What is a Data Scientist?

06 Feb 2017

Over the last year, I've been focusing at writing posts to help aspiring data scientists get into the field. I've been sharing as much as possible about my learnings. During this period of time, I realized there was a fundamental problem amongst people wanting to get into data science: the definition of a "data scientist" is quite fuzzy and it's causing a lot of confusion.

Some people will tell you a data scientist is a person that focuses only on building predictive models using statistics and machine learning. Other people will have a much broader definition. Where is the *truth*? What do companies expect when they hire a *data scientist*?

In this post, I will summarize in broad strokes what the main views around the data scientist definition is. A lot of ink has been spilled trying to define what a data scientist is. My goal here is not to come up with yet a new definition. I rather want to summarize the main and most acknowledged ones.

A Data Scientist is just a Statistician

> "Data scientist is a sexed up term for a statistician" -

> Nate Silver

Nate Silver is an eminent statistician. He became known from the general public during the 2008 U.S. Presidential election when he successfully predicted the winner of 49 of the 50 states. He has been ranked amongst the 100 most influential people by Time in 2009[1].

On the term "data scientist", Silver once said that a "data scientist is a sexed up term for a statistician. Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician"[2].

Silver is not the only one to believe that "data scientist" is just another word for an existing job title. Gil Press, a contributor on Forbes, argued that data science is just another word for "business analytics"[3].

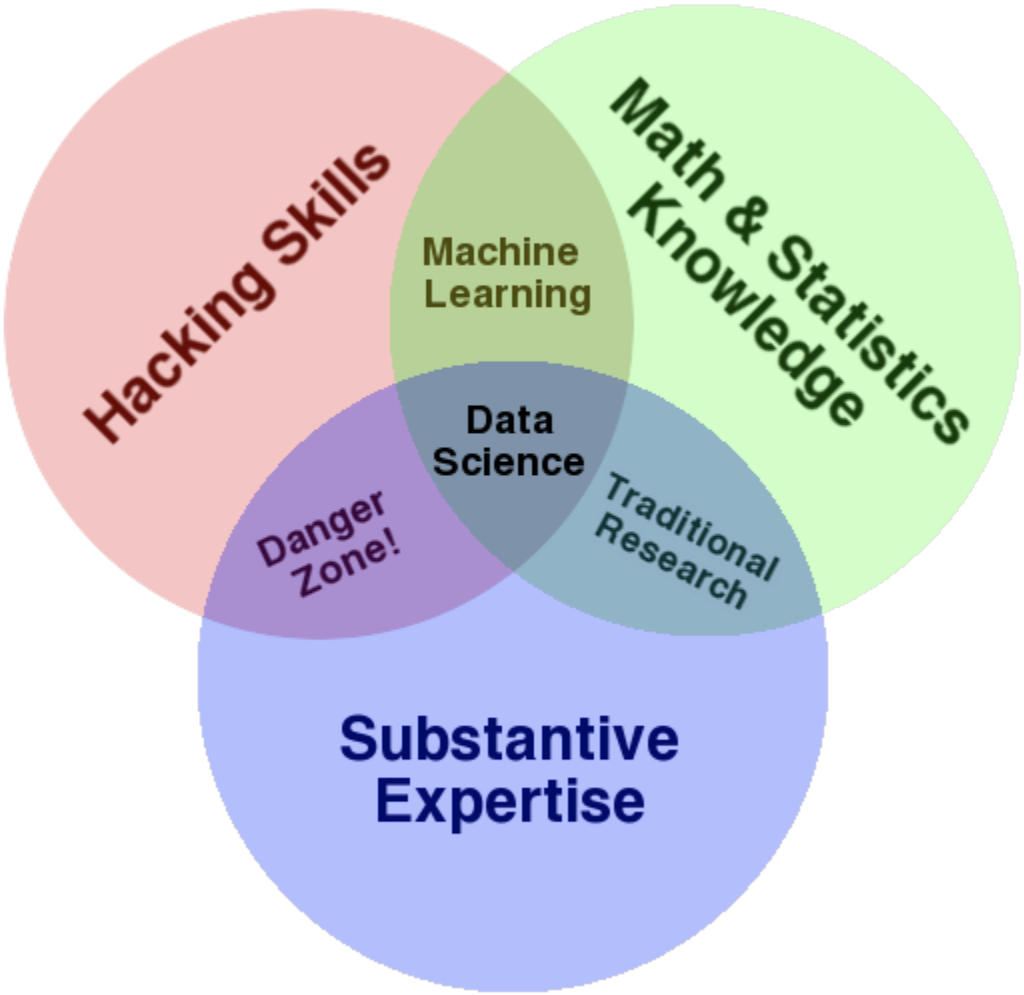
While the "data scientist" term is controversial, it seems to be here to stick. Over the last 5 years, the term has been increasingly used in job postings[4]. More and more people and businesses use the term. Nowadays, many prominent figures in the data community describe themselves as data scientists.

Conway's Venn Diagram

> "The primary colors of data: hacking skills, math

> and stats knowledge, and substantive expertise" - Drew Conway

Drew Conway is a well-known data scientist and the author of Machine Learning for Hackers. Conway presented one of the most widely acknowledged definitions of a data scientist. In his famous Venn Diagram [5], Conway presents a data scientist as an individual that is at the intersection of three skills: *hacking*, *mathematics and statistics* and *substantive expertise*.



The *hacking skills* refer to the computer science skills. Data is digital. In order to efficiently manipulate the data, you need to have some programming skills. You need to be comfortable at

the command line, be able to manipulate files of different formats, program algorithms that will modify the data, etc. According to Conway, "this does not require a background in computer science—in fact—many of the most impressive hackers [he has] met never took a single CS course"[6].

The *Math and Statistics knowledge* refer to the mathematical abilities required to properly analyze and infer things about the data. On this topic, Conway says "this is not to say that a PhD in statistics in required to be a competent data scientist, but it does require knowing what an ordinary least squares regression is and how to interpret it"[6].

The *Substantive Knowledge* is the knowledge specific to the area where data science is applied. It is often referred to as "domain knowledge". For example, if you are applying data science to genome problems, you should have "substantive knowledge" on that topic.

In that vein Brad Schumitsch, a data scientist at Twitch, explains that data science, in his team, is composed of three main areas of knowledge: statistics, programming and product knowledge. When interviewed by Mixpanel on the subject of the "Data Scientist" job title, he said:

> "At Twitch, our data science team brings together three things: statistics, programming, and product knowledge. And we would never hire someone who wasn't strong in stats. You can be a great programmer, but if you don't know what Bayes Rule is, then we have an engineering department I can

> point you to."[7]

Type A and Type B

While Conway's definition of a data scientist is generally agreed on, there are divergences of opinions on the breath and depth that should be expected from a data scientist to have in each of those areas.

Michael Hochster, Head of Research at Pandora, defines a data scientist as someone "with some mix of coding and statistical skills who work on making data useful in various ways"[8].

What's special in Hochster definition is that he categorizes the data scientist in two categories: "Type A" and "Type B".

Type A for Analysis

> Type A Data Scientist: The A is for Analysis. This type is primarily concerned with making sense of data or working with it in a fairly static way. The Type A Data Scientist is very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on. [8]

Type B for Building

> Type B Data Scientist: The B is for Building. Type B

> Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data "in production." They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results). [8]

As you can see, the depth of each knowledge area is different between the *types* of data scientist. The *Type A* data scientist is expected to have much stronger skills on the statistical side. The *Type B* is expected to be a "very strong coder" that will push code in production.

This categorization of the different *types* of data scientists is a recurring theme. Raj Bandyopadhyay, Director of Data Science at Springboard, believes there are two types of data science roles: the product one and the insights one[9]. While the categorization is slightly different, the concept remains the same: there are different types of data scientists out there with different strengths and weaknesses in each of the core knowledge areas.

In Summary

Across those different definitions, it's clear that a data scientist combines multiple skill sets: statistics, hacking and domain knowledge. The data scientist should have sufficient knowledge in all three of those areas.

The depth in each of those knowledge areas will vary depending to whom you talk. That's probably why the definition can be fuzzy.

Each company will have their very own expectations of what is a data scientist.

A data scientist working on search algorithms at Google will have a different profile than a data scientist working on understanding product usage of an e-commerce website. Both will have the same title, but their training will be fundamentally different. The one working at Google will most likely have a Ph.D and have very advanced knowledge in computer science and maths. The second one will have good domain knowledge about the e-commerce space but might not have a formal training in maths or computer science.

There is a whole spectrum of *data scientists*. You need to know where you stand in this spectrum.

1. The World's Most Influential People – Time Magazine ↩
2. What I Need From Statisticians - Nate Silver ↩
3. Data Science: What's The Half-Life Of A Buzzword? - Forbes ↩
4. "Data Scientist" Job Trends ↩
5. The Data Science Venn Diagram - Drew Conway ↩
6. bitly's Hilary Mason on "What is A Data Scientist?" - Forbes ↩ ↩2
7. This is the difference between statistics and data science - Mixpanel ↩
8. What is Data Science? - Quora ↩ ↩2 ↩3
9. Do I need a Masters/PhD to become a data scientist? - Quora ↩

Subscribe to get a weekly email that will get you one step forward to becoming a data scientist.

Subscribe

Every week, you will get fresh content in your inbox. I cover the topics of Data Analysis in Python, Applied Statistics, Machine Learning, SQL etc.

Related Posts

Profiling a Dataset of Craft Beers 23 Apr 2017

Scraping for Craft Beers 17 Jan 2017

What I Learned Implementing a Classifier from Scratch in Python 04 Jan 2017