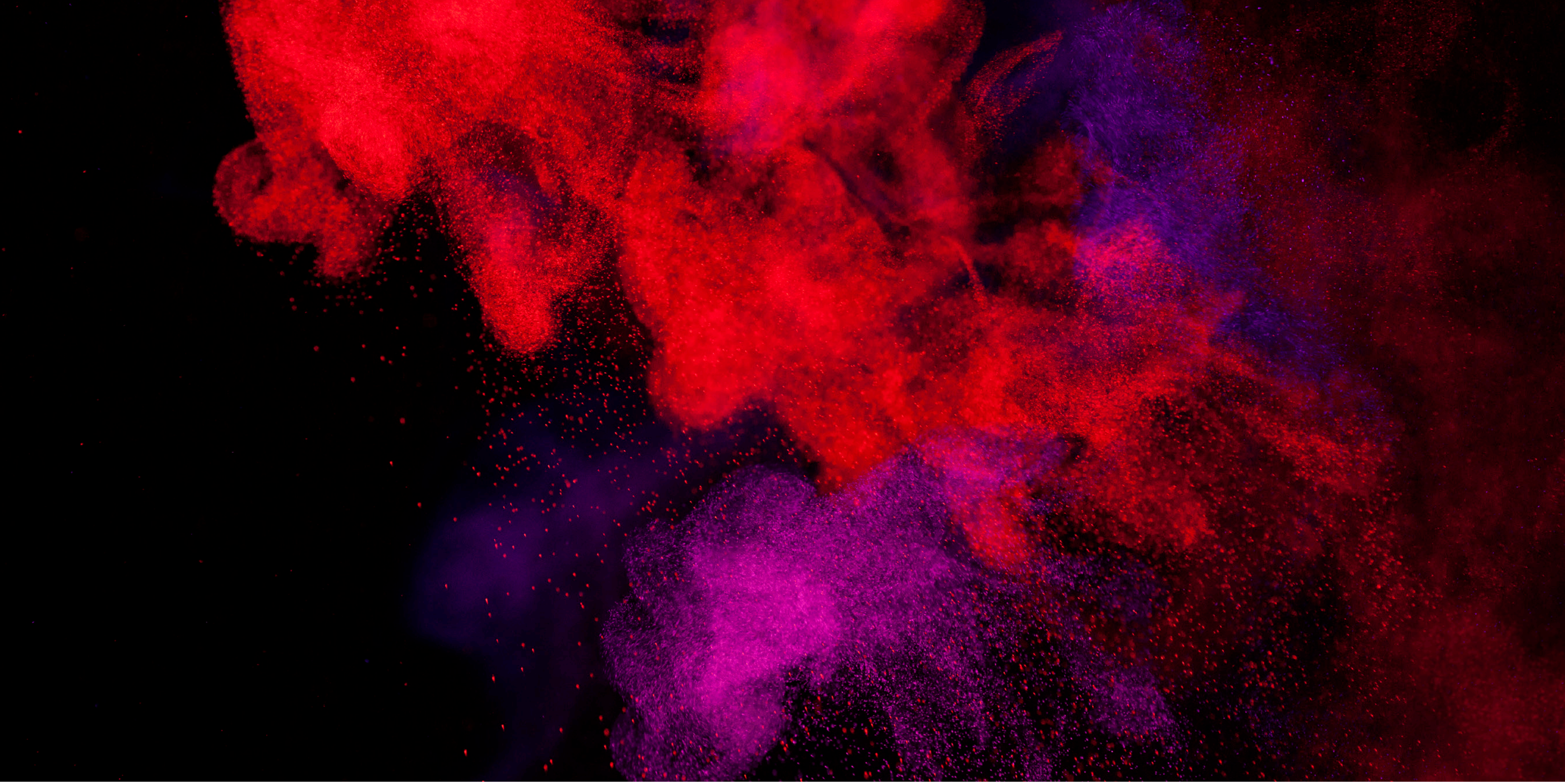


Mastering the art of data science

How to craft cohesive teams that create business results

IBM Institute for Business Value

Executive Report

Business analytics and strategy

IBM Big Data and Analytics

The IBM Big Data and Analytics practice integrates management consulting expertise with the science of analytics to enable leading organizations to succeed. IBM is helping clients realize the full potential of big data and analytics by providing them with the expertise, solutions and capabilities needed to infuse intelligence into virtually every business decision and process; empower more immediate and certain action by capitalizing on the many forms of data and insights that exist; and develop a culture of trust and confidence through a proactive approach to security, governance and compliance. For more information about IBM Big Data and Analytics offerings from IBM, visit www.ibm.biz/bigdataandanalytics

Filling a critical role intelligently

Data scientists have become increasingly invaluable to senior executives around the world, helping to inform decision-making processes as they seek to grow revenues, increase operational efficiency, improve impact analyses, reduce cost and increase operational efficiency. Successful data scientists work with business leaders to understand the challenges, define analytic solutions and analyze data to uncover innovations and insights. Organizations that want to build or expand a data science team in today's environment, where qualified candidates are in short supply, need to understand how to proceed intelligently. This report provides expert advice on how to build a strong team with relevant experience, a solid academic foundation, varied backgrounds and skills, and an ability to envision and work toward outcomes.

Executive summary

Data science is a hot topic in the C-suites of today's digital age. Executives leading digital disruptions recount tales of valuable insights uncovered through the application of data science to solve complex business problems, and their global audiences often appear spellbound. But executives wanting to join the quest for deep, data-driven insights and value need to understand the requirements — and the risks — when building a powerful data sciences program.

McKinsey & Company projects that demand for deep analytical professionals could exceed the supply by 140,000 to 190,000 positions in the United States alone, noting that this supply constraint will be global. The research firm cites the difficulty of producing this type of talent, estimating that it takes "years of training in the case of someone with intrinsic mathematical abilities."[1]

The demand for data scientists is so great the recruiting company Glassdoor ranks it as the top job in the United States, giving it high marks for both job satisfaction and career opportunities.[2] One recent analysis of LinkedIn's global database, which many consider representative of the marketspace, found more than 60,000 job openings for data scientists.[3] However, another analysis found only 11,400 professionals worldwide with the required skills.[4] Moreover, while there has been "impressive growth" in the number of data scientist positions, at least 52 percent of all LinkedIn's self-identified data scientists have earned that title within the past four years.[5]

Good data scientists **apply an art and a science to deliver** the insights required to solve organizations' biggest challenges.

Data scientists are **in high demand, and competition is fierce** for the relatively small number of qualified candidates.

Data scientists are well-compensated professionals, and **they are expensive to hire and costly to retain**.

These are well-compensated professionals, both expensive to hire and costly to retain. The U.S. Department of Commerce found that, on average, wages were 68 percent higher for workers in data occupations than for all private workers, and data scientists are among the most well-compensated of all data workers.[6]

This supply-versus-demand imbalance creates a risk for executives looking to hire. Insights aimed at solving an organization's biggest challenges – operational optimization, revenue generation, innovation — require the art and science that a true data scientist can bring to bear; the wrong approach or flawed analysis from an inexperienced or inadequately trained data scientist can have catastrophic consequences (see Figure 1).[7]

Hiring a data scientist is just one step in creating a successful data science program capable of delivering beguiling solutions. Achieving those results requires that executives also think about organizational structures, tools, access and outcomes, too.

As a trio, we have logged decades as hands-on data scientists tackling strategically significant business challenges, consulting with hundreds of client teams and training thousands of aspiring data scientists, both inside and outside of IBM.

These experiences have led to a healthy set of lessons learned, and in this IBM Institute for Business Value report, we offer our perspective on the key factors you should consider when seeking to create an effective data sciences program — one that centers on a skilled data scientist creating the right amount of data science artistry to deliver those dazzling results.

Seek professionals with relevant experience

By IBM's definition, data scientists work with business leaders to solve business problems by understanding, preparing and analyzing data to predict emerging trends and provide recommendations to optimize business results. Woven into that definition are five key characteristics of data scientists that we believe are critical to their success:

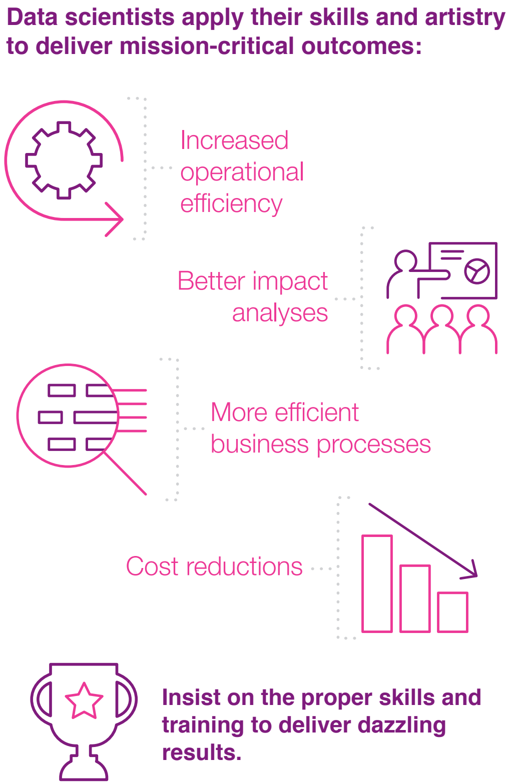
Keen business acumen: An understanding of the organizational business strategy and execution; the ability to listen to domain experts, quickly grasp the underlying business process and gain an understanding of how it works; expertise in converting a business problem into an analytical solution; and experience in business transformation.

Deep analytics knowledge: The ability to determine the appropriate analytics technique for addressing classes of business problems; and a clear understanding of both basic and advanced data mining techniques, ranging from regression analysis, cluster analysis, decision trees, neural networks and Bayesian machine learning methods to optimization, simulation and stochastic analysis.

Advanced software knowledge: The ability to determine the appropriate software packages to run; experience with key tools such as SPSS Modeler, SPSS Statistics, SAS, R, Python; the ability to design, develop and apply appropriate computational techniques to solve business problems; and the ability to create repeatable, automated processes.

21st century data management skills: An understanding of key internal and external data sources and how they are gathered, stored and retrieved; experience manipulating large volumes of data, both structured and unstructured, native and non-native; familiarity with massively parallel platforms; familiarity with tools such as SQL, NoSQL and Hadoop; HDFS infrastructure such as Pig, Hive, Hue, Sqoop, Hbase and Flume; and accelerators such as PureData or Exadata; and data analysis languages like Groovy.

Figure 1
Organizations should look to experienced and qualified data scientists to tackle their most significant challenges



Data scientists apply their skills and artistry to deliver mission-critical outcomes:

Increased operational efficiency

Better impact analyses

More efficient business processes

Cost reductions

Insist on the proper skills and training to deliver dazzling results.

Visionary and storytelling skills: The ability to effectively deploy analytics into the business to create value; the ability to assist senior executives in reimagining business processes; and the ability to leverage machine learning, artificial intelligence and cognitive solutions to automate prescriptive actions and continual learning.

These five characteristics are table stakes; by contrast, the traits that quickly differentiate a *great* data scientist from a *good* one are soft skills: Curiosity, scientific thinking, communication and visualization skills. Data scientists rely on these softer skills to form relationships within the organization, collaborate with stakeholders and make effective presentations.

Curiosity helps the data scientist delve into areas that others may not have explored or that the business had not considered in order to look at business challenges from a unique perspective and to expand the organization's thinking. A focus on the scientific thinking about the problem, versus just the tools and data, helps the data scientist see the big picture and formulate a plan of execution. Communication and visualization skills embody the art of data science: The ability to simplify the complexity of data science into a vision of implementable actions and forecasted outcomes and made real through a variety of data animation and visualization techniques, charts and graphs.

Professionals who excel at all of these skills are extremely rare (so rare, in fact, that many call them "unicorns").[8] Instead of hunting for mythical beings, executives should prioritize which of these characteristics are most crucial given their particular organization and seek candidates with the most suitable mix. They can then augment any gaps with targeted expertise within the larger data science team.

Demand a strong academic foundation to achieve deep science

Data science is a profession that is fundamentally based on a strong academic background. Executives should look for candidates with academic training in a quantitative discipline, such as statistics, operations research, machine learning, informatics, econometrics or physics. An analysis of LinkedIn's global database found that 80 percent of data scientists who include their education have earned a graduate degree, with 38 percent listing a doctoral and 42 percent a master's as the highest degree attained (see Figure 2).[9]

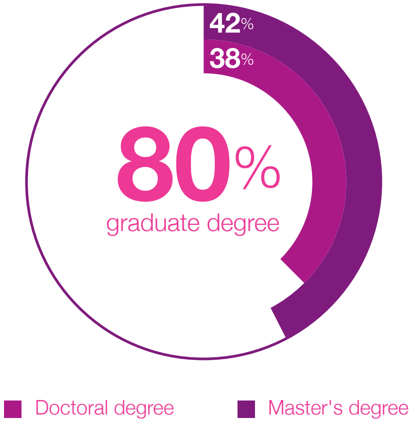
An undergraduate degree in a quantitative discipline provides a good foundation for junior-level members of the data science team, but additional coursework in linear algebra, applied statistics and machine learning may be required. Those with advanced degrees in these fields, in our experience, are ready to deploy on data science projects.

An emerging — and we think alarming — trend is that organizations are re-badging business intelligence (BI) analysts as data scientists, using cursory courses and tool training to fill gaps. We believe thisis a very risky stop-gap measure that executives should approach with caution. While it is true that most trained analysts can perform exploratory or rudimentary data science at the function or department level, we do not believe they should replace a trained data scientist's work with organizational imperatives. We feel that these newly designated data scientists are usually underprepared and unable to provide value-producing solutions to strategic business challenges, and they are unlikely to be able to lead a full-fledged data science team.

Data scientists have an important role to play as pioneers, interpreting the latest technology and mathematical concepts and making them routine. Many of these projects take substantial risks in either the data they utilize, the mathematical methods they deploy or the business problems they target — risks that could lead management to bet on poorly formulated models. A data scientist with an advanced education hedges these risks.

Figure 2

Data scientists tend to be very highly educated



Source: Stitch, Inc. 2016.

Lessons learned on the front line

The trick is knowing the best approach

A data scientist overheard two statisticians chatting about a marketing challenge.

The statisticians had created a heuristic to repeatedly loop through a data set until it appeared to have optimally assigned marketing treatments to geographic units.[10] Unfortunately, this process took more than 18 hours to run, completely tying up their computer, and — in the end — the results did not meet business objectives.

Within a few hours, the data scientist crafted an integer program to make the assignments.[11] Not only was the solution guaranteed to be optimal, it ran in under one second on the same hardware, making it possible to iterate and innovate to develop even more robust marketing campaigns.

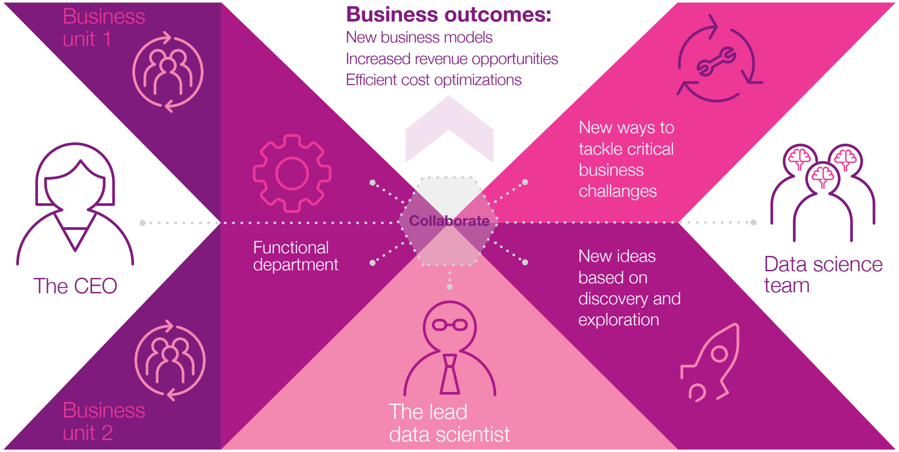
Encourage a disciplined approach to deliver greater value

Data scientists need to deeply understand the business challenges confronting the organization. We believe success with data science derives from a disciplined, collaborative approach, both in the top-down organizational processes and in the science applied to challenges.

The role of the data scientist is two-fold: First, to understand the business challenges the organization faces and define analytic solutions to automate or optimize processes; and second, to work with the broader data sciences team to uncover innovations and insights using a discovery-based analysis of the data. Organizations need to foster a tight coupling of data scientists with business leaders and subject matter experts (SMEs) to optimize the value they can create in either role (see Figure 3).[12]

Figure 3

Collaboration can enhance processes and improve outcomes


Business unit 1
Business outcomes:
New business models
Increased revenue opportunities
Efficient cost optimizations
New ways to tackle critical business challanges
The CEO
Functional department
Collaborate
Data science team
New ideas based on discovery and exploration
Business unit 2
The lead data scientist

Source: IBM methodology.

Data scientists work best in open, collaborative relationships with business leaders, facilitated through a business-driven governance system that provides structure to both the funding and project prioritization processes. A well-functioning process helps organizations develop innovation pipelines from the business that the data science teams can then utilize.[13]

Lessons learned on the front line

Good data science requires both analytic breadth and targeted depth

Data scientists owe it to their employers (and clients, if consultants) to be broadly skilled in a variety of mathematical and analytics techniques. If the data scientist's skills are not broad enough, he risks applying the wrong technique to the problem at hand. For example, a statistician was analyzing millisecond sensor data from a smart car in an attempt to build a model to represent the behavior of the car. The statistician intended to model the behavior as a linear program. When a data scientist asked the statistician what the objective function and constraints were and the significance of the problem, given that the time increment was in milliseconds, the statistician didn't have answers.

After the data scientist described agent-based simulation to the statistician, he agreed that it made more sense. Because the statistician didn't have any experience with that mathematical approach, he asked the data scientist to recommend someone who could step in and move the project forward to completion.

In addition to being broadly versed in many skills, the data scientist should be deep in at least a few targeted techniques, often tailored to the organization's data ecosystem and industry. This balance of skills will make the data scientist the go-to expert to define the analytics approach required to solve the business problem at hand.

Lessons learned on the front line

Start with the business problem and validate it throughout the process

Executives for a major U.S. insurance company hired a group of outside analysts to determine which paid claims were most likely fraudulent. A team of junior analysts was assigned, and they instead forecasted the total number of fraudulent claims. The executives, understandably, were disappointed; the solution was simply of no value to the insurer because it provided no insight into how to quickly identify probable fraudulent claims — an outcome the company needed to reduce losses by optimizing the use of limited investigative resources.

A data scientist corrected the situation by solving the appropriate problem. She then moved the client forward on the correct path to detecting and preventing fraud before paying any claims.

A frequent lament among strong data science teams is that once their success becomes evident, demand for their services swiftly outpaces available mental, physical and monetary resources. A structured prioritization approach places utilization decisions firmly with the business leaders, leaving the data scientist to focus on the business problem.

Good data science teams use methodologies that start with an understanding of the business challenge, then move from data collection and preparation to modeling, evaluation and deployment. The cross-industry standard process for data mining — CRISP-DM — follows this outline. While CRISP-DM might not perfectly fit every situation, it is intended to be flexible. That flexibility is one of its strengths, making it a good place to start when planning a data science project.

The most import step is the first one: Understanding the business challenge. This knowledge helps to ensure that the team attacks the *correct* problem, which is the key to achieving the optimal mix of data, mathematical technique and artistry required to meet business objectives.

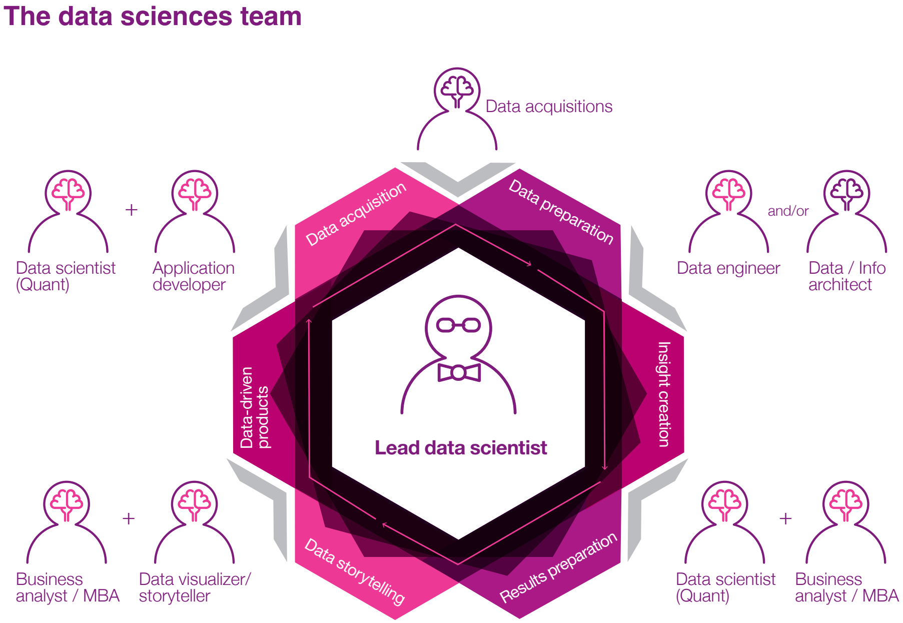
While significant industry experience is not the first qualification for a data scientist, there is no substitute for it. Simply understanding industry jargon can be beneficial, and there are often nuances in the data that the data scientist might be aware of only because of his or her experience within an industry. Executives can work around a data scientist’s lack of industry experience by fostering a strong partnering model with business analysts and SMEs. In time, the in-depth examination of the data and working knowledge of the business gained on each new challenge will help these professionals grow relevant expertise.

Create data science teams with varied backgrounds and skills

Data science is often described as a team sport, and with good reason. While the data scientist usually receives the bulk of the attention, a supporting team with specialized capabilities helps to ensure the broad and varied skills needed to create solutions (see Figure 4).[14]

Figure 4

Collaboration is key to the success of any data sciences team



Source: IBM methodology.

Lessons learned on the front line

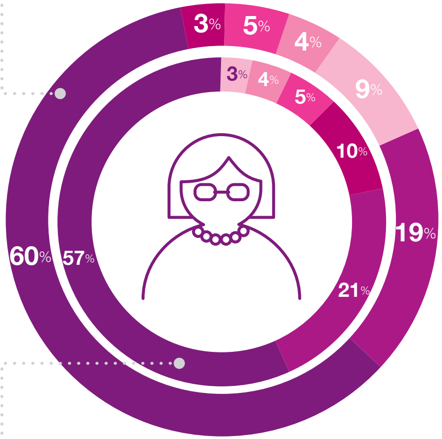
Tap into ideas from another industry

For data scientists, applicable industry experience can be very beneficial. But sometimes, so is the outside industry experience that a data scientist brings to bear on a business problem. For example, one data scientist who had used a data envelopment analysis technique to evaluate hospitals applied the same approach to evaluating independent agencies for a regional property and casualty insurer, unlocking more than USD 100 million in net-new written premiums.

Figure 5

Data scientists are an expensive resource, so organizations should strive to use them effectively

Data preparation accounts for about 80% of the work of data scientists



But the majority consider it the least enjoyable part of the work

- ■ Building training sets
- ■ Cleaning and organizing data
- ■ Collecting data sets
- ■ Mining data for patterns
- ■ Refining algorithms
- ■ Other

Source: CrowdFlower, 2016.

All too often, organizations make the mistake of believing the data scientist can or should perform every step of a data project. For example, data scientists are often encumbered with the additional role of the data engineer, which means they then have to spend their limited time and energy discovering, organizing, cleaning and sorting data.[15] While these are tasks most data scientists can perform, we consider it an expensive way — both because the resource is costly and their time is limited — to get the job done.

Data science is data-driven, and we find that too many organizations devote insufficient time and resources to achieve their objectives. A recent survey of data science professionals found that data collection, cleansing and organizing consumes almost 80 percent of their time; meanwhile, more than three out of four data scientists reported these tasks to be the least enjoyable (see Figure 5).[16] Executives would do well to consider these statistics when deciding whether to hire skills to augment the data science team; in today's low-supply, high-demand marketplace, a qualified data scientist has many other employment options available.

In our experience, it is not so much about the data you have, but the data you don't have that determines the success of a project. Data scientists put forth considerable efforts to derive data synthetically in areas that cannot be observed directly. Rather than spending time cleaning data, the better data scientists focus on intelligent data transformations, which can turn ordinary data into insightful data.

Seek professionals who can envision and realize outcomes

After understanding, transforming, modeling and evaluating data, data scientists must be able to pivot and effectively visualize and communicate the insights they have derived from the work. Here is where soft skills can make or break a data scientist's ability to successfully communicate and build consensus among decision makers. Increasingly, data science teams are employing English majors to help ensure that written communications throughout are flawless.

Another culprit that can doom a project is poor visualization of the data, the analysis or the anticipated outcomes. The lessons learned from such failures have made data visualization one of the more sought-after data roles, with 40 percent of organizations citing the need in 2015, up from 27 percent in 2014 (see Figure 6).[17]

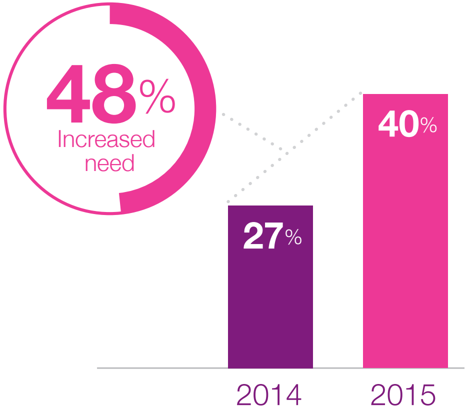
Acceptance and implementation of a data science initiative can transform the way a company works or the way the customer experiences a company. Yet, even after deploying and operationalizing the models a data scientist has envisioned — steps that can take as long as a solution's data transformations and mathematics — the journey is not over.

When a data science team achieves a breakthrough, that breakthrough is seldom the end state. In our experience, the achievement typically launches an iterative cycle of continuous improvement and exploration into other areas.

Figure 6

Organizations are recognizing the importance of accurate visualization of data, analysis and outcomes

Data visualization is one of the more sought-after data roles



Source: "Analytics: The upside of disruption." IBM Institute for Business Value. October 2015.

Lessons learned on the front line

If you don't document it, it never happened

A newly hired chief data scientist arrived at an organization and learned that the organization had completed "more than 15 highly successful data science projects" over the previous year. However, the data scientist soon realized that there was minimal documentation for only three of the projects, and none for the remaining 12. One of the projects reportedly saved the company billions of dollars, but there was no record of its implementation or any traceability to the purported savings. And, as is inevitable, the data scientist on that project was no longer with the company. As a result, gone was any methodology, experience, lesson learned or insight. Gone, too, was the potential for the organization to realize similar savings in the future, since it could not replicate the success. This is not a viable knowledge-management strategy. Organizations need to ensure that the methodology enforces thorough documentation along the way.

Ready or not? Ask yourself these questions

Data scientists are in high demand, and competition is fierce for the relatively small number of candidates who are truly qualified to do the job well. If you seek to build or expand a strong data science team within your organization, you must proceed intelligently. After all, you will likely entrust the team with analyses that can have far-reaching consequences for your organization. The following questions will help prepare you to confront a competitive marketplace and land a highly functional team capable of delivering the intended value.

- What organizational conundrums and opportunities have been left unanswered due to lack of capable analysis?
- When evaluating data scientist candidates, do you weigh the professional's soft skills, including communication, scientific thinking and curiosity?
- Are structures in place to govern, prioritize and manage requests for data science activities?
- Does your data science team have adequate resources capable of the variety of tasks needed?
- Can you identify a pattern that explains why analytic projects fail to gain acceptance or achieve results?
- Are you prepared to act upon the analysis you receive from the data scientists you select to perform this important job?

Authors

Mark Grabau is the Chief Data Scientist for the IBM North America Financial Services Sector. He has more than 24 years of experience building advance analytics models supporting the marketing and operational functions of large financial services, retail, travel and government organizations. Mark recently developed internal and client training courses for data scientists both within IBM and around the globe. Mark can be reached at mgrabau@us.ibm.com.

Dr. Emily Plachy is a Distinguished Engineer in the IBM Chief Data Office and leads cognitive initiatives for data and information governance. Her role has spanned management and technical, as well as development, advanced technology, research, sales, consulting and transformation. She has experience in multiple industries – telecommunications, utilities, banking, consumer products, retail, pharmaceutical, healthcare and petroleum. Emily can be reached at eplachy@us.ibm.com.

Dr. Michael Haydock currently serves as an IBM Fellow and Chief Scientist in the IBM Cognitive & Analytics Services Practice, specializing in the areas of customer and supply chain intelligence. Michael's role within the practice is to develop innovative application capabilities that leverage large-scale computing technologies appropriate for massive data manipulation and advanced numerical analysis in business settings, where time to critical decision provides IBM clients with a key competitive advantage. He can be reached at mhaydock@us.ibm.com.

Related publications

“Analytics: The upside of disruption.” IBM Institute for Business Value. October 2015. ibm.com/business/value/2015analytics/

“Analytics: The speed advantage.” IBM Institute for Business Value. October 2014. ibm.com/business/value/2014analytics/

“Analytics: A blueprint for value.” IBM Institute for Business Value. October 2013. ibm.com/business/value/ninelevers

Contributors

Raphael Ezry, Executive Sponsor
Rebecca Shockley, IBM Institute for Business Value.

For more information

To learn more about this IBM Institute for Business Value study, please contact us at iibv@us.ibm.com. Follow @IBMIBV on Twitter, and for a full catalog of our research or to subscribe to our monthly newsletter, visit: **ibm.com**/iibv.

Access IBM Institute for Business Value executive reports on your mobile device by downloading the free “IBM IBV” apps for phone or tablet from your app store.

The right partner for a changing world

At IBM, we collaborate with our clients, bringing together business insight, advanced research and technology to give them a distinct advantage in today’s rapidly changing environment.

IBM Institute for Business Value

The IBM Institute for Business Value, part of IBM Global Business Services, develops fact-based strategic insights for senior business executives around critical public and private sector issues.

Notes and sources

[1] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. “Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. McKinsey&Company. May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[2] Glassdoor. “25 Best Jobs in America.” 2016. https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

[3] “The top 3 qualities of a great data scientist.” Big Data Analytics News. April 21, 2015. http://bigdataanalyticsnews.com/top-3-qualities-great-data-scientist/

[4] Pierson, Lillian. “The State of Data Science.” Stitch, Inc. 2016. https://rjmetrics.com/resources/reports/the-state-of-data-science/

[5] Ibid.

[6] Hawk, William, Regina Powers, and Robert Rubinovitz. “The Importance of Data Occupations in the U.S. Economy.” U.S. Department of Commerce. Economics and Statistics Administration. ESA Issue Brief #01-15. March 12, 2015.

[7] Dataversity Education, LLC. Dataversity 2015 survey, “Business Intelligence versus Data Science.” www.dataversity.net/

[8] Purcell, Brandon, Srividya Sridharan, Megan Doerr, and Tyler Thurston. “The Forrester Wave: Customer Analytics Solutions, Q1 2016.” Forrester Research, Inc. March 7, 2016. https://www.forrester.com/report/The+Forrester+Wave+Customer+Analytics+Solutions+Q1+2016/-/E-RES128785#endnote5.

[9] Pierson, Lillian. “The State of Data Science.” Stitch, Inc. 2016. https://rjmetrics.com/resources/reports/the-state-of-data-science/

[10] Merriam-Webster defines a heuristic as “involving or serving as an aid to learning, dicovery, or problem-solving by experimental and especially trial-and-error methods. September 2016. http://www.merriam-webster.com/dictionary/heuristic

[11] Wikipedia defines an integer programming problem as “a mathematical optimization or feasibility program in which some or all of the variables are restricted to be integers.” August 29, 2016. https://en.wikipedia.org/wiki/Integer_programming

[12] IBM methodology.

13 Balboni, Fred; Glenn Finch; Cathy Rodenbeck Reese; and Rebecca Shockley. “Analytics: A blueprint for value, Converting big data and analytics insights into results.” IBM Institute for Business Value. October 2013. http://www-935.ibm.com/services/us/gbs/thoughtleadership/ninelevers/

14 IBM methodology.

15 Walker, Michael. “Data Scientists vs. Data Engineers.” Data Science Central. July 2, 2013.

16 2016 Data Science Report. CrowdFlower. http://visit.crowdflower.com/data-science-report.html

17 Finch, Glenn; Stephen Davidson; Dr. Pierre Haren; Jerry Kurtz; and Rebecca Shockley. “Analytics: The upside of disruption.” IBM Institute for Business Value. October 2015. www.ibm.biz/2015analytics; Finch, Glenn; Stephen Davidson; Christian Kirschniak; Marcio Weikersheimer; Cathy Rodenbeck Reese; and Rebecca Shockley. “Analytics: The speed advantage.” IBM Institute for Business Value. October 2014. www.ibm.biz/2014analytics.

Route 100
Somers, NY 10589
Produced in the United States of America
October 2016

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The information in this document is provided “as is” without any warranty, express or implied, including without any warranties of merchantability, fitness for a particular purpose and any warranty or condition of non-infringement. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

This report is intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. IBM shall not be responsible for any loss whatsoever sustained by any organization or person who relies on this publication.

The data used in this report may be derived from third-party sources and IBM does not independently verify, validate or audit such data. The results from the use of such data are provided on an “as is” basis and IBM makes no representations or warranties, express or implied.

Please Recycle

IBM®