Big Data Applications and Opportunities for Librarians and Information Professionals

# The Accidental Data Scientist

Amy Affelt

Foreword by Thomas H. Davenport

# The Accidental Data Scientist

# Praise for *The Accidental Data Scientist*

"A generation ago, librarians and information specialists could be forgiven for not knowing the ways that the internet would disrupt their institutions and their professional lives. They don't have that excuse today as they face the next revolutions in the emergence of Big Data and the Internet of Things. Amy Affelt has spelled out the implications here in a volume that is as practical as it is wise."

—Lee Rainie, Director, Pew Research Center Internet Project

"If you're a librarian or information scientist, this book will introduce you to the key concepts and terminology you need to understand Big Data."

—Daniel Tunkelang, Head of Search Quality, LinkedIn

"A practical, easy-to-understand guide that lets librarians and information professionals leverage their transferable skills into Big Data knowledge and become expert guides on a career path that promises to be hot for years to come."

—Deborah Hunt, Past President, Special Libraries Association, and Director, Mechanics' Institute Library

"Librarians' skills, values, and knowledge around information management make us the perfect fit for the needs of the Big Data movement. Amy Affelt's timely book shows the how and why for librarians in an interesting and understandable way."

—Bonnie Tijerina, Fellow, Data and Society Research Institute

"A much-needed call to action to ensure that librarians retain their essential role as guides, curators, and knowledgeable experts as every aspect of our lives becomes increasingly data driven. I highly recommend this book as essential reading for anyone about to jump down the Big Data rabbit hole!"

—Rick Smolan, co-creator, *The Human Face of Big Data*

"A pressing call to action for librarians to engage with Big Data. … The rich sources Affelt draws upon enrich and expand our understanding of the world of Big Data and how we, as information professionals, can play an important role in a world of exponentially increasing amounts and varieties of data."

—Kathryn J. Deiss, Content Strategist, Association of College & Research Libraries, American Library Association

"Part librarian manifesto, part how-to guide, *The Accidental Data Scientist* makes the convincing case that librarians have been playing many critical roles expected of the modern and much-hyped data scientist, emphasizing the often undervalued importance of data verification and data integrity."

—Dr. Cathy O'Neil, Director, The Lede Program, Columbia University Graduate School of Journalism; data scientist, author, and daily blogger at mathbabe.org

Big Data Applications and Opportunities for Librarians and Information Professionals

# The Accidental Data Scientist

**Amy Affelt**

**Foreword by Thomas H. Davenport**

Information Today, Inc.
Medford, New Jersey

*First Printing*

***The Accidental Data Scientist***



**Library of Congress Cataloging-in-Publication Data**

Affelt, Amy L., 1970-
The accidental data scientist : big data applications and opportunities for librarians and information professionals / by Amy L. Affelt.
pages cm
Includes bibliographical references and index.
ISBN 978-1-57387-511-0
1. Librarians--Effect of technological innovations on. 2. Library science--Vocational guidance. 3. Information science--Vocational guidance. 4. Big data. 5. Data libraries. 6. Database searching. 7. Electronic information resource literacy. I. Title.
Z682.35.T43A38 2015
020.23--dc23

2014037859

Printed and bound in the United States of America

President and CEO: Thomas H. Hogan, Sr.
Editor-in-Chief and Publisher: John B. Bryans
Project Editor: Theresa Cramer
Associate Editor: Beverly Michaels
Production Manager: Norma Neimeister
Book Designer: Jennifer Burmood
Cover Designer: Ashlee Caruolo **www.infotoday.com**

For my father, Gerald, my mother, Carol,
and my husband, Michael

# Contents

# Figures and Tables

# Introduction

When I first read Thomas Davenport's October, 2012, *Harvard Business Review* article about the "new" profession of "data scientists," which he described as "high-ranking professionals with the training and curiosity to make discoveries in the world of big data,"[1] alarm bells went off in my mind. Training in discovery? Data? Curiosity? It sounded like all the crucial elements of librarianship to me. I started to panic as I recalled past information industry game-changers to which librarians came late.

When the task of searching company and industry news began to move from proprietary, fee-based databases to internet websites I felt that, in many cases, librarians missed the opportunity to remain front and center as the professionals who are best at finding, evaluating, and most importantly, presenting content, regardless of platform. Even though we have high-level expertise at locating information, librarians somehow sat back and watched as the general public became well-versed in what had been our exclusive domain. Seemingly overnight everyone became what a lot of librarians called themselves at the time: online searchers. Our business is information, and all of a sudden there was an explosion of information readily available to anyone with a phone line. Somehow that translated to information being free, and the internet became a threat to professionals who are experts at finding and applying information, and more importantly, knowledge.

Why did librarians fall prey to a phenomenon we could have easily turned around and used to our advantage? Why were we not first out of the gate to caution people that the internet is not free and its information is not always reliable? Why didn't we position ourselves as critically necessary to finding information that will never appear in a Google search results list?

Our expertise as a profession was suddenly under fire. People were asking, "Why do we need librarians when people can just find things themselves on the internet?" This line of inquiry was especially pervasive when budgets and financial expenditures were discussed. Other professions did not have this problem. Consider insurance agents. Stalwart insurance companies like State Farm and Allstate, along with upstarts such as Geico, are thriving in the internet marketplace which allows them to reach consumers more directly than ever before. When you can spend 15 minutes buying a policy online, it would seem that the insurance agent would be obsolete. The same scenario holds true for even supposed dinosaur industries such as travel agencies. But instead of giving up and finding new lines of work, insurance and travel agents have rebranded themselves as the authorities in their respective professions, catering to people with complex insurance and travel needs—and providing a level of convenience that many clients still find worthwhile.

We also found ourselves on the outside when the field of Knowledge Management appeared. Strategic knowledge—and internal strategic knowledge in particular (that is, knowledge that is proprietary or uniquely possessed by individuals within a firm)—is one of the most valuable assets possessed by a corporation. Yet, without highly skilled individuals who specialize in locating, documenting, organizing, and curating that knowledge, that information is worthless.

Although many librarians and information professionals now work in knowledge management, when the field first rose to prominence, enterprises first looked to computer programmers and information technology professionals when hiring experts to develop and staff their KM initiatives and work groups. In a similar vein, storing, indexing, and making records and digital assets easily findable are information intensive activities and a natural fit with library and information science skills. However, records

departments and digital asset management departments are usually staffed by non-librarians.

Why are librarians and their skills not first and foremost in the minds of senior executives when they look to get involved in these new fields? How is it that we keep missing these opportunities to transfer our skill sets and expand our career opportunities into cutting edge technologies? I find it extremely frustrating when our skills and abilities are co-opted in this way. What is unique about our profession that we allow this to happen?

That brings us back to the data scientists, who will be working with what is, of course, now known in the popular lexicon as Big Data. As I read more and more about Big Data, the same thought keeps recurring in my mind: I don't want history to repeat itself. Who are these data scientists who are so well-versed in determining which data to collate, use, and analyze? Who understands how to separate the wheat from the chaff, and parse out quality information in order to enable strategic decision-making that will advance the mission of the organization? Who are the people who can thoroughly and critically review data and carefully use it to truly affect an organization's bottom line? Librarians!

I am on a personal mission to ensure that librarians and information professionals are first and foremost in the minds of those making hiring decisions for Big Data plans, positions, and projects. It is my hope that after reading this book, you will join me in that mission.

Regardless of your age, gender, profession, or location—whether or not you are aware of it—Big Data plays a prominent role in your life. The sheer amount of data that is being collected—from every possible source, and of every possible variety—has exploded in recent years, and it is growing at an unfathomable rate. Literally every click on every website can be considered data that a third party might want to monitor or collect. Social media behavior, such as Facebook posts and Twitter tweets, is collated and stored. Data

gleaned from smartphone usage—including everything from geospatial location data assigned to pictures to the choice of internet browsers and search engines used when surfing the web—is tracked and compiled by telecommunications carriers.

Big Data is also collected independent of an individual's internet usage on a personal computer, tablet, or smartphone. Sensors on everyday objects such as vending machines, global positioning systems, public transit turnstiles, exercise equipment, Automatic Teller Machines, bar code scanners, and a seemingly endless host of other devices collect a constant stream of data known as the "Internet of Things." By 2020, there will be 26 billion units collecting Internet of Things data, according to the Gartner Group[2].

Our entertainment choices also add to the flow of Big Data. Netflix, satellite, and cable television companies closely observe viewing patterns and decision-making in order to predict future subject matter interest. They are using this data to delve into a new arena of provider-produced programming. Traditionally, data has been collected by marketers for use in bolstering sales, but now revelations about the National Security Agency have revealed that data regarding internet usage patterns and behavior are also being collected by the United States and foreign governments.

For corporations, law firms, governments, and nonprofits, Big Data presents many obstacles and many opportunities. It is an unprecedented way to track consumer behavior, so revelatory that it is impossible to ignore, but at the same time, it requires massive amounts of storage capability, programming, and statistical expertise, as well as careful attention to the legalities and privacy issues that can arise. The promise—and the peril—of Big Data is foremost in the minds of directors and CEOs. They are wondering how to find the golden needles in the haystacks of information that will affect their relevancy and revenue, and ultimately, their ability to fulfill the mission of their organizations and strategic plans for the

future. They are looking to hire these data scientists, which Davenport has called "the sexiest job of the 21st century."

Big Data has become a ubiquitous, yet nebulous, term. News websites, blogs, Twitter feeds, and social media are fast and loose in characterizing almost all data as Big Data. In addition, there are almost as many different definitions of Big Data as there are references to it. Not only are there sessions focused on Big Data at conferences of all stripes, but there are now entire conferences on Big Data itself. The applications are interesting, the predictions based on it are endless, and the future ramifications of its insights can be both mind-numbing and mind-boggling.

With myriad streams of data running in different directions, the average individual can easily become lost and confused. Traditional mathematicians, statisticians, and computer programmers are expert at processing this data, but guides are needed to navigate this new information superhighway of data to carve out the best route, avoid dead ends, and ensure that important detours and landmarks are not missed along the way. The best route is the data that lands the client, wins the case, or obtains the patent for the firm. The dead ends are full of data that is collected but may not be reliable, reputable, or relevant. Most important of all are the detours and landmarks that the "numbers people" might miss; the important but subtle data that is often overlooked but that contains clues for solving the business problem. It is the data that other firms might not recognize as important, but is the very data that gives one company an edge that helps it to catapult over the competition.

Librarians can be expert guides along the Big Data superhighway. Basic skills that are part and parcel of every Library and Information Science (LIS) curriculum are the competencies that are needed in order to navigate the Big Data rabbit hole. Reference interviewing is one of the keys to LIS programs. In fact, it is usually one of the first courses that we take when beginning an LIS program. Almost from day one, we learn how to ask the right questions

of our clients in order to get at the real question, or what they really need to know. As any librarian can attest, it is often not what is initially asked. The reference interview is key when discussing a research project involving data. We need to know the exact issue that is being considered in order to determine which data is best to obtain and analyze.

Indexing and abstracting skills are also imperative when working with Big Data. Datasets compiled by corporations can be extremely valuable, not only for use in research and analysis, but also as proprietary intellectual property. However, if that data is not easy to locate using controlled yet intuitive vocabulary, it is virtually worthless. Librarians, being well-versed in metadata and taxonomy skills, are the best people to organize these datasets using the terminology of the industry in which we work, allowing easy retrieval of the datapoints critical to a research project.

The most important reason that librarians and information professionals should be hired to play key roles in Big Data is that we are excellent at storytelling, and storytelling is one of the most important—if not the most important—ways of understanding the insights uncovered by a Big Data project.

Data, on its own, can be extremely confusing. When the amount of data is more massive than any you've previously worked with and is in unusual formats, decision makers can find it difficult to understand the insights uncovered. When massive and complex datasets are being used for problem solving and decision making, librarians are the people who are best able to carefully consider the data and explain what it does and does not show.

We will highlight patterns in the data, and might caution against using them to make predictions. We will find coincidences in the data, while cautioning that correlation is not always the same as causation. We will place all of these observations in the context of other non-data factors that may or may not affect conditions. How will we do this in a way that makes sense to our stakeholders?

We will do what we do best. We will write a compelling narrative that engages decision makers and helps them to understand the data in the context of real solutions to real problems, which is fundamental to our mission as librarians.

Whether you are a public librarian helping someone to navigate the Affordable Care Act, or a corporate librarian analyzing the pros and cons of a multi-million dollar deal, we all engage in the same activity. We solve problems using information, and our ability to solve those problems is stellar thanks to the skills that we bring to the table. The format of the information we use will always remain secondary to our ability to turn information into knowledge that answers questions, which allows individuals, companies, and governments to make better decisions.

A major part of this book consists of a discussion of cutting-edge and innovative Big Data projects undertaken within industries of all types and by myriad organizations. These projects and initiatives are the best way to understand how Big Data can be used and applied. I have discussed these applications in numerous conference presentations, and there is a commonality among all the audiences I've encountered. Everyone wonders, while all of these projects and initiatives are extremely interesting and compelling, how do we actually work with and use Big Data in our everyday jobs as librarians and information professionals? This book will show you how.

A large portion of this text consists of real-life reference scenarios that could play out in any library or information center. They are situation-based research problems and challenges that not only require the use of data, but also require critical thinking and analytical skills used in tandem with creativity, and the ability to communicate in a clear, concise way to reach a successful solution.

The title of this book is The Accidental Data Scientist, but the ability of librarians and information professionals to work in data science is not random or serendipitous. When we find ourselves working with data, we know that this circumstance did not arise

by chance. Further, the career opportunities afforded to us in this field should not be unexpected. A LIS degree has always provided those who choose to pursue it with a highly diverse skill set that is transferable to a wide variety of jobs and settings. Indeed, Special Libraries Association surveys have found that its members have hundreds of different job titles; yet their collective mission remains the same: to solve problems and improve the quality of life by locating and harnessing reliable information and transforming it into concrete knowledge at the point of need. This book is for you if you:

- Find yourself wondering "What's the big deal about Big Data?"
- Want to acquire a rudimentary understanding of the terminology and language needed to "speak Big Data"
- Are interested in working with Big Data in order to enhance your core skills and move into nontraditional roles in either your current field or a new one
- Are a LIS student who wants to learn about job opportunities in Big Data and who needs to know about the coursework that will help you get there
- Would like to experiment with free or low-cost Big Data software and algorithms
- Need to determine which types of information-intensive projects require the use of data and how you can locate the data you need
- Would like to incorporate data analysis into everyday reference requests but are not certain how to get started
- Want to learn how to separate "junk data" from high quality, citable data
- Want to learn about cutting-edge Big Data projects and initiatives that will, for better or worse, affect the future of each one of us

- Want to find out which skills to acquire to ensure that your resume gets noticed when you apply for data science jobs
- Would like to be able to demonstrate to management how your skill set in this area can contribute value to your organization

As a profession, we have largely survived the disruptions of the ubiquity of internet searching and late starts in working in KM, records, and digital asset management. As many librarians can attest, the information overload that resulted from the explosion of the World Wide Web did indeed fulfill predictions that end-user searching would make our skills even more important. Early database providers such as LexisNexis, in bids to convince information center managers to put their software on end-user desktops, always stated that in doing so, librarians would be making themselves indispensable. These end users would need help choosing sources, determining search syntax, and reviewing results. They promised that allowing our requestors access to databases would result in more in-depth research project requests, which would allow us to showcase our creativity and our skills in analytics and writing. As someone who worked in online database research during this nascent stage, I can assure you that that is exactly how the situation evolved. The same will be true of our roles in the Big Data explosion.

On April 9, 2014, a *Wall Street Journal* headline screamed, "Get Familiar with Big Data Now—Or Face 'Permanent Pink Slip.'"[3] Reading it was "déjà vu all over again" (hat tip to Yogi Berra). Threatening headlines that portend our demise are nothing new to librarians. Back in 1957, in the movie *Desk Set*[4], television station librarian Katharine Hepburn fretted that she would lose her job to a computer. The merits of "going Dewey-free," have been debated since the 1960s and continue today. In the late 1990s and early 2000s, there was a ubiquitous siren call that the internet and ebooks would eliminate libraries altogether. For librarians, Big Data is

another tool in our arsenal that we can use to expand our career opportunities and better position ourselves as key employees in our organizations.

We need to learn to unleash the power of our ability and embrace the inevitable growth of data and see this Big Data not as a disruption, but instead as the development of a new field for which we are uniquely positioned not only to survive but to thrive. Big Data is not a new field for us, but since it is a new field for others, we can embrace it as such. The trick is viewing Big Data—and the perils and promise that come with it—with a realistic understanding of what it can and cannot do. This book will help you to master that understanding and show you how to help others in your organization reach a better understanding as well.

Whether you choose to learn completely new skills, or apply those that you already use in new and revolutionary ways, I'm glad you're here. Thank you for joining me on the accidental journey to becoming a data scientist. Together, we can turn Big Data into not just better data, but the best data for whatever situation we find ourselves working in—accidentally or otherwise.

## Endnotes

1. Thomas H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, October 2012. Accessed November 26, 2013, http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
2. "Gartner Says the Internet of Things Will Transform the Data Center," Gartner website, March 18, 2014, accessed May 30, 2014, http://www.gartner.com/newsroom/id/2684915
3. Nikki Waller, "Get Familiar with Big Data Now, Or Face 'Permanent Pink Slip,'" *Wall Street Journal*, April 9, 2014, accessed April 9, 2014 http://online.wsj.com/news/articles/SB10001424052702304819004579489541746990638?mg=reno64-wsj
4. *Desk Set,* Internet Movie Database, accessed April 9, 2014, http://www.imdb.com/title/tt0050307/

*Disclaimer: The views and opinions expressed in this book are the author's and do not necessarily reflect the views or opinions of her employer or the publisher.*

Chapter 1

# Big Data: Everything Old is New Again

## Big Data Hits the Big Time

Big Data has become an often used—and possibly over-used—term not only in the scientific and industry press, but also in the popular media. It began as a term of art used by computer code writers and mainframe network administrators, but very quickly became a commonplace phrase with which even the most casual consumers of mainstream media are now familiar. One of the first indications that Big Data had hit the big time was when the *Oxford English Dictionary (OED)* added Big Data to its quarterly update in June 2013.[1] The *OED's* goal is to "tell the history of the English language," adding new words, expressions, and phrases which its editors find significant.[2] The addition caused a surge in discussion of Big Data, as Oxford updates are always newsworthy, but their definition is a bit disappointing: "data sets that are too large and complex to manipulate or interrogate with standard methods or tools."[3] As I will discuss throughout this book, Big Data is much more rich, valuable, and interesting than something thus characterized as gigantic statistics that are rendered useless unless unconventional tools and programming methodologies are applied to them.

While *OED* recognition is a huge indicator of the relevance of a new buzzword, I wanted to explore the term Big Data as used by journalists writing for consumers of mainstream media. Therefore, to try to understand the frequency with which the term is used and how rapidly it has been adopted by journalists, I decided to conduct a historical search for the phrase in a comprehensive news and

information database. I chose to use Dow Jones Factiva to conduct this search because of its range of sources (over 1,000), the countries and languages represented (over 200 and 28, respectively), and the fact that it contains an archive going back almost 35 years.[4] My initial search of all publications in Factiva on November 27, 2013, for the phrase "Big Data" resulted in 75,514 articles returned. This set of articles established the fact that Big Data is definitely both a term of art and a casual phrase used by journalists with an assumption that it does not need to be defined when it appears in a story.

Next, I wanted to find the tipping point for the rise in frequency of use of the term "Big Data." One might expect that there would be a certain date after which Big Data would be used on a daily basis by journalists all over the world. As you will see in Figure 1.1, Factiva was able to show that over 50,000 of the original 75,514 articles were from 2012–2013.

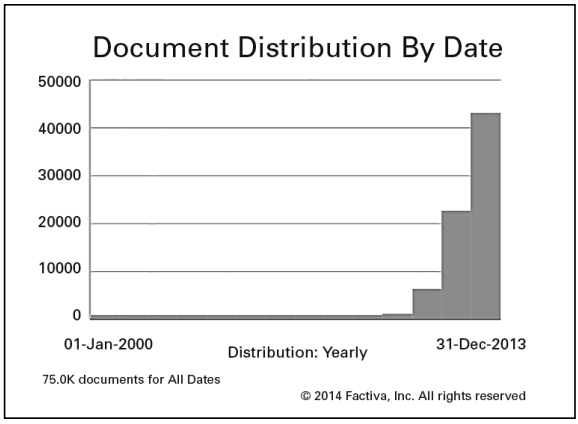



Figure 1.1 Frequency of the term Big Data: Document distribution by date

In 2011 there were 6,010 articles mentioning Big Data. Between 2012 and 2013, however, the number of articles mentioning Big Data nearly doubled (22,510 in 2012 and 43,122 between January 1 and November 27 of 2013.) The exact tipping point can be debated,

but these search results demonstrate that the use of the term Big Data exploded in 2012 and became very widespread in 2013.

The *OED* and Dow Jones aside, what really convinced me of the commonplace usage of the term Big Data was when I first read it in the most unexpected of places, my hometown local newspaper. The *News Tribune* of LaSalle, Illinois serves a cluster of rural communities ranging in population from a few hundred to a few thousand.[5] Typical articles discuss the yearly corn crop and the current price of gasoline. However, in the past six months I have seen more than one article with a headline mentioning Big Data. Granted, these articles were written in the wake of revelations regarding the United States' National Security Agency monitoring initiatives and focused on cybersecurity and privacy concerns, but I can say with confidence that Big Data has truly hit the big time if it is of interest to rural Midwesterners.

This omnipresence, however, hasn't lessened the confusion regarding how to define Big Data, or how to understand it. In a *Wall Street Journal* poll of prominent business executives that asked, "Which Buzzwords Would You Ban in 2014?" one of the respondents chose "Big Data," stating that "A lot of companies talk about it but not many know what it is."[6]

It is only a slight exaggeration to state that there are almost as many definitions of Big Data as there are datapoints in Big Data initiatives. A Harris Interactive study conducted in 2012 found that 28 percent of C-suite executives surveyed defined Big Data as "massive growth in transaction data." Twenty-four percent of respondents agreed with the statement, "it consists of new technologies that address the volume, variety, and velocity changes of the data itself." Nineteen percent responded that it "refers to requirements to store and archive data for regulatory compliance," while eighteen percent viewed it as a phenomenon involving "the rise in new data sources," such as social media, mobile, and apps.[7]

## You Know It When You See It

As I write and talk about Big Data, a common question I hear is "What exactly is Big Data?" My first answer is always a tongue-in-cheek reference to United States Supreme Court Justice Potter Stewart's definition of obscenity in *Jacobellis v. Ohio,* 378 U.S. 174 (1964), "You know it when you see it."[8] Jacobellis v. Ohio involved a movie theater's rights under the First Amendment to show a film *(The Lovers)* that was banned by the State of Ohio, which judged it to be "obscene." The court overturned lower court decisions banning the film, ultimately determining that the movie was not obscene. The justices were unable to agree on a definition of obscenity, much like executives are now unable to agree on a definition of Big Data.

This disagreement resulted in the case yielding four separate opinions, the most famous of which is by Justice Potter Stewart. Justice Stewart ultimately concurred that the U.S. Constitution protected all material with the exception of "hard-core pornography." Justice Stewart wrote that he would not even attempt to try to define "the kinds of material" that would be included in that category. Echoing some of the frustrations that can arise in trying to define Big Data, he stated that perhaps he "might never succeed in intelligibly doing so." This admission was followed by the famous quote that has become a popular turn of phrase for all things vague: "I know it when I see it."

If the data involved in a particular project is massive, untidy, in an unusual format, of dubious origin, or is a type of data that previously went uncollected, I am certain that you too, will know it when you see it. Librarians have a long history with data sets big and small, making them the Justice Stewart of a burgeoning industry—ready to spot Big Data when they see it.

## Old Friends in a New Package

Raw data has always been an integral part of the work of librarians and information professionals in all types of professional settings.

Public, academic, and special libraries are similar in that data is used in almost every role in the institution. It is used in the reference department to answer most queries presented. Data from the circulation desk provides metrics that are used to plan future orders, gauge subject matter interest, and quantify library usage. Catalogers, indexers, and abstracters in the technical services department translate the written word into datapoints that are used to quickly locate and retrieve needed materials. Systems librarians work with the myriad data involved in managing libraries' computer systems—everything from the number of clicks on the library's website or intranet, including differentiation of the clicks on different types of content, to the number of logins by individuals by time of day, falls under their purview. Administrators use data for everything from budgeting and strategic planning to goal setting and employee performance appraisal.

Special librarians working in business and finance can probably remember the days of Lotus 1-2-3, a pioneering spreadsheet software package that had a maximum capacity of 65,536 rows per sheet.[9] I would love to have a dollar for every time I crashed Lotus 1-2-3 because the dataset I was working with was too large. Big Data is a new term coined by the media and technology industries, but it is not a new concept to librarians. We have always worked with large amounts of data. So how is Big Data different from the data with which we have always worked?

The data that librarians worked with in the past was, for the most part, stored in relational databases and was largely in traditional formats. The main tenet of this data is that it was viewed in the context of past activities. For example, at the reference desk, this meant looking up historical facts or events. At the circulation desk and in the systems department, we tracked patrons' past usage. Technical services staff made taxonomy decisions based on previously determined coding, and administrators looked to historical data to plan for the future.

In contrast, Big Data is often viewed in the context of the future. The data itself may be tabulated in real time, but it is evaluated after it is quantified, and it is used to make predictions about the future and to help users map out pathways to solve problems and avoid past mistakes.

The McKinsey Global Institute issued a study in May, 2011, that looked at Big Data within the context of five industries: healthcare, government, retail, manufacturing, and geography. It was determined that 15 or 17 industry sectors in the United States currently have more data stored per company than the total storage of the Library of Congress. McKinsey also estimated that the amount of data is growing by 40 percent per year, and will increase 44 fold between 2009 and 2020. While 5 percent of this data is traditional, 95 percent of it is internally stored and is of an unstructured nature.[10] This data consists of items such as:

- Server log files created by employees using computer hardware in organizations
- Content generated by members of social media such as Twitter and Facebook (industry reports estimate over 2 billion registered users of these sites, generating over 8 terabytes of data on a daily basis)[11]
- Digital images, whether they are uploaded by individuals posting on platforms such as Instagram, or by third-party devices such as police and security cameras
- Smartphone geospatial location data (18 percent of Americans own a smartphone)[12]
- "Internet of Things" data
- Highly personal data, such as that obtained through the U.S. Department of Homeland Security's "Trusted Traveler" program

- Random data that can, at first glance, seem inconsequential (data that was "previously dropped on the floor")[13]
- Video, where the lack of controlled vocabulary and taxonomy, coupled with a dearth of tools for visual and image search, make locating specific videos a hit-or-miss activity

## The Proliferation of Social Data

When Chloe Sladden, director of content and programming at Twitter, declared Twitter "the new newswire," at Stanford University's 2012 Future of Media conference,[14] it was more of a future prediction than a truism. Fast forward to 2014, however, and that statement could not be more accurate. A 2013 Pew Research Center survey found that 31 percent of participants questioned had abandoned a traditional news outlet such as a newspaper or magazine (both print and online) because it "no longer provides the news and information they are accustomed to."[15] This does not mean these respondents are less interested in news itself. Indeed, when data from weather satellites is mashed up with Twitter streams, insightful information on developing conditions can be discovered in real time. For example, if a weather satellite indicates there is a storm beginning in one part of the world, analysis of tweets from users in that location can be used to track the impact of the storm on the local population and the veracity of weather predictions.[16] Although Twitter keeps a tight rein on its usage statistics and number of registered users, in 2013 it reported 232 million active users, almost double the number reported in 2012.[17]

This number is certain to increase significantly in light of a recent decision by the U.S. Securities and Exchange Commission to allow publicly traded companies to disclose material findings and announce market-moving and other time-sensitive information, such as earnings guidance, via social media outlets.[18] This

announcement proved to be a major game-changer because the "first mention" of a company's activities can now be revealed via Twitter, which means that Twitter will become a must-have tool for anyone who needs to keep track of the financial markets—everyone from librarians and information professionals working in business or finance to investment bankers, hedge fund managers, and corporate counsel.[19]

The influx of tweets from this whole new set of users will vastly add to the archive of Big Data housed in the Twitter archives, and it is these archival tweets and other social media postings that are sometimes the hardest data to locate. If the items are changed or deleted, archival search methods using internet search engines must be used to locate them, and these efforts are met with varying degrees of success. Another confounding factor is erroneous tweets from hacked or fake accounts.

In an infamous episode, on April 23, 2013, the Associated Press's Twitter account was hacked and used to falsely tweet about an explosion at the White House and President Obama being hurt. The Tweet read, "Breaking: Two Explosions in the White House and Barack Obama is injured." Fortunately, the hackers were discovered very quickly and conventional news media reported the story.

A close look at the original tweet showed obvious inconsistencies. There was no corroboration. Most reputable news organizations refer to President Obama as "President Obama," not "Barack Obama." "Breaking" was not typed in all caps, which is standard in AP tweets, and there was no attribution such as "sources report" or "officials say."[20]

The problem for researchers lies in how false tweets like these will be curated. Will they be available to be searched for and retrieved in the future? If they are found, will we know they were false tweets from hacked accounts? Will we need to build separate data storage archives for tweets, with flags for real and fake? Figuring all of that out is definitely a job for a data scientist!

## The Five V's of Big Data

The Gartner Group characterizes Big Data by "the three V's": Volume, Velocity, and Variety.[21] Volume refers to the sheer amount of data being collected. McKinsey described Big Data's massive volume as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."[22] This description implies that for corporations looking to capitalize on all that Big Data has to offer in terms of meeting the needs of their clients in order to boost revenue, new (and expensive) computer storage infrastructure will have to be implemented.

The velocity of Big Data refers to the speed at which the data is being created, transferred, delivered, and collected. Twitter is an excellent illustration of Big Data's velocity. Tweets are delivered in real time, creating an almost limitless reservoir of user-generated content. Tweets from corporations and sponsored tweets add to this archive. The same is true of mobile device activity. Regardless of the content being created, as users deliver information at the time of their choosing, they are constantly adding new data to whichever platform—Facebook, Gmail, Instagram, or just about any other app—they choose to populate. Readers of a certain age might remember that television and radio broadcasts used to be released on thirty-second delay, which provided a short window of opportunity to censor the content. The delay is long gone now, as transactions are recorded immediately, whether they originate from a smartphone, desktop, tablet, or strategically placed sensor. The phrase "thirty-second delay" is now obsolete, sitting on a dusty shelf with other relics like "broken record." (note: online card catalogs are still called card catalogs)

As I previously discussed, the sky is the limit with regard to variety, or the third V of Big Data. In compiling the preceding list of formats, I took a chance, knowing that the list will probably be outdated by the time this book goes to print. It is difficult even to imagine the format, shapes, and types of data that will exist in the

future. We can speculate that mobile usage will play a huge role in generating the content, and indeed, mobile internet traffic doubled between 2012 and 2013,[23] but the exact shape it will take is anyone's guess. To apply an often-used expression, the only constant in the variety of Big Data is the fact that it is ever changing, with new formats being added almost overnight.

### Two V's for Librarians and Info Pros

In addition to the three V's described by Gartner, there are two additional V's that are also applicable to Big Data. These two descriptors are of particular interest to librarians and information professionals because they present true opportunities for us to apply librarianship skills in working with Big Data.

Verification of Big Data refers to the process by which librarians and information professionals analyze data sources and retrieval systems to determine data quality. When working with data, we can apply the same evaluation skills we use when judging the integrity of any kind of information and its sources in order to help our constituents determine whether or not data is "clean," or uncorrupted from its original source, and whether the producer or distributor of the data is reputable and can be cited as a reference. The sidebar on page 21 is a Data Verification Consideration Checklist to use when evaluating data.

The data discrepancy check is the most critical step in the verification process and one for which librarians are particularly skilled. It cannot be overemphasized that in many cases a seemingly infinitesimal percentage change can result in vastly different conclusions in the analysis of the data. Librarians understand this and can serve as critical voices of caution and skepticism when working on data project teams.

It is well documented that what seem to be small issues can have tragic life-or-death consequences. For example, the January 28, 1986, Space Shuttle Challenger disaster—in which the space shuttle broke apart 73 seconds into flight, killing all seven crew

### Data Verification Consideration Checklist

1. What is the source of the data? Is it a government or nonprofit? Is the data being purchased from a vendor? If it is from a vendor that is unfamiliar to you, investigate the vendor's reputation. Have others used data from the source and found that it was flawed? Additionally, conduct searches in the literature for the industry in which the project is being conducted to see if other research cites the vendor as a source. By identifying other projects that have used the same source you will not only be able to establish the source's credibility, but you might also discover other ways in which the data you are looking at can be applied.
2. Is the data being used in the original format in which it was downloaded? Is it programmed or transposed? If so, discuss the programming process with the requestor in order to ensure that the underlying data is not inadvertently being changed.
3. Is it possible that other datapoints being used will affect the data that you are examining? If so, discuss these issues with the project manager to ensure that there are not any conflicts that would make the data inapplicable.
4. Are there other sources of the same data? If so, look at the data from these other sources as well. Are the numbers the same or different? If they are different, how great are the differences? Review any discrepancies in data with the project manager.

members aboard—was caused by the failure of an O-ring seal in the rocket booster, which disabled joints and ultimately led to the failure of the external fuel tank and the breakup of the orbiter.[24] In 1985, engineers from the National Aeronautics and Space Administration's (NASA) Marshall Space Flight Center wrote scientific

papers stating that joints sealed by the O-rings should be built with an additional three inches of steel, which would have reinforced these joints and prevented them from rotating, thus averting the chain of breakdowns that ultimately led to the disaster.[25]

Unfortunately, this finding was not communicated strongly enough to halt flights and prompt a redesign of the shuttle, but the Marshall engineers knew critical data when they saw it. To a non-engineer, three inches does not seem very long. To a NASA engineer, however, three inches is huge. I am not certain that librarians would have been able to stop the NASA powers-that-be from going forward with the launch, but I am certain that if they were involved in these projects, they would have emphasized the importance of these additional three inches of steel.

Problems with data quality do not need to have such dramatic effects to lead to serious problems for companies that can affect revenue and the bottom line. Data that is formatted in different ways can result in duplication or missed connections. For example, in a dataset of customer information used for sales and marketing, transposition of numbers in a street address or inconsistent entry of customer names (some entries containing middle initials, and some not, or the random use of nicknames, for example) can lead to incomplete results and failure to reach current or potential clients.

It is difficult for a company to have accurate revenue projections and earnings forecasts if it is working with data that is inconsistent. It can also be quite costly for companies to work with flawed customer data. Depending on the size of the customer base, potentially thousands or even hundreds of thousands of dollars in mailing costs could be wasted if the communications are being sent to incorrect customer names or addresses.

The thorough examination of possible data flaws can be one of the main responsibilities of librarians working on data projects. The statisticians and project managers using the data will make the final determinations regarding the data's usability, but librarians

can play the important role of providing the analysis of the data quality. The preceding Data Verification Consideration Checklist is a template that can be used as a starting point in every project when deciding which data to use. For information professionals, source and data quality is one of the utmost concerns, and because the evaluation of sources is second nature to us it is easy to forget that not everyone has this focus—or is even aware that sources should be judged and scrutinized. It is critically important that, when working on Big Data projects, we keep the above concerns first and foremost in our own minds, and in the minds of our constituents when we discuss project data with them.

The fifth V of Big Data, and the second that showcases unique librarianship skills, is value. Deriving true value from data is very difficult for three reasons: it is challenging, it is expensive, and it is risky.

A June 2013 Gartner study identified the following challenges in working with Big Data, based on a survey of 720 IT and business leaders:[26]

- Determining how to get value from the data
- Determining data strategy
- Hiring data scientists
- Integrating new platforms into existing IT architecture
- IT infrastructure issues

I found this list to be extremely helpful when considering possible roles for librarians who want to work with Big Data. Deriving value from data is challenging primarily because of the reasons outlined in the verification checklist above. If data is falsified, corrupted, fabricated, or simply not applicable to the project at hand, it has zero value. It is also possible for the data to have a negative value. Depending upon the amount of time spent working with the data (not to mention any money that may have been spent to

purchase it), an organization could find itself significantly impacted by a poor choice in data and data source.

Purchasing and using the wrong data has both direct and indirect effects on a company's bottom line. In real terms, the money used to buy the data is wasted, along with the time of the consultants working with the data. Also, collective time is lost when corrupted data needs to be discarded and the entire project team needs to start back at the beginning of the process to find new and different data to use.

There are also less obvious but equally damaging effects of corrupted data. The dispiriting effects of redoing a project, or even killing it due to a lack of reliable data, can seriously damage morale among team members at all levels. It can lead to questioning the potential success of future projects, or even to an exodus of workers, if serious questions remain regarding the company's use of trustworthy data. Careful scrutiny of data by embedded data project team librarians prior to the "point of no return" (that is, the point in the project after which the above-mentioned wasted money and time cannot be recovered) can help an organization to avoid these issues altogether.

Harnessing value from Big Data is expensive. Companies looking to begin Big Data initiatives face significant start-up costs in both data management and analysis. First, the company must choose and invest in one of two major storage platforms: the data warehouse, or the Hadoop cluster.[27] The characteristics, similarities, and differences of these platforms will be discussed in greater detail in Chapter 2, but in terms of cost, it is estimated that a Hadoop cluster can cost around $1 million, with its distribution architecture having a similar annual cost, while an enterprise data warehouse can cost anywhere between $10 million and $100 million.[28] If these platforms precipitate changes to existing information technology (IT) infrastructure, additional costs will be incurred.

Next, computer programmers and statisticians must be hired, IT staff need to be trained in Big Data and its security, and professional library staff who can be tasked with working in an embedded situation with the data project teams need to be added. Recruitment costs, salary, compensation and benefits packages, along with the cost of providing continuing education opportunities for these highly specialized employees, are all Big Data cost considerations that add to the expense of Big Data initiatives. Depending on the number of data scientists that need to be hired, it is also possible that the company will need to acquire additional office space, another expense that would add to the cost of the initiative.

The third aspect of the fifth V that needs to be considered is the risk inherent in working with Big Data. Companies are taking a huge risk when investing significant amounts of time, money, and human capital to begin Big Data initiatives. Companies need to see a rapid return-on-investment (ROI) in order to justify these costs. Although the research generally has found that a large payoff will be realized (International Data Corporation projects that in 2015, revenue from Big Data will be $16.9 billion, up from $3.2 billion in 2010[29]), these monetary growth predictions employ Big Data itself in tabulating the revenue numbers. As with anything in life, it is entirely possible that a disruption will occur and the outcome will be completely different, with ROI numbers much lower than expected.

Perhaps the biggest risk for companies working with Big Data is the question of ownership of the final product. Because of the huge investment made, companies rightly want to be able to claim work products as proprietary intellectual property protected by copyright. However, because our courts sometimes move more slowly than the technological advances taking place in our society, legal precedent and judicial opinion in the arena of Big Data product ownership remain murky. When making protection decisions, the U.S. Copyright Office relies upon a 1991 U.S. Supreme Court ruling in *Feist Publications Inc. v. Rural Telephone Services Co.,* 499 U.S.

340 (1991), that information on its own (one single datapoint would be included in this category) is not a copyrightable fact.[30] Writing for the majority, Justice Sandra Day O'Connor stated that a "spark" or a "minimal degree" of creativity has to be applied to information to qualify it for protection by copyright.[31]

Do similar datapoints grouped together into a database contain that spark of creativity necessary to deem them the property of their creator? The European Union (EU) has passed the Database Directive, a law that extends copyright protection to such databases, but the United States has lagged behind in extending this protection.[32] Even if a database is granted copyright protection, that protection might not extend to users of the database who access single datapoints in separate downloads.[33]

Librarians can seize the opportunities inherent in the challenges presented by the need to verify data and scrutinize its value. But before we can become indispensible to Big Data teams, we need to acquaint ourselves with the basic terminology and technology that make this industry tick.

## Endnotes

1. Gil Press, "Big Data News: A Revolution Indeed," *Forbes*, June 18, 2013, accessed November 29, 2013 http://www.forbes.com/sites/gilpress/2013/06/18/big-data-news-a-revolution-indeed/
2. The OED Today, http://public.oed.com/the-oed-today/
3. Ibid.
4. Factiva website, accessed November 29, 2013, http://www.dowjones.com/factiva/sources.asp
5. *NewsTribune*, http://newstrib.com/
6. Elana Zak, "Which Buzzwords Would You Ban in 2014," *Wall Street Journal*, January 2, 2014, accessed January 2, 2014, http://online.wsj.com/news/articles/SB10001424052702304325004579295143935713378

7. "Small and Midsize Companies Look to Make Big Gains with Big Data," ENP Newswire, June 27, 2012.
8. Jacobellis v. Ohio, 378 U.S. 184 (1964)
9. Limitations of 1-2-3 for Windows, accessed November 27, 2013, http://www-01.ibm.com/support/docview.wss?uid=swg27003548
10. McKinsey Global Institute, Big Data: The Next Frontier for Innovation, Competition, and Productivity, May 2011, accessed November 26, 2013, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
11. Mitesh Agarwal, "Capitalizing on Future Opportunities," Dataquest, May 6, 2012, accessed 10/2/14, http://www.dqindia.com/dataquest/news/154183/capitalizing-future-opportunities.
12. Kenny Olmstead, Jane Sasseen, Amy Mitchell, and Tom Rosenstiel, "The State of the News Media 2012," The Pew Research Center's Project for Excellence in Journalism, March 19, 2012, accessed November 26, 2013, http://www.pewresearch.org/2012/03/19/state-of-the-news-media-2012/
13. Joab Jackson, "Five Things CIOs Should Know About Big Data," CIO, May 22, 2012, accessed November 26, 2013, http://www.infoworld.com/d/business-intelligence/5-things-cios-should-know-about-big-data-193090
14. Kenny Olmstead, Jane Sasseen, Amy Mitchell, and Tom Rosenstiel, "Digital: News Gains Audience but Loses Ground in Chase for Revenue," The State of the News Media 2012, accessed October 6, 2014, http://stateofthemedia.org/2012/digital-news-gains-audience-but-loses-more-ground-in-chase-for-revenue/
15. Jodi Enda and Amy Mitchell, "Americans Show Signs of Leaving a News Outlet, Citing Less Information," The State of the News Media 2013, March 18, 2013, accessed November 26, 2013, http://stateofthemedia.org/2013/special-reports-landing-page/citing-reduced-quality-many-americans-abandon-news-outlets/
16. Judith Hurwitz, Alan Nugent, Dr. Fern Harper, and Marcia Kaufman, *Big Data for Dummies* (Hoboken: Wiley, 2013), 208.
17. Jim Edwards, "Twitter's Dark Pool," Business Insider, November 6, 2013, accessed, November 26, 2013, http://www.businessinsider.com/twitter-total-registered-users-v-monthly-active-users-2013-11
18. Jessica Holzer and Greg Bensinger, "SEC Embraces Social Media," *Wall Street Journal*, April 2, 2013, accessed November 26, 2013, http://online.wsj.com/news/articles/SB10001424127887323611604578398862292997352
19. Amy Affelt, "Market Moving News via Social Media: Hazards Ahead," *Online Searcher* (July/August 2013): 16.
20. Ibid.
21. Gartner IT Glossary, accessed November 26, 2013, http://www.gartner.com/it-glossary/big-data/
22. McKinsey Global Institute, Big Data.

23. Brian X. Chen, "U.S. Mobile Internet Traffic Nearly Doubled This Year," *New York Times*, December 23, 2013, accessed January 2, 2014, http://bits.blogs.nytimes.com/2013/12/23/u-s-mobile-internet-traffic-nearly-doubled-this-year/?_r=0
24. "Report of the Presidential Commission on the Space Shuttle Challenger Accident," U.S. Government Printing Office : 1986 0 -157-336, accessed November 29, 2013, http://er.jsc.nasa.gov/seh/explode.html
25. Ibid.
26. John Jordan, "The Risks of Big Data for Companies," *Wall Street Journal*, October 20, 2013, accessed November 30, 2013, http://online.wsj.com/news/articles/SB10001424052702304526204579102941708296708
27. "Big Data: What Does It Really Cost?" WinterCorp Special Report, accessed November 30, 2013, http://www.asterdata.com/big-data-cost/
28. John Bantleman, "The Big Cost of Big Data," *Forbes*, April 16, 2012, accessed November 30, 2013, http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/
29. Agarwal, "Capitalizing on Future Opportunities."
30. Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991)
31. Ibid.
32. Timothy Denny Greene, "What Do Yoga Poses and Big Data Have in Common?" Mondaq Business Briefing, October 9, 2012, accessed November 30, 2013, http://www.mondaq.com/unitedstates/x/200010/Copyright/What+Do+Yoga+Poses+and+Big+Data+Have+in+Common
33. Ibid.

# About The Author

Photograph by Hoang Lam

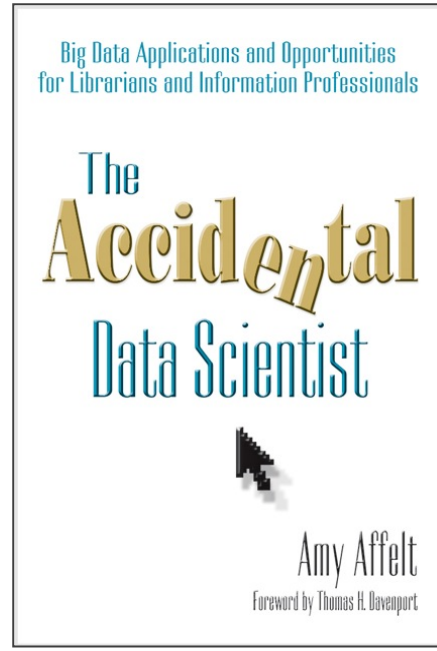
Amy Affelt grew up in LaSalle, Illinois, a rural town of 10,000 people. Some of her most cherished childhood memories are of walking to LaSalle's Carnegie Public Library with her mother. She loved to stand at the desk with a stack of books she had selected, waiting for them to be stamped by the librarian.

Affelt traveled to the "big city" of Chicago to attend the University of Illinois, where, in just three short years, she earned a bachelor's degree in history and was named to Phi Beta Kappa. Perusing the help wanted ads, she felt a sense of urgency to have a job title to look for, and Librarian seemed like a good fit with her research skills, so she went on to earn a master's degree in Library and Information Science from what is now Dominican University. She has spent her entire professional career in economic consulting, working with PhD economists who testify as experts in litigation. She conducts research and analysis to reinforce the points they need to make. She also manages a group of other librarians who do the same.

Affelt is a well-known author and conference presenter on topics such as adding value to information, evaluating information integrity and quality, and marketing of information services. She is also very active in the Special Libraries Association (SLA), having served as the chair of its Future Ready Toolkit initiative and as chair of the Leadership and Management Division.

When she is not reading and studying about Big Data, Affelt enjoys running, cycling, yoga, sailing, ballet, skiing, and cooking, as well as foreign and independent film. Never one to let too much grass grow under her feet, she indulges her passion for travel at every opportunity.

If you enjoyed reading this chapter of *The Accidental Data Scientist*, please visit our bookstore to pre-order your copy.




Information Today, Inc.