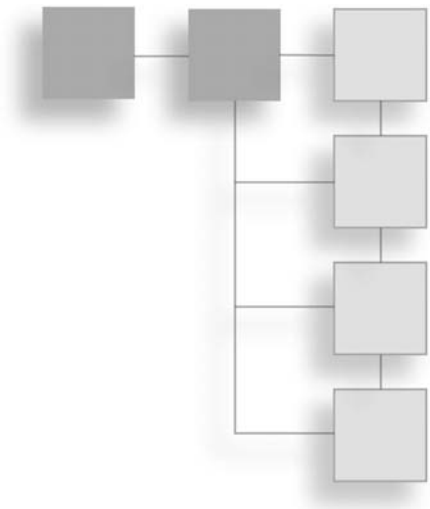
For your lifelong learning solutions, visit **cengageptr.com**.

Visit our corporate website at **cengage.com**.

Printed in the United States of America
1 2 3 4 5 6 7 16 15 14

To my daughter Stephanie and my son Ben; you are my inspiration each and every day and the joy of my life. To my mother Nana Duffey; my profound gratitude for teaching me critical thinking skills from a very early age and providing me with a strong, lifelong education and living example of exemplary ethics.

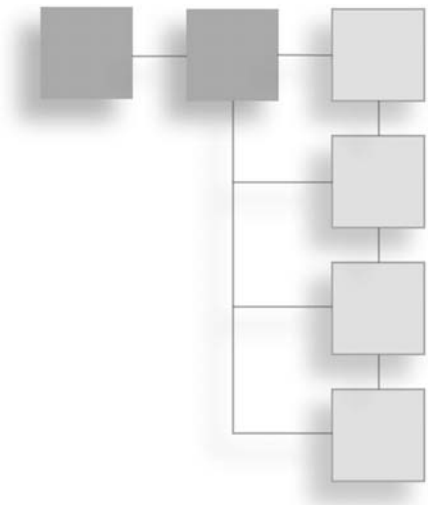
Acknowledgments



First and foremost, I would like to thank the publishing team at Cengage Learning PTR for their hard work and patience during this time. Specifically, thank you Stacy Hiquet for publishing our text. Thank you Heather Hurley for facilitating this book and being the marvelous person that you are. Your professionalism is second to none but your demeanor is an absolute joy. Thank you Kezia Endsley for your calm management of the editing and deliverables, despite the challenges I inadvertently presented to you. Your many talents, eternal patience, and wise guidance were invaluable to this effort. You are by far the best editor I have had the pleasure of working with and I sincerely hope to have the honor of working with you again one day. Thank you Richard Santalesa for your thorough tech review and insightful suggestions. You are and always will be the greatest legal resource and tech editor a writer can ever have, not to mention the truest of friends. Thank you to the entire team at Cengage.

Many thanks also to my family for their patience and support during this time. Spending long and seemingly unending hours finishing the book is not only hard on the authors, but our families as well. A special thanks to two of my brothers—Steven Duffey and John Duffey—for pitching in to create screenshots and such to exacting specs to help speed completion of the book on such a tight deadline.

About the Authors



Pam Baker is a noted business analyst, tech freelance journalist and the editor of the online publication and e-newsletter, FierceBigData. Her work is seen in a wide variety of respected publications, including but not limited to *Institutional Investor* magazine, *ReadWriteWeb,* CIO (paper version), CIO.com, *Network World*, *Computerworld*, *IT World*, *LinuxWorld, iSixSigma*, and *TechNewsWorld*. Further she formerly served as a contracted analyst for London-based VisionGain Research and Evans Data Corp, headquartered in Santa Cruz, California. She has also served as a researcher, writer, and managing editor of *Wireless I.Q.* and *Telematics Journal* for ABI Research, headquartered in New York.

Interested readers can view a variety of published clips and read more about Pam Baker and her work on these websites: Mediabistro Freelance Marketplace at http://www.mediabistro.com/PamBaker and the Internet Press Guild at http://www.netpress.org/ipg-membership-directory/pambaker. There are also numerous professional references on her LinkedIn page at http://www.linkedin.com/in/pambaker/.

She has also authored numerous ebooks and several of the dead tree variety. Six of those books are listed on her Amazon Author Central page. Further, Baker co-authored two books on the biosciences for the Association of University Technology Managers (AUTM), a global nonprofit association of technology managers and business executives. Those two books were largely funded by the Bill and Melinda Gates Foundation.

Among other awards, Baker won international acclaim for her documentary on the paper-making industry and was awarded a Resolution from the City of Columbus, Georgia, for her news series on the city in *Georgia Trend Magazine*. The only other author to receive such recognition from the city was the legendary Carson McCullers. Baker is a member of the National Press Club (NPC) and the Internet Press Guild (IPG). You can follow or chat with her on Twitter where her handle is @bakercom1 or on Google + at google .com/+PamBaker. You can also reach her through the contact form at FierceBigData where she is the editor (see http://www.fiercebigdata.com/).

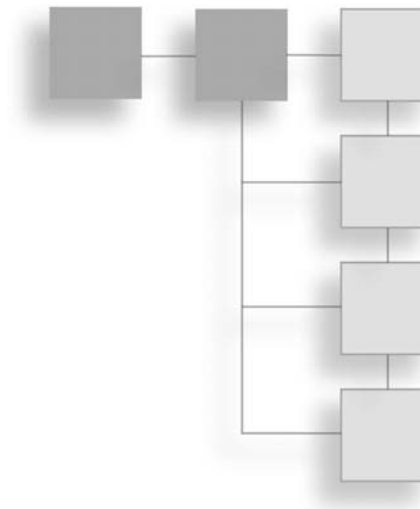
Bob Gourley is a contributing writer in *Data Divination*. He wrote a significant portion of the chapter on use cases in the Department of Defense and Intelligence Community given that is the area he focuses on in his work with big data. Gourley also wrote the chapter on Empowering the Workforce. He is the editor in chief of CTOvision.com and is the founder and Chief Technology Officer (CTO) of Crucial Point LLC, a technology research and advisory firm.

Bob was named one of the top 25 most influential CTOs in the globe by *InfoWorld*, and one of DC's "Tech Titans" by *Washingtonian*. Bob was named one of the "Top 25 Most Fascinating Communicators in Government IT" by the Gov2.0 community GovFresh. In 2012 Bob was noted as "Most Influential on Twitter for Big Data" by *Forbes*.

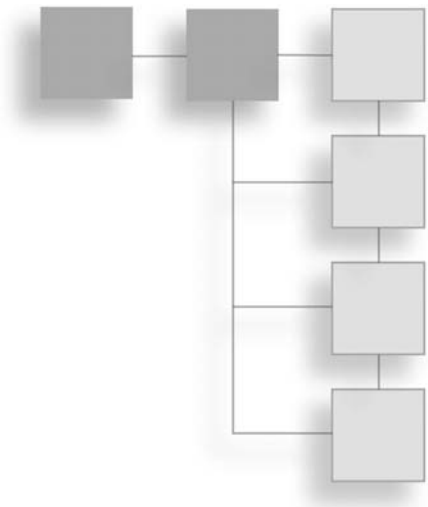
Bob holds three master's degrees including a master of science degree in scientific and technical intelligence from Naval Postgraduate School, a master of science degree in military science from USMC university, and a master of science degree in computer science from James Madison University. Bob has published over 40 articles on a wide range of topics and is a contributor to the January 2009 book titled *Threats in the Age of Obama*. His blog, CTOvision, is now ranked among the top federal technology blogs by *WashingtonTech*.

Bob is a founding member and member of the board of directors of the Cyber Conflict Studies Association, a non-profit group focused on enhancing the study of cyber conflict at leading academic institutions and furthering the ability of the nation to understand the complex dynamics of conflict in cyberspace. You can follow and chat with him on Twitter where his handle is @bobgourley. You can also find him on Twitter as @AnalystReport and @CTOvision and online at http://ctovision.com/pro.

Contents



Introduction



Amidst all the big data talk, articles, and conference speeches lies one consistently unanswered question: What can we actually do with big data? Sure, the answer is alluded to frequently but only in the vaguest and most general terms. Few spell out where to begin, let alone where to go with big data from there. Answers to related questions—from how to compute ROI for big data projects and monetize data to how to develop a winning strategy and ultimately how to wield analytics to transform entire organizations and industries—are even rarer. That's why *Data Divination* was written—to answer all of those most pressing questions and more from a high-level view.

This Book Is for You If

If you are interested in the business end of big data rather than the technical nuts and bolts, this book is for you. Whether your business is a one-man operation or a global empire, you'll find practical advice here on how and when to use big data to the greatest effect for your organization. It doesn't matter whether you are a data scientist, a department head, an attorney, a small business owner, a non-profit head, or a member of the C-Suite or company board, the information contained within these pages will enable you to apply big data techniques and decision-making to your tasks.

Further, many of the chapters are dedicated to use cases in specific industries to serve as practical guides to what is being and can be done in your sector and business. Ten industries are addressed in exquisite detail in their own chapters. There you'll find use cases, strategies, underlying factors, and emerging trends detailed for the governments,

department of defense and intelligence community, security, healthcare, small businesses and farms, transportation, energy, retail, banking and insurance, and manufacturing sectors. However, it is a mistake to read only the chapter on your own industry, as changes wrought by big data in other industries will also affect you, if they haven't already.

If there is one thing that big data is shaping up to be, it is a catalyst of disruption across the board. Indeed, it is helping meld entire industries in arguably the biggest surge of cross-industry convergence ever seen. It therefore behooves you to note which industries are converging with yours and which of your customers are reducing or eliminating a need for your services entirely. It's highly likely that you'll find more than a few surprises here in that regard.

Strategy Is Everything

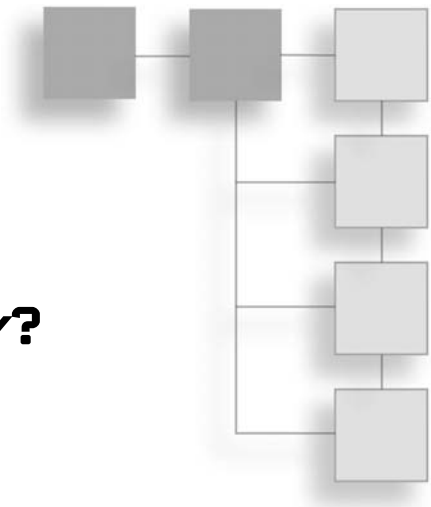
Data Divination is about how to develop a winning big data strategy and see it to fruition. You'll find chapters here dedicated to various topics aimed at that end. Included in these pages are the answers to how to calculate ROI; build a data team; devise data monetization; present a winning business proposition; formulate the right questions; derive actionable answers from analytics; predict the future for your business and industry; effectively deal with privacy issues; leverage visualizations for optimum data expressions; identify where, when, and how to innovate products and services; and how to transform your entire organization.

By the time you reach the end of this book, you should be able to readily identify what you need to do with big data, be that where to start or where to go next.

There are some references to tools here, but very few. Big data tools will age out over time, as all technologies do. However, your big data strategies will arch throughout time, morphing as needed, but holding true as the very foundation of your business. Strategy then is where you need to hold your focus and it is where you will find ours here.

From your strategy, you will know what tools to invest in and where and how you need to use them. But more than helping you pick the right tools and to increase your profits, your strategy will see you through sea changes that are approaching rapidly and cresting on the horizon now. The changes are many and they are unavoidable. Your only recourse is to prepare and to proactively select your path forward. We do our best to show you many of your options using big data in these pages to help you achieve all of that.

Chapter 1



What Is Big Data, Really?

One would think that, given how the phrase "big data" is on the tip of nearly every tongue and top of mind for most, everyone knows what big data is. That's not quite the case. Although there is a technical definition of sorts, most people are unsure of where the defining line is in terms of big versus regular data sizes. This creates some difficulty in communicating and thinking about big data in general and big data project parameters in particular.

This chapter considers the different interpretations of the meaning of the term "big data."

Technically Speaking

As discussed in more detail in the next chapter, big data does not mean more of the same data, simply boosting gigabytes to terabytes, although obviously it includes the expected growth of existing data sets. Rather, big data is a collection of data sets, some structured and some unstructured, some "onboarded" from physical sources to online sets, some transactional and some not, from a variety of sources, some in-house and some from third parties. Often it is stored in a variety of disparate and hard-to-reconcile forms. As a general rule, big data is clunky, messy, and hard, if not impossible, as well as significantly expensive, to shoe-horn into existing computing systems.

Furthermore, in the technical sense there is no widely accepted consensus as to the minimum size a data collective must measure to qualify as "big." Instead the technical world favors a definition more attuned to data characteristics and size relative to current computing capabilities.

You'll commonly hear big data defined as "containing volume, velocity, and variety" which is the three-legged definition coined by a 2001 Gartner (then Meta) report. These days, some people throw in a fourth "v,"—veracity—to cover data quality issues too.

But in essence big data is whatever size data set requires new tools in order to compute. Therefore, data considered big by today's standards will likely be considered small or average by future computing standards.

That is precisely why attaching the word "big" to data is unfortunate and not very useful. In the near future most industry experts expect the word big to be dropped entirely as it fails to accurately describe anything essential to the concept. For what makes "big data" truly valuable are the "big connections" it makes possible—between people, places, and things—that were previously impossible to glean in any coherent fashion.

Even so, there are those who try to affix a specific size to big data, generally in terms of terabytes. However this is not a static measurement. The measure generally refers to the amount of data flowing in or growing in the datacenter in a set timeframe, such as weekly. Conversely, since data is growing so quickly everywhere, at an estimated rate of 2,621,440 terabytes daily according to the Rackspace infographic in Figure 1.1, a static measurement for a "big data" set is frequently meaningless after a very short time. (This infographic can also be found online at http://www.rackspace.com/blog/exploring-the-universe-of-big-data-infographic/.)

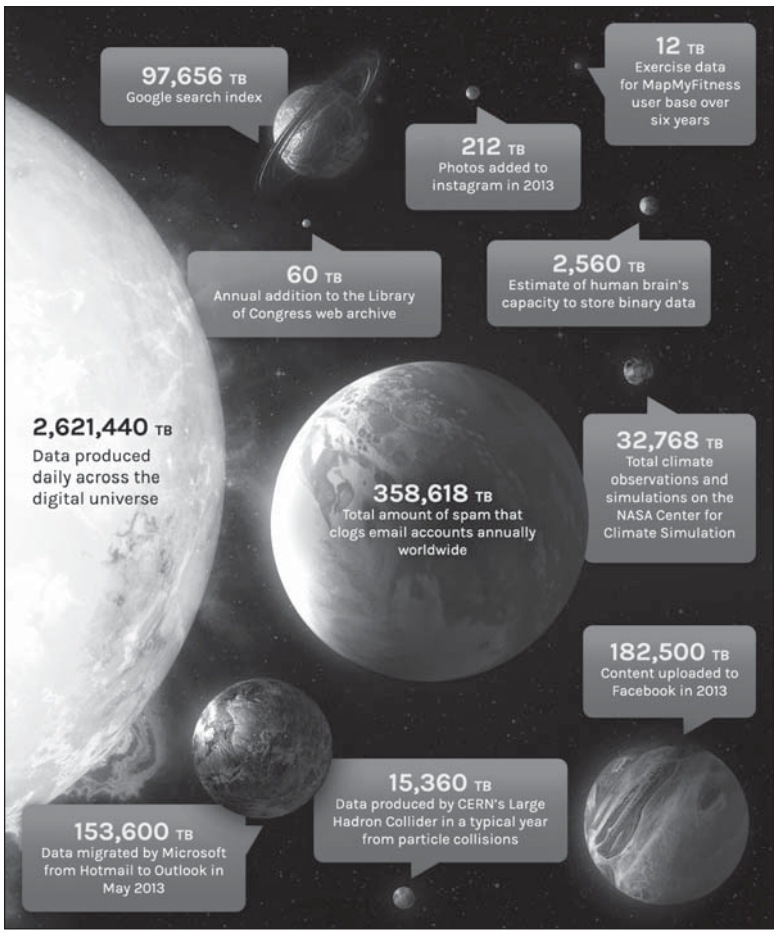


Figure 1.1
This interactive infographic has a counter at the top that shows how many terabytes of data were added to the digital universe since the user opened the infographic. The speed in which the counter counts gives you a good idea of just how fast data is growing overall. By rolling a mouse over the different planets, the user reveals the size of data in different categories relative to the size of all data generated (represented here as the sun) such as in email spam, in the Google search index, and in Facebook.

Source: Infographic courtesy of Rackspace. Concept and research by Dominic Smith; design and rendering by Legacy79.

Already we know that bigger data is coming. Data sets so big that we don't yet have a measuring term for it. But until then we'll use the measurements we do have: first up is zettabytes and then yottabytes. To give you an understanding of the magnitude of a yottabyte, consider that it equals one quadrillion gigabytes or one septillion bytes—that is a 1 followed by 24 zeroes. Consider Figure 1.2 for other ways to visualize the size of a yottabyte.

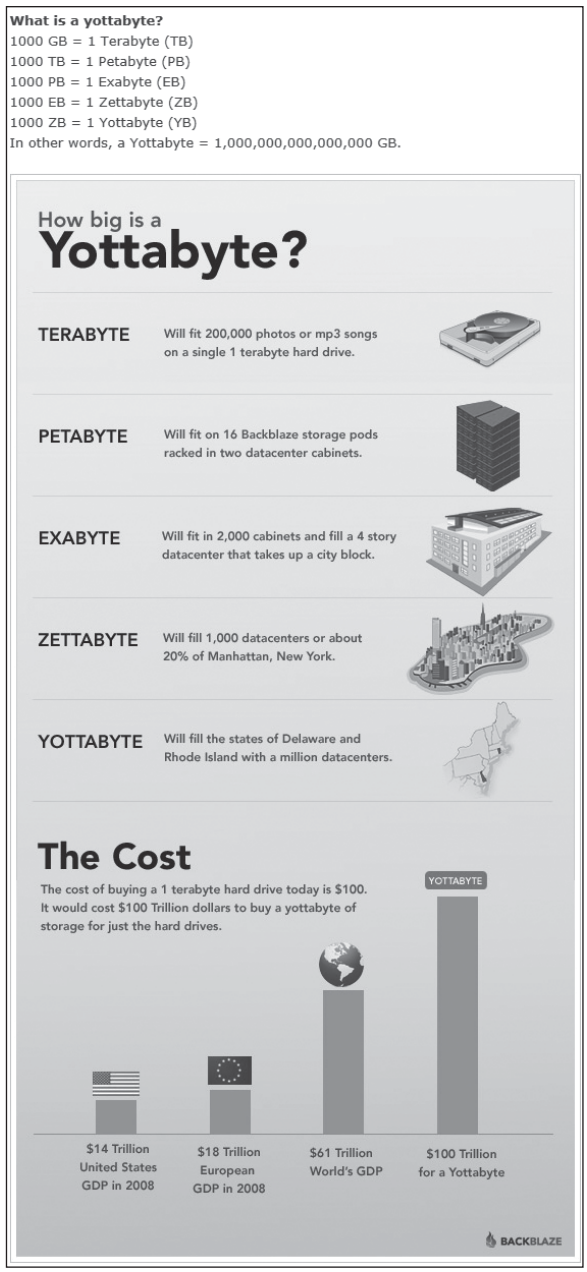


Figure 1.2
This graphic and accompanying text visualize the actual size of a yottabyte.
Source: Backblaze; see http://blog.backblaze.com/2009/11/12/nsa-might-want-some-backblaze-pods/.

As hard as that size is to imagine, think about what comes next. We have no word for the next size and therefore can barely comprehend what we can or should do with it all. It is, however, certain that extreme data will arrive soon.

Why Data Size Doesn't Matter

Therefore the focus today is primarily on how best to access and compute the data rather than how big it is. After all, the value is in the quality of the data analysis and not in its raw bulk.

Feel confused by all this? Rest assured, you are in good company. However, it is also a relief to learn that many new analytic tools can be used on data of nearly any size and on data collections of various levels of complexities and formats. That means data science teams can use big data tools to derive value from almost any data. That is good news indeed because the tools are both affordable and far more capable of fast (and valuable) analysis than their predecessors.

Your company will of course have to consider the size of its data sets in order to ultimately arrange and budget for storage, transfer, and other data management related realities. But as far as analytical results, data size doesn't much matter as long as you use a large enough data set to make the findings significant.

What Big Data Typically Means to Executives

Executives, depending on their personal level of data literacy, tend to view big data as somewhat mysterious but useful to varying degrees. Two opposing perceptions anchor each endpoint of the executive viewpoint spectrum. One end point views big data as a reveal all and tell everything tool whereas the other end of the spectrum sees it is simply as a newfangled way to deliver analysis on more of the same data they are accustomed to seeing in the old familiar spreadsheet. Even when presented with visualizations, the second group tends to perceive it, at least initially, as another form of the spreadsheet.

There are lots of other executive perceptions between these two extremes, of course. But it is useful for your purposes here to consider the two extremes—omniscience and spreadsheet upgrade—in order to quickly assess the executive expectations. This will better prepare you to deliver data findings in the manner most palatable and useful to your individual executives.

The "Data Is Omnipotent" Group

For the first group, it may be necessary to explain that while big data can and does produce results heretofore not possible, it is not, nor will it ever be, omniscience as is often depicted in many movies. In other words, data, no matter how huge and comprehensive, will never be complete and rarely in proper context. Therefore, it cannot be omnipotent.

This group also tends to misunderstand the limitations of predictive analytics. These are good tools in predicting future behavior and events, but they are not magical crystal balls that reveal a certain future. Predictive analytics predict the future assuming that current conditions and trends continue on the same path. That means that if anything occurs to disrupt that path or significantly change its course, the previous analysis from predictive analytics no longer applies. This is an important distinction that must be made clear to executives and data enthusiasts. Not only so that they use the information correctly but they also understand that their role in strategizing is not diminished or replaced by analytics, but greatly aided by it.

Further, most big data science teams are still working on rather basic projects and experiments, learning as they go. Most are simply unable to deliver complex projects yet. If executives have overly high initial expectations, they may be disappointed in these early stages. Disappointment can lead to executive disengagement and that bodes ill for data science teams and business heads. This can actually lead to scrapping big data projects and efforts all together. Therefore, it's important to properly and realistically manage executive expectations from the outset.

On the upside, executives in this group may be more open to suggestions on new ways to use data and be quicker to offer guidance on what information they most need to see. Such enthusiastic involvement and buy-in from executives is incredibly helpful to the initiative.

The "Data Is Just Another Spreadsheet" Group

At the other extreme end of the spectrum, the second group is likely to be unimpressed with big data beyond a mere nod to the idea that more data is good. This group views big data as a technical activity rather than as an essential business function.

Members of this executive group are likely to be more receptive to traditional visualizations, at least initially. To be of most assistance to this group of executives, ask outright

what information they wish they could know and why. Then, if they answer, you have a solid and welcomed way to demonstrate the value of the company's big data efforts by presenting exactly what was needed but heretofore missing.

If they can't or don't answer the question, work proactively to find ways to demonstrate the value of data analysis in ways that are meaningful to those executives.

Expect most executives to have little interest in how data is *cooked*—gathered, mixed, and analyzed. Typically they want to know its value over the traditional ways of doing things instead.

Whether executives belong to one of these two extreme groups or are somewhere in between, it is imperative to demonstrate the value of big data analysis as you would in any business case and/or present ongoing metrics as you would for any other technology.

However, your work with executives doesn't end there.

Big Data Positioned in Executive Speak

Although data visualizations have proven to be the fastest and most effective way to transfer data findings to the human brain, not everyone processes information in the same way. Common visualizations are the most readily understood by most people, but not always. Common visualizations include pie charts, bar graphs, line graphs, cumulative graphs, scatter plots, and other data representations used long before the advent of big data.

The most common of all is the traditional spreadsheet with little to no art elements. Figure 1.3 shows an example of a traditional spreadsheet.

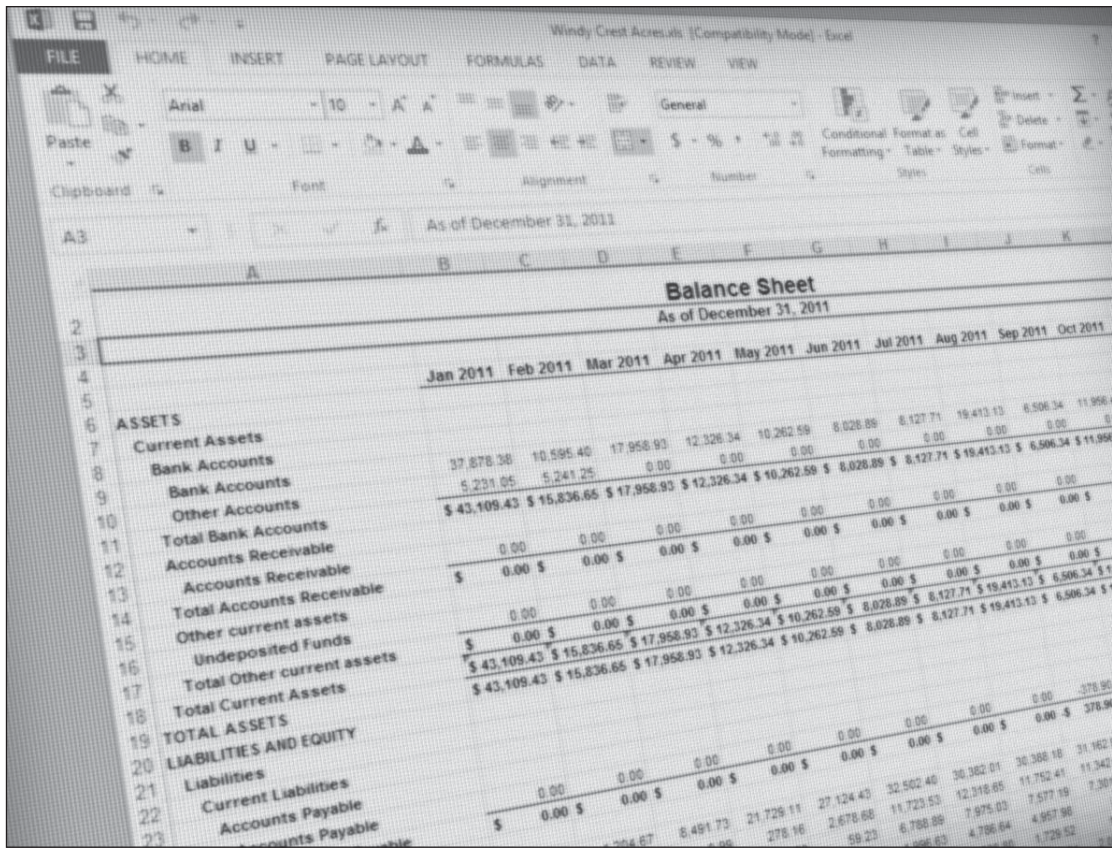


Figure 1.3
An example of a traditional spreadsheet with little to no art elements.
Source: Pam Baker.

Newer types of visualizations include interactive visualizations wherein more granular data is exposed as the user hovers a mouse or clicks on different areas in the visual; 3D visualizations that can be rotated on a computer screen for views from different angles and zoomed in to expose deeper information subsets; word clouds depicting the prominence of thoughts, ideas, or topics by word size; and other types of creative images.

Figure 1.4 is an example of an augmented reality image. Imagine using your phone, tablet, or wearable device and seeing your multi-dimensional data in an easy-to-understand form such as in this VisualCue tile. In this example, a waste management company is understanding the frequency, usage, and utility of their dump stations.

Figure 1.5 shows an example of a word cloud that quickly enables you to understand the prominence of ideas, thoughts, and occurrences as represented by word size. In this example, a word cloud was created on an iPad using the Infomous app to visualize news from several sites like FT, *Forbes, Fortune, The Economist,* The Street, and Yahoo! Finance. The size of the word denotes its degree of topic prominence in the news.

Figure 1.4
Augmented reality visualization. Imagine using your phone, tablet, or wearable device and seeing your multi-dimensional data in an easy to understand form such as in this VisualCue tile. In this example, a waste management company is understanding the frequency, usage, and utility of their dump stations.
Source: VisualCue™ Technologies LLC. Used with permission.