Bridging the Gap Between Statistics and Other Data Sciences: Where's the Bridge? Where's the Gap?

Randy Bartlett, Blue Sigma Analytics, 105 Shelbourne Lane, Phoenixville, PA 19460
Fred Hulting, General Mills, Inc, 9000 Plymouth Road N, Minneapolis, NN 55247
Mark Lancaster, Northern Kentucky University, Highland Heights, KY 41099
Jason Brinkley, Abt Associates, Inc, 5001 South Miami Blvd, Durham, NC 27703

Abstract

The recent rise in popularity of data science has been met with mixed feedback within the statistics community. What is clear is that there currently is a community of data scientists who are not engaged with the statistics community and see themselves as existing in a different space. Is there a need for better integration with that community? Does that integration already exist for some? Is there common ground to be found and how big is the gap between data science and applied statistics? Is data science the future of statistics? Are we adequately preparing students to enter into a world where data science is on the rise and occasionally at odds with statistics? How much emphasis should be placed on statistics within data science programs? How much emphasis should be placed on data science within statistics programs? How should companies organize to best leverage opportunities in both areas? This panel discussed both business, government, and academic settings and key topics from this discussion are listed here.

Key Words: Data Science; Applied Statistics; Analytics; Culture; Future

1. Introduction

A recent panel was put together with a variety of experts working in both the data science and applied statistics realms. Our team included five volunteers coming from areas of academia, industry, and government work. The panel discussion was well received and some of the important thought points from some of our volunteers have summarized key points and ideas here.

2. Introduce yourself, and describe what data science looks like in your role. What part of data science would you consider to be NOT a part of applied statistics?

2.1 Randy Bartlett

I am one of a large segment of corporate statisticians on the front lines of this challenge to the relevance of our profession. I would say that I am neo classically trained—the first part of my technical training. I was anointed with alphas and empowered in software: Ph.D. & M.S. Department of Statistics, Texas A&M University; B.S. Computer Science & Statistics, Iowa State University.

Like my peers, I completed much of my training in the field—a ten year 'post doc.' I needed to learn more methodology; how to use my current client's software to solve problems in text books; more techniques as they became applicable; and professional skills like talking to clients; leading other quants; organizing teams; planning; making stat-based decisions; etc. I learned a great deal from software manuals, conferences/workshops, veteran statisticians and econometricians, trial and error, and Toastmasters (Competent Communicator & Competent Leader). These discoveries would go into a book 'A Practitioner's Guide to Business Analytics' (McGraw-Hill, 2013), which is the practice of statistical data science AKA applied statistics. Meanwhile, I needed to master whatever domain within which I was applying statistics—banking, pharma, manufacturing, retail, etc.

I earned the Ph.D. with the idea of going to the clinical side of big pharma, yet I graduated when the industry was not hiring. Instead, my formative technical years were in the government and the banking industry. I worked on MEPS (Medical Expenditure Panel Survey) with John Sommers, Ph.D. and Steve Cohen, Ph.D. I found that a great deal of sampling statistics know-how is not in the literature.

At Citigroup/The Associates, Sears, and Wells Fargo, I worked with a number of bright minds, especially Isaac Abiola, Ph.D. and Vidyut Vashi, Ph.D. The datasets were extensive (some might say Big) for that time—size is relative, relative to your needs and capabilities. We were trying to predict customer behavior using new types of predictive models and by adapting models and diagnostics built for coefficient estimation. Through trial and error, we found how to better apply various types of models AND the value of Statistical Qualifications (CH 7), Statistical Diagnostics (CH 8), and Statistical Review (CH 9) in creating competitive advantage. We called our hard-learned methodological and technical flourishes, Best Statistical Practice. We did not publish any of it. Unknown to us, it was around this time that C.F. Jeff Wu, Ph.D. started using the term, 'data science,' to describe applied statistics, which is what we and thousands of others were all about.

Meanwhile, I could not help noticing better ideas for leading technical teams (CH 2 & 4); making decisions (CH 3); organizing resources (CH 5); and planning how to apply statistics to an organization (CH 6). These strategic areas were rounded out by there more tactical ones: Data Collection (CH 10), Data Software (CH 11), and Data management (CH 12). These areas would be recurring themes throughout my career.
I spent the next ten years working in various industries, collecting the best ideas from my generation on how to compete using applied statistics. Eventually, I put this into a book 'A Practitioner's Guide to Business Analytics' (McGraw-Hill, 2013), like a note in a bottle, so that the next generation can find their way off the deserted island and bypass a generation of time-consuming trial and error.

As I am a consultant, I vie for work in the open market—such as it is; competing against pretenders (braggarts, liars, and thieves), and some of the best applied statisticians in the world. I recognize the value of certifications for protecting a profession from knockoffs: CAP® (INFORMS) & PSTAT® (ASA). About four years ago, there were vigorous discussions on the internet: LinkedIn, KD Nuggets, et al. about how irrelevant applied statistics is now that … nothing. Nothing had changed, except IT. Now IT needed a place to expand and their software jingles brazenly claimed that they were the answer to everything. Part of this everything was data analysis and their initial claims there did not fare very well. They were rebuffed by this field named statistics. Their response was to deny and mischaracterize applied statistics, which was in the way of their fabricated value proposition.

A number of trolls and their sock puppet allies joined in; extolling their role as arbitrators of what is and is not statistics. Their ignorance of statistics demonstrated why they were excluded from some statistics job or speaking opportunity on their road to becoming a troll. Meanwhile, many statisticians working in other parts of the statistics realm were so easily buffaloed that they started writing about how these fields are so different and how we can help these new data scientists learn statistics—ever the experts in all matters of statistics, even those with which they are wholly ignorant.

It would be unjust not to mention that many in IT have all along been savvy enough to recognize that their talking heads were shameless talking fools.

In the references (Bartlett, Gray, Mitchell-Guthrie, Luker, & Speidel), you will find some responses to the horse manure of 'celebrated' IT talking heads. In my Statistical Denial series, I tore the wings off of all of the insidious myths up to that time. New ones are always brewing. One goes that data sciences are a step beyond applied statistics because they go further, they are multidisciplinary! They use stat, math, and software. The reader is left to wonder just how this is different from applied statistics, econometrics, psychometrics, et al., which use stat, math, and software. Another new claim is that we can see the leap forward provided by data science by juxtaposing the curriculums of Data Science vs Statistics. However, this just shows that the statistics curriculum has not kept up with applied statistics.

If nothing else, my career has taught me that applied statistics is about one thing … people. Most non-statisticians still believe in us and our value proposition and if we hold firm, we will hold on.

IT data science or data management should be separate from applied statistics AKA statistical data science. Applied statistics is now one of the so-called data sciences. Applied statistics deals with extracting information from numbers with uncertainly, as it always has. We define it by its problems, not its tools. The remainder of the data sciences can go into data management and/or 'other' if you like.

2.2 Fred Hulting

I currently lead a diverse, international, knowledge services organization within the R&D organization of a large, international, food company. My team includes both statistics and computing professionals working from the US and India. I also serve on corporate advisory teams for our connected data and data science initiatives. My education includes a B.S. in Statistics and Computer Science from UC Davis, and M.S. and Ph.D. degrees in Statistics from Iowa State University

My company is experiencing the impact of the growth of data science within industry, and working to understand the similarities and differences between that discipline and our current statistics practices. I'd like to begin by sharing three points that will shape my comments throughout the session today.

The first is that I will not try to define either the term "data science" or "applied statistics." There is little I can add to that ongoing debate. Instead, I will draw conclusions about these fields by reflecting on the work being done in my company by self-described "data scientists" and "applied statisticians."

Second, there is a lot of heterogeneity in what falls under the heading of "applied statistics." Statisticians play many different roles in academia, government, and industry, working on a huge variety of applications, and drawing from a large toolbox of methodology, theory, and computing technology. This is true within our company as well. Any comparison between applied statistics and other disciplines needs to recognize this fact.

Finally, in the context of the problems and challenges faced by my company, the actual work of data scientists and applied statisticians is quite similar, and could be described as "analytics." I use this term to refer broadly to the use of data, data analysis, and modeling to influence and execute successful business decisions. Analytics work shares a common framework, with connected DATA sources being prepared for an ANALYSIS that produces the necessary insights that are then fed to a DELIVERY mechanism that will spur ACTION to solve a problem.

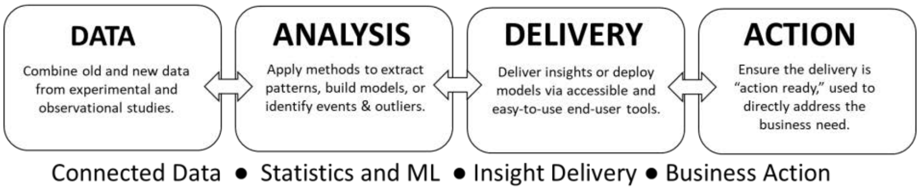
DATA
Combine old and new data from experimental and observational studies.
ANALYSIS
Apply methods to extract patterns, build models, or identify events & outliers.
DELIVERY
Deliver insights or deploy models via accessible and easy-to-use end-user tools.
ACTION
Ensure the delivery is "action ready," used to directly address the business need.
Connected Data ● Statistics and ML ● Insight Delivery ● Business Action

Figure 1: A framework for the components of an analytics process.

This shared purpose and framework has allowed our statisticians and data scientists to come together as a community, and identify the similarities and some differences, in the work that they do.

What distinguishes the work of "data scientists" from the work of more traditional "applied statisticians"? I will call out three areas that I see in my company:

Application Areas: Our data scientists primarily working in those areas where the data sources are very large (e.g. the 50 billion rows in our consumer data foundation) or where there is a need to create models, and deliver predictions, at scale. So, for example, we see these roles forming in our central IT organization and our Marketing and Consumer Insights organization. In contrast, we still see traditional statistical approaches in our R&D and regulatory areas, where experimentation and risk assessment activities dominate. Between these ends are a variety of business areas where we see a blend of these practitioners and their approaches.

Practitioner Skill Sets: The data scientists in the company have enriched our organization with a variety of new talents, enabling us to operationalize the analytics process at a large scale. These individuals are typically stronger in areas that are less familiar to many of statisticians; namely, new tools and technology for data management and insight delivery, and algorithmic approaches to building predictive models ("machine learning").

Perspectives on Data: The work of the data scientists tends to be "data centric." That is, the he applications tend to be driven by the availability of data, and the analytics usually focus on drawing insights about the data set "as is". In contrast, the statistics applications view data as coming from a real-world process and either: generate new data directly through active experimentation; or recognizes how the underlying process generated the data.

In presenting these differences, I will not get into value judgements between the work of data scientists and statisticians. The first two areas of difference – application and skills sets – could apply even within the varied work of "applied statisticians." The last area of difference is probably one that most differentiates the work of these practitioners, and seems to be driven primarily by the application and the needs of the business.

2.3 Mark Lancaster

There are methods/algorithms in the realm computer science that are highly applicable to data science, but fall short of being part of statistics. An example of this would be the field of association rules mining (aka market basket analysis) and the *a priori* algorithm, where the goal is to identify what purchasing associations can be found for a given set of items. Being able to efficiently search and identify the rules falls into the realm of data science, but that portion of the analysis doesn't have a statistical component until you want to test hypotheses gleaned from the rule search.

3. Has the rise of popularity of data science been a real threat to the statistics community or are these threats more perception?

3.1 Randy Bartlett

At the moment the chaos is a threat to applied statisticians in the field, in the tradition of Six Sigma, yet far worse. Our career opportunities have been limited in several industries; just look at the job descriptions. The curriculum for newly minted statisticians has not kept pace with changes in applied statistics. IT has a powerful marketing arm telling our would-be customers that IT knows how to analyze their data and/or that applied statistics is now irrelevant because we have them. Most IT professionals are fine; there some who believe that they can perform statistics using Oracle and similar tall tales. Meanwhile, some statisticians have been buffaloed into believing some of these claims and they are drastically underestimating what applied statisticians are already doing in the field. We are working on every type of data analysis problem known. We are better trained to use data mining/machine learning tools; we just need to get those projects.

3.2 Fred Hulting

Building on my comments from the previous question, my own view is that the work of our data scientists fits under the very broad and heterogenous activities of 'applied statistics. In fact, I have heard others refer to data science as "applied statistics for technology." I interpret this to mean that data science has expanded the field of applied statistics by tackling new applications, driving innovation in analysis methods, and expanding the toolset that can drive results for the business.

But even under this view, the fact is that the work of data scientists has largely developed outside of the usual societies and forums that define the statistics profession. Hence the need to ask this question.

I do think there are challenges (threats). These are being driven by several changes over the last twenty years or so. The first is the greater availability of data of all types; data is easy to collect, and cheap to store. The second is the greater awareness, fueled by the media, that data and data analysis can play a significant role in solving problems and guiding decisions. Finally, there is the emergence of new business models that benefit from data availability and algorithmic approaches to modeling and prediction at scale (online retail, finance, etc.)

These changes have shaped the differences I discussed above. New applications and new skills sets were required to cope with these changes. With the high demand for such skills, it is not surprising that supply has focused on delivering the requisite skill sets; skill sets often seem incomplete from the perspective of a statistician. Of course, many statisticians have complementary skills gaps. Closing these gaps is a key to overcoming the challenges faced by the statistics profession.

3.3 Mark Lancaster

I believe that the popularity of data science has enhanced the need and desire for statistical training, so that threats are more perception. There will always be a need to understand and quantify the effects of randomness in our data. The real threat is in the lack of training of our statisticians on the tools and techniques use to data science. Applied statisticians and consultants would most likely agree that knowing non-statistical methods (such as various data processing, cleaning, and extraction methods) are a very important to industry, and that these are usually not taught at an undergraduate level. Having these skills, in addition to being a statistician, would minimize the threat of "data science."

4. It has been suggested that degree doesn't matter in Data Science, if that becomes a commonly accepted idea, what impact will this have on the statistics community?

4.1 Randy Bartlett

Training is what matters. Obviously, applied statisticians require far more training than they are receiving from their degree programs and statistics departments should open themselves to feedback. Table 1 from 'A Practitioner's Guide To Business Analytics' provides some guidance into how well our degree programs are covering the six areas on the left.

Table 1: Progress from Academic Training

	BS in Quant (uncommon)	**MS in Quant**	**PhD in Quant**	**MBA**	**Post-Academic Training**
Analytics-Based Decision Making	30–50%	40–60%	+10–50%	+0–30%	Remainder of Training
Client Interaction	0–10%	20-30%	+20–60%	+30–60%	Remainder of Training
Analytics (Basic Algorithms, Statistics, and Mathematics)	Plenty	Plenty	Plenty	+30–100%	NA
Advanced Analytics	30–60%	30–60%	+5–20%	+0%	Remainder of Training
Statistical Software	60–80%	40–80%	+0–100%	+0–20%	Remainder of Training
Theory	200–300%	0–500%	+100–500%	+0%	Never Happen

Furthermore, these problems with the profession are exacerbated by society's low statistics literacy. Universities have an opportunity to train the customers/partners of applied statisticians to recognize false claims and distortions in statistics.

4.2 Fred Hulting

In industry, there has been a general shift in our hiring processes towards deemphasizing any specific degree, and instead assessing the collective education, experience and skills of prospective employees. Within my own team we have hired individuals with varied degrees, who have had statistics coursework and applied experience. Still they exhibit the talent and understanding to apply statistics to get business results. Regardless of the degree, or where you were trained, the focus is on the ability to be successful at the job. As I stated above, many of our data scientists were hired because of their ability to do a certain type of work.

In the future, I believe that both data science and applied statistics will evolve to become even more of a team sport. Given the various types of expertise that are required to successfully implement analytics in a large company, we need to assemble teams that supplement generalists with deeper expertise in several disciplines. For example, one of the business functions in our company has a high-performing team with clearly identified roles of data stewards, data scientists, statisticians, and data visualizers (nicely paralleling the structure in Figure 1). This is a good example of how we need to work differently.

4.3 Mark Lancaster

Statisticians have seen this behavior before in the university setting, where various degree programs (business, psychology, etc.) offer their own upper-level statistical methods course, instead of requiring a course from the offerings of the university's statistics program. If this becomes commonly accepted in data science, then I feel this will have positive impact on the statistics community. There will always be a need for true statistical experts, who understand the theory and can make the appropriate adjustments to methods when required. My belief is that data science is a team sport, and that no one individual will have all the skills necessary for all data science problems.

5. What can (and should) our academic institutions do to help keep young statisticians competitive in this growing arena? What can (and should) our professional organizations do to provide advocacy for statisticians? Are our potential perceptions and biases holding us back from making progress?

5.1 Randy Bartlett

Train the applied statisticians for the field—this means more methodology and more software training; and possibly professional skills like leadership, project management, organization, et al. (can be at ASA). Open up to feedback; survey former students. Rethink your publication-centric incentive structure. There is something very wrong with fields like physics, for which most of it graduates can not find a job in physics.

ASA should survey applied statisticians about how to enhance its value proposition to meet their 'non-publishing' needs. They are the right place for continuing education, especially professional skills. Adding the PSTAT was a huge step forward.

I second Fred and Mark's ideas below.

5.2 Fred Hulting

I will defer to my academic colleagues on the education component, particularly for our new graduates.

But I do want to comment on the continuing education of the mid- to late-career workforce. In an environment of change, many practitioners are facing the need to acquire new skills to stay competitive in their companies. They are being asked to do new types of work; or the are being directly asked to become "data scientists." Statistical societies and academic institutions need to consider how they can deliver the educational approaches, and certifications, that this group will require.

There is also a role for our industry practitioners to work with corporate leaders to advocate for the "big tent" approach to analytics work and analytics professionals, focusing on building teams with diverse skill sets who can collaborate to deliver the results that are needed to drive results.

Finally, while there are indeed many challenges for statisticians facing tremendous change, there is also great opportunity. The growing demand for individuals that can solve problems and drive innovation through the thoughtful use of data and analysis is a positive change. Whether we call it "analytics, statistics," or "data science," it is an exciting time for our profession!

5.3 Mark Lancaster

I think that colleges/universities can keep young statisticians competitive by organizing data analysis challenges like ASA's DataFest http://ww2.amstat.org/education/datafest . They should also keep current on the newest computing methods and software that the industry uses to work with large data sets. It used to be that a knowledge of Excel and/or SAS was enough, but now industry wants statisticians and data scientists with exposure to R, Python, SQL/noSQL, RapidMiner, Tableau, etc. The local ASA chapters can also engage young statisticians by inviting practicing statisticians from regional industries and universities for seminars.

Acknowledgements

The authors would like to thank our other panelists, Andrew Ekstrom and Edward Boone. In addition, special thanks to the ASA Section on Statistical Consulting for Sponsoring the Panel.

References

Bartlett, Randy, Statistical Denial Series, http://tinyurl.com/ybbwfpan

Bartlett, Randy. 2013. A Practitioner's Guide To Business Analytics. McGraw-Hill. http://amzn.to/YGhXzv

Gray, Kevin, 'What the Heck is "Data Science" Anyway?,' https://www.linkedin.com/pulse /what-heck-data-science-anyway -kevin-gray

Mitchell-Guthrie, Polly, 'Analytics, OR, data science and machine learning: what's in a name?,' http://blogs.sas.com/content/s ubconsciousmusings/2016/05/31/ data-science-and-machine-learn ing/

Luker Jr, Bill. BIG DATA IS DEAD. LONG LIVE BIG DATA.
March 7, 2013, http://analyticscluster.com/?p=511

Speidel, Thomas. 'Time to Embrace a New Identity?,' Amstat News http://magazine.amstat.org/blo g/2014/10/01/statview-oct14/