DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING

Abstract :

Companies have realized they need to hire data scientists, academic institutions are scrambling to put together data science programs, and publications are touting data science as a hot—even ‘‘sexy’’—career choice. However, there is confusion about what exactly data science is, and this confusion could lead to disillusionment as the concept diffuses into meaningless buzz. In this article, we argue that there are good reasons why it has been hard to pin down exactly what is data science.

One reason is that data science is intricately intertwined with other important concepts also of growing importance, such as big data and data-driven decision making. Another reason is the natural tendency to associate what a practitioner does with the definition of the practitioner’s field; this can result in overlooking the fundamentals of the field. We believe that trying to define the boundaries of data science precisely is not of the utmost importance.

We can debate the boundaries of the field in an academic setting, but in order for data science to serve business effectively, it is important (i) to understand its relationships to other important related concepts, and (ii) to begin to identify the fundamental principles underlying data science. Once we embrace (ii), we can much better understand and explain exactly what data science has to offer. Furthermore, only once we embrace (ii) should we be comfortable calling it data science. In this article, we present a perspective that addresses all these concepts. We close by offering, as examples, a partial list of fundamental principles underlying data science.

Introduction:

With vast amounts of data now available, companies in almost every industry are focused on exploiting data for competitive advantage. The volume and variety of data have far outstripped the capacity of manual analysis, and in some cases have exceeded the capacity of conventional databases. At the same time, computers have become far more powerful, networking is ubiquitous, and algorithms have been developed that can connect datasets to enable broader and deeper analyses than previously possible. The convergence of these phenomena has given rise to the increasingly widespread business application of data science. Companies across industries have realized that they need to hire more data scientists. Academic institutions are scrambling to put together programs to train data scientists. Publications are touting data science as a hot career choice and even “sexy.”[1] However, there is confusion about what exactly is data science, and this confusion could well lead to disillusionment as the concept diffuses into meaningless buzz. In this article, we argue that there are good reasons why it has been hard to pin down what exactly is data science. One reason is that data science is intricately intertwined with other important concepts, like big data and data-driven decision making, which are also growing in importance and attention. Another reason is the natural tendency, in the absence of academic programs to teach one otherwise, to associate what a practitioner actually does with the definition of the practitioner’s field; this can result in overlooking the fundamentals of the field. At the moment, trying to define the boundaries of data science precisely is not of foremost importance. Data-science academic programs are being developed, and in an academic setting we can debate its boundaries. However, in order for data science to serve business effectively, it is important (i) to understand its relationships to these other important and closely related concepts, and (ii) to begin to understand what are the fundamental principles underlying data science. Once we embrace (ii), we can much better understand and explain exactly what data science has to offer. Furthermore, only once we embrace (ii) should we be comfortable calling it data science.

In this article, we present a perspective that addresses all these concepts. We first work to disentangle this set of closely interrelated concepts. In the process, we highlight data science as the connective tissue between data-processing technologies (including those for “big data”) and data-driven decision making. We discuss the complicated issue of data science as a field versus data science as a profession. Finally, we offer as examples a list of some fundamental principles underlying data science.

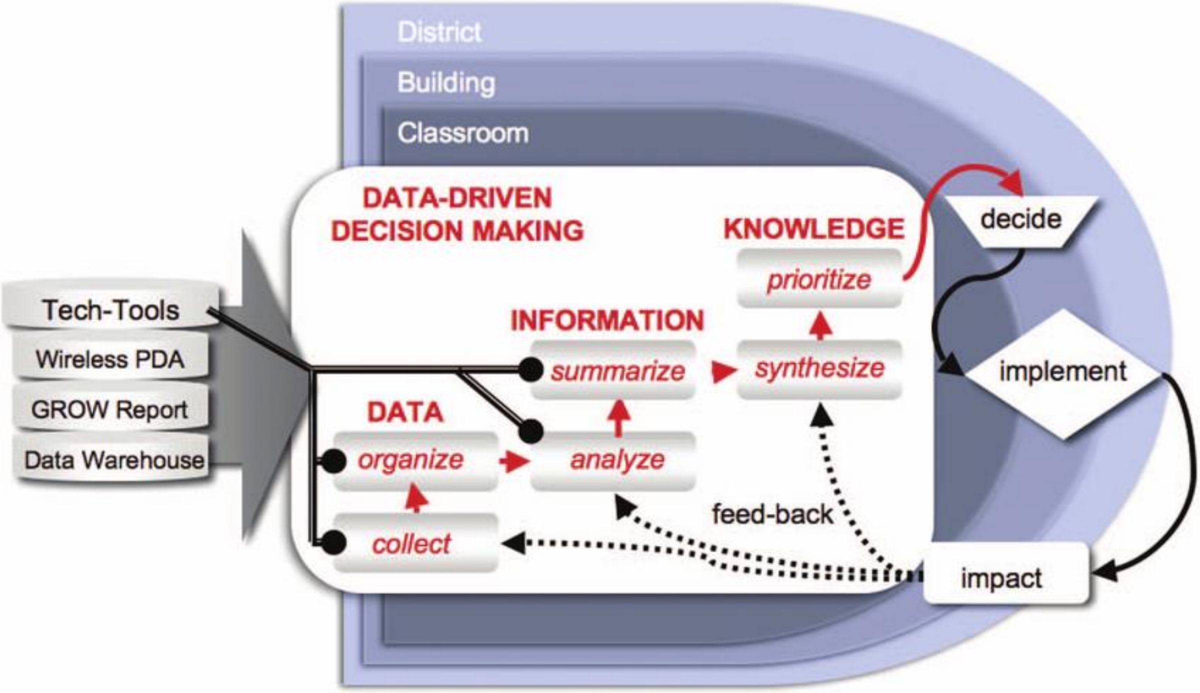
Data Science :

At a high level, data science is a set of fundamental principles that support and guide the principled extraction of infor   mation and knowledge from data. Possibly the most closely related concept to data science is data mining—the actual extraction of knowledge from data via technologies that in   corporate these principles. There are hundreds of different data-mining algorithms, and a great deal of detail to the methods of the field. We argue that underlying all these many details is a much smaller and more concise set of fundamental principles. These principles and techniques are applied broadly across functional areas in business. Probably the broadest business applications are in marketing for tasks such as targeted marketing, online advertising, and recommendations for cross-selling. Data science also is applied for general customer relationship management to analyze customer behavior in order to manage attrition and maximize expected customer value. The finance industry uses data science for credit scoring and trading and in operations via fraud detection and work force management.

Major retailers from Wal-Mart to Amazon apply data science throughout their businesses, from mar keting to supply-chain management. Many firms have differ entiated themselves strategically with data science, sometimes to the point of evolving into data-mining companies. But data science involves much more than just data-mining algorithms. Successful data scientists must be able to view business problems from a data perspective. There is a fun damental structure to data-analytic thinking, and basic principles that should be understood. Data science draws from many "traditional" fields of study. Fundamental prin ciples of causal analysis must be understood. A large portion of what has traditionally been studied within the field of statistics is fundamental to data science.

Methods and methology for visualizing data are vital. There are also particular areas where intuition, creativity, common sense, and knowledge of a partic ular application must be brought to bear. A data-science perspective provides practitioners with structure and principles, which give the data scientist a framework to systematically treat problems of extracting useful knowledge from data.

Data Science and Data-Driven Decision Making :

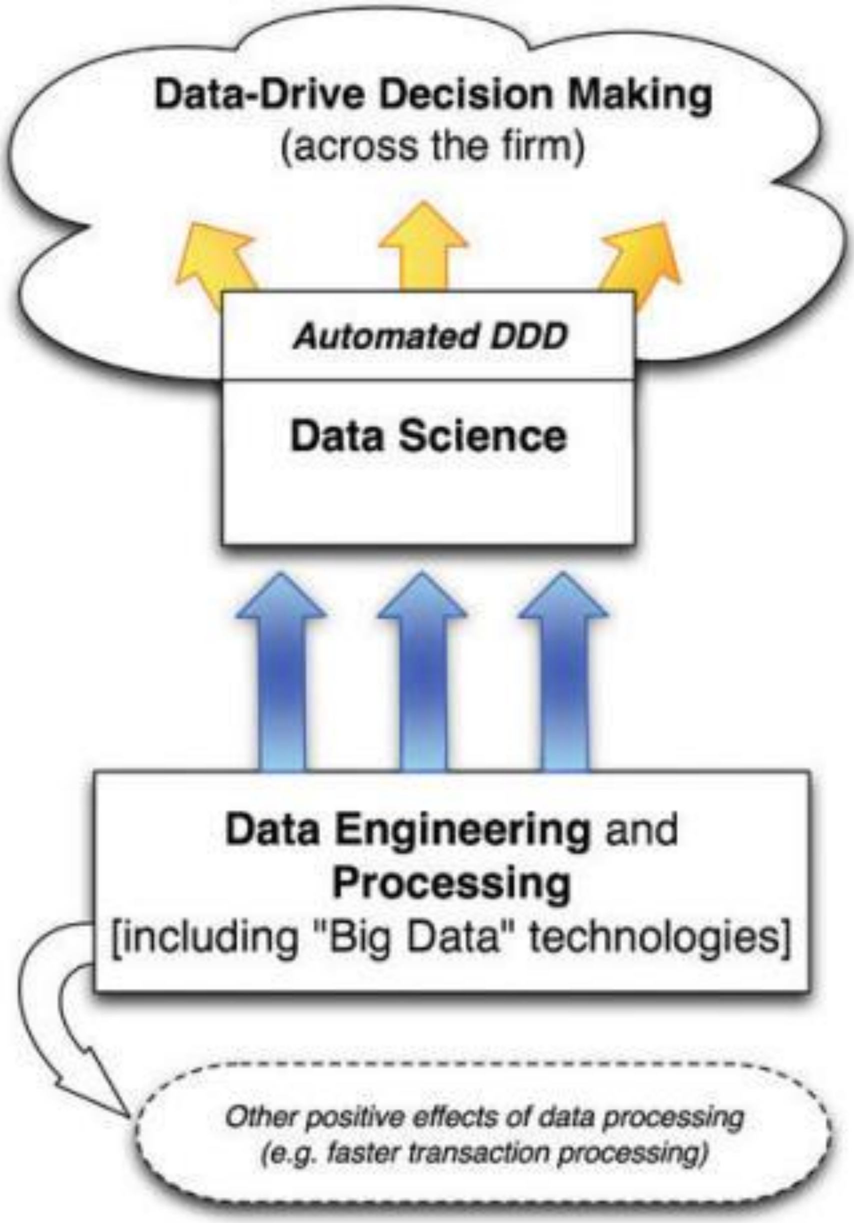


Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data. For the perspective of this article, the ultimate goal of data science is improving decision making, as this generally is of paramount interest to business. Figure 1 places data science in the context of other closely related and data-related processes in the organization. Let's start at the top. Data-driven decision making (DDD)3 refers to the practice of basing decisions on the analysis of data rather than purely on intuition. For example, a marketer could select advertisements based purely on her long experience in the field and her eye for what will work. Or, she could base her selection on the analysis of data regarding how consumers react to different ads. She could also use a combination of these approaches.

DDD is not an all-or nothing practice, and different firms engage in DDD to greater or lesser degrees. The benefits of data-driven decision making have been demonstrated conclusively. Economist Erik Brynjolfsson and his colleagues from MIT and Penn's Wharton School recently conducted a study of how DDD affects firm performance.3 They developed a measure of DDD that rates firms as to how strongly they use data to make decisions across the company.

They show statistically that the more data-driven a firm is, the more productive it is—even controlling for a wide range of possible confounding factors. And the differences are not small: one standard deviation higher on the DDD scale is associated with a 4–6% increase in productivity. DDD also is correlated with higher return on assets, return on equity, asset utilization, and market value, and the relationship seems to be causal**.** Our two example case studies illustrate two different sorts of decisions: (1) decisions for which "discoveries" need to be made within data, and (2) decisions that repeat, especially at massive scale, and so decision making can benefit from even small increases in accuracy based on data analysis. The Wal Mart example above illustrates a type-1 problem. Linda Dillman would like to discover knowledge that will help Wal Mart prepare for Hurricane Frances's imminent arrival. Our churn example illustrates a type 2 DDD problem. A large telecommunications company may have hundreds of millions of customers, each a candidate for defection. Tens of millions of customers have contracts expiring each month, so each one of them has an increased likelihood of defection in the near future. If we can improve our ability to estimate, for a given customer, how profitable it would be for us to focus on her, we can potentially reap large benefits by applying this ability to the millions of customers in the population. This same logic applies to many of the areas where we have seen the most intense application of data science and data mining: direct marketing, online advertising, credit scoring, financial trading, help-desk management, fraud detection, search ranking, product recommendation, and so on.

Data Processing and "Big Data" :


Data-Drive Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and
Processing
[including "Big Data" technologies]
Other positive effects of data processing
(e.g. faster transaction processing)

Despite the impression one might get from the media, there is a lot to data processing that is not data science. Data engineering and processing are critical to support data-science activities, as shown in Figure 1, but they are more general and are useful for much more. Data-processing technologies are important for many business tasks that do not involve extracting knowledge or data-driven decision making, such as efficient transaction processing, modern web system processing, online advertising campaign management, and others.

''Big data'' technologies, such as Hadoop, Hbase, CouchDB, and others have received considerable media attention recently. For this article, we will simply take big data to mean datasets that are too large for traditional data-processing systems and that therefore require new technologies. As with the traditional technologies, big data technologies are used for many tasks, including data engineering. Occasionally, big data technologies are actually used for implementing data mining techniques, but more often the well-known big data technologies are used for data processing in support of the data-mining techniques and other data-science activities, as represented in Figure 1.

Conclusion :

Underlying the extensive collection of techniques for mining data is a much smaller set of fundamental concepts comprising data science. In order for data science to flourish as a field, rather than to drown in the flood of popular attention, we must think beyond the algorithms, techniques, and tools in common use. We must think about the core principles and concepts that underlie the techniques, and also the systematic thinking that fosters success in data-driven decision making. These data science concepts are general and very broadly applicable. Success in today's data-oriented business environment requires being able to think about how these fundamental concepts apply to particular business problems—to think data-analytically. This is aided by conceptual frameworks that themselves are part of data science. For example, the automated extraction of patterns from data is a process with well defined stages. Understanding this process and its stages helps structure problem solving, makes it more systematic, and thus less prone to error. There is strong evidence that business performance can be improved substantially via data-driven decision making,3 big data technologies,4 and data-science techniques based on big data.9,10 Data science supports data-driven decision making—and sometimes allows making decisions automatically at massive scale—and depends upon technologies for "big data" storage and engineering. However, the principles of data science are its own and should be considered and discussed explicitly in order for data science to realize its potential.

References :

1. Davenport T.H., and Patil D.J. Data scientist: the sexiest job of the 21st century. Harv Bus Rev, Oct 2012.

2. Hays C. L. What they know about you. N Y Times, Nov. 14, 2004.

3. Brynjolfsson E., Hitt L.M., and Kim H.H. Strength in numbers: How does data-driven decision making affect firm performance? Working paper, 2011. SSRN working paper. Available at SSRN: http://ssrn.com/abstract = 1819486.

4. Tambe P. Big data know-how and business value. Working paper, NYU Stern School of Business, NY, New York, 2012.

5. Fusfeld A. The digital 100: the world's most valuable startups. Bus Insider. Sep. 23, 2010.

6. Shah S., Horne A., and Capella´ J. Good data won't guarantee good decisions. Harv Bus Rev, Apr 2012.

7. Wirth, R., and Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000, pp. 29–39.

8. Forsythe, Diana E. The construction of work in artificial intelligence. Science, Technology & Human Values, 18(4), 1993, pp. 460–479.

9. Hill, S., Provost, F., and Volinsky, C. Network-based marketing: Identifying likely adopters via consumer networks. Statistical Science, 21(2), 2006, pp. 256–276.

10. Martens D. and Provost F. Pseudo-social network targeting from consumer transaction data