Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field

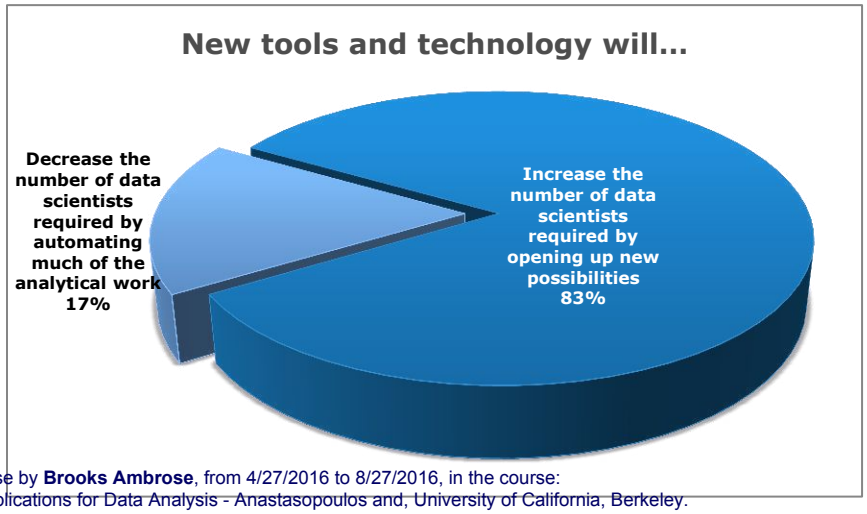
Introduction

As the cost of computing power, data storage, and high-bandwidth Internet access and have plunged exponentially over the past two decades, companies around the globe recognized the power of harnessing data as a source of competitive advantage. But it was only recently, as social web applications and massive, parallel processing have become more widely available that the nescient field of data science revealed what many are becoming to understand: that data is the new oil,[i] the source for corporate energy and differentiation in the 21st century. Companies like Facebook, LinkedIn, Yahoo, and Google are generating data not only as their primary product, but are analyzing it to continuously improve their products. Pharmaceutical and biomedical companies are using big data to find new cures and analyze genetic information, while marketers leverage the same technology to generate new customer insights. In order to tap this newfound wealth, organizations of all sizes are turning to practitioners in the new field of data science who are capable of translating massive data into predictive insights that lead to results.

Data science is an emerging field, with rapid changes, great uncertainty, and exciting opportunities. Our study attempts the first ever benchmark of the data science community, looking at how they interact with their data, the tools they use, their education, and how their organizations approach data-driven problem solving. We also looked at a smaller group of business intelligence professionals to identify areas of contrast between the emerging role of data scientists and the more mature field of BI. Our findings, summarized here, show an emerging talent gap between organizational needs and current industry capabilities exemplified by the unique contributions data scientists can make to an organization and the broad expectations of data science professionals generally.

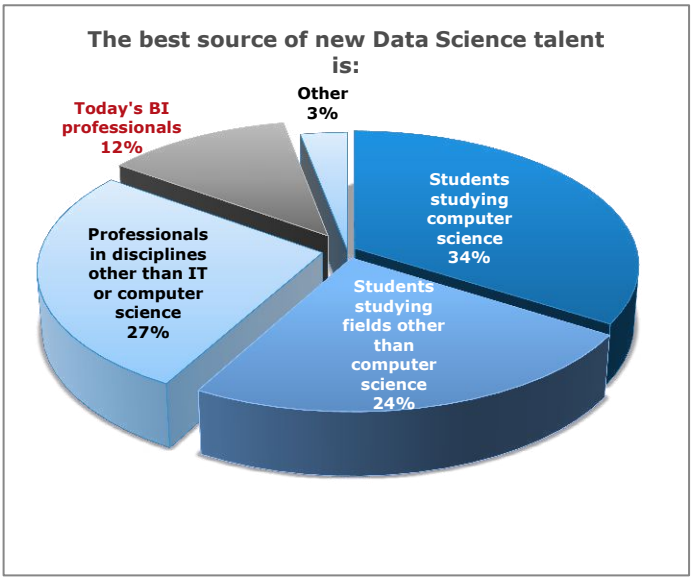
The Emerging Talent Gap

For the past two decades, business and political leaders have sounded a warning call about the shortage of trained computer scientists, programmers, and engineers necessary to continue to advance a high tech economy. Less noticed has been the coming gap in analytics professionals needed to use all of the data created by these high tech systems. In 2009, Google Chief Economist Hal Varian predicted that the statisticians will be the next sexy job[ii], and in our own survey, 83% of Data Scientists felt that new technology would increase the demand for data scientists, and 64% believe that it will outpace the supply of available talent.



This makes intuitive sense for a young field. Technological trends tend to follow a cycle, where the initial opportunity leads to ever increasing

demand for a certain set of skills, while later demand wanes as many of those initial skills are automated by even newer tools. Consider, for instance, the way many data processing and network management jobs that used to require legions of computer operators are now handled by automated monitoring tools. Data science is still in its very early phase, with the amount of data exploding and the right tools to process them just becoming available.



Although data science is generating new opportunities, our capacity to train new data scientists is not keeping up, and nearly two-thirds of respondents foresee a looming shortfall in the number of data scientists over the next five years. This aligns with other research, including a recent McKinsey Global Institute study that predicts a shortage of 190,000 data scientists by the year 2019[iii]. And when our respondents were asked where the best source for talent was, few looked to today's business intelligence professional. Instead, nearly two-thirds looked for today's university students.

Who is the Data Scientist?

Although the term data science has been around for decades – indeed, most scientists' use data of some form – the term data scientist in its current context is relatively new, frequently credited to DJ Patil, who started the data science team at LinkedIn.[iv] But as a new term, the field is still very much in flux, and without evidence about the practitioners, we're left to speculate about what it may mean. In our survey, we allowed users to self-identify as "data science professionals," in order to avoid conflicts over terminology in job titles. In this section we'll attempt to define the data scientist by comparing them with the previous big player in the analytics space, business intelligence professionals.

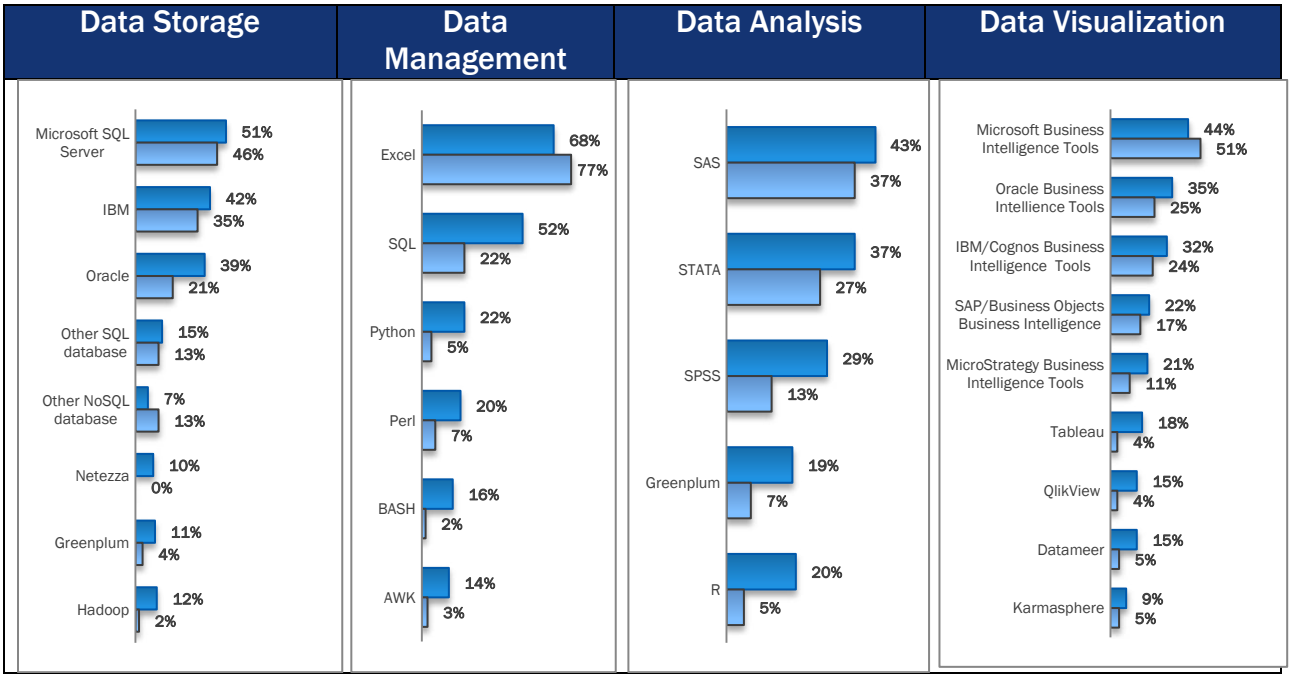
***Jim Asplund**, Chief Scientist at Gallup Consulting, is a data scientist focused on evaluating the role that human perception has on everything from disease conditions and GDP to worker productivity and consumer behavior. He works with massive data sets linking perception with actual behavior, and micro and macroeconomic outcomes. His work has isolated emotional factors that are most highly related to outcomes organizations care about.*

Twenty years ago, business intelligence was itself a new term, just emerging to take over the various database management and decisions support functions within an organization. As the field grew rapidly in the 90s, it also coalesced around a smaller number of tools, more consistent expectations for talent, better training, and more rigorous organizational standards. As our data demonstrates, data scientists are currently going through that transition,

It may be helpful to think of data science and business intelligence as being on two ends of the same spectrum, with business intelligence focused on managing and reporting existing business data in order to monitor or manage various concerns within the enterprise. In contrast, data science applies advanced analytical tools and algorithms to generate predictive insights and new product innovations that are a direct result of the data.

The need for rigorous scientific training was born out in our research on data scientists, and paints a clear distinction between data scientists and BI professionals. The most popular undergraduate degree for BI professionals was in business at 37% - more than the next three categories combined. In contrast, the most popular degree for data science professionals was computer science (24%), followed closely by engineering (17%) and the hard sciences (11%). We also found that data science professionals were over 2.5 times more likely to have a master's degree, and over 9 times more likely to have a doctoral degree as business intelligence professionals.

The data science toolkit is more varied and more technically sophisticated than the BI toolkit. While most BI professionals do their analysis and data processing in Excel, data science professionals are using SQL, advanced statistical packages, and NoSQL databases. Further, although big-data tools like Hadoop, and advanced visualization tools like Tableau are just starting to emerge in the data science world, they are almost unseen in the business intelligence world. Open Source tools, like the R statistics package, Python, and Perl, are each used by one in five data science professionals, but around one in twenty BI professionals.



An important facet of data science is the ability to run experiments on data, as evidenced by DJ Patel's description of how they built the "people you may know" function at LinkedIn:

> *It would have been easy to turn this into a high-ceremony development project that would take thousands of hours of developer time, plus thousands of hours of computing time to do massive correlations across LinkedIn's*

membership. But the process worked quite differently: it started out with a relatively small, simple program that looked at members' profiles and made recommendations accordingly. Asking things like, did you go to Cornell? Then you might like to join the Cornell Alumni group. It then branched out incrementally. In addition to looking at profiles, LinkedIn's data scientists started looking at events that members attended. Then at books members had in their libraries. The result was a valuable data product that analyzed a huge database -- but it was never conceived as such. It started small, and added value iteratively. It was an agile, flexible process that built toward its goal incrementally, rather than tackling a huge mountain of data all at once.

Without free access to the data and the ability to run tests without bureaucratic hurdles, LinkedIn would never have been able to add this critical feature. In the Data Science Survey, we found that data scientists were nearly twice as likely as business intelligence professionals to have this freedom, but that most organizations are falling short on this critical characteristic.

Broadening the Data Science Community

While Data Science is most often associated with Big Data, it is important to consider the host of other professions and roles that deem their work to be data science. This includes people from fields as diverse as Market Research, Financial Analysis, Information Technology, Management Consulting, Marketing and Media, Academia, Social Research, Demographic and Census Research and the Intelligence Community – it is no wonder this segment is difficult to define

As part of our study, we asked respondents for their job titles, and in the ranks of many business analysts, consultants, and analytics managers, we also found graphic designers, physicians, and research scientists, including one fisheries biologist. Successful data science often requires working with teams, including specialists who are able to divide labor and collaborate on the final outcome.

The absence of clear boundaries defining data science, and the many people co-opting the term for their own, is a good thing for the burgeoning function. It creates more interest in data science, both to support organizational decision-making, as well as to attract more talent into the field of data science. Raising the profile of both the function

Data Scientist Profile

Usama Fayyad

Former Chief Data Officer at Yahoo, and currently CEO or Chairman at 3 mid-stage start-up companies

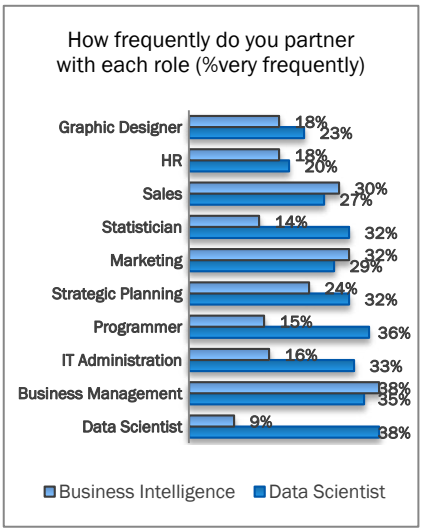
As chief data officer and executive VP, Research & Strategic Data Solutions at Yahoo! Inc. Fayyad was responsible for Yahoo!'s overall data strategy, policies, systems, and managing the Company's data processing/analytics infrastructure. He also oversaw Yahoo! Research in building the premier scientific research organization to develop the new sciences of the Internet, on-line marketing, and innovative interactive applications.

Prior to this, Fayyad co-founded and led the DMX Group, a data mining and data strategy company that was acquired by Yahoo! in 2004.

Fayyad earned his Ph.D. in engineering from the University of Michigan, Ann Arbor, and also holds BSE's in both electrical and computer engineering.

Fayyad serves as an ideal example of a "Big Data Scientist" – taking his statistical smarts and turning them into viable business entities and an executive role in a Fortune 500 organization. In a recent keynote at KAUST (King Abdulla University of Science and Technology), Usama spoke to the power of data to help define the social web and how businesses can use this data to their own advantage – and that without the expertise of data scientists, businesses will not be able to use the data at their disposal (particularly big data).

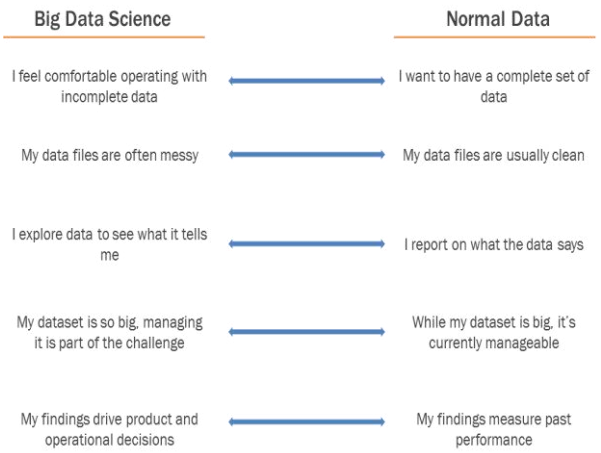
and the role is critical in organizations making the most out of the explosion of data they have access to.



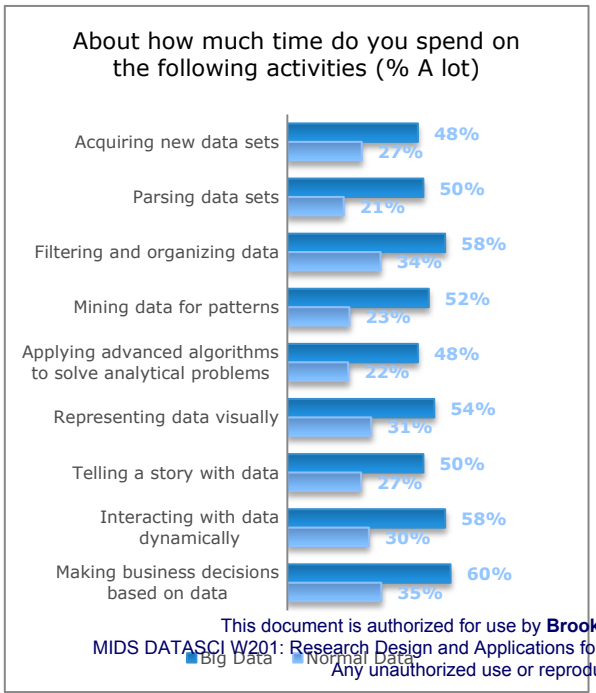
In addition to data science representing a broad sample of the population, we also found that data scientists are not lone actors, but work on teams that frequently partner with diverse roles in an organization. On average, a data science professional has one additional frequent partner than the average business intelligence professional, and will partner more frequently with statisticians, programmers, IT administrators, and most importantly other data scientists to gather, organize, and make use of their data.

The "Big Data" Data Scientist

Perhaps the greatest emerging opportunity in data science is "big data" – the ability to analyze massive data sets generated by web logs, sensor systems, and transaction data in order to identify insights and derive new data products. In addition to asking whether respondents self-identified as data scientists, our survey also constructed a psychometric screener to identify data scientists who worked on big data, based on background research into emerging expectations for big data, most notably O'Reilly Media VP Mike Loukides's article "What is Data Science?"[v] Based on these five items, we created a subsample that includes only respondents who answered all five questions on the "Big Data" side of the spectrum. Four-fifths of big data scientists classifed themselves as data scientists, and the remaining 20% cosnidered themselves business intelligence professionals.



Our findings showed that the emerging big data scientist is distinctly different from other data professionals. For instance, nearly half of big data scientists use R, despite the fact that it is only used by only 13% of other practitioners. They are also twice as likely to use a big data storage tool like Hadoop, Greenplum, or Netezza. Big data scientists are also remarkably educated – 40% have a master's degree, and an additional 17% have a doctorate. Over 90% have at least a college education.



Big data is also even more of a team sport. Half of big data scientists partner very frequently with a Data Scientist, Statistician, or Programmer – nearly twice the rate of the normal data group. They are

also more likely to partner with frequently with business management, but are interestingly no more likely to partner with IT administration.

Finally, big data scientists touch data in more ways. They are twice as likely as those working with normal data to work across the data life cycle, everywhere from acquiring new data to business decision making, and around half spend a lot of time on each of these activities.

Organizational Implications

In order to remain competitive in the world of data science, companies need to create organizational cultures that are conducive to data-driven decision making. First, they need to expand their view on the possibilities when hiring data scientists, and look outside business degrees, and even computer science, to find practitioners with the intellectual curiosity and technical depth to solve big data problems, with academic concentrations in the hard sciences, statistics, and mathematics. Data scientists use a variety of tools, but also recognize skill gaps as a barrier to adoption. Rather than hiring for experience with a certain toolkit, companies should invest in on-the-job training with their chosen set of emerging technologies.

Once companies have brought in the right talent, they need to create an environment conducive to effective data science. That means building high-performing, cross-functional teams that include a variety of roles, including programmers, statisticians, and graphic designers, and aligning them to directly support interested business decision makers. They should also loosen restrictions on data in the enterprise, allowing employees to more freely run data-driven experiments. Finally, data scientists should be given free access to run experiments on data, without bureaucratic obstacles, so that they can rapidly translate their own intellectual curiosity into business results.

Our Methodology

The EMC Data Science Community Survey interviewed 497 data scientists and business intelligence professionals from around the world, including deliberate samples in the United States, India, China, the United Kingdom, Germany, and France. 465 respondents were collected through a partnership with Toluna, one of the world's premier online panel providers. All Toluna participants were pre-screened for information technology decision making authority, and further screened as either data science professionals or business intelligence professionals. An additional 25 responses came from participants in the 2011 Data Science Summit, and six through publication by Kaggle, an online contest community for data scientists. All groups were asked the same questions with the same screeners.

---

[i] http://gerdleonhard.typepad.com/files/wef_ittc_personaldatanewasset_report_2011.pdf
[ii] http://www.nytimes.com/2009/08/06/technology/06stats.html
http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
[iv] http://radar.oreilly.com/2011/09/building-data-science-teams.html
[v] http://radar.oreilly.com/2010/06/what-is-data-science.html