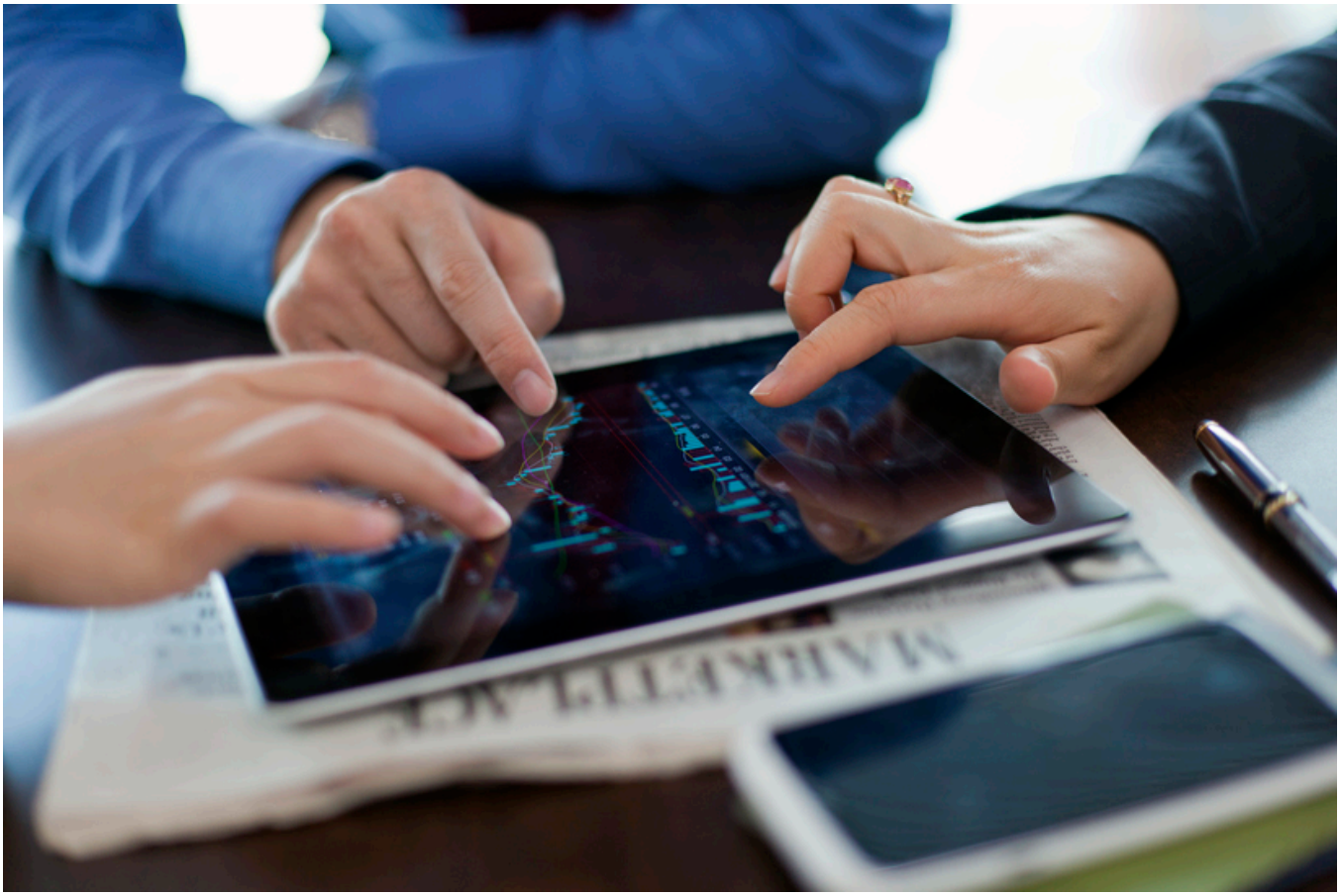
It Seems Like Anyone Can Be a Data Scientist... but Is It True?



Is data science too easy? *originally appeared on* Quora: *the place to gain and share knowledge, empowering people to learn from others and better understand the world.*

Answer **by** Maurice Ewing, Trained and Led Data Science teams in over 50 countries, on Quora:

No, data science is not easy. It's just unshaped and not "professionalized."

By this I mean there are no standard sets of tools, no educational curricula, no certifying bodies, nor any specific career paths that lead to becoming a data scientist; yet all the essential bits are there—and they're not easy to acquire, assemble or apply well.

Yes, one can learn R and Hadoop and "claim" to be a data scientist, but that's far from the truth. By comparison, one can also take a few medical classes and claim to be a doctor or watch a few courtroom TV shows and claim to be a lawyer. The difference is that the disciplines of medicine and law are "professionalized." As a result, they are able to guard their gates by setting standards on who can call themselves a "doctor" or a "lawyer." In data science, we cannot do that as of yet.

Insofar as R and Hadoop, they're just part of the data science toolkit. They don't constitute "data science" any more than a scalpel constitutes "surgery." In the same way that physics relies upon mathematics, data science relies upon statistical tools for handling large and small data sets, structured and unstructured data, etc. But the mathematics of physics is not a substitute for scientific thinking, analysis, approach or method—and neither are Hadoop and R substitutes for understanding behaviour in data.

Statistics, specifically, is concerned largely with methods for testing hypotheses using data; thus, before one can constructively use Hadoop or R, one needs to know statistics and know it well. Because, unlike statistics—which is concerned largely with testing the hypotheses and stops there—data science focuses on the implications of systematic departures from hypotheses (as evidenced by statistical tests) and the bigger conclusions we can make as a result of those departures.

Moreover, apart from data science requiring a cumulative knowledge of numerous tools or sub-disciplines, like statistics, R, Hadoop, etc., one must be able to bring those tools to bear in answering important business questions and achieving business outcomes, neither of which initiative directly follows from knowledge of the tools. That ability, skill, experience or talent, is what's brought to the table by the data scientist, enabling him or her to justifiably call himself or herself, a "data scientist."

This leads me to believe that the real question here is, "Can anyone BE a data scientist?" And to that I would say no, not at all, for the very reasons I just mentioned. In my experience, not even your top CS or STEM majors from a top school can easily become good data scientists, without additional training in it, and some personal factors. Apart from its multidisciplinary nature, data science requires a deep love of the divergence between observed reality in data and the prediction of mathematical models. To do that, one needs something more than just a mastery of tools. One needs a love for imperfection.

I've been in this field almost 20 years, from back before the term "data science" existed, so I've seen many things. In fact I believe data science excellence requires a number of years in actually applying it before one can truly understand data, how it behaves, how different models work, backwards and forwards, etc. Yet, most importantly, excellence requires making mistakes and understanding mistakes, along with appreciating the variations between observed and predicted reality. Thus, I affectionately call data science the science for imperfect people, like myself.

I say that half-jokingly. In truth, I believe all good science is for imperfect people—people who become curious, not angry, when they see imperfection and variation. STEM majors that cannot stand imperfection and variation will never make good scientists or good data scientists, much like bigots cannot make for good neighbours. Why? Because the world we live in is imperfect and variable and its beauty lies in that imperfection and variability. Moreover, iignorance – not knowledge – drives science, and imperfection and variation are the hallmarks of ignorance.

Thus, while I love mathematics, I don't find it incredibly interesting beyond a certain level for one simple reason: it always works out. In this one and only way, I found myself a kindred spirit with one of my former teachers—John Nash (the "Beautiful Mind" Nash). I once asked him why he didn't stay in mathematics instead of shifting to economics. Nash replied, "Because mathematics is too easy." Now I can't say I shared that realisation (I mean, seriously?), but after exploring a number of mathematical disciplines I have come to the conclusion that it's just too "perfect" for imperfect me.

Data science, by contrast, takes these perfect models and builds them against real data generated by human beings, animals and human beings that sometimes behave like animals. These creatures rarely exhibit behavior that results in closed form systems and solutions. In other words, data science cuts to the heart of how we operate as humans in our surrounding world.

We have expectations (i.e., mental models) that usually depart from reality. When pursuing our goals, we exhibit the drama or comedy of that departure. Thus, in doing data science we're doing something truly Shakespearean way: We are beautifully characterising, with numbers, the drama (or comedy) of human behaviour!

Okay, so now that I've given my Good Will Hunting speech, let me offer a specific example. My typical consulting job involves independently validating suites of commercial models produced by a client's or a consulting team's data scientists. Done properly, I employ a toolkit of validation and sampling techniques (small sample, non-parametric, weighted/unweighted, etc.) that I apply in a kind of exploratory and stress test fashion, like a CSI. (I throw that in to give me sex appeal.) However, because of my experience, I can usually see what's gone wrong with the models even before undertaking any formal tests.

Now I'm no genius, but even when the models are exceptionally complex, I can do this, sometimes even more easily. Thus, I've spotted the problems in highly non-linear models containing well over 100 variables (which, itself, is often the problem!) But all

that comes from experience in seeing mistakes, making mistakes and in understanding reality versus forecasted perfection. If there's a divergence, I get all excited about it. Moreover, I've got business sense and executive experience, so I well understand that the "right" answer is often the one that must support some business outcome or objective.

Thus, in all these data science adventures, I've commonly observed two things: 1) I'm usually right (or no one would hire me again) and, 2) 99% of the people I am validating (STEMs, who usually have PhDs in physics, mathematics, astrophysics, etc.) didn't see it. Thus, diplomacy becomes a necessary and added dimension to the data science toolkit, as one often has to reveal bad news!

Moreover, it's important to understand that most of the times, the developers didn't make any serious mistakes. Their models just don't do what they were expected to do or ignore the business realities they are being paid to observe. So that's when I'm called back in to train them in trying to see what I saw so they can later see it for themselves.

And all that's tough! Sometimes it's like trying to describe the taste of honey to someone that's never tasted it. Of course, that's not meant, at all, to sound condescending. I'm just coming back to the point that there is an essential, learning-in-the-field aspect to data science that is above and beyond the importance of learning tools like R and Hadoop.

But that brings me back, full circle, to the confusion surrounding one being able to "call" one's self a data scientist, even if one doesn't possess the entire toolkit, the experience and the love of imperfection and variability. As a budding profession, data science has a lot of work to do. We need a more standardised multidisciplinary curriculum, implemented by people with field and business experience (not just academics) and perhaps a professional body or two that can guard the gates.

Until then, people that call the shots in high places will continue to hire whatever STEM or CS major knows Hadoop and R and is willing to work for cheap. That obfuscates things and probably frustrates them as well. Because, in truth, it's much more difficult and complicated than that, and so is data science.

This question *originally appeared on* Quora. *the place to gain and share knowledge, empowering people to learn from others and better understand the world. You can follow Quora on* Twitter, Facebook, and Google+. More questions: