: What is a data scientist?

by Daniel J. Power

Editor, DSSResources.COM

A practitioner of the emerging discipline of data science is called a data scientist. Data science is **not** the science of data, rather the term refers to a more sophisticated and systematic analysis of data. Many bloggers and commentators who use both terms are extending the word "science" well beyond how physical and social scientists typically use that term. The definition of science focuses on systematic study involving observation, experimentation, and data collection to try to understand a phenomenon. Loosely the phrase data science fits the definition. We need to ask: "Have the knowledge and skill requirements and the task expanded to justify training and preparing a new category of scientists?"

Data analysis increasingly requires advanced educational preparation. As a descriptive title and marketing moniker some are calling the new M.S. and Ph.D. degrees "data science" programs and the emerging professionals are hence called data scientists.

Analyzing data requires the expertise expected in many scientific disciplines. The increasing complexity of large data sets from nontraditional sources requires expanded expertise in statistical analysis, data retrieval and management, hypothesis generation and testing, data presentation and interpretation, and report writing.

According to Malcolm Chisholm, "There's a joke running around on Twitter that the definition of a data scientist is 'a data analyst who lives in California.'"

In an excellent overview blog post/column at O'Reilly Radar (http://radar.oreilly.com/), Mike Loukides (2010) addressed the question "What is data science?". He examined the technologies, the companies and the skill sets associated with data science and its experts, the data scientist. Loukides asserted that "data science enables the creation of data products". Data science practiced in companies is intended to create value. As far as the job, Loukides explains "data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others". The story is what managers and others pay money to hear or read.

In **Understanding Big Data**, IBM researchers claim "A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart 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. Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization."

Anjul Bhambhri argues "Many organizations today solve the 'data silos' problem by storing large volumes of decision support data in warehouses... Data Scientists should always have this in mind, in order to avoid creating new 'silos' in the enterprise."

See these articles on Forbes.com for definitions of a data scientist from leading experts in the field:

What is a Data Scientist?: Michael Rappa, NC State University. Rappa asserts "The term 'data scientist' is a useful one, because it captures people's imagination."

IBM's Anjul Bhambhri on "What Is a Data Scientist?", he envisions a data scientist as a change agent.

TIBCO Spotfire's Michael O'Connell on "What is a Data Scientist?" He identifies the followings important skills of a data scientist: "Think analytically, rigorously, and systematically about a business problem and come up with a solution that leverages the available data."

Tableau Software's Pat Hanrahan on "What is a Data Scientist?" Hanrahan argues that the "definition of 'data scientist' could be broadened to cover almost everyone who works with data in an organization."

LinkedIn's Monica Rogati on "What is a Data Scientist?" Rogati notes "By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights."

LinkedIn's Daniel Tunkelang on "What is a Data Scientist?" Tunkelang quotes Hilary Mason, chief scientist at bit.ly "a data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

EMC Greenplum's Steven Hillion on "What Is a Data Scientist?" Hillion states data scientists are “analytically-minded, statistically and mathematically sophisticated data engineers who can infer insights into business and other complex systems out of large quantities of data.”

Amazon's John Rauser on "What Is a Data Scientist?" According to Rauser, the ideal data scientist is “someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience.”

Stuart Firestein, in his book **Ignorance: How It Drives Science**, states "Aggregating facts is useless if you don’t have a context to interpret them." He is trying to be provocative. A data scientist must have a "thoughtful ignorance" and identify gaps in a community’s understanding and seek to resolve them".

Jim Harris in the OCDQ Blog argues "Big data is useless if you don’t have a business context to interpret it." Data Scientists supposedly use Big Data and create a context and story that is useful.

Fern Halper discusses what's different about Big Data analytics. She notes 1) the "infrastructure you use to process and store the data will be different", 2) algorithms may need to be changed, 3) analytics will be applied close to the sources of big data, 4) more analytics will be programmed, for example using location data from your phone automatically targeting a promotion, 5) "big data analytics can be hypothesis-free", 6) "big data analytics can use lots of attributes".

Halper's last two points that big data analytics can be hypothesis-free and use lots of attributes highlight the need for professional data scientists and the statistical nightmare of "unscientific" analysis of Big Data. Some people rationalize the virtues of hypothesis-free, data driven analysis, but the dangers of letting the data drive the analysis are often ignored. Correlating many variables is only part of the problem. Using so many variables will lead to false positive results. The errors from multiple tests of significance can lead to many problems.

According to https://datajobs.com/what-is-data-science, a "defining personality trait of data scientists is they are deep thinkers with intense intellectual curiosity. Data science is all about being inquisitive – asking new questions, making new discoveries, and learning new things."

Data scientists should be prepared to perform 3 primary tasks -- DAD (cf., Data Science Central, 2013):

Discover: Find, identify the sources of good data, and the metrics. Sometimes request the data to be created (work with data engineers, business analysts).

Access: Access the data. Sometines via an API, a web crawler, an Internet download, a database access or sometimes in-memory within a database.

Distill: Extract meaning from data, decision relevant information to increase ROI and take actions (such as determining optimum bid prices in an automated bidding system). Distilling involves exploring the data,
cleaning the data, and refining and summarizing. Distilling often involves statistical analyses and may involve use of other analytical techniques.

In summary, a data scientist is a person who has the knowledge and skills to conduct sophisticated and systematic analyses of data. A data scientist extracts insights from data sets for product development, and evaluates and identifies strategic opportunities. He/she will become the expert for digital metrics within the company. He/she will have a unique blend of thoughtful, data-driven curiosity and pragmatism, and clearly know the difference between game-changing ideas and time "sucking" activities (added 2/26/2014).

Do we need data scientists? **Yes**. Do we know how many we need? Do we know what they will exactly do? and Do we know who they will work for to analyze data? **No**.

References

Bhambhri, A. "IBM's Anjul Bhambhri on What Is a Data Scientist?" Forbes, 2/16/2012 at URL http://www.forbes.com/sites/danwoods/2012/02/16/ibms-anjul-bhambhri-on-what-is-a-data-scientist/

Chisholm, M. "Data Management is Based on Philosophy, Not Science," Information Management, MAY 1, 2012 at URL http://www.information-management.com/newsletters/data-scientist-philosophy-Hayek-Collingwood-IAU-10022399-1.html

Difference between data engineers and data scientists, Data Science Central, October 17, 2013 at URL http://www.datasciencecentral.com/profiles/blog/show?id=6448529:BlogPost:111381&commentId=6448529:Comment:111576

Firestein, S. Ignorance: How It Drives Science, Oxford University Press, 2012 at URL http://www.oup.com/us/catalog/general/subject/Medicine/Neuroscience/?view=usa&ci=9780199828074

Halper, F. "What's Really Different about Big Data Analytics?" TDWI, April 30, 2013, at URL http://tdwi.org/Articles/2013/04/30/Big-Data-Analytics.aspx?j=186992

Harris, J. "What is the Philosophy of Data Science?" OCDQ Blog, March 18, 2013 at URL http://www.information-management.com/blogs/what-is-the-philosophy-of-data-science-10024090-1.html

Loukides, M. "What is data science?" O'Reilly Radar, June 2, 2010 at URL http://radar.oreilly.com/2010/06/what-is-data-science.html

What is a data scientist? at http://www-01.ibm.com/software/data/infosphere/data-scientist/

Author: Daniel Power
Last update: 2014-09-26 06:07