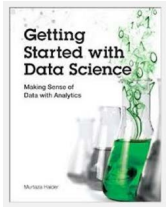
Course Text Book: 'Getting Started with Data Science'
Publisher: IBM Press; 1 edition (Dec 13 2015) Print.
Author: Murtaza Haider



BIG DATA UNIVERSITY

Prescribed Reading: Chapter 1 Pg. 12-15

What Makes Someone a Data Scientist?

Now that you know what is in the book, it is time to put down some definitions. Despite their ubiquitous use, consensus evades the notions of big data and data science. The question, "who is a data scientist?" is very much alive and being contested by individuals, some of whom are merely interested in protecting their discipline or academic turfs. In this section, I attempt to address these controversies and explain why a narrowly construed definition of either big data or data science will result in excluding hundreds of thousands of individuals who have recently turned to the emerging field.

"Everybody loves a data scientist," wrote Simon Rogers (2012) in the *Guardian*. Mr. Rogers also traced the newfound love for number crunching to a quote by Google's Hal Varian, who declared that "the sexy job in the next ten years will be statisticians."

Whereas Hal Varian named statisticians sexy, it is widely believed that what he really meant were data scientists. This raises several important questions:

- What is data science?
- How does it differ from statistics?
- What makes someone a data scientist?

In the times of big data, a question as simple as, "What is data science?" can result in many answers. In some cases, the diversity of opinion on these answers borders on hostility.

I define *data scientist* as someone who finds solutions to problems by analyzing big or small data using appropriate tools and then tells stories to communicate her findings to the relevant stakeholders. I do not use the data size as a restrictive clause. A data below a certain arbitrary threshold does not make one less of a data scientist. Nor is my definition of a data scientist restricted to particular analytic tools, such as machine learning. As long as one has a curious mind, fluency in analytics, and the ability to communicate the findings, I consider the person a data scientist.

I define *data science* as something that data scientists do. Years ago, as an engineering student at the University of Toronto I was stuck with the question: What is engineering? I wrote my master's thesis on forecasting housing prices and my doctoral dissertation on forecasting homebuilders' choices related to what they build, when they build, and where they build new housing. In the civil engineering department, others were working on designing buildings, bridges, tunnels, and worrying about the stability of slopes. My work, and that of my supervisor, was not your traditional garden-variety engineering. Obviously, I was repeatedly asked by others whether my research was indeed engineering.

When I shared these concerns with my doctoral supervisor, Professor Eric Miller, he had a laugh. Dr. Miller spent a lifetime researching urban land use and transportation, and had earlier earned a doctorate from MIT. "Engineering is what engineers do," he responded. Over the next 17 years, I realized the wisdom in his statement. You first become an engineer by obtaining a degree and then registering with the local professional body that regulates the engineering profession. Now you are an engineer. You can dig tunnels; write software codes; design components of an iPhone or a supersonic jet. You are an engineer. And when you are leading the global response to financial crisis in your role as the chief economist of the International Monetary Fund (IMF), as Dr. Raghuram Rajan did, you are an engineer.

Professor Raghuram Rajan did his first degree in electrical engineering from the Indian Institute of Technology. He pursued economics in graduate studies, later became a professor at a prestigious university, and eventually landed at the IMF. He is currently serving as the 23rd Governor of the Reserve Bank of India. Could someone argue that his intellectual prowess is rooted only in his training as an economist and that the fundamentals he learned as an engineering student played no role in developing his problem-solving abilities?

Professor Rajan is an engineer. So are Xi Jinping, the President of the People's Republic of China, and

Alexis Tsipras, the Greek Prime Minister who is forcing the world to rethink the fundamentals of global economics. They might not be designing new circuitry, distillation equipment, or bridges, but they are helping build better societies and economies and there can be no better definition of engineering and engineers—that is, individuals dedicated to building better economies and societies.

So briefly, I would argue that data science is what data scientists do.

Others have much different definitions. In September 2015, a co-panelist at a meetup organized by BigDataUniversity.com in Toronto confined data science to machine learning. There you have it. If you are not using the black boxes that make up machine learning, as per some experts in the field, you are not a data scientist. Even if you were to discover the cure to a disease threatening the lives of millions, turf-protecting colleagues will exclude you from the data science club.

Dr. Vincent Granville (2014), an author on data science, offers certain thresholds to meet to be a data scientist.[29] On pages 8 and 9 in *Developing Analytic Talent* Dr. Granville describes the new data science professor as a non-tenured instructor at a non-traditional university, who publishes research results in online blogs, does not waste time writing grants, works from home, and earns more money than the traditional tenured professors. Suffice it to say that the thriving academic community of data scientists might disagree with Dr. Granville.

Dr. Granville uses restrictions on data size and methods to define what data science is. He defines a data scientist as one who can "easily process a 50-million-row data set in a couple of hours," and who distrusts (statistical) models. He distinguishes data science from statistics. Yet he lists algebra, calculus, and training in probability and statistics as necessary background "to understand data science" (page 4).

Some believe that big data is merely about crossing a certain threshold on data size or the number of observations, or is about the use of a particular tool, such as Hadoop. Such arbitrary thresholds on data size are problematic because with innovation, even regular computers and off-the-shelf software have begun to manipulate very large data sets. Stata, a commonly used software by data scientists and statisticians, announced that one could now process between 2 billion to 24.4 billion rows using its desktop solutions. If Hadoop is the password to the big data club, Stata's ability to process 24.4 billion rows, under certain limitations, has just gatecrashed that big data party.[30]

It is important to realize that one who tries to set arbitrary thresholds to exclude others is likely to run into inconsistencies. The goal should be to define data science in a more exclusive, discipline- and platform-independent, size-free context where data-centric problem solving and the ability to weave strong narratives take center stage.

Given the controversy, I would rather consult others to see how they describe a data scientist. Why don't we again consult the Chief Data Scientist of the United States? Recall Dr. Patil told the *Guardian* newspaper in 2012 that a "data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data." What is admirable about Dr. Patil's definition is that it is inclusive of individuals of various academic backgrounds and training, and does not restrict the definition of a data scientist to a particular tool or subject it to a certain arbitrary minimum threshold of data size.

The other key ingredient for a successful data scientist is a behavioral trait: curiosity. A data scientist has to be one with a very curious mind, willing to spend significant time and effort to explore her hunches. In journalism, the editors call it having the nose for news. Not all reporters know where the news lies. Only those who have the nose for news get the story. Curiosity is equally important for data scientists as it is for journalists.

Rachel Schutt is the Chief Data Scientist at News Corp. She teaches a data science course at Columbia University. She is also the author of an excellent book, *Doing Data Science*. In an interview with the *New York Times*, Dr. Schutt defined a data scientist as someone who is part computer scientist, part software engineer, and part statistician (Miller, 2013). But that's the definition of an average data scientist. "The best," she contended, "tend to be really curious people, thinkers who ask good questions and are O.K. dealing with unstructured situations and trying to find structure in them."