The difference between Statistics and Data Science

Ben Klemens

30 October 2014

An academic field is a social network built around an accepted set of methods.

Economics has grown into the study of human decision making in all sorts of aspects. At this point, nobody finds it weird that some of the most heavily-cited papers in the Economics journals are about the decision to commit crimes or even suicide. These papers use methods accepted by economists, writing down utility functions and using certain mathematical techniques to extract information from these utility functions. Anthropologists also study suicide and crime, but using entirely different methods. So do sociologists, using another set of tools. To which journal you submit your paper on crime depends on matching the methods you use to the methods readers will be familiar with, not on the subject.

[A notational digression: I hate the term 'data science'. First, there's a general rule (that has exceptions) that anybody who describes what they're doing as "science" is not a scientist—self-labelling like that is just trying too hard. And are we implying that other scientists don't use data? Is it the data or the science that statisticians are lacking? Names are just labels, but I'll hide this one under an acronym for the rest of this. I try to do the same with the United States DHS.]

I push that the distinction is about the set of salient tools because I think it's important to reject other means of cleaving apart the Statistics and DS networks. Some just don't work well and some are as odious as any other *our people do it like this, but the other people do it like this* kind of generalizations. These are claims about how statisticians are too interested in theory and too willing to assume a spherical cow, or that DSers are too obsessed with hype[1] and aren't careful with hypothesis testing. Hadley[2] explains that "...there is little work [in Statistics] on developing good questions, thinking about the shape of data, communicating results or building data products" which is a broad statement about the ecosystem that a lot of statisticians would dispute, and a bit odd given that he is best known for building tools to help statisticians build data products. It's not hard to find people who say that DS is more applied than Stats, which is an environment issue that is hard to quantify and prone to observation bias. From the comment thread of this level-headed post[3]: "I think the key differentiator between a Data Scientist and a Statistician is in terms of accountability and commitment."

Whatever.

We can instead focus on characterizing the two sets of tools. What is common knowledge among readers of a Stats journal and what is common knowledge among readers of a DS journal?

It's a subjective call, but I think it's uncontroversial to say that the abstract methods chosen by the DSers rely more heavily on modern computing technique than commonly-accepted stats methods, which tend to top out in computational sophistication around Markov Chain Monte Carlo.

[1] `http://simplystatistics.org/2014/10/28/why-i-support-statisticians-and-their-resistance-to-hype/`
[2] `http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics`
[3] `http://www.datasciencecentral.com/profiles/blogs/data-science-the-end-of-statistics`

One author[4] went to the extreme of basically defining DS as the practical problems of data shunting and building Hadoop clusters. I dispute that any DSer would really accept such a definition, and even the same author effectively retracted his comment[5] a week later after somebody gave him an actual DS textbook.

If you want to talk about tools in the sense of using R versus using Apache Hive, the conversation won't be very interesting to me but will at least be a consistent comparison on the same level. If we want to talk about generalized linear models versus support vector machines, that's also consistent and closer to what the journals really care about.

The basic asymmetry that the price of admission for using DS techniques is greater computational sophistication will indeed have an effect on the people involved. If we threw a random bunch of people at these fields, those who are more comfortable with computing will sort themselves into DS and those less comfortable into Stats. We wind up with two overlapping bell curves of computing ability, such that it is not challenging to find a statistician-DSer pair where the statistician is a better programmer, but in expectation a randomly drawn DSer writes better code than a randomly drawn statistician. So there's one direct corollary of the two accepted sets of methods.

Three Presidents of the ASA[6] wrote on the Stats vs DS thing, and eventually faced the same technical asymmetry:

> Ideally, statistics and statisticians should be the leaders of the Big Data and data science movement. Realistically, we must take a different view. While our discipline is certainly central to any data analysis context, the scope of Big Data and data science goes far beyond our traditional activities.

This technical asymmetry is a real problem for the working statistician, and statisticians are increasingly fretting about losing funding[7]—and for good reason. Methods we learned in Econ 101 tell us that an unconstrained set leads to an unambiguously (weakly) better outcome than a constrained set.

If you're a statistician who is feeling threatened, the policy implications are obvious: learn Python[8]. Heck, learn C—it's not that hard, especially if you're using my C textbook, whose second edition was just released[9] (or *Modeling with Data*, which this blog is ostensibly based on). If you have the grey matter to understand how the F statistic relates to SSE and SSR, a reasonable level of computing technique is well within your reach. It won't directly score you publications (DSsers can be as snobby about how writing code is a "mere clerical function" as the statisticians and US Federal Circuit can be), but you'll have available a less constrained set of abstract tools.

If you are in the DS social network, an unconstrained set of tools is still an unambiguous improvement over a constrained set, so it's worth studying what the other social network takes as given. Some techniques from the 1900s are best left in the history books, but now and then you find ones that are exactly what you need—you won't know until you look.

By focusing on a field as a social network built around commonly accepted tools, we see that Stats and DS have more in common than differences, and can (please) throw out all of the bigotry

[4] http://andrewgelman.com/2013/11/14/statistics-least-important-part-data-science/

[5] http://andrewgelman.com/2013/11/19/22182/

[6] http://magazine.amstat.org/blog/2013/06/01/the-asa-and-big-data/

[7] http://normaldeviate.wordpress.com/2013/04/13/data-science-the-end-of-statistics/

[8] https://duckduckgo.com/?q=learn%20python

[9] http://www.amazon.com/exec/obidos/redirect?link_code=ur2&camp=1789&tag=caltechdivini-20&creative=9325&path=tg/detail/-/1491903899/qid=1120157199/sr=8-1/ref=pd_bbs_ur_1

that comes with searching for differences among the people or whatever environment is prevalent this week. What the social networks will look like and what the labels are a decade from now is not something that we can write a policy for (though, srsly, we can do better than "data science"). But as individuals we can strive to be maximally inclusive by becoming conversant in the techniques that the other social networks are excited by.

Next time, I'll have more commentary derived from the above definition of academic fields, then it'll be back to the usual pedantry about modeling technique.