What is data science

Introduction

What can data science do? What characteristics distinguish data science from previous scientific discovery paradigms? What are the methods for conducting data science? What is the impact of data science? This chapter offers initial answers to these and related questions. A companion chapter (Brodie, 2018b) addresses the development of data science as a discipline, as a methodology, as well as data science research and education. Let's start with some slightly provocative claims concerning data science.

Data science has been used successfully to accelerate discovery of probabilistic outcomes in many domains. Piketty's (2014) monumental result on wealth and income inequality was achieved through data science. It used over 120 years of sporadic, incomplete, observational economic data, collected over ten years from all over the world (Brodie, 2014b). What is now called computational economics was used to establish the correlation, with a very high likelihood (0.90), that wealth gained from labor could never keep up with wealth gained from assets. What made front page news worldwide was a second, more dramatic correlation that there is a perpetual and growing wealth gap between the rich and the poor. This second correlation was not derived by data analysis but is a human interpretation of Piketty's data analytic result. It contributed to making Capital in the 21st Century the best-selling book on economics, but possibly the least read. Within a year, the core result was verified by independent analyses to a far greater likelihood (0.99). One might expect that further confirmation of Piketty's finding would be newsworthy; however, it was not as the more dramatic rich-poor correlation, while never analytically established had far greater appeal. This illustrates the benefits and risks of data science.

Frequently, due to the lack of evidence, economic theories fail. Matthew Weinzierl, a leading Harvard University economist, questions such economic modelling in general saying, "that the world is too complicated to be modelled with anything like perfect accuracy" and "Used in isolation, however, it can lead to trouble" (Economist, February 2018). Reputedly, Einstein said: "Not everything that counts can be counted. Not everything that's counted, counts". The hope is that data science and computational economics will provide theories that are fact-based rather than based on hypotheses of "expert" economists (Economist, January 2018) leading to demonstrably provable economic theories, i.e., what really happened or will happen. This chapter suggests that this hope will not be realized this year.

Many such outcomes2 have led to verified results through methods outside data science. Most current data analyses are domain specific, many even specific to classes of models, classes of analytical methods, and specific pipelines. Few data science methods have been generalized outside their original domains of application, let alone to all domains (to illustrated in a moment). A rare and excellent exception is a generic scientific discovery method over scientific corpora (Nagarajan et. al.,

2015) generalized from a specific method over medical corpora developed for drug discovery (Spangler et. al., 2014) that is detailed later in the chapter.

It is often claimed that data science will transform conventional disciplines. While transformations are underway in many areas, including supply chain management3 (Waller and Fawcett, 2013) and chemical engineering (Data Science, 2018), only time and concrete results will tell the extent and value of the transformations. The companion chapter On Developing Data Science (Brodie, 2018b) discusses with the transformation myth.

While there is much science in many domain-specific data science activities, there is little fundamental science that is applicable across domains. To warrant the designation data science, this emerging paradigm requires fundamental principles and techniques applicable to all relevant domains, just as the scientific principles of the scientific method apply across many domains. Since most data science work is domain specific, often model- and method-specific, data science does not yet warrant the designation as a science.

This chapter explores the current nature of data science, its qualitative differences with its predecessor scientific discovery paradigms, its core value and components that, when mature, would warrant the designation data science. Descriptions of large-scale data science activities referenced in this chapter apply, scaled down, to data science activities of all sizes, including increasingly ubiquitous desktop data analytics in business.