Promoting the Practice and Profession of Statistics®

ASA Statement on *The Role of Statistics in Data Science*

August 8, 2015

The rise of *data science*, including *big data* and *data analytics,* has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) *Database Management* enables transformation, conglomeration, and organization of data resources; (ii) *Statistics and Machine Learning* convert data into knowledge; and (iii) *Distributed and Parallel Systems* provide the computational infrastructure to carry out data analysis.

Certainly, data science intersects with numerous other disciplines and areas of research. Indeed it is difficult to think of an area of science, industry, commerce, or government that is not in some way involved in the data revolution. But it is databases, statistics, and distributed systems that provide the core pipeline. At its most fundamental level, we view data science as a mutually beneficial collaboration among these three professional communities, complemented with significant interactions with numerous related disciplines. For data science to fully realize its potential requires maximum and multifaceted collaboration among these groups.

Statistics and machine learning play a central role in data science. Framing questions statistically allows us to leverage data resources to extract knowledge and obtain better answers. The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes and to quantify uncertainty in their answers. A statistical framework allows researchers to distinguish between causation and correlation and thus to identify interventions that will cause changes in outcomes. It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do all of this using algorithms that exhibit predictable and reproducible behavior. In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources. Simply put, statistical methods allow researchers to accumulate knowledge.

For statisticians to help meet the considerable challenges faced by data scientists requires a sustained and substantial collaborative effort with researchers with expertise in data organization and in the flow and distribution of computation. Statisticians must engage them, learn from them, teach them, and work with them. Engagement must occur at all levels: with individuals, groups of researchers, academic departments, and the profession as a whole. New problem-solving strategies are needed to develop "soup to nuts" pipelines that start with managing raw data and end with user-friendly efficient implementations of principled statistical methods and the communication of substantive results. Statistical education and training must continue to evolve—the next generation of statistical professionals needs a broader skill set and must be more able to engage with database and distributed systems experts. While capacity is increasing within existing and innovative new degree programs, more is needed to meet the massive expected demand. The next generation must include more researchers with skills that cross the traditional boundaries of statistics, databases, and distributed systems; there will be an ever-increasing demand for such "multi-lingual" experts.

Working with statisticians, departments of statistics, and other professional societies, the American Statistical Association (ASA) is well positioned to help formulate discussion around the role of statistics in data science, to navigate the way forward in this quickly evolving environment, and to provide forums for communication and collaboration among data scientists, including statisticians and non-statisticians alike. The ASA aims to facilitate collaboration between statisticians and other data scientists and thus enable them to achieve more than they could on their own.