HOW YOUR BUSINESS WORKS
information management
November 11, 2013

Data Science and Data Scientists: What's in a Name?

By Todd Saunders

Once again, people in the information management field are confronted with a new term surrounded by lots of excitement and promise, but also lots of hype and ambiguity. The term du jour is data science. I will attempt to give a reasonable definition for data science and describe the capabilities possessed by a data scientist. I'll also attempt to compare and contrast data science with business intelligence and data analysis, which often seem to be confused with (or considered synonymous with) with data science.

The Hype

The hype is nearly impossible to miss. Large numbers of articles and blogs, often with provocative titles, have been published. For example, back in 2009 Mike Driscoll wrote an article, "The Three Sexy Skills of Data Geeks." And about a year ago, the Harvard Business Review published an article titled, "Data Scientist: The Sexiest Job of the 21st Century." Let's face it, in the technical fields we don't see titles like that very often! The website KDNuggets, which focuses on the data mining community, reported that, "The big demand for analytics, data mining, and data science professionals led to a significant jump in their salaries in 2013." And many colleges and universities are beginning to offer courses and degrees in data science. (For a sample list see: http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees).

This illustrates that there is at least some validity to the hype. Companies are spending money on data science, people are being hired and paid generous salaries to be data scientists, and individuals are investing their own time and money to take courses and earn degrees to become data scientists. But what exactly is data science and what is a data scientist?

What is a Data Scientist?

First, we might want to consider where the term originated. A very nice article that was published in Forbes Magazine titled, "A Very Short History of Data Science" gives a good overview of the origins of data science (both the term and the practice). The article mentions that one of the first occurrences of the term "data science" appeared in a book published in 1974 by Peter Naur titled, "Concise Survey of Computer Methods." The Forbes article then goes on to cite other articles, papers, books, and conferences over the subsequent years that mention data science. What is interesting is that the frequency of the use of the term increases significantly the last few years. While that last statement is purely anecdotal, I can attest that from my many years of working in the BI services industry that the term "data scientist" is certainly more prevalent now than at any time in the past.

This history of the term data science shows us that it typically referred to a combination of the usage of statistics, computers, and data along with some domain (i.e., business or a specific discipline) knowledge. This leads to what I think works well as a definition of data science (drawing heavily from the mission statement of International Association for Statistical Computing):

Data science is the linkage of traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.

This definition is somewhat broad, and we can have a rather lengthy discussion about each of the different components of the definition (I will touch on some of the components later in this article), but the real point is to provide sufficient

clarity for the term so that it is less likely to be confused with other terms such as business intelligence or data analysis.

And why is that important? Simply to be pragmatic. The point of a definition is efficient communication. The three terms that sometimes get confused – BI, data analysis, data science – all have relatively wide ranges of definitions and overlap. But what is important is that you and the person you are talking to have a similar understanding of whatever term the two of you are talking about. You can certainly find much more varied and detailed definitions of data science, but many suffer from what is called "over-fit" in the statistical world. Those definitions are so detailed and specific they tend to be more obfuscating than illuminating. I prefer to sacrifice a little bit of specificity in favor of practicality and usefulness.

What it comes down to is that if you have a data-related problem that needs to be solved, it helps to understand what skills would be needed to solve that problem and what type of person would have those skills so you can put the right person on the job. So what separates a data scientist from a business intelligence practitioner from a data analyst? To answer that question, let's take a slightly more detailed look at each role and what skills are required for each.

Data Analyst

Data analysts typically focus on the sources and uses of data. These people will identify where data is created and how it flows through the business processes, track the business rules that are applied to the data, and observe how the data is reported and used. They research and understand how different pieces of data are related and affect each other. They examine and monitor the quality of data (completeness, timeliness, consistency, accuracy), and often they build or support data models used in solutions within the organization. In addition to data models, data analysts often develop and work with data dictionaries and metadata. They also utilize software that can help them analyze data along the quality components. Rudimentary statistics (min, max, mean, count, average) are typically all that is needed for most data analyst tasks. However, they need a rather thorough understanding of the business and business processes to effectively analyze the data.

Business Intelligence Practitioner

A BI practitioner can cover a wide range of activities and skills. Since BI focuses on providing information to business users to support decision-making, the BI practitioner typically possesses understanding and skills involved in both business and technology. This type of person may leverage technical skills to combine data from multiple sources into an integrated environment (e.g., data warehouse or data mart) from which reports and analysis can be developed using reporting and visualization software. Or they may apply their knowledge of the data and the business to define and build reports and analysis that are meaningful to the business and can convey that information effectively to the business users. This type of person typically has computer skills that facilitate moving data among systems, building databases or using reporting and visualization software. And they usually have the ability to communicate their findings to business users. While the types of analyses they perform are usually more along the lines of descriptive statistics (current results, trends), they may incorporate results from statistical models that are provided to the BI solution in their business communications.

Data Scientist

What sets a data scientist apart from the other two categories is primarily the use of statistics. While some definitions of data science minimize (or at least reduce) the dependency on advanced statistical techniques, history seems to indicate that data science is a broadening of the capabilities of statisticians. While a data scientist has the capability to do much of the same work as a data analyst or BI practitioner, his or her primary focus is statistical analysis, be it predictive model development, machine learning or data mining. And what sets data scientists apart from pure statisticians is their computer programming ability used to manage large data sets along with their domain knowledge, which they use to guide their analysis.

Consider the following example to illustrate the above definitions. If, as a businessperson, you need to understand what your organization's definition of "current customer" and where that information resides, you probably need a data analyst. If you need to visualize and organize information and present it to an executive team, you probably need a business intelligence person. If you would like to predict future earnings based on past performance, current and forecasted economic conditions, social media sentiment, website activity and customer churn rates, you probably want a data scientist.

What's Changed?

So what happened recently that has made data science so popular? Why is a (relatively) new job description needed now when it wasn't needed nearly as much only a couple years ago? My assertion is that the emergence of big data has led to the data science movement.

Since the term "big data" is another relatively new and hyped term, allow me to give it a definition within the context of this article. When I refer to big data, I'm talking about the vast amounts and types of structured and unstructured data that is growing by leaps and bounds that can be captured and made available for analysis. Businesses and organizations now have access to enormous amounts of data that come in as and/or can be converted into a variety of forms: key value pairs, documents, object notation, text, free-form language, columnar data, etc. To manage all this data, many new tools and technologies have been developed, e.g., Hadoop, MongoDB, MapReduce, natural language processing, columnar databases, NoSQL, Pig, Hive, Impala, and many others.

Because statistical analysis relies heavily on the quality and structure of the underlying data being analyzed, statisticians started gaining skills with these new data management

technologies to allow them to structure and build the data sets they needed to perform the desired statistical analysis. The data structures and technologies also provided more capability to analyze extremely large sets of data like clickstream data, astronomy data, health care data, and weather data.

That is where the data scientist most clearly stands out: While the statistician typically carried SAS or SPSS in their tool chest, data scientists carry the statistical packages (SAS, SPSS, R) along with programming and data manipulation languages (Python, PHP, Ruby, Java, MapReduce, C++, Hive, Pig, Impala, etc.). They know how to best write the code to execute machine learning algorithms based on the underlying data structure and volumes. And they know how to apply the domain context to the results.

Covering More Ground

One last caveat about data science: As I stated earlier, I view data science as a broadening of the practice of statistical analysis. This implies that data science initiatives are broader than statistical analysis initiatives. With traditional statistical analysis (and admittedly oversimplified), statisticians just needed a server loaded with their favorite stats software package and plenty of disk space, and they were good to go. Now that we have big data in the mix, a data science initiative needs to cover a lot more ground. The types and volumes of data need to be assessed to determine what technical architecture will be required. Will you build a HDFS system? NoSQL? Document store? Do you need large numbers of standard servers or a small number of beefy, multicore servers? What is the goal of the analysis? Who is responsible for finding, loading and assessing the data? Who will install the software? Who else needs to be involved with the initiative? The question list is extensive.

The point being made is that data science initiatives will probably require program and project management, business analysis and executive sponsorship and oversight in addition to the statistical analysis. It's not just one guy with a PC in a back room with a six-pack of Mountain Dew and a bag of Cheetos (… no disrespect). You might want your data scientist to cover the project management, solution design, architecture and presentation components of the initiative in addition to the actual analysis. Or you may want to consider a team approach, since it is rare to find one individual with all those capabilities.

Overlap and Divergence

I've attempted to show that BI, data analysis, and data science all serve valuable functions and provide substantial benefits to organizations by helping to turn data into information. Hence, there is a lot of overlap among these disciplines. Namely, data is being captured, structured, manipulated and analyzed in order to provide information and insight. Where these disciplines diverge is in regard to the tools they use, the types of analyses they perform, and the specific types of issues they address.

DISCIPLINE	TECHNOLOGIES	SKILLS	FOCUS
BUSINESS INTELLIGENCE	• ETL Tools / SQL • RDBMS • Reporting • Visualization	• Programming • Data Analysis • Data Modeling • Report Development • Basic Statistics • Technical Architecture • Business Analysis & Strategy • Presentation	• Information Delivery and Reporting • Data Visualization • Descriptive Statistics • Data Integration and Consolidation
DATA ANALYSIS	• Data Modeling Software • Diagraming Software • Documentation Software • SQL • Data Profiling Software	• Data Modeling • Business Analysis • Data Manipulation • Basic statistics	• Business Rules • Data Definitions and Lineage • Data Entity Relationships • Data Attributes • Data Structures • Sources and Targets of Data • Data Quality
DATA SCIENCE	• Statistics Software • Columnar Data • Map-Reduce • NoSQL • Programming Languages • Graphing/Charting Software	• Advanced Statistics • Programming • Business Analysis • Modern Data Management Technologies and Architectures	• Predictive Modeling • Advanced Statistical Analysis • Data Mining • Unstructured Data Management • Large data volumes • Research

Description: http://cdn.information-management.com/media/newspics/Saunders-Nov_fig1.jpg

---

Todd Saunders, Principal with CBIG Consulting, is responsible for overseeing the delivery of business intelligence, Big Data, and data warehousing solutions and consulting services for CBIG's West Region. Todd has over 23 years of management consulting experience with the last 15 years focusing on business intelligence and data warehousing. Todd began his consulting career with McKinsey & Co. before moving on to Coopers & Lybrand. In previous positions, Todd served as the National Vice President of BI for Braun Consulting and VP of Consulting Services for Quaero Inc. Todd holds a B.S. degree in Physics and Engineering Science from Manchester College, as well as an MBA in Finance and an MSEE in Quantum Electronics from the University of Illinois. Todd is a Certified Business Intelligence Professional and has served as a member of the faculty of The Data Warehousing Institute.