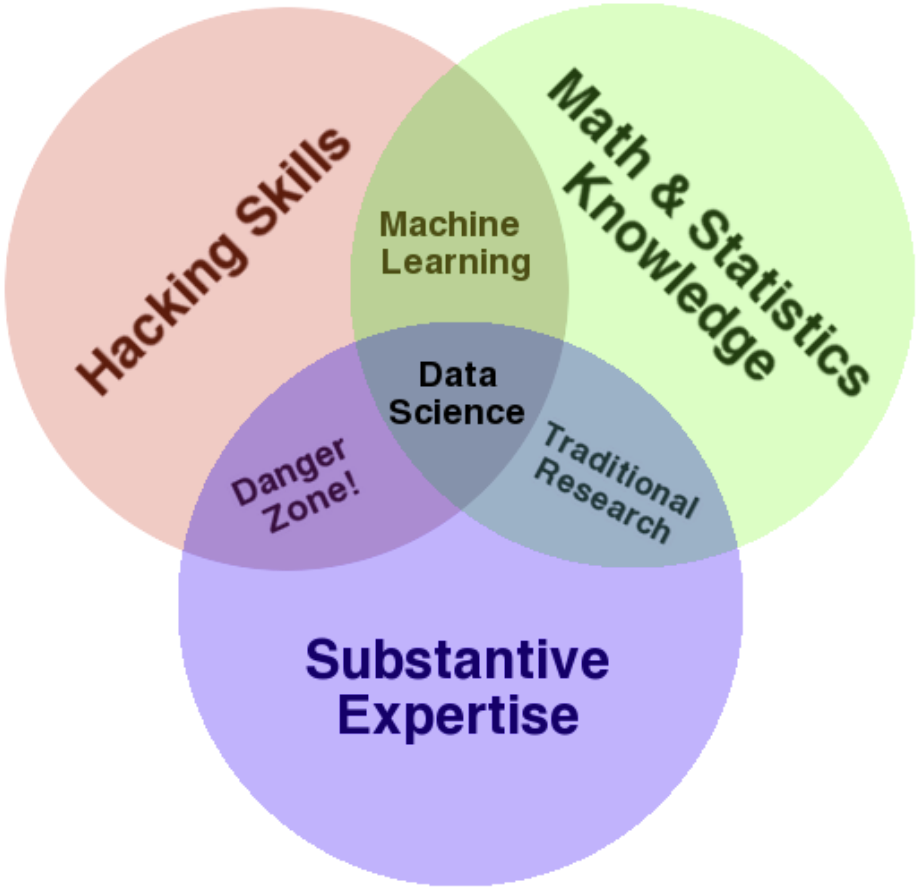
THE DATA SCIENCE VENN DIAGRAM

On Monday I—humbly—joined a group of NYC's most sophisticated thinkers on all things data for a half-day unconference to help O'Reily organize their upcoming Strata conference. The break out sessions were fantastic, and the number of people in each allowed for outstanding, expert driven, discussions. One of the best sessions I attended focused on issues related to teaching data science, which inevitably led to a discussion on the skills needed to be a fully competent data scientist.

As I have said before, I think the term "data science" is a bit of a misnomer, but I was very hopeful after this discussion; mostly because of the utter lack of agreement on what a curriculum on this subject would look like. The difficulty in defining these skills is that the split between substance and methodology is ambiguous, and as such it is unclear how to distinguish among hackers, statisticians, subject matter experts, their overlaps and where data science fits.

What is clear, however, is that one needs to learn a lot as they aspire to become a fully competent data scientist. Unfortunately, simply enumerating texts and tutorials does not untangle the knots. Therefore, in an effort to simplify the discussion, and add my own thoughts to what is already a crowded market of ideas, I present the Data Science Venn Diagram.



How to read the Data Science Venn Diagram

The primary colors of data: **hacking skills, math and stats knowledge, and substantive expertise**

- On Monday we spent a lot of time talking about "where" a course on data science might exist at a university. The conversation was largely rhetorical, as everyone was well aware of the inherent interdisciplinary nature of the these skills; but then, why have I highlighted these three? First, none is discipline specific, but more importantly, each of these skills are on their own very valuable, but when combined with only one other are at best simply not data science, or at worst downright dangerous.
- For better or worse, data is a commodity traded electronically; therefore, in order to be in this market you need to speak hacker. This, however, does not require a background in computer science—in fact—many of the most impressive hackers I have met never took a single CS course. Being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically; these are the hacking skills that make for a successful data hacker.
- Once you have acquired and cleaned the data, the next step is to actually extract insight from it. In order to do this, you need to apply appropriate math and statistics methods, which requires at least a baseline familiarity with these tools. 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.
- In the third critical piece—substance—is where my thoughts on data science diverge from most of what has already been written on the topic. To me, data plus math and statistics only gets you machine learning, which is great if that is what you are interested in, but not if you are doing data science. Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods. On the flip-side, substantive expertise plus math and statistics knowledge is where most traditional researcher falls. Doctoral level researchers spend most of their time acquiring expertise in these areas, but very little time learning about technology. Part of this is the culture of academia, which does not reward researchers for understanding technology. That said, I have met many young academics and graduate students that are eager to bucking that tradition.
- Finally, a word on the hacking skills plus substantive expertise danger zone. This is where I place people who, "know enough to be dangerous," and is the most problematic area of the diagram. In this area people who are perfectly capable of extracting and structuring data, likely related to a field they know quite a bit about, and probably even know enough R to run a linear regression and report the coefficients; but they lack any understanding of what those coefficients mean. It is from this part of the diagram that the phrase "lies, damned lies, and statistics" emanates, because either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created. Fortunately, it requires near willful ignorance to acquire hacking skills and substantive expertise without also learning some math and statistics along the way. As such, the danger zone is sparsely populated, however, it does not take many to produce a lot of damage.

I hope this brief illustration has provided some clarity into what data science is and what it takes to get there. By considering these questions at a high level it prevents the discussion from degrading into minutia, such as specific tools or platforms, which I think hurts the conversation.

I am sure I have overlooked many important things, but again the purpose was not to be speific. As always, I welcome any and all comments.

Cross-posed at dataists