On Advancing Data Science

Michael L. Brodie
Computer Science and Artificial Intelligence Lab, Database Group
mlbrodie@mit.edu

ABSTRACT

Data Science, a new discovery paradigm, is potentially one of the most significant advances of the early 21st century. Originating in scientific discovery, it is being applied to every human endeavor for which there is adequate data. While remarkable successes have been achieved, far greater claims are made. Risks and challenges abound. The science underlying *data science* has yet to emerge. Maturity is more than a decade away. This claim is based firstly on observing the centuries-long developments of its predecessor paradigms – empirical, theoretical, and Jim Gray's *Fourth Paradigm of Scientific Discovery* [9] (aka data-intensive, computational, procedural, eScience); and secondly on my studies of over 100 data science use cases, several data science-based startups, and, on my scientific advisory role for two Data Science Research Institutes (DSRIs) [10][11] that requires understanding the opportunities, state of the art, and research challenges for the emerging discipline of data science. Essential questions for a DSRI are: *What is data science*? *What is world-class data science research*? And *How should data science be developed*?

1. The Emergence of Data Science

Data science activities have emerged in research labs in most universities and national research labs. For example, many Harvard University departments had one or more groups conducting data science research and offered multiple data science degrees and certificates. In March 2017, the Harvard Data Science initiative was established to coordinate these activities. This pattern has repeated at over 100 major universities worldwide, resulting in over 100 Data Science Research Institutes (DSRIs) being established since 2015 – themselves just emerging. The creation of 100+ DSRIs in approximately two years, all heavily funded by governments and by partner industrial organizations, indicates strength in the belief of the potential of data science not just as a new discovery paradigm, but as a basis for scientific, humanistic, business and economic advances.

2. The Importance of Collaboration

Collaboration is an emerging challenge in data science not only at the scientific level but also at the strategic and organizational levels. Analysts report that most early industry big data deployments failed due to a lack of domain-business-analytics-IT collaboration[8]. Most of the 100+ DSRIs involve combining departments or groups, each pursing data science in their domain, into a higher level DSRI. A rather large example is the Fraunhofer Big Data Alliance, which might be termed a DSRI of DSRIs, that describes itself as: "The Fraunhofer Big Data Alliance consists of 30 institutes bundling their cross-sector competencies. Their expertise ranges from market-oriented big data solutions for individual problems to the professional education of data scientists and big data specialists."

A DSRI should strive for higher-level, scientific and strategic goals, such as contributing to data science (i.e., the science underlying data science) in contrast with contributions made in specific domains by partner groups. A DSRI must be more than the sum of the parts. To do so, how should a DSRI operate or be organized so as to encourage collaboration and achieve higher-level goals.

While data science is inherently multi-disciplinary, hence collaborative, in nature, most data scientists and practitioners lack training in collaboration and are motivated to focus on domain objectives. Why would a bioinformaticist (bioinformatician) attempt to establish a data science method that goes beyond her requirements, especially as it requires a deep understanding of many other domains such as deep learning? Collaboration is also a significant organizational challenge specifically for the 100+ DSRIs that were formed as a federation of organizations each of which conduct data science in different domains. Like the bioinformaticist, each organization has its own objectives, budget, and investments in funding and intellectual property. In such an environment, how does a DSRI establish strategic directions and set research objectives? Through a DSRI Chief Scientific Officer [6]?

3. The State of Data Science

Data science is following the pattern of emergence of new domains that involve multiple disciplines, as did computer science. Computer science emerged as a discipline in the early 1940's [4] and in academia in the early 1960's. The first Department of Computer Sciences in the United States was established at Purdue University in October 1962. At that time computer science departments were being established in most universities. Computer science

involves many disciplines, principally mathematics and engineering and the initial application domains, the physical and natural sciences: physics, chemistry, biology, geology, seismology, astronomy, oceanography, and meteorology. These domains raised the first "*grand challenge* problems that demand massive high-speed computations, performed on new generations of massively parallel computers with new kinds of algorithms."[4] As a result, academic computer science departments emerged from different departments in different universities: mathematics, engineering, and the natural sciences. Initially, each department focused on computing as applied to or in its discipline, hence depending on its origins, computer science in a given department was an extension of the original discipline. As will be the case in data science, there was lots of "science" underlying the application of mathematics in computing, of engineering in computing, and of each natural science in computing; otherwise there was little "science" in "computer science"; that is, there were few computer science principles and techniques (models, methods, infrastructure) that were applicable to all domains of computing. Over the 1970s and 1980, computer science emerged as a distinct discipline, as seen in the establishment of computer science departments independent of the original mathematics, engineering, and science departments. As with data science, computer science draws on many disciplines, hence might be considered a multi-discipline that requires interdisciplinary collaboration.

Since its establishment as an independent discipline, computer science has continually evolved as its principles and techniques have advanced together with application to new domains. By the late 1990's, computer science was fundamental in library science; management science; economics; medicine and biology; psychology, cognitive, and behavioral sciences; linguistics, philosophy, and the humanities. In the early 2000's, a new scientific discovery paradigm, labelled *data science*, emerged from the application of existing and new analytical techniques to massive amounts of data in specific scientific domains using massive computing power. By 2010, these techniques were applied to every human endeavor for which data was available.

Data science is at the same evolutionary stage as was computer science in the early 1960s. Departments of data science with degrees or certificates have been established in academia worldwide over the last few years. The current accelerated rates of change, massive expectations, and marketing and profit motives of industry, analysts, and academia, have gone beyond mere departments to DSRIs with large government and industry endowments. Most of the 100+ DSRIs in major universities and industries worldwide are less than two years old. Yet, like computer science in the 1960's, the science of data science has yet to emerge. A qualitative difference from 1960's computer science is that data science is being applied to all human activities, and explicitly to social, political, and economic issues that directly involve ethics and have the potential of social, political, and economic disruption. While data science offers the potential is accelerating discovery of potential solutions to cancer and global warming, it's potential power can be and has already been used negatively to "nudge" major elections[2] and to control society[1][3].

The power and potential of Big Data and Data Science adds urgency to the question of advancing data science. Data scientists and DSRIs should address the questions *What is Data Science*?[7] and *How might the discipline of data science be developed*?[6]

4. REFERENCES

[1] Big data and government: China's digital dictatorship – Worrying experiments with a new form of social control, The Economist, Dec 17th 2016

[2] Big data and the Future of Democracy - The Matrix world behind the Brexit and the US Elections, Eurasia Diary, Sept. 2, 2017. Original: Hannes Grassegger, Mikael Krogerus, "Ich habe nur gezeigt, dass es die Bombe gibt." Das Magazin, Dec. 2016 (Switzerland); reprinted, The Data That Turned the World Upside Down, Motherboard, January 2017

[3] Big data, meet Big Brother: China invents the digital totalitarian state - The worrying implications of its social-credit project, The Economist, Dec 17th 2016

[4] Computer science: the discipline by Peter J. Denning, 1999. This article attempts to define computer science as a scientific discipline, its principal subdivisions, and the relationship of computer science to other disciplines.

[5] M. Braschler, T. Stadelmann, K. Stockinger (Eds.), "Applied Data Science - Lessons Learned for the Data-Driven Business", Berlin, Heidelberg: Springer, expected 2018

[6] M.L. Brodie, Necessity is the Mother of Invention: On Developing Data Science, to appear in [5]

[7] M.L. Brodie, What is Data Science? to appear in [5]

[8] Predictions 2016: The Path from Data to Action for Marketers: How Marketers Will Elevate Systems of Insight. Forrester Research, November 9, 2015.

[9] The Fourth Paradigm: Data-Intensive Scientific Discovery Edited by Tony Hey, Stewart Tansley, and Kristin Tolle, Microsoft Research, 2009

[10] Insight Center for Data Analytics, Ireland.

[11] Swinburne Data Science Research Institute, Melbourne, Australia