July 15, 2015

Data Science

What Is It and How Is It Taught?

By Hans De Sterck, Chris Johnson

Effectively combining computational and data science, CSE15 invited speaker Anna Michalak provided surprising statistics on greenhouse gases, derived in part from an innovative study in which Argentinian researchers used plastic tanks strapped to the backs of cows to estimate methane production.
REUTERS/Marcos Brindicci

The term "big data" has become ubiquitous. People who can wrangle big data are called data scientists. According to a number of sources, there is a growing need for people trained as data scientists. But what is data science? Is data science its own field, or is it an interdisciplinary mix of computer science, mathematics and statistics, and domain knowledge? Or might it really be what statisticians have been doing all along? Because data science at scale involves large-scale computation, what is the relation between data science and computational science?

At the 2015 SIAM Conference on Computational Science and Engineering, we convened what turned out to be a very lively panel* to discuss the current and future status of data science, its relation to computational science, opportunities for data and computational scientists, and the education of future data scientists.

Here are some surprising things we heard from the panelists:

- data science is not statistics;
- data science should be taught by computer scientists;
- in five years, every domain of science and engineering will center on data science;
- in ten years, all of data science will be applying machine learning; and
- data science is not new—it's just the other side of the coin of computational science and engineering.

In the ensuing discussion with the audience, every statement made by a panelist was countered by at least one opposing response, and we heard many additional ideas about data science. For example: Although machine learning and data mining are often taught in computer science departments, many in the audience pointed to the mathematical and statistical foundations and developments that underlie progress in these fields. So statisticians and mathematicians will certainly also teach data science!

Classical Science versus Modern Data Science: A Clash of Paradigms?

One of the panelists characterized modern data science as a paradigm in which black-box statistics-based models from data mining and machine learning are unleashed on large data sets to analyze and predict complex phenomena. Presumably, this approach implicitly uncovers the "rules" that govern the behavior of the complex systems. It is undeniable that significant (and sometimes spectacular) progress is made via machine learning approaches in several areas of investigation: image and speech recognition, automatic translation, fraud detection, online recommendation, and business analytics for the retail sector, among others. This work exploits the availability of unprecedented volumes of data in areas in which quantitative data and models were virtually nonexistent until very recently.

The situation is different in much of science and engineering, however. In these fields, there is a crucial role for models based on first principles that have been tremendously successful in modeling physical reality. Experimental data has always been essential for validating these models. But the tremendous growth in the volume of scientific data from such sources as high-bandwidth experiments and observations, extreme-scale simulations, and large networks of sensors opens exciting new possibilities. The availability of this data is enabling great progress in quantifying uncertainties in the results produced by physics-based models, and in making these models truly predictive [1].

Incorporating vast amounts of data into scientific applications often relies on statistical techniques, including, for example, Bayesian inference of model parameters and stochastic simulation, which have become essential tools in computational science. Similarly, visual analytics techniques developed for science and engineering often have broad applicability in other areas of data science [4]. The development of all these techniques, including parallel algorithms and implementations that are efficient at scale, constitute key contributions to data science, enabling data-driven scientific discovery. Models based on first principles are essential components of systems that extract valuable insights from massive scientific data, insights that tend to go far beyond what can be recovered by black-box statistical modeling alone.

The Synergy Between Computational Science and Data Science

Given that data science and computational science overlap significantly, both in the expertise required and in the methodologies used, one panelist claimed that data science is just the other side of the coin of computational science: Computational science and data science are both rooted in solid foundations of mathematics and statistics, computer science, and domain knowledge, and this common core can be exploited in educational programs that will prepare the computational and data scientists of the future [2]. Indeed, many computational science competencies translate directly to the analysis of massive data sets at scale with high-end computing infrastructure. As complementary interdisciplinary endeavors, both data science and computational science suffer from the entrapments created by disciplinary boundaries. To provide rigorous, multifaceted educational preparation for the growing ranks of computational and data scientists needed to optimally advance scientific discovery and technological development in the years to come, universities will need to implement new multidisciplinary structures.

Data Science Research at CSE15: Estimating Greenhouse Gas Emissions from Human Activity

As to the synergy between data science and model-based computational science, Anna Michalak of the Carnegie Institute and Stanford provided a case in point in her invited presentation, “Statistical and Computational Challenges of Constraining Greenhouse Gas Budgets.” The main question she addressed was how we can properly quantify the amount of greenhouse gases released in the atmosphere as a result of human activity [3]. Self-reporting by local governments and tracking of inventories, she pointed out, have been shown to lead to inaccurate estimates.

Michalak described how continuous measurements of atmospheric concentrations at a large set of locations can be used to estimate surface fluxes through the solution of large stochastic inverse problems. In data for 2013, 10^6 fluxes were estimated from 10^5 observations, resulting in large computational problems with dense 10^6 x 10^6 matrices. She showed how efficient algorithms and large-scale implementations for matrix multiplication and posterior covariance make it possible to solve such problems. Citing studies in which extensive data and physics-based models were combined with advanced mathematical and statistical algorithms and large-scale computing, she revealed that U.S. anthropogenic methane emissions were actually 50% higher than EPA estimates. Emissions from cattle were shown to be nearly double what inventories suggest, and oil and gas emissions five times as high as data reported in the reference international database on global greenhouse gas emissions [3]. A great example of computational and data science at work!

Data Science Education

So how do we teach data science? One of the panelists described graduate programs at the University of Utah that include a Big Data Certificate and MS and PhD programs in Computing with a Data Management and Analysis track [5]. These programs, which are centered in the Utah School of Computing, focus on fundamentals that include databases, algorithms, data mining, machine learning, statistics, and visualization.

Among other models is a new undergraduate program at Virginia Tech, Computational Modeling and Data Analytics [6], which is organized by an interdisciplinary group of faculty with majority representation from mathematics and statistics. This genuinely new degree mainly comprises new classes, starting with a year-long 12-credit course called Integrated Quantitative Sciences that is team taught by a mathematician and a statistician. Covering topics from multivariable calculus, differential equations, linear algebra, and basic probability and statistics, this course is intended to provide a solid foundation for an education in data science. In a sense, creating such a program offers the opportunity to rethink curricula on classical topics like calculus that have at many institutions not seen substantial change throughout most of the past century. It is a significant investment, but it appears to pay off—the first year the program was offered, with minimal advertising, almost 90 freshman applicants at Virginia Tech specified it as their first-choice major, and more than 200 as their second choice.

Another interesting example of a data science program is the Data Engineering and Science Initiative at Georgia Tech [7]. Degree programs offered include a one-year MS in analytics, and MS and PhD programs with a data focus in CSE and biotech fields. The MS in analytics is offered jointly by the School of Computational Science and Engineering (College of Computing), the School of Industrial & Systems Engineering (College of Engineering), and the Scheller College of Business. About a quarter of the extensive set of courses are offered by the School of CSE, with the focus on computational algorithms and high-performance analytics.

The Utah, Virginia Tech, and Georgia Tech programs are just three of a quickly growing number of data science-related programs in the U.S. An overview of almost 100 U.S. master's programs in analytics and data science (when this article was written) can be found in [8]. Enrollment began in 2014 or 2015 for more than half of these programs. Not surprisingly, due to the recent successes of business intelligence applications in many sectors of the economy, including retail and banking, more than 40 business analytics programs hosted by business schools appear on the list. Many of these are professional degrees (estimated tuition costs are also listed in [8]) with a more applied focus, but business analytics is quickly evolving from its roots in data warehousing and standard statistical methods to more sophisticated approaches in terms of algorithms and computation. About 15 of the other programs in [8] are hosted by colleges of arts and sciences, another 15 by engineering and computer science colleges, and, reflecting the interdisciplinary nature of data science, an additional 15 are organized by interdisciplinary data-focused institutes or consortia of colleges. The picture on the international scene is similar: For example, 51 programs in data science in Europe (about 20 of them in the UK), mostly at the master's level, are listed in [9].

Data Science and Computational Science: Synergy with a Bright Future?

It is clear that the data tsunami is only increasing in intensity and that the current focus on data analytics will not easily fade. The data revolution is shaping up to become one of the great new quantitative endeavors of our time, and, as in all quantitative fields, mathematics is poised to play an important role. The synergy between data science and computational science makes it clear that educational programs in areas like "computational and data science" or "mathematics of data and computation" hold significant promise for interdisciplinary success. Readers wishing to offer thoughts and insights on data science and how to teach it are invited to join the discussion at the companion blogpost [10] to this article.

References

[1] M. Adams et al., *Report of the National Academies Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification*, National Academies Press, Washington, DC, 2012.
[2] J. Chen et al., *Synergistic Challenges in Data-Intensive Science and Exascale Computing*, DOE ASCAC Data Subcommittee Report, Office of Science, Department of Energy, 2013.
[3] S.M. Miller et al., *Anthropogenic emissions of methane in the United States*, Proc. Natl. Acad. Sci., 110.50 (2013): 20018–20022.
[4] P.C. Wong et al., *The top 10 challenges in extreme-scale visual analytics*, IEEE Comput. Graph., 32 (2012), 63.

Notes
[5] http://www.cs.utah.edu/bigdata/
[6] http://www.science.vt.edu/ais/cmda/
[7] http://bigdata.gatech.edu/, http://www.analytics.gatech.edu/
[8] http://analytics.ncsu.edu/?page_id=4184
[9] http://www.kdnuggets.com/education/index.html
[10] http://blogs.siam.org/data-science-what-is-it-and-how-to-teach-it/

*The slides from the panel are available at http://www.sci.utah.edu/~chris/Data-Science- Panel-CSE15.

Hans De Sterck is a professor of computational mathematics and scientific computing in the Department of Applied Mathematics at the University of Waterloo. Chris Johnson is director of the Scientific Computing and Imaging Institute and a Distinguished Professor of Computer Science at the University of Utah.