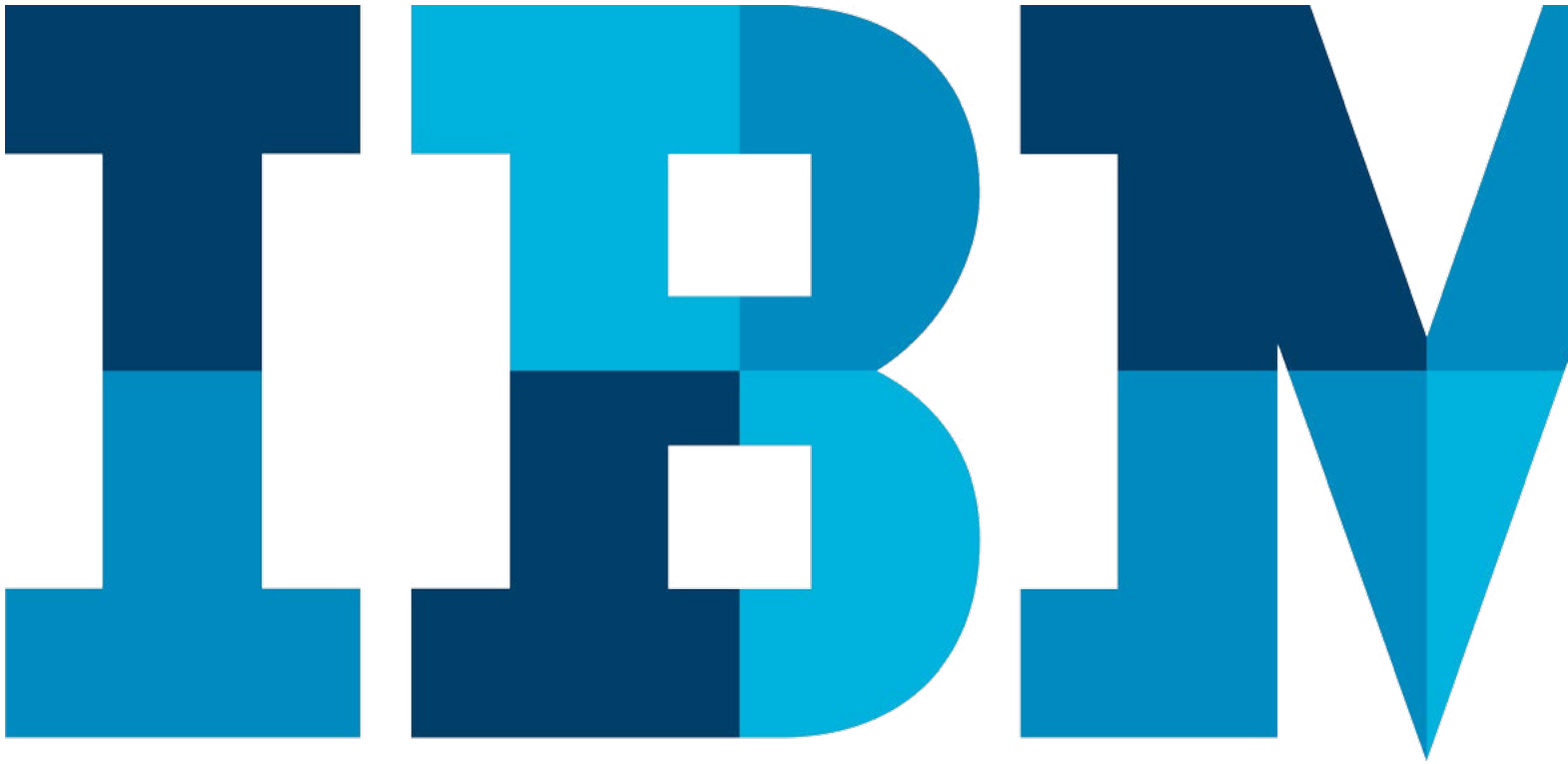
IBM Analytics

A glimpse inside the mind of a data scientist

Must-have skills? Daily challenges? Find out what actual data scientists really think about their critical role in data science



IBM

1 Introduction

2 What tasks consume most of your time?

3 What are some of the challenges you face on a day-to-day basis?

4 What skills are the most valuable for data scientists?

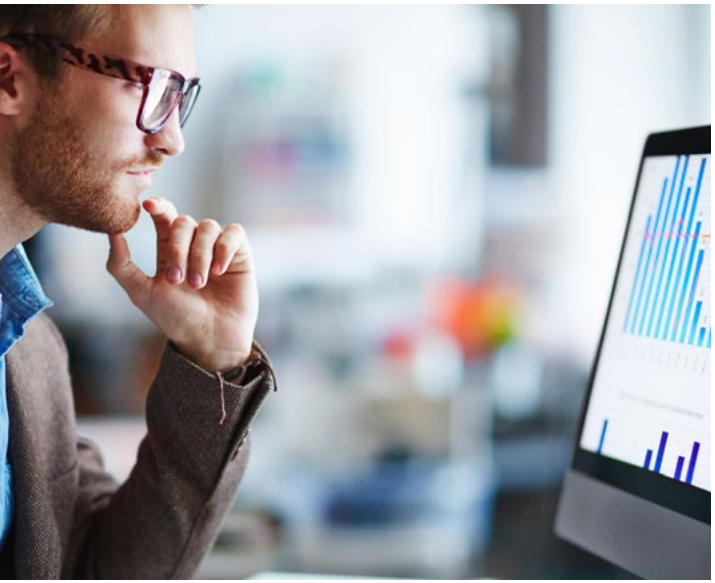
5 What has been your experience in working across teams to drive an end result?

6 What advice would you give other companies that are trying to get value from their data?

7 Embracing the big data challenge

Introduction

It's all well enough for an organization to collect every slice of data it can reach, but having more data doesn't mean you'll automatically get better insights. First, you have to figure out what you want from your data—you have to find its value.



Enter the data scientist. As chief data handlers and strategists, they are tasked with transforming volumes of data into actionable insights, enabling your organization to strengthen customer relationships, improve service delivery and drive new opportunities.

For the practitioners in the trenches, data science is definitely not a buzzword, and it's anything but simple. Data science is a highly complex discipline with a monumental task. Making data sing involves mastering statistics, math and programming to get a pool of results, and then extracting insights by applying the same business acumen—and gut instinct—that drives many executive-level decisions.

Even though there is a drastic shortage of resources and demands on their time stretch them to the limit, most data scientists love their jobs. According to a recent survey by CrowdFlower,[1] more than one-third (35 percent) of respondents gave their job the highest mark possible. And about half (47 percent) ranked their job a 4 out of 5. In other words, more than 80 percent of data scientists are happy with their work.

To get an inside look at what drives these big data professionals, we spoke with several data scientists to get their perspectives on their greatest obstacles, where they spend the majority of their time and what skills they find most valuable. Read on and see if their opinions sound familiar—or if there are a few pitfalls you need to watch out for.

What tasks consume most of your time?

Jack Burgess, Security Operations Tools Developer, Telstra

"Data preparation tends to be one of my most time-intensive activities. It's extremely critical, but an onerous task. It also leaves less time to actually analyze the data and deliver new insights to decision makers.

"I also spend quite a bit of time figuring out how to get access to the right data when I need it. Often data is siloed across different lines of business, or we have to gain permission to access the data. In an ideal world, we would have all the data in the company available to us in the optimum format. This would allow us to explore the sources we want to analyze, ask the questions we want to ask and focus on our core competency."

Michael Schmidt, Data Scientist/Founder, Nutonian

"Gaining access to data and getting it into the proper format can be extremely time-consuming. Availability issues can range from not having sufficient volume or variety of data, to having extremely inconsistent or 'dirty' data, where the effort to clean, filter or repair is so monumental that it increases the risk beyond what is tolerable. Once you have that data, you then need to find and interpret what it means. It's a significant effort to set that up to make sure you're getting the best results possible. That's an area I'd love to see more tools to make that easier and faster. It's a tedious process but we're starting to see some progress here."

Jeff Jonas, IBM Fellow and Chief Scientist, Context Computing

"Complex data types yield large volumes of information to be analyzed. But it's not the amount of data per se that consumes the most time; it's getting it in the right format, augmenting it and figuring out what information might be missing. It's an ongoing process that we have to perform again and again.

"While machines can help with some of this effort, a large portion of the work depends on the human ability to theorize, interpret and explore the problem and the potential solution at a deeper level. With the right tools to streamline this task, we could spend a lot more time on actual modeling and driving value from the data."

Key takeaways

- **Don't leave data preparation for last.** Poor-quality data leaves data scientists with a big burden of cleanup work. Data governance solutions have automated capabilities to cleanse, organize and manage data, helping to avoid the "Is this data ready to use?" question and associated delays.
- **Accessible data is usable data.** Data that's hidden in silos or buried in archives can't add value without intensive search efforts. Master data management and other information integration tools help bring data out of the shadows and make it easier to find.

What are some of the challenges you face on a day-to-day basis?

Michael Schmidt, Data Scientist/Founder, Nutonian

"One of the biggest challenges as a data scientist is applying the domain expertise to solve a problem. We have a plethora of algorithms and techniques to get value from data, but we need solutions to help us apply that to applications—to connect the dots from the statistics to the business opportunity.

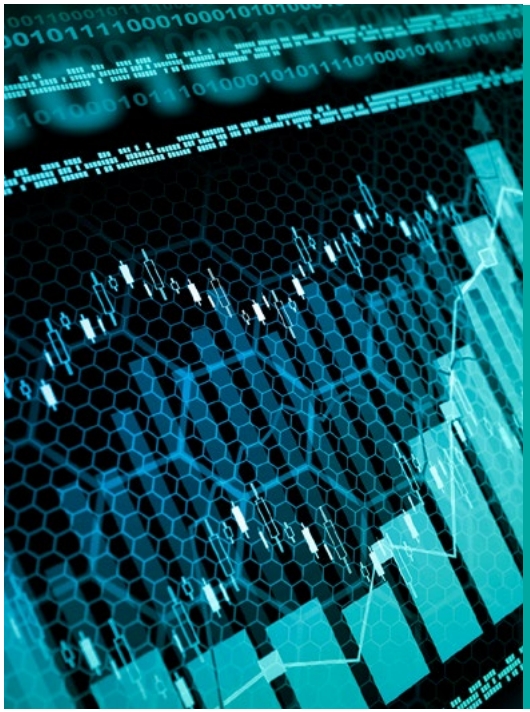
"Solving problems and predicting outcomes using sophisticated models require both an understanding of the capabilities, tools and techniques behind data science and the ability to get out from behind the keyboard and ask questions to inform the data process. Interpreting the problem is as much an art as it is a science."

Andy Gants, Principal Data Scientist, Spare5

"One of the larger challenges I faced in my current job was that the probability and statistical estimation tools that I used previously in my earth science research were the same tools, but they do not necessarily perform the same way on these new problems that involve user and answer estimation, quality estimation in these intelligent crowdsourcing problems. So the tools are the same, but the application of those tools is different. Also, I didn't have a software development department that I was iterating with in terms of implementing analyses into specific software features. Learning how to perform with the software development department proved to be quite a challenge—but a fun one."

Roman Schindlauer, Program Manager, Dato

"One of the biggest obstacles to analytical productivity is refining and formatting the data required for high-quality analytics. The lack of a universal or standardized programming language specifically geared to the data science domain doesn't make it any easier. Even with the best tools today, there is really not a good way around cleaning up the data manually. It's a continuous cycle of collecting and cleaning data and trying to figure out if it will yield significantly relevant insights. Or will you need to go back and change the parameters or massage the data more? I think we're getting to a point where the tooling support is helping with that, but it still requires a great deal of manual manipulation."



Key takeaway

- **Make data quality management a team effort.** Data quality is not just an IT issue or a data science issue—it's an enterprise issue. Everyone from sales and marketing to finance benefits from good-quality, well-curated information. Set up consistent, cross-enterprise governance policies and employ departmental data stewards to keep everyone focused on maintaining data quality.

What skills are the most valuable for data scientists?

Benjamin Skrainka, Principal Data Scientist, Galvanize

"Data scientists need expertise within multiple disciplines. You need to be good at databases. You need some knowledge of software engineering. You need to know some machine learning. And you need to know some statistics.

"At the same time, I think curiosity is important. Data scientists are inquisitive. They are continuously exploring, asking questions, doing what-if analyses questioning existing assumptions and processes. They will always be learning and thinking about what new technologies are out there that will help them be efficient and help the business succeed. While there are a lot of great tools available, there is no substitute for thinking."

Cliff Click, Chief Technology Officer, Neurensic

"Data scientists need a good blend of domain knowledge and a blend of business expertise. They need to be extremely inquisitive and relentless at figuring out how to solve a particular problem. That means digging into different approaches and alternatives—not just building models and running algorithms, but also interpreting the results to drive new business opportunities."

Jorge Castañón, Lead Data Scientist, IBM

"Creativity is a key element of data science. You need to have the technical background, but you also need to be curious enough to explore at a deeper level. You have to be able to go to the next layer, go deeper and explore. A skilled data scientist explores and examines data from multiple disparate sources. They simply do not collect and report on data, but also look at it from many angles, determine what it means and then recommend ways to apply the findings."

Jonathan Dinu, Vice President of Academic Excellence, Galvanize

"One of the key attributes that distinguishes today's data scientist is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Data scientists often become the liaison between IT and C-level executives. Therefore, they need to be able to speak both languages and understand the hierarchy of data; they can't just be the data expert."

Key takeaways

- **Be more than a data expert:** While data scientists can possess the necessary technical aptitude and quantitative skills, many today are actually business domain specialists who have deep understanding and keen insights into the business problems they're analyzing and modeling.
- **Back up convictions with courage:** Being able to present findings and clearly state the impact and implications of decisions related to those findings takes confidence—especially if executives aren't yet convinced of the value of analytics.

What has been your experience in working across teams to drive an end result?

Bob Lackey, Customer Success, Alteryx

"Data science is a sharing-oriented discipline. In my experience, I've found that if I have a question, there's always someone out there with an answer—whether it's inside or outside my organization. I have never had a time when I felt like someone was holding back information. In my company, we constantly engage with other teams in coming up with ideas and collaborating to solve problems. There is a great deal of knowledge sharing that is borne of that sort of interaction and teamwork."

Andrew Huynh, Data Engineer, Funding Circle

"Changes are happening so fast that we have to rely on diverse input for us to work together and drive better business outcomes. One of the things that attracted me to data science was the diversity of people you work with in addition to the diversity of techniques you might use. It really helps to have that collaboration across domains, across subjects, across techniques and with other people."

Jason Hill, Senior Big Data Engineer/Scientist, CA Technologies

"The way that we handle it is to have everybody on one team. Traditionally, an engineer wrote the code and a data scientist developed the algorithm, each siloed in their own area. Now they know each other's role and work together. We have data scientists that can write code and work with the engineer to develop algorithms."

Key takeaways

- **Use *all* the data:** The technology to collect and analyze massive volumes of business data is available, and companies should exploit it to their benefit. The ability to use all their data—and use it for growth and competitive advantage—will be determined by the level of organizational influence exerted by the data scientist to make insights from the data actionable and productive.
- **Data science is an art:** Successful organizations will have a robust data science culture, enterprising data scientists who can influence C-level decision makers, and the right combination of business intelligence and analytical capabilities.

Give us an example of something you've worked on that has added value to the business.

Jason Hill, Senior Big Data Engineer/Scientist, CA Technologies

"One thing that becomes increasingly important with big data is keeping track of how analytics applications are performing. Our data scientists have provided a lot of value by building up metrics that enable us to measure what's going on in an application when we push a code change—so we can rapidly identify whether it helps or hurts. If you're serving an application to a customer and you look at the client-side response times of that application and you see something change, that's pretty major."

Jorge Castañón, Lead Data Scientist, IBM

"A few months ago, we worked with BlocPower to build a model that estimates energy usage of buildings in American inner cities. In addition, we developed an application that scores the model and visualizes the results. As a result, BlocPower can identify the most energy-inefficient buildings, and estimate the reduction on greenhouse gas emissions for a building that acquires a new energy strategy. It is extremely satisfying to be part of a project that helps the world be greener."

Key takeaway

- **Set them free:** Give data scientists the freedom to work together with engineers and IT on problems that may not strictly seem like data science work. They may well surprise you by coming up with solutions to difficult or long-standing challenges.

What advice would you give other companies that are trying to get value from their data?

Bob Lackey, Customer Success, Alteryx

"As computational power continues to improve and people are better connected, I think you'll see amazing things being created as companies continue to be transformed. For me, data science and business success go hand in hand.

"I think the value is simply understanding that data is an asset like anything else, and it's an asset with a certain lifetime value attached to it—if you don't use it, you lose it. Therefore, you want to make sure you use it as efficiently as possible and as widely as possible to drive maximum business value."

Jacques Roy, Technical Sales, IBM

"Having people with the right skills is equally as important as having the right technology. Because the data is only as good as the value it provides—so you need people with the expertise to find the insights in the data. Building out a data scientist role or data science team will foster collaboration among the organization and provide 'champions of data' who can derive maximum business value from the organization's data."

Andy Gants, Principal Data Scientist, Spare5

"A major obstacle many organizations face in extracting value from data is properly preparing the data and getting it ready for analysis. Bad data will yield bad results. So you need to consider a more formalized approach to how you access and refine your data, especially as new sources of data, including the Internet of Things, continue to generate even larger volumes. For example, a federated approach to big data, which is taking the analytics where the data resides, can be faster and more cost-effective than storing all the data inside a data warehouse."

Jeff Jonas, IBM Fellow and Chief Scientist, Context Computing

"To gain an edge, organizations need to be able to make sense of what they're observing while it's happening. They need to be more event-driven—as data comes in, they need to be able to look at it, and take actions with it right away. This is where streaming analytics can play a vital role, allowing you to get the data coming in, analyze it on the fly and discard what you don't need or store it for future use."

Sally Macki, Senior Business Analyst, Pacific Gas & Electric

"Depending on the industry, different types of data are more critical to the organization than others. These various types of data need to be located and analyzed in order to obtain critical insights. Companies today are already competing on higher-quality decisions. If Company A can make a better decision than Company B, Company A wins."

Key takeaways

- **To deal with big data, you have to get it, keep it and make it sing:** Fully exploiting the opportunity presented by big data involves creating a value chain that helps address the challenges of acquiring data, evaluating its value, distilling it, building models (both manually and automatically), analyzing the data, creating applications and changing business processes based on what is discovered.
- **Be a voice for change:** The data scientist can (and should) play a key role in advocating for a dynamic, information-focused view on business growth. Enterprises will need to cast a wide net for these individuals, and once they get them, empower them with the right tools and a healthy data science culture.

Embracing the big data challenge

"In the data era, every problem is a data problem. If you reframe your mind-set around that concept, you really start to see the kind of impact data science can have."

—Rob Thomas, Vice President, Product Development, IBM Analytics

For data scientists to fully exploit the opportunity presented by big data, they need to invest time in building their skills and working with tools and technology that are going to help them overcome these obstacles. IBM is focused on helping data scientists get better and faster at their jobs. Since 2001, IBM has supported an academic initiative targeted at training future data science professionals—and last year it vowed to train one million data scientists through events, meetups, courses, content, advancements in the open source community and more. IBM also has more than 1,000 academic partnerships worldwide focused on big data and analytics.

IBM provides the critical infrastructure today's data scientists need to maximize business insights. IBM big data solutions allow users to store, manage and analyze data across numerous sources while making data accessible to business analysts, data scientists and IT users.

To remain prepared and equipped to be a solid team player, you need to improve your stats by building knowledge and enhancing your skills. IBM is one of many companies that is helping data science professionals get better and faster at doing their jobs.

Any rapidly growing field needs professionals with new skills and expertise. IBM has been an active supporter of the data science community, and plans to continue this support in online and in-person educational forums. Through events, meetups, courses, content, contributions to the open source community and more, IBM supports today's soon-to-be and current data professionals, helping them prepare to meet the high-volume, high-speed data demands of the future.

Ready to boost your data science skills? Here are resources to get you started:

- Build your data science skills with Big Data University
- Get started with the tools you need with the IBM Data Science Experience

IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
July 2016

[1] 2016 Data Science Report, CrowdFlower

CDM12349-USEN-00