ASA *Amstat News* ASA Community The World of Statistics STAT*tr@k*

STATS.org This is Statistics

AMSTATNEWS

The Membership Magazine of the American Statistical Association

- Home
- About
  - Submission Instructions
- Editorial Calendar
- PDF Archives
  - 2008 Amstat News
  - 2009 Amstat News
  - 2010 Amstat News
  - 2011 Amstat News
  - 2012 Amstat News
  - 2013 Amstat News
  - 2014 Amstat News
  - 2015 Amstat News
  - 2016 Amstat News
  - 2017 Amstat News
  - 2018 Amstat News
- Advertise
- Statisticians in History
  - Celebrating Women in Statistics

Search the archive...

Home » A Statistician's View, Departments

Data Science: The Evolution or the Extinction of Statistics?

1 January 2016 3,789 views 5 Comments

Jennifer Lewis Priestley is a professor of applied statistics and data science at Kennesaw State University, where she is the director of the Center for Statistics and Analytical Services. She oversees the PhD program in advanced analytics and data science and teaches courses in applied statistics at the undergraduate, master's, and PhD levels.

As a discipline, I think statistics—and by association statisticians—are going through a midlife crisis. Just look around a typical university. Where is statistics housed? Mathematics? The business school? Engineering? Humanities? All of the above? Who are we? This crisis of identity has been accelerated by this new term "data science." Is it a discipline? Is it an application of statistics? Is it an application of computer science? Is it a buzz word just having its moment?

I agree with Tommy Jones in his *Amstat News* article, "The Identity of Statistics in Data Science," when he says the "…conversation around data science betrays an anxiety about our identity."

As the director of one of the country's first PhD programs in data science and a professor of statistics, I believe data science is the full-length mirror we have needed to hold up in front of our discipline for a long time so we can examine how we look from multiple angles. As any middle-aged woman will tell you, full-length mirrors contribute to anxiety.

As we turn in front of this mirror, there are angles that are not working for us. Theoretical statistics is increasingly a bastion of academia. While there will always be a need for PhD theoretical statisticians in universities, a BS in theoretical statistics—defined by derivations of theorems and execution of formulas completely by hand with no experience with real data—does not prepare undergraduates to work in a 21st-century economy. And, as most theoretical statisticians will tell you, if someone does want to pursue a PhD in statistics, they are better served pursuing an undergraduate degree in mathematics.

The other side of this angle is “business statistics” (“statistics-for-students-who-could-not-handle-the-math-in-real-statistics” in most universities), where students work with Excel spreadsheets characterized by 100 rows and three columns and they generate means and standard deviations—and in the advanced course, pivot tables. These courses also do a huge disservice to students and, similarly, do not prepare graduates to work in a 21st-century economy.

The 100% theoretical approach to statistics and the “statistics lite” approach are both bad for our discipline because of the same issue—data. Neither approach prepares students to work with real-world data. If you scan typical job advertisements, any position hiring a statistician will likely include required skills such as programming, analytical software experience (e.g., SAS), database management (e.g., SQL), and writing and communication. This is because the days of being a “data diva” are over—statisticians are expected in most companies to have some ability to extract, transport, load, clean, analyze, model, and “tell the story” of their results. This is particularly true in small companies. And even if they don’t have to do all points in the chain for every project, developing a working knowledge of how data are collected, stored, extracted, cleaned … makes for better models … and more complete communication of results.

But as we pivot in the mirror, data science is also allowing us to show off angles of our discipline that are sorely needed—by everyone. At my university, we have an MS in applied statistics. We will have companies from diverse disciplines such as health care, retail, finance, and energy recruiting the same student. Why would companies from such different domains be interested in the same student? Because they are all trying to solve the same problem! They are all trying to translate massive amounts of data into meaningful information to solve a problem and then explain the solution to their boss or their client.

This set of requirements is almost ubiquitous—and it’s certainly multidisciplinary. I think it’s a point of evolution for our discipline and has become the definition of the 21st-century statistician—converting data into information to solve problems or discover patterns and then telling the story. More than any other academic discipline, statistics (applied statistics) is needed by every other discipline. To use a dated phrase (we are in midlife after all), “our dance card is full.”

Again, a brief example from my own university. We have an undergraduate minor (not a BS) in applied statistics. This minor requires students to take five elective courses in applied statistics. This minor is not required for any undergraduate on campus. And yet, in any given semester, we have well more than 100 undergraduate students who have declared a formal minor in applied statistics. These students come from all the colleges across campus and from dozens of departments. We have biology majors sitting next to sociology majors sitting next to finance majors all solving the same problems. It’s the most popular minor field of study in the history of the university. Again—multidisciplinary. All of a sudden, everyone wants to study applied statistics.

So what about data science? Who are these people, and how are they different from us?

The definitions of data science are converging around the intersection of mathematics, statistics, and computer science—with some area of application (e.g., finance, biology, political science). I have heard data scientists referred to equally as “the computer scientist who was the best of his peers in his statistics courses” and “the statistician who was the best of his peers in his computer science courses.”

I referenced that I am an applied statistician running a PhD program in analytics and data science. While data scientists can do a great many things I can’t do—mainly in the areas of coding, API development, web scraping, and machine learning—they would be hard pressed to compete with a PhD student in statistics in supervised modeling techniques or variable reduction methods. Earlier this year, an article on the Simply Statistics blog, “Why Big Data Is in Trouble: They Forgot About Applied Statistics,” highlighted the issue of how a rush to the excitement of machine learning, text mining, and neural networks missed the importance of basic statistical concepts regarding the behavior of data—including variation, confidence, and distributions. Which lead to bad decisions.

So, where does this leave us statisticians? I believe data science is good for us. In fact, it’s great for us. People need us in new and exciting ways—to help them translate the data into information to tell a story. The “science of data” is becoming a nascent discipline that is lifting all boats. That nascent scientific discipline needs us.

5 Comments »

★★★★☆ (**1** votes, average: **4.00** out of 5)

- ***Vincent Granville*** *said:*

  Read my article on a combinatorial fast, efficient algorithm for feature selection using predictive power to jointly select variables: it is the data science approach to variable reduction and generation. Likewise, supervised modeling – which also belongs to machine learning – is not foreign to data scientists. Read about my automated indexation / tagging algorithm, used for taxonomy creation or cataloguing: it performs clustering on n data points in O(n), and can cluster billions of web pages in very little time. It is also used to turn unstructured data into structured data.