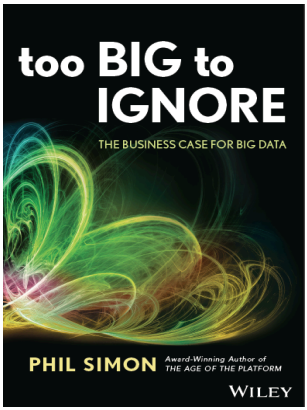
Introduction from *Too Big to Ignore: The Business Case for Big Data* by Phil Simon. Published by John Wiley & Sons, March 2013.

For more information, go to www.philsimon.com



Introduction: This Ain't Your Father's Data

Throughout history, in one field after another, science has made huge progress in precisely the areas where we can measure things—and lagged where we can't.

—Samuel Arbesman

Car insurance isn't a terribly sexy or dynamic business. For decades, it has essentially remained unchanged. Nor is it an egalitarian enterprise: while a pauper and a millionaire pay the same price for a stamp ($0.45 in the United States as of this writing), the car insurance world works differently. Some people just pay higher rates than others, and those rates have at least initially very little to do with whether one is a "safe" driver, whatever that means. Historically, many if not most car insurance policies were written based on very few independent variables: age, gender, zip code, previous speeding tickets and traffic violations, documented accidents, and type of car. As I found out more than twenty years ago, a newly licensed, seventeen-year-old guy in New Jersey who drives a sports car has to pay a boatload in car insurance for the privilege—even if he rarely drives above the speed limit, always obeys traffic signals, and has nary an accident on

his record. Like just about every kid my age, I wasn't happy about my rates. After all, I was an "above average" driver, or at least I liked to think so. Why should I have to pay such exorbitant fees?

Of course, we all can't be above average; it's statically impossible. Truth be told, I'm sure that back then I occasionally didn't come to a complete stop at every red light. While I've never been arrested for DUI, to this day I don't always obey the speed limit. (Shhh . . . don't tell anyone.) When I'm driving faster than the law says I should, I sometimes think of the famous George Stigler picture of Milton Friedman taken in the mid-twentieth-century. Friedman was paying a speeding ticket with, paradoxically, a big smile on this face. Why such joy? Because Friedman was an economist and, as such, he was rational to a fault. In his view of the world, the time that he regularly saved by exceeding the speed limit was worth more to him than the risk and fine of getting caught. To people like Friedman and me, speeding is only a simple expected value calculation: Friedman sped because the rewards outweighed the risks. When a cop pulled him over, he was glad to pay the fine. But I digress.

So why do most car insurance companies base their quotes and rates on relatively simple variables? The answer isn't complicated, especially when you consider the age of these companies. Allstate opened its doors in 1931. GEICO was founded in 1936, and the Progressive Casualty Insurance Company set up shop only one year later. Think about it: seventy-five years ago, those primitive models represented the best that car insurance companies could do. While each has no doubt tweaked its models since then, old habits die hard, as we saw with Art Howe and Billy Beane in the Preface. For real change to happen, somebody needs to upset the applecart. In this way, car insurance is like baseball.

BETTER CAR INSURANCE THROUGH DATA

The similarities between the ostensibly unrelated fields of baseball and car insurance don't end there. Much like the baseball revolution pioneered by Billy Beane, car insurance today is undergoing a fundamental transformation. Just ask Joseph Tucci. As the CEO at data storage behemoth EMC Corporation, he knows a thing or thirty about data. On October 3,

2012, Tucci spoke with Cory Johnson of Bloomberg Television at an Intel Capital event in Huntington Beach, California. Tucci talked about the state of technology, specifically the impact of Big Data and cloud computing on his company—and others.[1] At one point during the interview, Tucci talked about advances in GPS, mapping, mobile technologies, and telemetry, the net result of which is revolutionizing many businesses, including car insurance. No longer are rates based upon a small, primitive set of independent variables. Car insurance companies can now get much more granular in their pricing. Advances in technology are letting them answer previously unknown questions like these:

- Which drivers routinely exceed the speed limit and run red lights?
- Which drivers routinely drive dangerously slow?
- Which drivers are becoming less safe—even if they have received no tickets or citations? That is, who used to generally obey traffic signals but don't anymore?
- Which drivers send text messages while driving? (This is a big no-no. In fact, texting while driving [TWD] is actually considerably more dangerous than DUI.[2] As of this writing, fourteen states have banned it.)
- Who's driving in a safer manner than six months ago?
- Does a man with two cars (a sports car and a station wagon) drive each differently?
- Which drivers and cars swerve at night? (This could be a manifestation of drunk driving.)
- Which drivers checked into a bar using FourSquare or Facebook and drove their own cars home (as opposed to taking a cab or riding with a designated driver)?

Thanks to these new and improved technologies and the data they generate, insurers are effectively retiring their decades-old, five-variable underwriting models. In their place, they are implementing more contemporary, accurate, dynamic, and data-driven pricing models. For instance, in 2011, Progressive rolled out Snapshot, its Pay As You Drive (PAYD) program.[3] PAYD allows customers to voluntarily install a tracking device in their cars that transmits data to

Progressive and possibly qualifies them for rate discounts. From the company's site:

> How often you make hard brakes, how many miles you drive each day, and how often you drive between midnight and 4 a.m. can all impact your potential savings. You'll get a Snapshot device in the mail. Just plug it into your car and drive like you normally do. You can go online to see your latest driving details and projected discount.

Is Progressive the only, well, progressive insurance company? Not at all. Others are recognizing the power of new technologies and Big Data. As Liane Yvkoff writes on CNET, "State Farm subscribers self-report mileage and GMAC uses OnStar vehicle diagnostics reports. Allstate's Drive Wise goes one step further and uses a similar device to track mileage, braking, and speeds over 80 mph, but only in Illinois."[4]

So what does this mean to the average driver? Consider two fictional people, both of whom hold car insurance policies with Progressive and opt in to PAYD:

- Steve, a twenty-one-year-old New Jersey resident who drives a 2012, tricked-out, cherry red Corvette
- Betty, a forty-nine-year-old grandmother in Lincoln, Nebraska, who drives a used Volvo station wagon

All else being equal, which driver pays the higher car insurance premium? In 1994, the answer was obvious: Steve. In the near future, however, the answer will be much less certain: *it will depend on the data*. That is, vastly different driver profiles and demographic information will mean less and less to car insurance companies. Traditional levers like those will be increasingly supplemented with data on drivers' individual patterns. What if Steve's flashy Corvette belies the fact that he always obeys traffic signals, yields to pedestrians, and never speeds? He is the embodiment of safety. Conversely, despite her stereotypical profile, Betty drives like a maniac while texting like a teenager.

In this new world, what happens at rate renewal time for each driver? Based upon the preceding information, Progressive happily discounts Steve's previous insurance by 60 percent but triples Betty's

renewal rate. In each case, the new rate reflects new—and far superior—data that Progressive has collected on each driver.

Surprised by his good fortune, Steve happily renews with Progressive, but Betty is irate. She calls the company's 1-800 number and lets loose. When the Progressive rep stands her ground, Betty decides to take her business elsewhere. Unfortunately for Betty, she is in for a rude awakening. Allstate, GEICO, and other insurance companies have access to the same information as Progressive. All companies strongly suspect that Betty is actually a high-risk driver; her age and Volvo only tell part of her story—and not the most relevant part. As such, Allstate and GEICO quote her a policy similar to Progressive's.

Now, Betty isn't happy about having to pay more for her car insurance. However, Betty *should* in fact pay more than safer drivers like Steve. In other words, simple, five-variable pricing models no longer represent the best that car insurance companies can do. They now possess the data to make better business decisions.

Big Data is changing car insurance and, as we'll see throughout this book, other industries as well. The revolution is just getting started.

POTHOLES AND GENERAL ROAD HAZARDS

Let's stay on the road for a minute and discuss the fascinating world of potholes. Yes, potholes. Historically, state and municipal governments have had a pretty tough time identifying these pesky devils. Responsible agencies and departments would often scour the roads in search of potholes and general road hazards, a truly reactive practice. Alternatively, they would rely upon annoyed citizens to call them in, typically offering fairly generic locations like "on Main Street, not too far from the 7-Eleven." In other words, there was no good automatic way to report potholes to the proper authorities. As a result, many hazards remained unreported for significant periods of time, no doubt causing car damage and earning the ire of many a taxpayer. Many people agree with the quote from acerbic comedian Dennis Miller, "The states can't pave [expletive deleted] roads."

Why has the public sector handled potholes and road hazards this way? For the same reason that car insurance companies relied upon

very few basic variables when quoting insurance rates to their customers: in each case, it was the best that they could do at the time.

At some point in the past few years, Thomas M. Menino (Boston's longest-serving mayor) realized that it was no longer 1950. Perhaps he was hobnobbing with some techies from MIT at dinner one night. Whatever his motivation, he decided that there just had to be a better, more cost-effective way to maintain and fix the city's roads. Maybe smartphones could help the city take a more proactive approach to road maintenance. To that end, in July 2012, the Mayor's Office of New Urban Mechanics launched a new project called Street Bump, an app that

> allows drivers to automatically report the road hazards to the city as soon as they hear that unfortunate "thud," with their smartphones doing all the work.
>
> The app's developers say their work has already sparked interest from other cities in the U.S., Europe, Africa and elsewhere that are imagining other ways to harness the technology.
>
> Before they even start their trip, drivers using Street Bump fire up the app, then set their smartphones either on the dashboard or in a cup holder. The app takes care of the rest, using the phone's accelerometer—a motion-detector—to sense when a bump is hit. GPS records the location, and the phone transmits it to a remote server hosted by Amazon Inc.'s Web services division.[5]

But that's not the end of the story. It turned out that the first version of the app reported far too many false positives (i.e., phantom potholes). This finding no doubt gave ammunition to the many naysayers who believe that technology will never be able to do what people can and that things are just fine as they are, thank you. Street Bump 1.0 "collected lots of data but couldn't differentiate between potholes and other bumps."[6] After all, your smartphone or cell phone isn't inert; it moves in the car naturally because the car is moving. And what about the scores of people whose phones "move" because they check their messages at a stoplight?

To their credit, Menino and his motley crew weren't entirely discouraged by this initial setback. In their gut, they knew that they were

on to something. The idea and potential of the Street Bump app were worth pursuing and refining, even if the first version was a bit lacking. Plus, they have plenty of examples from which to learn. It's not like the iPad, iPod, and iPhone haven't evolved over time.

Enter InnoCentive Inc., a Massachusetts-based firm that specializes in open innovation and crowdsourcing. (We'll return to these concepts in Chapters 4 and 5.) The City of Boston contracted InnoCentive to improve Street Bump and reduce the number of false positives. The company accepted the challenge and essentially turned it into a contest, a process sometimes called *gamification*. InnoCentive offered a network of 400,000 experts a share of $25,000 in prize money donated by Liberty Mutual.

Almost immediately, the ideas to improve Street Bump poured in from unexpected places. Ultimately, the best suggestions came from

- A group of hackers in Somerville, Massachusetts, that promotes community education and research
- The head of the mathematics department at Grand Valley State University in Allendale, Michigan
- An anonymous software engineer

The result: Street Bump 2.0 is hardly perfect, but it represents a colossal improvement over its predecessor. As of this writing, the Street Bump website reports that 115,333 bumps have been detected. What's more, it's a quantum leap over the manual, antiquated method of reporting potholes no doubt still being used by countless public works departments throughout the country and the world. And future versions of Street Bump will only get better. Specifically, they may include early earthquake detection capability and different uses for police departments.

Street Bump is not the only example of an organization embracing Big Data, new technologies, and, arguably most important, an entirely new mind-set. With the app, the City of Boston was acting less like a government agency and more like, well, a progressive business. It was downright refreshing to see.

Crowdsourcing roadside maintenance isn't just cool. Increasingly, projects like Street Bump are resulting in substantial savings. And the public sector isn't alone here. As we've already seen with examples

like Major League Baseball (MLB) and car insurance, Big Data is transforming many industries and functions within organizations. Chapter 5 will provide three in-depth case studies of organizations leading the Big Data revolution.

RECRUITING AND RETENTION

In many organizations, Human Resources (HR) remains the redheaded stepchild. Typically seen as the organization's police department, HR rarely commands the internal respect that most SVPs and Chief People Officers believe it does. I've seen companies place poor performers in HR because they couldn't cut it in other departments. However, I've never seen the reverse occur (e.g., "Steve was horrible in HR, so we put him in Finance."). For all of their claims about being "strategic partners," many HR departments spend the majority of their time on administrative matters like processing new hire paperwork and open enrollment. While rarely called *Personnel* anymore (except on *Mad Men*), many HR departments are anachronistic: they operate now in much the same way as they did four decades ago.

My own theory about the current, sad state of HR is as follows: As a general rule, HR folks tend not to make decisions based upon data. In this way, HR is unique. Employees rely almost exclusively on their gut instincts and corporate policy. What if employees in other departments routinely made important decisions sans relevant information? Absent data, the folks in marketing, sales, product, and finance wouldn't command a great deal of respect either. W. Edwards Deming once said, "In God we trust, all others must bring data." Someone forgot to tell this to the folks in HR, and the entire function suffers as a result.

I wrote a book on botched IT projects and system implementations, many of which involved HR and payroll applications. Years of consulting on these types of engagements have convinced me that most employees in HR just don't think like employees in other departments. Most HR people don't seek out data in making business decisions or even use the data available to them. In fact, far too many HR folks actively try to *avoid* data at all costs. (I've seen HR directors manipulate data to justify their decision to recruit at Ivy League schools, despite the fact that trying to hire Harvard and Yale alumni didn't make the

slightest bit of financial sense.) And it's this lack of data—and, in that vein, a data mind-set—that has long undermined HR as a function. As we'll see throughout this book, however, ignoring data (big or small) doesn't make it go away. Pretending that it doesn't exist doesn't make it so. In fact, Big Data can be extremely useful, even for HR.

As the *Wall Street Journal* recently reported,[7] progressive and data-oriented HR departments are turning to Big Data to solve a long-vexing problem: how to hire better employees and retain them. It turns out that traditional personality tests, interviews, and other HR standbys aren't terribly good at predicting which employees are worth hiring—and which are not. Companies like Evolv "utiliz[e] Big Data predictive analytics and machine learning to optimize the performance of global hourly workforces. The solution identifies improvement areas, then systematically implements changes to core operational business processes, driving increased employee retention, productivity, and engagement. Evolv delivers millions of dollars in operational savings on average for each client, and guarantees its impact on operating profitability."[8]

Millions in savings? Aren't these just lofty claims from a start-up eager to cash in on the Big Data buzz? Actually, no. Consider some of the specific results generated by Evolv's software, as shown in Table I.1.

The lesson here is that Big Data can significantly impact each area of a business: its benefits can touch every department within an organization. Put differently, Big Data is too big to ignore.

Table I.1 Big Data Improves Recruiting and Retention

Employee Problem	Big Data Solution
Compensation	Caesars casino found that increasing pay within certain limits had no impact on turnover.
Attrition	Xerox found that experience was overrated for call-center positions. What's more, overly inquisitive employees tended to leave soon after receiving training.
Sick Time	Richfield Management tests applicants for opinions on drugs and alcohol. The company found that those who partake in "extracurricular" activities are more prone to get into accidents.

HOW BIG IS BIG? THE SIZE OF BIG DATA

How big is the Big Data market? IT research firm Gartner believes that Big Data will create $28 billion in worldwide spending in 2012, a number that will rise "to $34 billion in 2013. Most of that spending will involve upgrading 'traditional solutions' to handle the flood of data entering organizations from a variety of sources, including clickstream traffic, social networks, sensors, and customer interactions; the firm believes that a mere $4.3 billion in sales will come from 'new Big Data functionality.'"[9] For its part, consulting firm Deloitte expects massive Big Data growth, although precisely "estimating the market size is challenging."[10]

High-level projections from top-tier consulting firms are all fine and dandy, but most people won't be able to get their arms around abstract numbers like these. The question remains: just how big exactly is Big Data? You might as well ask, "How big is the Internet?"* We can't precisely answer these questions; we can only guess. What's more, Big Data got bigger in the time that it took me to write that sentence. The general answer is that Big Data is really big—and getting bigger all the time. Just look at these 2011 statistics on videos from website monitoring company Pingdom:

- **1 trillion**: The number of video playbacks on YouTube
- **140**: The number of YouTube video playbacks per person on Earth
- **48**: Hours video uploaded to YouTube every minute
- **82.5**: Percentage of the U.S. Internet audience that viewed video online
- **201.4 billion**: Number of videos viewed online per month (October 2011)
- **88.3 billion**: Videos viewed per month on Google sites, including YouTube (October 2011)
- **43**: Percentage share of all worldwide video views delivered by Google sites, including YouTube[11]

* For a stunning infographic on "A Day in the Life of the Internet," see www.mashable.com/2012/03/06/one-day-internet-data-traffic.

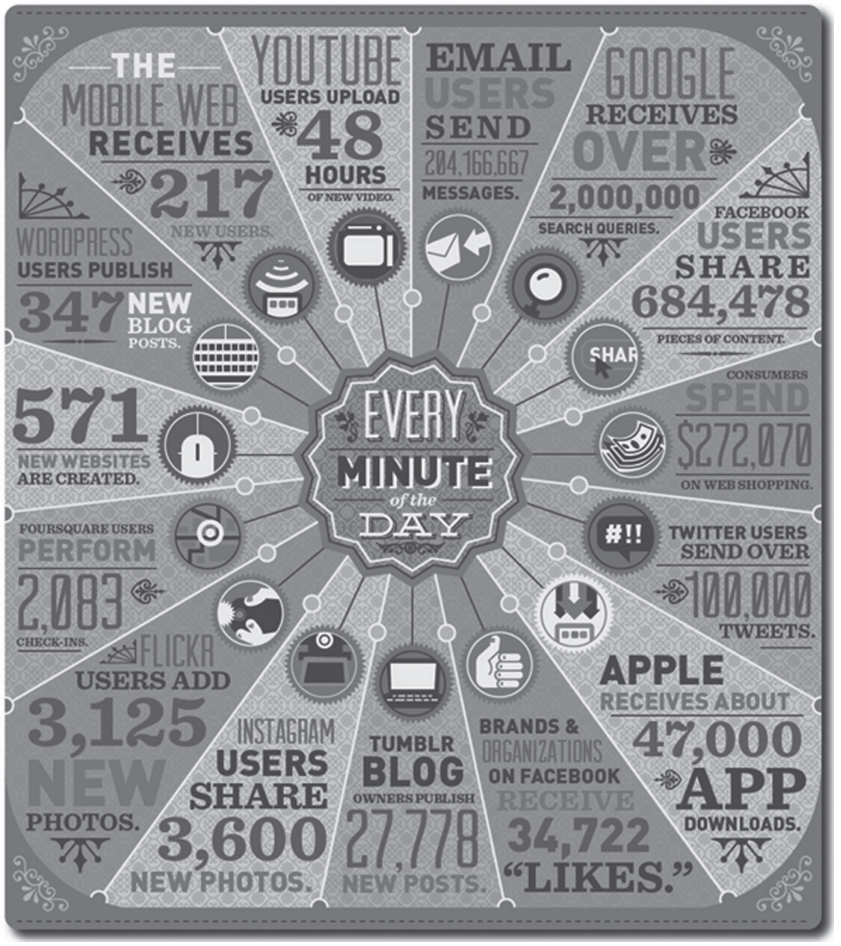


Figure I.1 The Internet in One Minute
Source: Image courtesy of Domo; www.domo.com

If those numbers seems abstract, look at the infographic in Figure I.1 to see what happens on the Internet every minute of every day.

As of 2009, estimates put the amount of data on the entire World Wide Web at roughly to 500 *exabytes*.[12] (An exabyte equals one million terabytes.) Research from the University of California, San Diego, reports that in 2008, Americans consumed 3.6 *zettabytes* of information,[13] a number that no doubt increased in subsequent years. (A zettabyte is equal to 1 billion terabytes.) You get my drift: Big Data is

really big—and it's constantly expanding. Cisco estimates that, in 2016, 130 exabytes of data will travel through the Internet each year.[14]

WHY NOW? EXPLAINING THE BIG DATA REVOLUTION

We are at the beginning of an exciting time in the enterprise IT world. CIOs surveyed place Big Data at or near the top of their highest priorities for 2013 and beyond.[15] Right now, Big Data is just beginning. It is in the nascent stages of Gartner Research's oft-used Hype Cycle.[16] Without question, some people believe that the squeeze from Big Data will not be worth its juice.

For its part, the global management consulting firm McKinsey has boldly called Big Data "the next frontier for innovation, competition, and productivity."[17] You'll get no argument from me, but reading that statement should give you pause. Why now? After all, something as big as the Big Data Revolution doesn't just happen overnight. It takes time. Nor does a single, discrete event give rise to a trend this, well, big. Rather, Big Data represents more of an evolution than a Eureka moment. So what are some of the most important reasons for the advent and explosion of Big Data? This is not intended to be a comprehensive list. In the interest of brevity, here are the most vital factors:

- The always-on consumer
- The plummeting of technology costs
- The rise of data science
- Google and Infonomics
- The platform economy
- The 11/12 watershed: Sandy and politics
- Social Media and other factors

Let's explore each one.

The Always-On Consumer

I wasn't around in the early 1800s, but I can't help but think that most people were pretty patient back then. While I'm oversimplifying here,

the Internet has made many citizens of industrialized countries pretty impetuous. It seems that most of us have too little idle time, far too many choices, and way too many things going on. Consider the following questions:

- Consumers who can watch a Netflix streaming video on their smartphones and tablets without buffering anywhere in the country are more likely to do so—and consume more data in the process. What would be the effect on Netflix if that same video constantly froze, like it often does on current airplane Wi-Fi connections?
- What if people had to wait six hours for their two-minute videos to upload to YouTube?
- What's the first thing that most people do when their airplane touches down (present company included)?
- Are most people going to remember to tweet something when they go back on the grid?

The answers to these questions should be obvious. As we'll see throughout this book, Big Data is largely consumer driven and consumer oriented. Moreover, the levels of data consumption we see today were simply impossible ten years ago—even if services like YouTube, Facebook, and Twitter had existed. Yes, data storage costs were just too high back then, but how many consumers really thought about that? In other words, the Big Data Revolution did not exclusively hinge upon lower corporate data storage prices. That was a necessary but insufficient condition, as was the arrival of the web. Neither immediately or directly triggered Big Data. Rather, Big Data was truly born when consumer technologies such as cell phones (and then smartphones), cloud computing, and broadband connections reached critical mass. Absent these improvements, innovations, and advents, we would not be hearing about Big Data. In its own way, each made generating, storing, and accessing data faster, easier, and more convenient for the masses. The always-on consumer represents arguably the biggest reason that the current Data Deluge is happening now. (Chapter 8 will have much more to say about the adjacent possible.)

The Plummeting of Technology Costs

For years now, consumers have been able to stay connected to the web from anywhere. In the process, they have consumed and generated unfathomable amounts of content and data. Many if not most consumers aren't aware of the precipitous drop in data storage costs—and others don't give it a second thought. Still, it's hard to overstate the impact of this on Big Data. If data storage costs had remained at 2000 levels, our world would be dramatically different. We would consume and generate far less data. Sites like YouTube may not exist—and they certainly wouldn't be nearly as popular.

We know from recent history that even small, non-zero fees represent a source of considerable economic friction. Tiny fees can have a disproportionately deterring effect on commercial behavior. In his 2009 book *Free: The Future of a Radical Price*, Chris Anderson writes about an Amazon promotion for free shipping on book orders of $25 (EU) or more across Europe. (In effect, this promotion would incentivize customers to order a second book. Most books cost less than $25.) Because of a programming error, though, Amazon inadvertently charged its French customers a nominal shipping fee of 1 franc (about 20 cents). The token charge ultimately yielded disproportionate and extremely telling results: significantly lower sales in France relative to other European countries. When Amazon fixed the error, French customers behaved like the rest of their European brethren.* The bottom line: price matters.

Back in June 2009, Anderson wrote a *Wired* piece "Tech Is Too Cheap to Meter,"[18] and those words today ring truer now than they did nearly four years ago. As I write this, I can buy a one terabyte (TB) external hard drive on eBay for $79. Fifteen years ago, I could expect to pay much, much more for such a device—if I could even find one. It's simple economics, really. If data storage costs had not dropped to the same extent, people would not be consuming and generating so much data (i.e., Big Data would not have happened). Just look at the carrier data usage rates among those with "all you

* Interesting postscript to the story: Amazon was sued for violating an obscure 1981 French law prohibiting such a promotion. A few years later, France joined the EU and that law disappeared.

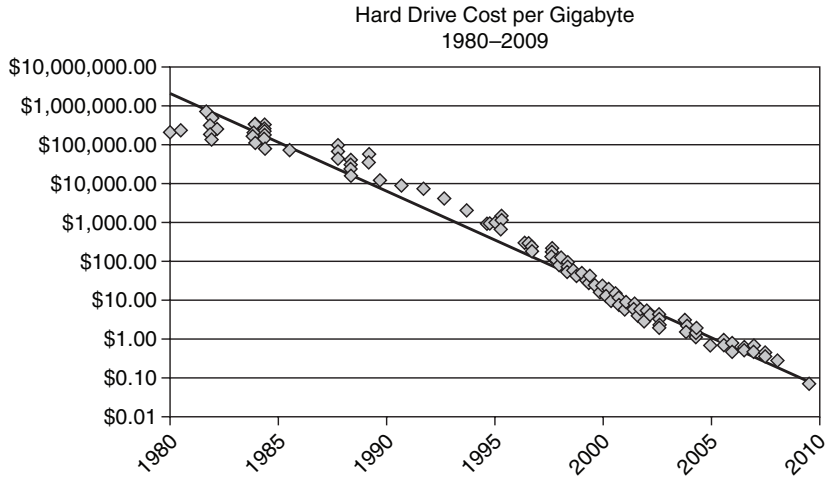


Figure I.2 The Drop in Data Storage Costs
Source: Matthew Komorowski

can eat" plans compared to those with strict data limits. Guess which folks use more data?

Whether online or off, data storage costs today are orders of magnitude cheaper than they were 30 years ago.

Figure I.2 is just the visual representation of Kryder's Law: magnetic disk areal storage density is increasing incredibly quickly, at a pace much faster than oft-cited Moore's Law. What's more, there's no end in sight to this trend—and data storage costs are not the only critical expenses to plummet over the past five years with respect to Big Data. Technologies like near field communication (NFC), radio-frequency identification (RFID), nanotechnology, and sensors have matured and become more commercially viable and affordable. To varying extents, each has allowed organizations to collect more data. Chapter 4 will have much more to say on these topics.

The Rise of Data Science

Five years ago, if you had asked 100 random people to tell you about a hot new field, I very much doubt that you would have heard any mention of the terms *data* and *science*. Since that time, though, the term

data scientist[*] has quietly begun to enter the business vernacular—and today it's white hot. In October 2010, *Harvard Business Review* called it the "sexiest job of the 21st century."[19] The reasons are simple: Big Data is blowing up, and supply and demand of practitioners is out of whack. More specifically, demand for data scientists far exceeds its available supply. Management consulting McKinsey predicts a severe "shortage of talent necessary for organizations to take advantage of Big Data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills . . . with the know-how to use the analysis of Big Data to make effective decisions."[20] As far as demand goes, myriad organizations are looking to leverage Big Data, but to do that they need plenty of help—and there aren't that many experienced data scientists out there. As a result, data scientists can pretty much write their own tickets these days.

But just what does a data scientist *do*, actually? While still evolving, data science encompasses a diverse set of fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high-performance computing (HPC). While some overlap exists between the modern data scientist and a traditional statistician, they are not one and the same. By tapping into these varied disciplines, data scientists are able to extract meaning from data in innovative ways. Not only can they answer the questions that currently vex organizations, they can find better ones to ask.

As is the case with Big Data, there's no one generally accepted definition of the term. I happen to find IBM's definition of the *data scientist* complete and thought provoking:

> What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems. They will pick *the right problems* that have the most value to the organization. [Emphasis mine.]

[*] D.J. Patil and Jeff Hammerbacher coined the term *data science* in 2008. At the time, they were the leads of data and analytics efforts at LinkedIn and Facebook, respectively.

> The data scientist role has been described as "part analyst, part artist." Anjul Bhambhri, vice president of Big Data products at IBM, says, "A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."
>
> Whereas a traditional data analyst may look only at data from a single source—a CRM system, for example—a data scientist will most likely explore and examine data from multiple disparate sources. The data scientist will sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem. A data scientist does not simply collect and report on data, but also looks at it from many angles, determines what it means, then recommends ways to apply the data.[21]

I particularly like Bhambhri's insight that a data scientist is part analyst, part artist. The job entails—in fact, *requires*—a high degree of human judgment, especially with respect to selecting and defining the problem. That is, it's downright misleading to suggest that data scientists are slaves to computers, automation, and data.

It turns out that data science need not be the exclusive purview of data scientists. Companies such as Datahero, Infogram, and Statwing are trying to make analytics accessible even to laypersons.[22] Today, there are even data science dog-and-pony shows. For instance, in May 2012, I attended Greenplum's second annual Data Science Summit[23] while performing some of the preliminary research for this book. (Impressed by what I saw, I wrote a piece for *Huffington Post* titled "Big Data Goes Mainstream."[24]) I left that conference convinced that data science is here to stay. If anything, it's only going to get bigger.

Google and Infonomics

It's been a gradual process, but we've reached a point at which most progressive thinkers, leaders, and organizations acknowledge the fact that data is in fact a business asset—perhaps their *greatest* asset. In this

regard, Google has been a watershed. The company has made tens of billions of dollars by serving up relevant ads exactly when its users were looking to buy something. While the company keeps the secret sauce of its search algorithm under tight wraps, at its core Google software runs on widespread and contextual *data*. Take that away, and it's hard to imagine Google as we know it. But Google embodies a much larger trend: more companies are finally recognizing that their success hinges upon how well they understand their users and customers.* In other words, it's all about the data.

In the late 1990s, Gartner's Douglas Laney[25] conducted extensive research on the value of information and its management. Laney certainly wasn't the first person to contend that information is valuable, but he went much further than most pundits had. He believed that information met the definition of a formal business asset and should be treated as such. He coined the term *Infonomics* (a portmanteau of *information* and *economics*) to describe the study and emergent discipline of quantifying, managing, and leveraging information. Its principles can be stated as follows:

- Information is an asset with value that can be quantified.
- Information should be accounted for and managed as an asset.
- Information's value should be used for budgeting IT and business initiatives.
- The realized value of information should be maximized.

To be sure, *Infonomics* remains an obscure term, in no small part due to the proprietary nature of Laney's research. Regardless of its popularity as a proper or recognized field, many people and companies have been unknowingly practicing Infonomics for decades. Case in point: Billy Beane.

On a typical corporate balance sheet, you'll find assets like cash and cash equivalents, short-term investments, receivables, inventory, and prepaid expenses. I have yet to see one, however, that lists an "information" bucket. Laney would like to see that change, and he has developed formal information asset valuation models. He even opened the nonprofit Center for Infonomics in 2010. Whether explicitly

* Amazon is another excellent example here.

quantified on a balance sheet or not, companies like Google and Facebook prove that data is exceptionally valuable.

If, like most learned folks, you believe that information is a business asset, then by definition Big Data inheres potentially enormous value. If you believe that data is a problem to be minimized, good luck surviving.

The Platform Economy

Without plugging my previous book too much here, it's imperative to at least briefly mention the role that platforms have played in the Big Data Revolution. Companies like Amazon, Apple, Facebook, and Google have all contributed to—and benefited from—the data avalanche in many ways. (There are some potentially nefarious effects to this trend, but that's a subject for Chapter 7.) More than the Gang of Four, other companies like LinkedIn, Twitter, and Salesforce.com have embraced platform thinking.

The impact of platforms on Big Data is hard to overstate. For instance, as smart as Mark Zuckerberg is, the idea of a social network preceded him. Early social networks like Classmates.com, MySpace, and Friendster all had transformative potential. For different reasons, however, those sites did not evolve into true platforms, effectively limiting their growth. From technical perspectives, they could not support anything near one billion users—or even a fraction of that number. For its part, in 2002 Friendster was extremely clunky. Despite its frequent problems, the site sported more than 3 million users soon after its launch. Friendster could only get so big. Back then no one wanted to visit a site that was down half of the time. The same is true now.

Give Zuckerberg credit for understanding how ubiquitous social networks and platforms could become, at least as they continued to offer a compelling user experience and rarely went down. Like Jeff Bezos, Larry Page, and Sergei Brin, Zuckerberg understood the impact of network effects on platforms. Put simply, bigger and wider platforms can support more users—and more users mean more data.

Many people forget that Apple's first iPhone launched in 2007 without the AppStore. The AppStore officially "opened" in July 2008 as an update to iTunes. In this sense, the original iPhone acted more like a traditional cell phone than the portable computer it ultimately became. As anyone who has paid attention over the past four years knows, the AppStore was nothing less than a game-changer. The expression "there's an app for that" entered the vernacular as developers from around the globe hurried to build an astonishing array of games, productivity tools, and so on.

Opening an AppStore only gets a company—even Apple—so far. Without the requisite tools to build many interesting apps (and some truly awful ones), developers would not have spent so much time and effort on them. As a result, apps would not have taken off—at least as quickly and to the same extent that they did. Application programming interfaces (APIs) and software development kits (SDKs) gave developers the tools to populate the AppStore with myriad offerings, and other companies like Google, Samsung, Microsoft, Facebook, and RIM followed Apple's lead. The result: further growth of Big Data.

The 11/12 Watershed: Sandy and Politics

We may look back at early November 2012 as the period in which Big Data entered the zeitgeist. In that short period, two watershed events took place that may well have ushered in the era of Big Data.

Before Halloween, meteorologists at the National Weather Service (NWS) had been keeping their eye on a bunch of clouds in the Caribbean. Those clouds appeared to be heading toward the Northeastern United States. Nearly a week later, a NWS computer model predicted that the Caribbean weather system would morph into a "superstorm" after taking a "once-in-a-century" sharp turn into New Jersey.[26] Ultimately, Hurricane Sandy wrecked an estimated $50 billion of damage on the country,[27] making it the second-worst natural disaster behind Katrina. While computer models and data couldn't *prevent* Sandy and its carnage, the ability of NWS to predict such a rare event allowed millions to prepare, minimized its damage, and saved lives.

Just a few days later, author and statistician Nate Silver came under fire from many conservatives. Silver's crime? He boldly asserted

that Barack Obama had established himself as more than a 70 percent favorite to win the Presidential election[28] when many reputable polls at the time put the incumbent and Republican nominee Mitt Romney in essentially a dead heat. Toward the end of the campaign, Silver increased his confidence level to more than 90 percent. Outrage, claims of political bias, and mockery came from many old-school pundits. They questioned the wisdom in trusting the methods of a 34-year-old stats geek over tried-and-true polls. (The parallels between Silver and Billy Beane are obvious.)

Well, you know how this story turned out. Obama beat Romney by a comfortable margin. The wunderkind correctly predicted all 50 states,[29] and Silver's critics suddenly fell silent. The methodology behind his predictions isn't terribly important here,* but it has shown to be incredibly accurate. And Silver is no one-trick pony. His track record at predicting past elections is astonishing. In 2008, Silver predicted 49 of the 50 states in the presidential election. Then in 2010 Silver correctly forecasted:

- 92 percent of U.S. Senate races. (He missed Alaska, Nevada, and Colorado.)
- 95 percent of governor races. (He missed Illinois and Florida.)

Sales of Silver's recently released book *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't* exploded. We'll come back to the 2012 presidential election and the use of Big Data later in the book. For now, suffice it to say that the highly public nature of these two events served as a data wake-up call of sorts.

Social Media and Other Factors

I don't want to devote too much space to it here, but I'd be remiss if I didn't mention the impact of social media on Big Data. LinkedIn, Twitter, Facebook, and other sites are driving a great deal of the Big Data revolution. And let's not forget about enormous advances in data capturing technologies such as RFID and sensors, discussed in more detail in Chapter 4.

---

* If you're curious, go to http://fivethirtyeight.blogs.nytimes.com/methodology.

On a broader and more philosophical level, as a society, the past few years seem to have proven that we have a nearly insatiable demand to generate and consume data. Perhaps it's something in our DNA, but much of this may have to do with the price of data consumption—often near zero. In the infamous words of American writer Steward Brand, "On the one hand information wants to be expensive, because it's so valuable. The right information in the right place just changes your life. On the other hand, information wants to be free, because the cost of getting it out is getting lower and lower all the time. So you have these two fighting against each other."

Finally, the Big Data revolution has arrived because of economic need. (Chapter 8 will have a great deal more to say about this.) Aside from the Oakland A's and the City of Boston, this book will introduce many people and institutions using data and available technologies to innovate, reduce costs, and reach new customers.

CENTRAL THESIS OF BOOK

It's easy for cynics and naysayers to dismiss Big Data as just another fad or the latest technology jargon. You don't have to look too hard to find an old-school CIO who considers Big Data hooey. And Big Data is not alone in this regard. Plenty of executives still view social media as nothing more than a waste of time, with every minute spent on "The Twitter" representing a minute better spent elsewhere. And then there are the naysayers who don't buy into the cloud either. In their view, the squeeze just isn't worth the juice.

All of this is to be expected. To be fair, in the world of technology, many companies, services, products, and vendors arrive with enormous hype only to quickly disappear—as do their acolytes. (Microsoft's Zune and Windows Me, and Sony's MiniDisc, Pressplay, and MusicNet, are just a few of tech's many misses. The dot-com bust need not be rehashed here.) This is the nature of the tech beast. Because of the high failure rates or ephemeral nature of the next shiny new things, any new technology or application initially faces far more laggards than early adopters—even the technologies that ultimately prove successful and important. This is doubly true in the enterprise,

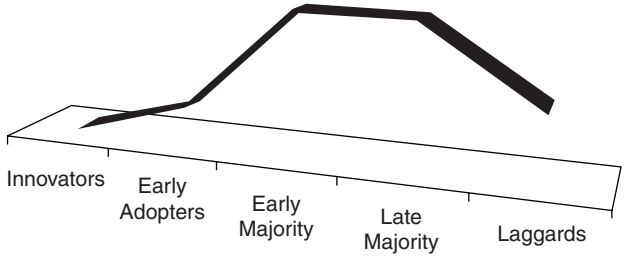


Figure I.3 The Technology Adoption Life Cycle (TALC)

where conservative CIOs understandably don't want to bet the future of their organizations (and often their own careers) on a technology that may be a flash in the pan. And Big Data is certainly no exception to this rule. As I write this book, most organizations have yet to embrace it. Many CIOs believe or hope that it just goes away—at least until they retire, that is.

In this way, Big Data is just part of the Technology Adoption Life Cycle (TALC). TALC depicts the adoption of new technologies by group. In a nutshell, few people and organizations implement new technologies right after their introduction. This is represented graphically in Figure I.3.

For progressive organizations of all sizes, however, Big Data is anything but bunk. On the contrary, it represents an enormous opportunity for one simple reason: when done right, Big Data can yield superior information and unparalleled insights into a range of behaviors and even predict a few. With that information, everyone—employees, departments, organizations, scientists, and politicians—can make vastly better decisions.

When most of us make decisions, we tend to use heuristics and rules of thumb—not data. While this may be understandable in our personal lives, it's less so in our professional ones. As many books and studies have repeatedly shown, human beings are not terribly adroit at making good (read: rational) personal and business decisions. Economists like Daniel Kahneman, Dan Ariel, Steven Levitt, Stephen Dubner, and others have written books and done remarkable studies that effectively prove our generally poor decision-making abilities. In fact, behavioral economics has become quite the hot field primarily

because it attempts to explain what traditional economic theory cannot. Human beings are not entirely rational, even when presented with complete information. Against that backdrop, is it remotely surprising that we are even more deficient when presented with unprecedented amounts of incomplete data?

PLAN OF ATTACK

Today, organizations no longer operate in a world in which only traditional data types and sources matter. That ship has long sailed. The fact that many organizations choose to ignore other forms and sources of data doesn't make them any less important.

The timing of Big Data could not be more propitious. Big Data and its attendant tools allow organizations to interpret previously unimaginable amounts and types of data, and the most progressive organizations are harnessing significant value in the process. Yes, there is a signal to go along with that noise. Big Data allows organizations to find the potential gold in the petabytes of tweets, texts, Facebook likes, blog posts and related comments, podcasts, photos, videos, and the like.

At a high level, *Too Big to Ignore: The Business Case for Big Data* makes a compelling business case for Big Data. Chapter 1, "Data 101 and the Data Deluge," provides the requisite background on Big Data and defines key terms, such as structured versus unstructured and semi-structured data. Chapter 2, "Demystifying Big Data," goes deeper. Much like *cloud computing* and *Web 2.0,* definitions of Big Data run the gamut. The chapter examines the major characteristics of Big Data. Chapter 3, "The Elements of Persuasion: Big Data Techniques," examines the specific techniques people are using to understand Big Data.

Chapter 4, "Big Data Solutions," moves us from theory to practice. We live in a world rife with data not easily represented in—much less effectively analyzed by—old standbys. To make Big Data come alive, new tools are needed, and this chapter introduces some of the most prevalent ones. Chapter 5, "Case Studies: The Big Rewards of Big Data," looks in depth at several organizations in different industries that have successfully deployed these tools.

Chapter 6, "Taking the Big Plunge," offers some advice on getting started with Big Data. Chapter 7, "Big Data: Big Issues and Big Problems," looks at the flip side of Big Data. We'll see that it's not all puppy dogs and ice cream. With Big Data, there is big danger. The book concludes with Chapter 8, "Looking Forward: The Future of Big Data." I'll offer some predictions about where Big Data is going and why it will cease to be a luxury in the coming years.

WHO SHOULD READ THIS BOOK?

Too Big to Ignore is about the increasingly important topic of Big Data. It makes the business case that there's tremendous value to be gleaned from the volume, variety, and velocity of information currently streaming at us. The book has no one intended audience, and many groups of people will benefit from reading it, including:

- CEOs, CIOs, and senior leaders who want to understand the fuss about Big Data
- Employees at consulting firms and software vendors who want to educate their current and prospective clients about the sea change that is Big Data
- Professors, school deans, and other academics who want to prepare their students about to enter the Big Data world
- Other people generally interested in Big Data and what they can do with it

Those looking for a technical guide on how to implement specific Big Data applications should look elsewhere. This is *not* a tactical, how-to book.

SUMMARY

For a variety of reasons, we are in the midst of a revolution of sorts. The Data Deluge has arrived, and it's only betting bigger. Car insurance, government, and recruiting are just three areas being transformed by Big Data. Organizations of all sorts are embracing it as we speak. And we're just getting started.

With the requisite foundation firmly in place, Chapter 1 looks at Big Data in a historical context. It examines how enterprise data has evolved in the Computer and Internet ages, with a particular emphasis on new data types and sources.

NOTES

1. Tucci, Joseph, "EMC CEO Says Big Data to Transform Every Industry," October 3, 2012, www.bloomberg.com/video/emc-ceo-says-big-data-to-transform-every-industry-XZ6adCHaTp~448vjZG97Zw.html, retrieved December 11, 2012.
2. LeBeau, Philip, "Texting and Driving Worse Than Drinking and Driving," June 25, 2009, www.cnbc.com/id/31545004/Texting_And_Driving_Worse_Than_Drinking_and_Driving, retrieved December 11, 2012.
3. "Snapshot Common Questions," www.progressive.com/auto/snapshot-common-questions.aspx, retrieved December 11, 2012.
4. Yvkoff, Liane, "Gadget Helps Progressive Offer Insurance Discount." March 21, 2011, http://reviews.cnet.com/8301-13746_7-20045433-48.html, retrieved December 11, 2012.
5. "'Street Bump' App Detects Potholes, Alerts Boston City Officials," July 20, 2012, www.foxnews.com/tech/2012/07/20/treet-bump-app-detects-potholes-alerts-boston-city-officials/, retrieved December 11, 2012.
6. Ngowi, Rodrique, "App Detects Potholes, Alerts Boston City Officials," July 20, 2012, www.boston.com/business/technology/articles/2012/07/20/app_detects_potholes_alerts_boston_city_officials/, retrieved December 11, 2012.
7. Walker, Joseph, "Meet the New Boss: Big Data Companies Trade In Hunch-Based Hiring for Computer Modeling," September 20, 2012, http://online.wsj.com/article/SB10000872396390443890304578006252019616768.html, retrieved December 11, 2012.
8. www.evolvondemand.com/home/about.php, retrieved December 11, 2012.
9. Kolakowski, Nick, "Big Data Spending Will Hit $28 Billion in 2012: Gartner," October 17, 2012, http://slashdot.org/topic/bi/big-data-spending-will-hit-28-billion-in-2012-gartner/, retrieved December 11, 2012.
10. "Billions and Billions: Big Data Becomes a Big Deal," 2012, www.deloitte.com/view/en_GX/global/industries/technology-media-telecommunications/tmt-predictions-2012/technology/70763e14447a4310VgnVCM1000001a56f00aRCRD.htm, retrieved December 11, 2012.
11. "Internet 2011 in Numbers," January 17, 2012, http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/, retrieved December 11, 2012.
12. Wray, Richard, "Internet Data Heads for 500bn Gigabytes," May 18, 2009, www.guardian.co.uk/business/2009/may/18/digital-content-expansion, retrieved December 11, 2012.
13. Wu, Suzanne, "How Much Information Is There in the World?" February 10, 2011, http://news.usc.edu/#!/article/29360/How-Much-Information-Is-There-in-the-World, retrieved December 11, 2012.

14. Fitchard, Kevin, "Despite Critics, Cisco Stands by Its Data Deluge," February 14, 2012, http://gigaom.com/2012/02/14/despite-critics-cisco-stands-by-its-data-deluge/, retrieved December 11, 2012.
15. Kolakowski, Nick, "CIOs See Big Data as Big Deal by 2013: Survey," May 10, 2012, http://slashdot.org/topic/bi/cios-see-big-data-as-big-deal/, retrieved December 11, 2012.
16. "Research Methodologies, Hype Cycles," copyright 2012, www.gartner.com/technology/research/methodologies/hype-cycle.jsp, retrieved December 11, 2012.
17. Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Hung Byers, Angela, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," May 2011, www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, retrieved December 11, 2012.
18. Anderson, Chris, "Tech Is Too Cheap to Meter: It's Time to Manage for Abundance, Not Scarcity," June 22, 2009, www.wired.com/techbiz/it/magazine/17-07/mf_freer?currentPage=all, retrieved December 11, 2012.
19. Davenport, Thomas H.; Patil, D.J., "Data Scientist: The Sexiest Job of the 21st Century," October 2012, http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/, retrieved December 11, 2012.
20. Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Hung Byers, Angela, "Big data: The next frontier for innovation, competition, and productivity," May 2011, www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, retrieved December 11, 2012.
21. "What Is Data Scientist?," www-01.ibm.com/software/data/infosphere/data-scientist/, retrieved December 11, 2012.
22. Harris, Derrick, "5 Trends That Are Changing How We Do Big Data," November 3, 2012, http://gigaom.com/data/5-trends-that-are-changing-how-we-do-big-data/, retrieved December 11, 2012.
23. www.greenplum.com/datasciencesummit, retrieved December 11, 2012.
24. Simon, Phil, "Big Data Goes Mainstream," May 24, 2012, www.huffingtonpost.com/phil-simon/big-data-goes-mainstream_b_1541079.html, retrieved December 11, 2012.
25. www.gartner.com/AnalystBiography?authorId=40872, retrieved December 11, 2012.
26. Borenstein, Seth, "Predicting Presidents, Storms and Life by Computer," November 12, 2012, www.weather.com/news/nate-silver-predicting-presidents-storms-20121111, retrieved December 11, 2012.
27. Craft, Matthew, "Hurricane Sandy's Economic Damage Could Reach $50 Billion, Equecat Estimates," November 1, 2012, www.huffingtonpost.com/2012/11/01/hurricane-sandy-economic-damage_n_2057850.html, retrieved December 11, 2012.
28. Blodget, Henry, "Nate Silver: Obama's Odds of Winning Are Now Back Over 70%," October 25, 2012, www.businessinsider.com/who-will-be-president-2012, retrieved December 11, 2012.
29. Wu, Joyce, "The Nate Silver Effect," November 12, 2012, http://cornellsun.com/section/opinion/content/2012/11/12/nate-silver-effect, retrieved December 11, 2012.