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 labour 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 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 distil 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 Portillo 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 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 Mr. 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.

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. 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.

How to Find the Data Scientists You Need

1. Focus recruiting at the "usual 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 favoured by data scientists) and Python Interest Group (for PIGgies) are good places to start.
3. Search the 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 unlikely 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 an answer?
10. Ask candidates about their favourite 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)?

Thomas H. Davenport is a visiting professor at Harvard Business School, a senior adviser to Deloitte Analytics, and a coauthor of Judgment Calls (Harvard Business Review Press, 2012). D.J. Patil is the data scientist in residence at Greylock Partners, was formerly the head of data products at LinkedIn, and is the author of Data Jujitsu: The Art of Turning Data into Product (O'Reilly Media, 2012).