GETTING CONTROL OF BIG DATA
How vast new streams of information are changing the art of management

PAGE 59
Contents

Cover Artist
Chris Gash

SPOTLIGHT ON BIG DATA

60 Big Data: The Management Revolution To reap the benefits of big data, you may have to change your company's decision-making culture. Andrew McAfee and Erik Brynjolfsson

70 Data Scientist: The Sexiest Job of the 21st Century A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—and the competition for them is fierce. Thomas H. Davenport and D.J. Patil

78 Making Advanced Analytics Work for You Don't dismiss big data as mere hype—it really can transform your business. Here's how you can use it to stay competitive. Dominic Barton and David Court

ABOVE
Tamar Cohen
What About Willie?, 2008
silk screen collage on vintage book pages, 60" X 44"

HBR.ORG
What big data means for you and your company hbr.org/insights/big-data
Spotlight

Big Data: The Management Revolution 60
by Andrew McAfee and Erik Brynjolfsson

Data Scientist: The Sexiest Job of the 21st Century 70
by Thomas H. Davenport and D.J. Patil

Making Advanced Analytics Work for You 78
by Dominic Barton and David Court

Big Data

Businesses are collecting more data than they know what to do with. To turn all this information into competitive gold, they'll need new skills and a new management style.
ARTWORK Tamar Cohen

Happy Motoring, 2010, silk screen®
on vintage road map, 36" x 18"
Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.

by Andrew McAfee and Erik Brynjolfsson
"You can’t manage what you don’t measure."

There’s much wisdom in that saying, which has been attributed to both W. Edwards Deming and Peter Drucker, and it explains why the recent explosion of digital data is so important. Simply put, because of big data, managers can measure and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making and performance.

Consider retailing. Booksellers in physical stores could always track which books sold and which did not. If they had a loyalty program, they could tie some of those purchases to individual customers. And that was about it. Once shopping moved online, though, the understanding of customers increased dramatically. Online retailers could track not only what customers bought, but also what else they looked at; how they navigated through the site; how much they were influenced by promotions, reviews, and page layouts; and similarities across individuals and groups. Before long, they developed algorithms to predict what books individual customers would like to read next—algorithms that performed better every time the customer responded to or ignored a recommendation. Traditional retailers simply couldn’t access this kind of information, let alone act on it in a timely manner. It’s no wonder that Amazon has put so many brick-and-mortar bookstores out of business.

The familiarity of the Amazon story almost masks its power. We expect companies that were born digital to accomplish things that business executives could only dream of a generation ago. But in fact the use of big data has the potential to transform traditional businesses as well. It may offer them even greater opportunities for competitive advantage (online businesses have always known that they were competing on how well they understood their data). As we’ll discuss in more detail, the big data of this revolution is far more powerful than the analytics that were used in the past. We can measure and therefore manage more precisely than ever before. We can make better predictions and smarter decisions. We can target more-effective interventions, and can do so in areas that so far have been dominated by gut and intuition rather than by data and rigor.

As the tools and philosophies of big data spread, they will change long-standing ideas about the value of experience, the nature of expertise, and the practice of management. Smart leaders across industries will see using big data for what it is: a management revolution. But as with any other major change in business, the challenges of becoming a big data-enabled organization can be enormous and require hands-on—or in some cases hands-off—leadership. Nevertheless, it’s a transition that executives need to engage with today.

What’s New Here?

Business executives sometimes ask us, "Isn’t ‘big data’ just another way of saying ‘analytics’?" It’s true that they’re related: The big data movement, like analytics before it, seeks to glean intelligence from data and translate that into business advantage. However, there are three key differences:

Volume. As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the internet every second than were stored in the entire internet just 20 years ago. This gives companies an opportunity to work with many petabytes of data in a single data set—and not just from the internet. For instance, it is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions. A petabyte is one quadrillion bytes, or the equivalent of about 20 million filing cabinets’ worth of text. An exabyte is 1,000 times that amount, or one billion gigabytes.
Data-driven decisions are better decisions—it’s as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management.

Companies that were born digital, such as Google and Amazon, are already masters of big data. But the potential to gain competitive advantage from it may be even greater for other companies.

The managerial challenges, however, are very real. Senior decision makers have to embrace evidence-based decision making. Their companies need to hire scientists who can find patterns in data and translate them into useful business information. And whole organizations need to redefine their understanding of “judgment.”

**Velocity.** For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors. For instance, our colleague Alex “Sandy” Pentland and his group at the MIT Media Lab used location data from mobile phones to infer how many people were in Macy’s parking lots on Black Friday—the start of the Christmas shopping season in the United States. This made it possible to estimate the retailer’s sales on that critical day even before Macy’s itself had recorded those sales. Rapid insights like that can provide an obvious competitive advantage to Wall Street analysts and Main Street managers.

**Variety.** Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Many of the most important sources of big data are relatively new. The huge amounts of information from social networks, for example, are only as old as the networks themselves; Facebook was launched in 2004, Twitter in 2006. The same holds for smartphones and the other mobile devices that now provide enormous streams of data tied to people, activities, and locations. Because these devices are ubiquitous, it’s easy to forget that the iPhone was unveiled only five years ago, and the iPad in 2010. Thus the structured databases that stored most corporate information until recently are ill suited to storing and processing big data. At the same time, the steadily declining costs of all the elements of computing—storage, memory, processing, bandwidth, and so on—mean that previously expensive data-intensive approaches are quickly becoming economical.

As more and more business activity is digitized, new sources of information and ever-cheaper equipment combine to bring us into a new era: one in which large amounts of digital information exist on virtually any topic of interest to a business. Mobile phones, online shopping, social networks, electronic communication, GPS, and instrumented machinery all produce torrents of data as a by-product of their ordinary operations. Each of us is now a walking data generator. The data available are often unstructured—not organized in a database—and unwieldy, but there’s a huge amount of signal in the noise, simply waiting to be released. Analytics brought rigorous techniques to decision making; big data is at once simpler and more powerful. As Google’s director of research, Peter Norvig, puts it: “We don’t have better algorithms. We just have more data.”

**How Data-Driven Companies Perform**

The second question skeptics might pose is this: “Where’s the evidence that using big data intelligently will improve business performance?” The business press is rife with anecdotes and case studies that supposedly demonstrate the value of being data-driven. But the truth, we realized recently, is that nobody was tackling that question rigorously. To address this embarrassing gap, we led a team at the MIT Center for Digital Business, working in partnership with McKinsey’s business technology office and with our colleague Lorin Hitt at Wharton and the MIT doctoral student Heekyung Kim. We set out to test the hypothesis that data-driven companies would be better performers. We conducted structured interviews with executives at 330 public North American companies about their organizational and technology management practices, and gathered performance data from their annual reports and independent sources.

Not everyone was embracing data-driven decision making. In fact, we found a broad spectrum of attitudes and approaches in every industry. But across all the analyses we conducted, one relation-
Expertise from Surprising Sources

Often someone coming from outside an industry can spot a better way to use big data than an insider, just because so many new, unexpected sources of data are available. One of us, Erik, demonstrated this in research he conducted with Lynn Wu, now an assistant professor at Wharton. They used publicly available web search data to predict housing-price changes in metropolitan areas across the United States. They had no special knowledge of the housing market when they began their study, but they reasoned that virtually real-time search data would enable good near-term forecasts about the housing market—and they were right. In fact, their prediction proved more accurate than the official one from the National Association of Realtors, which had developed a far more complex model but relied on relatively slow-changing historical data.

This is hardly the only case in which simple models and big data trump more-elaborate analytics approaches.

Researchers at the Johns Hopkins School of Medicine, for example, found that they could use data from Google Flu Trends (a free, publicly available aggregator of relevant search terms) to predict surges in flu-related emergency room visits a week before warnings came from the Centers for Disease Control. Similarly, Twitter updates were as accurate as official reports at tracking the spread of cholera in Haiti after the January 2010 earthquake; they were also two weeks earlier.

ship stood out: The more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. In particular, companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors. This performance difference remained robust after accounting for the contributions of labor, capital, purchased services, and traditional IT investment. It was statistically significant and economically important and was reflected in measurable increases in stock market valuations.

So how are managers using big data? Let’s look in detail at two companies that are far from Silicon Valley upstarts. One uses big data to create new businesses, the other to drive more sales.

Improved Airline ETAs

Minutes matter in airports. So does accurate information about flight-arrival times: If a plane lands before the ground staff is ready for it, the passengers and crew are effectively trapped, and if it shows up later than expected, the staff sits idle, driving up costs. So when a major U.S. airline learned from an internal study that about 10% of the flights into its major hub had at least a 10-minute gap between the estimated time of arrival and the actual arrival time—and 30% had a gap of at least five minutes—it decided to take action.

At the time, the airline was relying on the aviation industry’s long-standing practice of using the ETAs provided by pilots. The pilots made these estimates during their final approach to the airport, when they had many other demands on their time and attention. In search of a better solution, the airline turned to PASSUR Aerospace, a provider of decision-support technologies for the aviation industry. In 2001 PASSUR began offering its own arrival estimates as a service called RightETA. It calculated these times by combining publicly available data about weather, flight schedules, and other factors with proprietary data the company itself collected, including feeds from a network of passive radar stations it had installed near airports to gather data about every plane in the local sky.

PASSUR started with just a few of these installations, but by 2012 it had more than 155. Every 4.6 seconds it collects a wide range of information about every plane that it “sees.” This yields a huge and constant flood of digital data. What’s more, the company keeps all the data it has gathered over time, so it has an immense body of multidimensional information spanning more than a decade. This allows sophisticated analysis and pattern matching. RightETA essentially works by asking itself “What happened all the previous times a plane approached this airport under these conditions? When did it actually land?”

After switching to RightETA, the airline virtually eliminated gaps between estimated and actual arrival times. PASSUR believes that enabling an airline to know when its planes are going to land and plan accordingly is worth several million dollars a year at each airport. It’s a simple formula: Using big data leads to better predictions, and better predictions yield better decisions.

Speedier, More Personalized Promotions

A couple of years ago, Sears Holdings came to the conclusion that it needed to generate greater value from the huge amounts of customer, product, and promotion data it collected from its Sears, Craftsman, and Lands’ End brands. Obviously, it would be valuable to combine and make use of all these data to tailor promotions and other offerings to customers, and to personalize the offers to take advantage of local conditions. Valuable, but difficult: Sears required about
eight weeks to generate personalized promotions, at
which point many of them were no longer optimal for
the company. It took so long mainly because the data
required for these large-scale analyses were both vo-
luminous and highly fragmented—housed in many
databases and “data warehouses” maintained by the
various brands.

In search of a faster, cheaper way to do its analytic
work, Sears Holdings turned to the technologies and
practices of big data. As one of its first steps, it set up a
Hadoop cluster. This is simply a group of inexpensive
commodity servers whose activities are coordinated
by an emerging software framework called Hadoop
(named after the toy elephant in the household of Doug
Cutting, one of its developers).

Sears started using the cluster to store incoming
data from all its brands and to hold data from existing
data warehouses. It then conducted analyses on
the cluster directly, avoiding the time-consuming
complexities of pulling data from various sources
and combining them so that they can be analyzed.
This change allowed the company to be much faster
and more precise with its promotions. According to
the company’s CTO, Phil Shelley, the time needed
to generate a comprehensive set of promotions
dropped from eight weeks to one, and is still drop-
ning. And these promotions are of higher quality, be-
cause they’re more timely, more granular, and more
personalized. Sears’s Hadoop cluster stores and pro-
cesses several petabytes of data at a fraction of the
cost of a comparable standard data warehouse.

Shelley says he’s surprised at how easy it has been
to transition from old to new approaches to data
management and high-performance analytics. Be-
cause skills and knowledge related to new data tech-
nologies were so rare in 2010, when Sears started the
transition, it contracted some of the work to a com-
pany called Cloudera. But over time its old guard of
IT and analytics professionals have become comfort-
able with the new tools and approaches.

The PASSUR and Sears Holding examples illus-
trate the power of big data, which allows more
accurate predictions, better decisions, and precise
interventions, and can enable these things at seem-
ingly limitless scale. We’ve seen big data used in sup-
ply chain management to understand why a carmak-
er’s defect rates in the field suddenly increased, in
customer service to continually scan and intervene
in the health care practices of millions of people, in
planning and forecasting to better anticipate online
sales on the basis of a data set of product character-
istics, and so on. We’ve seen similar payoffs in many
other industries and functions, from finance to mar-
ketinig to hotels and gaming, and from human re-
source management to machine repair.

Our statistical analysis tells us that what we’re see-
ing is not just a few flashy examples but a more
fundamental transformation of the economy. We’ve
become convinced that almost no sphere of business
activity will remain untouched by this movement.

**A New Culture of Decision Making**

The technical challenges of using big data are very
real. But the managerial challenges are even grea-
ter—starting with the role of the senior executive team.

**Muting the HIPPOs.** One of the most critical
aspects of big data is its impact on how decisions
are made and who gets to make them. When data
are scarce, expensive to obtain, or not available in
digital form, it makes sense to let well-placed people
make decisions, which they do on the basis of expe-
rience they’ve built up and patterns and relationships
they’ve observed and internalized. “Intuition”

---

Big data’s power does not erase the need for vision
or human insight.

---

is the label given to this style of inference and deci-
sion making. People state their opinions about what
the future holds—what’s going to happen, how well
something will work, and so on—and then plan ac-
cordingly. (See “The True Measures of Success,” by
Michael J. Moutbousin, in this issue.)

For particularly important decisions, these
people are typically high up in the organization, or
they’re expensive outsiders brought in because of
their expertise and track records. Many in the big
data community maintain that companies often
make most of their important decisions by relying
on “HIPPO”—the highest-paid person’s opinion.

To be sure, a number of senior executives are
genuinely data-driven and willing to override their
own intuition when the data don’t agree with it. But
we believe that throughout the business world today,
people rely too much on experience and intuition and not enough on data. For our research we constructed a 5-point composite scale that captured the overall extent to which a company was data-driven. Fully 39% of our respondents rated their companies at or below 3 on this scale.

**New roles.** Executives interested in leading a big data transition can start with two simple techniques. First, they can get in the habit of asking “What do the data say?” when faced with an important decision and following up with more-specific questions such as “Where did the data come from?,” “What kinds of analyses were conducted?,” and “How confident are we in the results?” (People will get the message quickly if executives develop this discipline.) Second, they can allow themselves to be overruled by the data; few things are more powerful for changing a decision-making culture than seeing a senior executive concede when data have disproved a hunch.

When it comes to knowing which problems to tackle, of course, domain expertise remains critical. Traditional domain experts—those deeply familiar with an area—are the ones who know where the biggest opportunities and challenges lie. PASSUR, for one, is trying to hire as many people as possible who have extensive knowledge of operations at America’s major airports. They will be invaluable in helping the company figure out what offerings and markets it should go after next.

As the big data movement advances, the role of domain experts will shift. They’ll be valued not for their HIPPO-style answers but because they know what questions to ask. Pablo Picasso might have been thinking of domain experts when he said, “Computers are useless. They can only give you answers.”

**Five Management Challenges**

Companies won’t reap the full benefits of a transition to using big data until they’re able to manage change effectively. Five areas are particularly important in that process.

**Leadership.** Companies succeed in the big data era not simply because they have more or better data, but because they have leadership teams that set clear goals, define what success looks like, and ask the right questions. Big data’s power does not erase the need for vision or human insight. On the contrary, we still must have business leaders who can spot a great opportunity, understand how a market is developing, think creatively and propose truly novel offerings, articulate a compelling vision, persuade people to embrace it and work hard to realize it, and deal effectively with customers, employees, stockholders, and other stakeholders. The successful companies of the next decade will be the ones whose leaders can do all that while changing the way their organizations make many decisions.

**Talent management.** As data become cheaper, the complements to data become more valuable. Some of the most crucial of these are data scientists and other professionals skilled at working with large quantities of information. Statistics are important, but many of the key techniques for using big data are rarely taught in traditional statistics courses. Perhaps even more important are skills in cleaning and organizing large data sets; the new kinds of data rarely come in structured formats. Visualization tools and techniques are also increasing in value. Along with the data scientists, a new generation of computer scientists are bringing to bear techniques for working with very large data sets. Expertise in the design of experiments can help cross the gap between correlation and causation. The best data scientists are also comfortable speaking the language of business and helping leaders reformulate their challenges in ways that big data can tackle. Not surprisingly, people with these skills are hard to find and in great demand. (See “Data Scientist: The Sexiest Job of the 21st Century,” by Thomas H. Davenport and D.J. Patil, in this issue.)

**Technology.** The tools available to handle the volume, velocity, and variety of big data have improved greatly in recent years. In general, these technologies are not prohibitively expensive, and much of the software is open source. Hadoop, the most commonly used framework, combines commodity hardware with open-source software. It takes incoming streams of data and distributes them onto cheap disks; it also provides tools for analyzing the data. However, these technologies do require a skill set that is new to most IT departments, which will need to work hard to integrate all the relevant internal and external sources of data. Although attention to technology isn’t sufficient, it is always a necessary component of a big data strategy.

**Decision making.** An effective organization puts information and the relevant decision rights in the same location. In the big data era, information is created and transferred, and expertise is often not where it is needed to be. The artful leader will create an organization flexible enough to minimize the “not
invented here” syndrome and maximize cross-functional cooperation. People who understand the problems need to be brought together with the right data, but also with the people who have problemsolving techniques that can effectively exploit them.

**Company culture.** The first question a data-driven organization asks itself is not “What do we think?” but “What do we know?” This requires a move away from acting solely on hunches and instinct. It also requires breaking a bad habit we’ve noticed in many organizations: pretending to be more data-driven than they actually are. Too often, we saw executives who spiced up their reports with lots of data that supported decisions they had already made using the traditional HIPPO approach. Only afterward were underlings dispatched to find the numbers that would justify the decision.

**Without question, many barriers to success remain.** There are too few data scientists to go around. The technologies are new and in some cases exotic. It’s too easy to mistake correlation for causation and to find misleading patterns in the data. The cultural challenges are enormous, and, of course, privacy concerns are only going to become more significant. But the underlying trends, both in the technology and in the business payoff, are unmistakable.

The evidence is clear: Data-driven decisions tend to be better decisions. Leaders will either embrace this fact or be replaced by others who do. In sector after sector, companies that figure out how to combine domain expertise with data science will pull away from their rivals. We can’t say that all the winners will be harnessing big data to transform decision making. But the data tell us that’s the surest bet.

Andrew McAfee is a principal research scientist at MIT’s Center for Digital Business and the author of *Enterprise 2.0* (Harvard Business School Press, 2009). Erik Brynjolfsson is the Schussel Family Professor at MIT’s Sloan School of Management and the director of its Center for Digital Business. They are the coauthors of *Race Against the Machine* (Digital Frontier Press, 2012).
Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."
Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people’s connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in. He could imagine that new features capitalizing on the heuristics he was developing might provide value to users. But LinkedIn’s engineering team, caught up in the challenges of scaling up the site, seemed uninterested. Some colleagues were openly dismissive of Goldman’s ideas. Why would users need LinkedIn to figure out their networks for them? The site already had an address book importer that could pull in all a member’s connections.

Luckily, Reid Hoffman, LinkedIn’s cofounder and CEO at the time (now its executive chairman), had faith in the power of analytics because of his experiences at PayPal, and he had granted Goldman a high degree of autonomy. For one thing, he had given Goldman a way to circumvent the traditional product release cycle by publishing small modules in the form of ads on the site’s most popular pages.

Through one such module, Goldman started to test what would happen if you presented users with names of people they hadn’t yet connected with but seemed likely to know—for example, people who had shared their tenures at schools and workplaces. He did this by ginning up a custom ad that displayed the three best new matches for each user based on the background entered in his or her LinkedIn profile. Within days it was obvious that something remarkable was taking place. The click-through rate on those ads was the highest ever seen. Goldman continued to refine how the suggestions were generated, incorporating networking ideas such as “triangle closing”—the notion that if you know Larry and Sue, there’s a good chance that Larry and Sue know each other. Goldman and his team also got the action required to respond to a suggestion down to one click.

It didn’t take long for LinkedIn’s top managers to recognize a good idea and make it a standard feature. That’s when things really took off. “People You May Know” ads achieved a click-through rate 30% higher than the rate obtained by other prompts to visit more pages on the site. They generated millions of new page views. Thanks to this one feature, LinkedIn’s growth trajectory shifted significantly upward.

A New Breed
Goldman is a good example of a new key player in organizations: the “data scientist.” It’s a high-ranking professional with the training and curiosity to make discoveries in the world of big data. The title has been around for only a few years. (It was coined in 2008 by one of us, D.J. Patil, and Jeff Hammerbacher, then the respective leads of data and analytics efforts at LinkedIn and Facebook.) But thousands of data scientists are already working at both start-ups and well-established companies. Their sudden appearance on the business scene reflects the fact that companies are now wrestling with information that comes in varieties and volumes never encountered before. If your organization stores multiple petabytes of data, if the information most critical to your business resides in forms other than rows and columns of numbers, or if answering your biggest question would involve a “mashup” of several analytical efforts, you’ve got a big data opportunity.

Much of the current enthusiasm for big data focuses on technologies that make taming it possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualization. While those are important breakthroughs, at least as important are the people with the skill set (and the mind-set) to put them to good use. On this front, demand has raced ahead of supply. Indeed, the shortage of data scientists is becoming a serious constraint in some sectors. Greylock Partners, an early-stage venture firm that has backed companies such as Facebook, LinkedIn, Palo Alto Networks, and Workday, is worried enough about the tight labor pool that it has built its own specialized recruiting team to channel talent to businesses in its portfolio. “Once they have data,” says Dan Portillo, who leads
A new role is fast gaining prominence in organizations: that of the data scientist. Data scientists are the people who understand how to fish out answers to important business questions from today’s tsunami of unstructured information. As companies rush to capitalize on the potential of big data, the largest constraint many face is the scarcity of this special talent.

No university programs have yet been designed to churn out data scientists, so recruiting them requires creativity. Look for achievers in any field with a strong data and computational focus, which might take you as far afield from business as experimental physics or systems biology.

Recognize, too, that the aspects of a job that will attract and retain a data scientist may differ from what makes other professionals happy.

Data scientists need autonomy but want to be “on the bridge,” responding to management issues with their managerial colleagues in real time. Money counts as a signal of value, but in a fast-evolving discipline, the ability to make one’s mark by working on the most intriguing problems and tapping into the richest data flows may count more.

that team, “they really need people who can manage it and find insights in it.”

Who Are These People?
If capitalizing on big data depends on hiring scarce data scientists, then the challenge for managers is to learn how to identify that talent, attract it to an enterprise, and make it productive. None of those tasks is as straightforward as it is with other, established organizational roles. Start with the fact that there are no university programs offering degrees in data science. There is also little consensus on where the role fits in an organization, how data scientists can add the most value, and how their performance should be measured.

The first step in filling the need for data scientists, therefore, is to understand what they do in businesses. Then ask, What skills do they need? And what fields are those skills most readily found in?

More than anything, what data scientists do is make discoveries while swimming in data. It’s their preferred method of navigating the world around them. At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data.

Data scientists realize that they face technical limitations, but they don’t allow that to bog down their search for novel solutions. As they make discoveries, they communicate what they’ve learned and suggest its implications for new business directions. Often they are creative in displaying information visually and making the patterns they find clear and compelling. They advise executives and product managers on the implications of the data for products, processes, and decisions.

Given the nascent state of their trade, it often falls to data scientists to fashion their own tools and even conduct academic-style research. Yahoo, one of the firms that employed a group of data scientists early on, was instrumental in developing Hadoop. Facebook’s data team created the language Hive for programming Hadoop projects. Many other data scientists, especially at data-driven companies such as Google, Amazon, Microsoft, Walmart, eBay, LinkedIn, and Twitter, have added to and refined the tool kit.

What kind of person does all this? What abilities make a data scientist successful? Think of him or her as a hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful—and rare.

Data scientists’ most basic, universal skill is the ability to write code. This may be less true in five years’ time, when many more people will have the title “data scientist” on their business cards. More enduring will be the need for data scientists to communicate in language that all their stakeholders understand—and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or—ideally—both.

But we would say the dominant trait among data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested. This often entails the associative thinking that characterizes the most creative scientists in any field. For example, we know of a data scientist studying a fraud problem who realized that it was analogous to a type of DNA sequencing problem. By bringing together those disparate worlds, he and his team were able to craft a solution that dramatically reduced fraud losses.
Perhaps it's becoming clear why the word "scientist" fits this emerging role. Experimental physicists, for example, also have to design equipment, gather data, conduct multiple experiments, and communicate their results. Thus, companies looking for people who can work with complex data have had good luck recruiting among those with educational and work backgrounds in the physical or social sciences. Some of the best and brightest data scientists are PhDs in esoteric fields like ecology and systems biology. George Roumeliotis, the head of a data science team at Intuit in Silicon Valley, holds a doctorate in astrophysics. A little less surprisingly, many of the data scientists working in business today were formally trained in computer science, math, or economics. They can emerge from any field that has a strong data and computational focus.

It's important to keep that image of the scientist in mind—because the word "data" might easily send a search for talent down the wrong path. As Porillo told us, "The traditional backgrounds of people you saw 10 to 15 years ago just don't cut it these days." A quantitative analyst can be great at analyzing data but not at subduing a mass of unstructured data and getting it into a form in which it can be analyzed. A data management expert might be great at generating and organizing data in structured form but not at turning unstructured data into structured data—and also not at actually analyzing the data. And while people without strong social skills might thrive in traditional data professions, data scientists must have such skills to be effective.

Roumeliotis was clear with us that he doesn't hire on the basis of statistical or analytical capabilities. He begins his search for data scientists by asking candidates if they can develop prototypes in a mainstream programming language such as Java. Roumeliotis seeks both a skill set—a solid foundation in math, statistics, probability, and computer science—and certain habits of mind. He wants people with a feel for business issues and empathy for customers.

Then, he says, he builds on all that with on-the-job training and an occasional course in a particular technology.

Several universities are planning to launch data science programs, and existing programs in analytics, such as the Master of Science in Analytics program at North Carolina State, are busy adding big data exercises and coursework. Some companies are also trying to develop their own data scientists. After acquiring the big data firm Greenplum, EMC decided that the availability of data scientists would be a gating factor in its own—and customers'—exploitation of big data. So its Education Services division launched a data science and big data analytics training and certification program. EMC makes the program available to both employees and customers, and some of its graduates are already working on internal big data initiatives.

As educational offerings proliferate, the pipeline of talent should expand. Vendors of big data technologies are also working to make them easier to use. In the meantime one data scientist has come up

How to Find the Data Scientists You Need

1. Focus recruiting at "suspect" universities (Stanford, MIT, Berkeley, Harvard, Carnegie Mellon) and also at a few others with proven strengths: North Carolina State, UC Santa Cruz, the University of Maryland, the University of Washington, and UT Austin.

2. Scan the membership rolls of user groups devoted to data science tools. The R User Groups (for an open-source statistical tool favored by data scientists) and Python Interest Groups (for PIGgies) are good places to start.

3. Search for data scientists on LinkedIn—they're almost all on there, and you can see if they have the skills you want.

4. Hang out with data scientists at the Strata, Structure:Data, and Hadoop World conferences and similar gatherings (there is almost one a week now) or at informal data scientist "meet-ups" in the Bay Area, Boston, New York, Washington, DC, London, Singapore, and Sydney.

5. Make friends with a local venture capitalist, who is likely to have gotten a variety of big data proposals over the past year.

6. Host a competition on Kaggle or TopCoder, the analytics and coding competition sites. Follow up with the most creative entrants.

7. Don't bother with any candidate who can't code. Coding skills don't have to be at a world-class level but should be good enough to get by. Look for evidence, too, that candidates learn rapidly about new technologies and methods.

8. Make sure a candidate can find a story in a data set and provide a coherent narrative about a key data insight. Test whether he or she can communicate with numbers, visually, and verbally.

9. Be wary of candidates who are too detached from the business world. When you ask how their work might apply to your management challenges, are they stuck for answers?

10. Ask candidates about their favorite analysis or insight, and how they are keeping their skills sharp. Have they gotten a certificate in the advanced track of Stanford's online Machine Learning course, contributed to open-source projects, or built an online repository of code to share (for example, on GitHub)?
Data scientists want to build things, not just give advice. One describes being a consultant as “the dead zone.”

with a creative approach to closing the gap. The Insight Data Science Fellows Program, a postdoctoral fellowship designed by Jake Klamka (a high-energy physicist by training), takes scientists from academia and in six weeks prepares them to succeed as data scientists. The program combines mentoring by data experts from local companies (such as Facebook, Twitter, Google, and LinkedIn) with exposure to actual big data challenges. Originally aiming for 10 fellows, Klamka wound up accepting 30, from an applicant pool numbering more than 200. More organizations are now lining up to participate. “The demand from companies has been phenomenal,” Klamka told us, “They just can’t get this kind of high-quality talent.”

Why Would a Data Scientist Want to Work Here?

Even as the ranks of data scientists swell, competition for top talent will remain fierce. Expect candidates to size up employment opportunities on the basis of how interesting the big data challenges are. As one of them commented, “If we wanted to work with structured data, we’d be on Wall Street.” Given that today’s most qualified prospects come from nonbusiness backgrounds, hiring managers may need to figure out how to paint an exciting picture of the potential for breakthroughs that their problems offer.

Pay will of course be a factor. A good data scientist will have many doors open to him or her, and salaries will be bid upward. Several data scientists working at start-ups commented that they’d demanded and got large stock option packages. Even for someone accepting a position for other reasons, compensation signals a level of respect and the value the role is expected to add to the business. But our informal survey of the priorities of data scientists revealed something more fundamentally important. They want to be “on the bridge.” The reference is to the 1960s television show Star Trek, in which the starship captain James Kirk relies heavily on data supplied by

Dr. Spock. Data scientists want to be in the thick of a developing situation, with real-time awareness of the evolving set of choices it presents.

Considering the difficulty of finding and keeping data scientists, one would think that a good strategy would involve hiring them as consultants. Most consulting firms have yet to assemble many of them. Even the largest firms, such as Accenture, Deloitte, and IBM Global Services, are in the early stages of leading big data projects for their clients. The skills of the data scientists they do have on staff are mainly being applied to more-conventional quantitative analysis problems. Offshore analytics services firms, such as Mu Sigma, might be the ones to make the first major inroads with data scientists.

But the data scientists we’ve spoken with say they want to build things, not just give advice to a decision maker. One described being a consultant as “the dead zone—all you get to do is tell someone else what the analyses say they should do.” By creating solutions that work, they can have more impact and leave their marks as pioneers of their profession.

Care and Feeding

Data scientists don’t do well on a short leash. They should have the freedom to experiment and explore possibilities. That said, they need close relationships with the rest of the business. The most important ties for them to forge are with executives in charge of products and services rather than with people overseeing business functions. As the story of Jonathan Goldman illustrates, their greatest opportunity to add value is not in creating reports or presentations for senior executives but in innovating with customer-facing products and processes.

LinkedIn isn’t the only company to use data scientists to generate ideas for products, features, and value-adding services. At Intuit data scientists are asked to develop insights for small-business customers and consumers and report to a new senior vice president of big data, social design, and marketing. GE is already using data science to optimize
the service contracts and maintenance intervals for industrial products. Google, of course, uses data scientists to refine its core search and ad-serving algorithms. Zynga uses data scientists to optimize the game experience for both long-term engagement and revenue. Netflix created the well-known Netflix Prize, given to the data science team that developed the best way to improve the company’s movie recommendation system. The test-preparation firm Kaplan uses its data scientists to uncover effective learning strategies.

Data scientists today are akin to Wall Street “quants” of the 1980s and 1990s. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren’t a lot of people with their combination of scientific background and computational and analytical skills.

Data scientists today are akin to Wall Street “quants” of the 1980s and 1990s. In those days people with backgrounds in physics and math streamed to investment banks and hedge funds, where they could devise entirely new algorithms and data strategies. Then a variety of universities developed master’s programs in financial engineering, which churned out a second generation of talent that was more accessible to mainstream firms. The pattern was repeated later in the 1990s with search engineers, whose rarefied skills soon came to be taught in computer science programs.

One question raised by this is whether some firms would be wise to wait until that second generation of data scientists emerges, and the candidates are more numerous, less expensive, and easier to vet and assimilate in a business setting. Why not leave the trouble of hunting down and domesticating exotic talent to the big data start-ups and to firms like GE and Walmart, whose aggressive strategies require them to be at the forefront?

The problem with that reasoning is that the advance of big data shows no signs of slowing. If companies sit out this trend’s early days for lack of talent, they risk falling behind as competitors and channel partners gain nearly unassailable advantages. Think of big data as an epic wave gathering now, starting to crest. If you want to catch it, you need people who can surf.

There is, however, a potential downside to having people with sophisticated skills in a fast-evolving field spend their time among general management colleagues. They’ll have less interaction with similar specialists, which they need to keep their skills sharp and their tool kit state-of-the-art. Data scientists have to connect with communities of practice, either within large firms or externally. New conferences and informal associations are springing up to support collaboration and technology sharing, and companies should encourage scientists to become involved in them with the understanding that “more water in the harbor floats all boats.”

Data scientists tend to be more motivated, too, when more is expected of them. The challenges of accessing and structuring big data sometimes leave little time or energy for sophisticated analytics involving prediction or optimization. Yet if executives make it clear that simple reports are not enough, data scientists will devote more effort to advanced analytics. Big data shouldn’t equal “small math.”

The Hot Job of the Decade

Hal Varian, the chief economist at Google, is known to have said, “The sexy job in the next 10 years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?”

If “sexy” means having rare qualities that are much in demand, data scientists are already there.
Making Advanced Analytics Work For You

A practical guide to capitalizing on big data
by Dominic Barton and David Court

Big data and analytics have rocketed to the top of the corporate agenda. Executives look with admiration at how Google, Amazon, and others have eclipsed competitors with powerful new business models that derive from an ability to exploit data. They also see that big data is attracting serious investment from technology leaders such as IBM and Hewlett-Packard. Meanwhile, the tide of private-equity and venture-capital investments in big data continues to swell.

The trend is generating plenty of hype, but we believe that senior leaders are right to pay attention. Big data could transform the way companies do business, delivering the kind of performance gains last seen in the 1990s, when organizations redesigned their core processes. As data-driven strategies take hold, they will become an increasingly important point of competitive differentiation. According to research by Andrew McAfee and Erik Brynjolfsson, of MIT, companies that inject big data and analytics into their operations show productivity rates and profitability that are 5% to 6% higher than those of their peers (see "Big Data: The Management Revolution" in this issue).

Even so, our experience reveals that most companies are unsure how to proceed. Leaders are understandably leery of making substantial investments in big data and advanced analytics. They’re convinced that their organizations simply aren’t ready. After all, companies may not fully understand the data they already have, or perhaps they’ve lost piles of money on data-warehousing programs that never meshed with business processes, or maybe their current analytics programs are too complicated or don’t yield insights that can be put to use. Or all of the above. No wonder skepticism abounds.

Many CEOs, too, recall their experiences with customer relationship management in the mid-1990s, when new CRM software products often prompted great enthusiasm. Experts descended on boardrooms promising impressive results if new IT systems were built to collect massive amounts of customer data.
It didn't turn out that way. Too many C-suites were blind to the practical implications of new CRM technologies—namely, that to capitalize on them, organizations would have to make complex process changes and build employees' skills. The promised gains in performance were often slow in coming, because the systems remained stubbornly disconnected from how companies and frontline managers actually made decisions, and new demands for data management added complexity to operations. To be fair, most companies eventually managed to get their CRM programs on track, but not before some had suffered sizable losses and several CRM champions had lost career momentum.

Given this history, we empathize with executives who are cautious about big data. Nevertheless, we believe that the time has come to define a pragmatic approach to big data and advanced analytics—one tightly focused on how to use the data to make better decisions.

In our work with dozens of companies in six data-rich industries, we have found that fully exploiting data and analytics requires three mutually supportive capabilities. (See the exhibit “How to Benefit from Big Data.”) First, companies must be able to identify, combine, and manage multiple sources of data. Second, they need the capability to build advanced analytics models for predicting and optimizing outcomes. Third, and most critical, management must possess the muscle to transform the organization so that the data and models actually yield better decisions. Two important features underpin those activities: a clear strategy for how to use data and analytics to compete, and deployment of the right technology architecture and capabilities.

Equally important, the desired business impact must drive an integrated approach to data sourcing, model building, and organizational transformation. That's how you avoid the common trap of staring with the data and simply asking what it can do for you. Leaders should invest sufficient time and energy in aligning managers across the organization in support of the mission.

1. Choose the Right Data

The universe of data and modeling has changed vastly over the past few years. The sheer volume of information, particularly from new sources such as social media and machine sensors, is growing rapidly. The opportunity to expand insights by combining data is also accelerating, as more-powerful, less costly software abounds and information can be accessed from almost anywhere at any time. Bigger and better data give companies both more-panoramic and more-granular views of their business environment. The ability to see what was previously invisible improves operations, customer experiences, and strategy. But mastering that environment means upsing your game, finding deliberate and creative ways to identify usable data you already have, and exploring surprising sources of information.

Source data creatively. Often companies already have the data they need to tackle business problems, but managers simply don't know how the information can be used for key decisions. Operations executives, for instance, might not grasp the potential value of the daily or hourly factory and customer-service data they possess. Companies can impel a more comprehensive look at information sources by being specific about business problems they want to solve or opportunities they hope to exploit. For example, a banking team that needed to improve the efficiency of its customer-service operations created a 360-degree view by combining information from ATM transactions, online queries, customer complaints, and so on. That allowed duplicative interactions to be identified, thereby reducing costs and streamlining the customer experience.

Managers also need to get creative about the potential of external and new sources of data. Social media are generating terabytes of nontraditional, unstructured data in the form of conversations, photos, and video. Add to that the streams of data flowing in from sensors, monitoring processes, and external sources that range from local demographics to weather forecasts. One way to prompt broader thinking about potential data is to ask, "What decl-
The trend toward big data is growing rapidly, and senior leaders can’t afford to dismiss it as hype.

Advanced analytics is likely to become a decisive competitive asset in many industries and a core element in companies’ efforts to improve performance. It’s a mistake to assume that acquiring the right kind of big data is all that matters. Also essential is developing analytics tools that focus on business outcomes and that are relevant and easy to use for everyone from the C-suite to the front lines. That requires transforming your organization’s culture and capabilities, not in a rush to action but in a deliberative effort to weave big data into the fabric of daily operations.

Senior executives can take the lead here. The CEO of one major packaged-goods company told us that he views data as a strategic asset whose value he takes into account when assessing potential acquisitions. But leaders at all levels must also be attuned to novel approaches to gathering and leveraging information. As business practices in the internet era continue to evolve, inspiration can often arise from a scan of the external environment. A corporate finance executive, for instance, might look to a company such as Kabbage, a start-up that supplies working capital to online businesses. To slash the time required to underwrite loans, Kabbage asks merchants to opt in to sharing their customer-feedback ratings, Facebook interactions, and electronic shipping records. Those with the strongest feedback and highest business volume receive greater financing.

Get the necessary IT support. Legacy IT structures may hinder new types of data sourcing, storage, and analysis. Existing IT architecture may prevent the integration of siloed information, and managing unstructured data often remains beyond traditional IT capabilities. Many legacy systems were built to deliver data in batches, so they can’t furnish continuous flows of information for real-time decisions.

Fully resolving these issues often takes years. However, business leaders can address short-term big data needs by working with CIOs to prioritize requirements. This means quickly identifying and connecting the most important data for use in analytics, followed by a cleanup operation to synchronize and merge overlapping data and then to work around missing information. Such short-term tactics may lead companies to vendors that focus on analytics services or emerging software. New cloud-based technologies may also offer ways to scale computing power up or down to meet big data demands cost-effectively. Together those approaches establish an IT infrastructure that propels innovation by facilitating collaboration, rapid analysis, and experimentation.

2. Build Models That Predict and Optimize Business Outcomes

Data are essential, but performance improvements and competitive advantage arise from analytics models that allow managers to predict and optimize outcomes. More important, the most effective approach to building a model rarely starts with the data; instead it originates with identifying the business opportunity and determining how the model can improve performance.

Unfortunately, not all model building follows this course. One approach that gets inconsistent results, for instance, is simple data mining. Corralling huge data sets allows companies to run dozens of statistical tests to identify submerged patterns, but that provides little benefit if managers can’t effectively use the correlations to enhance business performance. A pure data-mining approach often leads to an endless search for what the data really say.

One company followed a more targeted strategy to optimize complex product pricing. At its core was a model based on the historical price elasticity of its products, sales data, competitors’ responses, and other variables. To improve its chances of success, the company began the modeling process by positing which factors affected sales volumes (for instance, competitors’ pricing and promotions) and then asked what data and which model would best deliver insights that were useful for making business decisions. We have found that such hypothesis-led modeling generates faster outcomes and also roots...
models in practical data relationships that are more broadly understood by managers.

Remember, too, that any modeling exercise has inherent risk. Although advanced statistical methods indisputably make for better models, statistics experts sometimes design models that are too complex to be practical. For example, a predictive model with 30 variables may explain historical data with high accuracy, but managing so many variables will exhaust most organizations' capabilities. Companies should repeatedly ask, "What's the least complex model that would improve our performance?"

3. **Transform Your Company's Capabilities**

The lead concern expressed to us by senior executives is that their managers don't understand or trust big data-based models. One large retailer intended its model to optimize returns on advertising spending, but despite considerable investment, it wasn't being used. The reason soon became evident: The frontline marketers who made key decisions on ad spending didn't believe the model's results and had little familiarity with how it worked.

Many companies grapple with such problems, often because of a mismatch between the organization's existing culture and capabilities and the emerging tactics to exploit analytics successfully. In short, the new approaches don't align with how companies actually arrive at decisions, or they fail to provide a clear blueprint for realizing business goals. Tools seem to be designed for experts in modeling rather than for people on the front lines, and few managers find the models engaging enough to champion their use—a key failing if companies want the new methods to permeate the organization. Bottom line: Using big data requires thoughtful organizational change, and three areas of action can get you there.

**Develop business-relevant analytics that can be put to use.** Like early CRM misadventures, many initial implementations of big data and analytics fail simply because they aren't in sync with the company's day-to-day processes and decision-making norms. The aforementioned case of a company that aimed to optimize prices illustrates how to avoid those common pain points. The company started with an analytics task force that convened a series of meetings with pricing and promotions managers to better understand the types of decisions they made when setting prices—and how those choices ultimately affected revenue and customer retention. Model designers also inquired about the types of business judgments that managers make to align their actions with broader company goals. These conversations ensured that both pricing analytics and resulting scenario tools would complement existing decision processes. The modeling allowed the company to reach its ultimate goal: more-effective management of price and volume trade-offs as product launches proliferated.

**Embed analytics into simple tools for the front lines.** Managers need transparent methods for using the new models and algorithms on a daily basis. By necessity, terabytes of data and sophisticated modeling are required to sharpen marketing, risk management, and operations. The key is to separate the statistics experts and software developers from the managers who use the data-driven insights. One large industrial company, for instance, sought to better forecast workforce needs to reflect local market variations. Historically, as the company had tried to keep labor costs low, it had often found itself short-staffed in some markets, leading to significant overtime costs and service snafus.

To remedy the problem, the company convened a small working group of analysts and IT programmers who developed a series of predictive models that

---

**How to Benefit from Big Data**

To improve performance with advanced analytics, companies need to develop strengths in three areas.

1. **Multiple Data Sources**
   Creatively source internal and external data.
   Upgrade IT architecture and infrastructure for easy merging of data.

2. **Prediction and Optimization Models**
   Focus on the biggest drivers of performance.
   Build models that balance complexity with ease of use.

3. **Organizational Transformation**
   Create simple, understandable tools for people on the front lines.
   Update processes and develop capabilities to enable tool use.
forecast workforce availability on the basis of factors such as vacation time, absenteeism, and work rules in labor contracts. The models incorporated millions of new data points on thousands of employees across dozens of locations. But rather than providing managers with reams of data and complex models, they created a simple visual interface that highlighted projected workforce needs and necessary actions. Ultimately, that approach of using a simple tool to deliver complex analytics substantially improved workforce planning and reduced the need for new hires and overtime.

**Develop capabilities to exploit big data.**
Even with simple and usable models, most organizations will need to upgrade their analytical skills and literacy. Managers must come to view analytics as central to solving problems and identifying opportunities—to make it part of the fabric of daily operations. Efforts will vary depending on a company’s goals and desired timeline. Adult learners often benefit from a “field and forum” approach, whereby they participate in real-world, analytics-based workplace decisions that allow them to learn by doing.

At one industrial services company, the mission was to get basic analytics tools into the hands of its roughly 200 sales managers. Training began with an in-field assignment to read a brief document and collect basic facts about the market. Next managers met in centralized, collaborative training sessions during which they figured out how to use the tools and market facts to improve sales performance. They then returned to the field to apply what they had learned and, several weeks later, reconvened to review progress, receive coaching, and learn about second-order analysis of their data. This process enabled a four-person team to eventually build capabilities across the entire sales management organization.

Adjusting culture and mind-sets typically requires a multifaceted approach that includes training, role modeling by leaders, and incentives and metrics to reinforce behavior. One large consumer-products company applied such an approach successfully. It created a sophisticated program to improve the profitability of promotional spending with its retailers. The launch included training—led by company management—and a new promotions-analysis tool for sales representatives. However, after an initial whirlwind of activity, the program and use of the tool fizzled. The obstacle was that company incentives and reporting protocols for sales managers tracked sales and sales growth, not profits. As a result, the managers considered the profit-focused program to be bureaucratic overhead that was unrelated to their key sales goals. After a series of discussions with the managers, the company re-launched the program, offered new incentives for improving profits, and tailored reports to profit-related data. Although ongoing training and coaching was necessary, the efforts gradually produced a shift in mind-set such that the power of promotions analytics is now used to further the common goal of increasing profitability.

**THE ERA OF BIG DATA** is evolving rapidly, and our experience suggests that most companies should act now. But rather than undertaking massive overhauls of their companies, executives should concentrate on targeted efforts to source data, build models, and transform the organizational culture. Such efforts will play a part in maintaining flexibility. That nimbleness is essential, given that the information itself—along with the technology for managing and analyzing it—will continue to grow and change, yielding a constant stream of opportunities. As more companies learn the core skills of using big data, building superior capabilities may soon become a decisive competitive asset. 

October 2012 Harvard Business Review 85