ORIGINAL ARTICLE

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

Foster Provost[1] *and Tom Fawcett*[2]

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

[1]Leonard N. Stern School of Business, New York University, New York, New York.
[2]Data Scientists, LLC, New York, New York and Mountain View, California.

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*.

"PUBLICATIONS ARE TOUTING DATA SCIENCE AS A HOT CAREER CHOICE AND EVEN 'SEXY.'"

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 information 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 incorporate 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 workforce management. Major retailers from Wal-Mart to Amazon apply data science throughout their businesses, from marketing to supply-chain management. Many firms have differentiated 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 fundamental structure to data-analytic thinking, and basic principles that should be understood. Data science draws from many "traditional" fields of study. Fundamental principles 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 particular 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 in Action

For concreteness, let's look at two brief case studies of analyzing data to extract predictive patterns. These studies illustrate different sorts of applications of data science. The first was reported in the *New York Times*:

> Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons…predictive technology.
>
> A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.[2]

Consider *why* data-driven prediction might be useful in this scenario. It might be useful to predict that people in the path

of the hurricane would buy more bottled water. Maybe, but it seems a bit obvious, and why do we need data science to discover this? It might be useful to project the *amount of increase* in sales due to the hurricane, to ensure that local Wal-Marts are properly stocked. Perhaps mining the data could reveal that a particular DVD sold out in the hurricane's path—but maybe it sold out that week at Wal-Marts across the country, not just where the hurricane landing was imminent. The prediction could be somewhat useful, but probably more general than Ms. Dillman was intending.

It would be more valuable to discover patterns due to the hurricane that were not obvious. To do this, analysts might examine the huge volume of Wal-Mart data from prior, similar situations (such as Hurricane Charley earlier in the same season) to identify unusual local demand for products. From such patterns, the company might be able to anticipate unusual demand for products and rush stock to the stores ahead of the hurricane's landfall.

"FROM SUCH PATTERNS, THE COMPANY MIGHT BE ABLE TO ANTICIPATE UNUSUAL DEMAND FOR PRODUCTS AND RUSH STOCK TO THE STORES AHEAD OF THE HURRICANE'S LANDFALL."

Indeed, that is what happened. The *New York Times* reported that: "… the experts mined the data and found that the stores would indeed need certain products—and not just the usual flashlights. 'We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane,' Ms. Dillman said in a recent interview.' And the pre-hurricane top-selling item was beer.*'"[2]

Consider a second, more typical business scenario and how it might be treated from a data perspective. Assume you just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States. They are having a major problem with customer retention in their wireless business. In the mid-Atlantic region, 20% of cell-phone customers leave when their contracts expire, and it is getting increasingly difficult to acquire new customers. Since the cell-phone market is now saturated, the huge growth in the wireless market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own. Customers switching from one company to another is called *churn*, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs.

You have been called in to help understand the problem and to devise a solution. Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to prevent churn. Marketing has already designed a special retention offer. Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts. Specifically, how should MegaTelCo decide on the set of customers to target to best reduce churn for a particular incentive budget? Answering this question is much more complicated than it seems initially.

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

*Of course! What goes better with strawberry Pop-Tarts than a nice cold beer?

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.

"THE BENEFITS OF DATA-DRIVEN DECISION MAKING HAVE BEEN DEMONSTRATED CONCLUSIVELY."

The diagram in Figure 1 shows data science supporting data-driven decision making, but also overlapping with it. This highlights the fact that, increasingly, business decisions are being made automatically by computer systems. Different industries have adopted automatic decision making at different rates. The finance and telecommunications industries were early adopters. In the 1990s, automated decision making changed the banking and consumer-credit industries dramatically. In the 1990s, banks and telecommunications companies also implemented massive-scale systems for managing data-driven fraud control decisions. As retail systems were increasingly computerized, merchandising decisions were automated. Famous examples include Harrah's casinos' reward programs and the automated recommendations of Amazon and Netflix. Currently we are seeing a revolution in advertising, due in large part to a huge increase in the amount of time consumers are spending online and the ability online to make (literally) split-second advertising decisions.

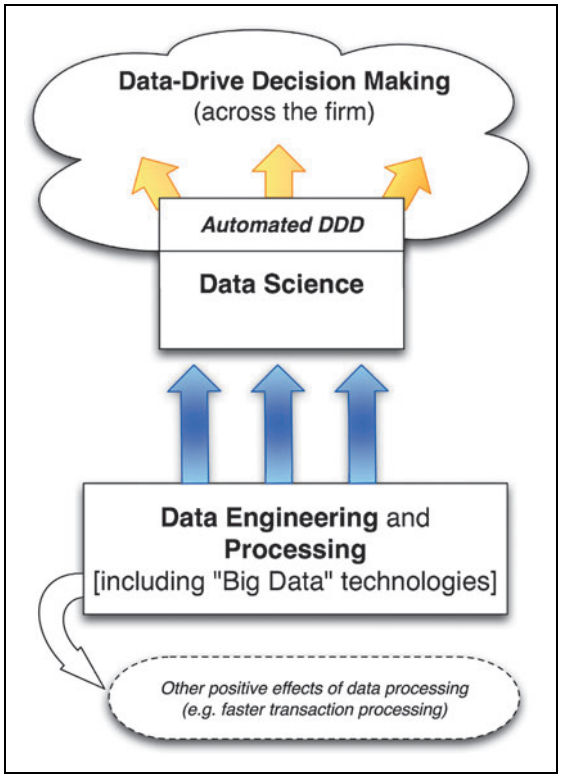

FIG. 1. Data science in the context of closely related processes in the organization.

Data Processing and "Big Data"

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.

Economist Prasanna Tambe of New York University's Stern School has examined the extent to which the utilization of big data technologies seems to help firms.[4] He finds that, after controlling for various possible confounding factors, the use of big data technologies correlates with significant additional productivity growth. Specifically, one standard deviation higher utilization of big data technologies is associated with 1–3% higher productivity than the average firm; one standard deviation lower in terms of big data utilization is associated with 1–3% lower productivity. This leads to potentially very large productivity differences between the firms at the extremes.

From Big Data 1.0 to Big Data 2.0

One way to think about the state of big data technologies is to draw an analogy with the business adoption of internet technologies. In Web 1.0, businesses busied themselves with getting the basic internet technologies in place so that they could establish a web presence, build electronic commerce capability, and improve operating efficiency. We can think of ourselves as being in the era of Big Data 1.0, with firms engaged in building capabilities to process large data. These primarily support their current operations—for example, to make themselves more efficient.

With Web 1.0, once firms had incorporated basic technologies thoroughly (and in the process had driven down prices) they started to look further. They began to ask what the web could do for them, and how it could improve upon what they'd always done. This ushered in the era of Web 2.0, in which new systems and companies started to exploit the interactive nature of the web. The changes brought on by this shift in thinking are extensive and pervasive; the most obvious are the incorporation of social-networking components and the rise of the "voice" of the individual consumer (and citizen).

"SIMILARLY, WE SHOULD EXPECT A BIG DATA 2.0 PHASE TO FOLLOW BIG DATA 1.0 … THIS IS LIKELY TO USHER IN THE GOLDEN ERA OF DATA SCIENCE."

Similarly, we should expect a Big Data 2.0 phase to follow Big Data 1.0. Once firms have become capable of processing massive data in a flexible fashion, they should begin asking: *What can I now do that I couldn't do before, or do better than I could do before?* This is likely to usher in the golden era of data science. The principles and techniques of data science will be applied far more broadly and far more deeply than they are today.

It is important to note that in the Web-1.0 era, some precocious companies began applying Web-2.0 ideas far ahead of the mainstream. Amazon is a prime example, incorporating the consumer's "voice" early on in the rating of products and product reviews (and deeper, in the rating of reviewers). Similarly, we see some companies already applying Big Data 2.0. Amazon again is a company at the forefront, providing data-driven recommendations from massive data. There are other examples as well. Online advertisers must process extremely large volumes of data (billions of ad impressions per day is not unusual) and maintain a very high throughput (real-time bidding systems make decisions in tens of milliseconds). We should look to these and similar industries for signs of advances in big data and data science that subsequently will be adopted by other industries.

Data-Analytic Thinking

One of the most critical aspects of data science is the support of data-analytic thinking. Skill at thinking data-analytically is important not just for the data scientist but throughout the organization. For example, managers and line employees in other functional areas will only get the best from the company's data-science resources if they have some basic understanding of the fundamental principles. Managers in enterprises without substantial data-science resources should still understand basic principles in order to engage consultants on an informed basis. Investors in data-science ventures need to understand the fundamental principles in order to assess investment opportunities accurately. More generally, businesses increasingly are driven by data analytics, and there is great professional advantage in being able to interact competently with and within such businesses. Understanding the fundamental concepts, and having frameworks for organizing data-analytic thinking, not only will allow one to interact competently, but will help to envision opportunities for improving data-driven decision making or to see data-oriented competitive threats.

Firms in many traditional industries are exploiting new and existing data resources for competitive advantage. They employ data-science teams to bring advanced technologies to bear to increase revenue and to decrease costs. In addition, many new companies are being developed with data mining as a key strategic component. Facebook and Twitter, along with many other "Digital 100" companies,[5] have high valuations due primarily to data assets they are committed to capturing or creating.[†] Increasingly, managers need to manage data-analytics teams and data-analysis projects, marketers have to organize and understand data-driven campaigns, venture capitalists must be able to invest wisely in businesses with substantial data assets, and business strategists must be able to devise plans that exploit data.

As a few examples, if a consultant presents a proposal to exploit a data asset to improve your business, you should be able to assess whether the proposal makes sense. If a competitor announces a new data partnership, you should recognize when it may put you at a strategic disadvantage. Or, let's say you take a position with a venture firm and your first project is to assess the potential for investing in an advertising company. The founders present a convincing argument that they will realize significant value from a unique body of data they will collect, and on that basis, are arguing for a substantially higher valuation. Is this reasonable? With an

[†]Of course, this is not a new phenomenon. Amazon and Google are well-established companies that obtain tremendous value from their data assets.

understanding of the fundamentals of data science, you should be able to devise a few probing questions to determine whether their valuation arguments are plausible.

On a scale less grand, but probably more common, data-analytics projects reach into all business units. Employees throughout these units must interact with the data-science team. If these employees do not have a fundamental grounding in the principles of data-analytic thinking, they will not really understand what is happening in the business. This lack of understanding is much more damaging in data-science projects than in other technical projects, because the data science supports improved decision making. Data-science projects require close interaction between the scientists and the business people responsible for the decision making. Firms in which the business people do not understand what the data scientists are doing are at a substantial disadvantage, because they waste time and effort or, worse, because they ultimately make wrong decisions. A recent article in Harvard Business Review concludes: "For all the breathless promises about the return on investment in Big Data, however, companies face a challenge. Investments in analytics can be useless, even harmful, unless employees can incorporate that data into complex decision making."[6]

> "FACEBOOK AND TWITTER, ALONG WITH MANY OTHER 'DIGITAL 100' COMPANIES, HAVE HIGH VALUATIONS DUE PRIMARILY TO DATA ASSETS THEY ARE COMMITTED TO CAPTURING OR CREATING."

Some Fundamental Concepts of Data Science

There is a set of well-studied, fundamental concepts underlying the principled extraction of knowledge from data, with both theoretical and empirical backing. These fundamental concepts of data science are drawn from many fields that study data analytics. Some reflect the relationship between data science and the business problems to be solved. Some reflect the sorts of knowledge discoveries that can be made and are the basis for technical solutions. Others are cautionary and prescriptive. We briefly discuss a few here. This list is not intended to be exhaustive; detailed discussions even of the handful below would fill a book.* The important thing is that we understand these fundamental concepts.

Fundamental concept: *Extracting useful knowledge from data to solve business problems can be treated systematically by following a process with reasonably well-defined stages.* The Cross-Industry Standard Process for Data Mining[7] (CRISP-DM) is one codification of this process. Keeping such a process in mind can structure our thinking about data analytics problems. For example, in actual practice one repeatedly sees analytical "solutions" that are not based on careful analysis of the problem or are not carefully evaluated. Structured thinking about analytics emphasizes these often underappreciated aspects of supporting decision making with data. Such structured thinking also contrasts critical points at which human intuition and creativity is necessary versus points at which high-powered analytical tools can be brought to bear.

Fundamental concept: *Evaluating data-science results requires careful consideration of the context in which they will be used.* Whether knowledge extracted from data will aid in decision making depends critically on the application in question. For our churn-management example, how exactly are we going to use the patterns that are extracted from historical data? More generally, does the pattern lead to better decisions than some reasonable alternative? How well would one have done by chance? How well would one do with a smart "default" alternative? Many data science evaluation frameworks are based on this fundamental concept.

Fundamental concept: *The relationship between the business problem and the analytics solution often can be decomposed into tractable subproblems via the framework of analyzing expected value.* Various tools for mining data exist, but business problems rarely come neatly prepared for their application. Breaking the business problem up into components corresponding to estimating probabilities and computing or estimating values, along with a structure for recombining the components, is broadly useful. We have many specific tools for estimating probabilities and values from data. For our churn example, should the *value* of the customer be taken into account in addition to the likelihood of leaving? It is difficult to realistically assess any customer-targeting solution without phrasing the problem as one of expected value.

Fundamental concept: *Information technology can be used to find informative data items from within a large body of data.* One of the first data-science concepts encountered in business-analytics scenarios is the notion of finding correlations. "Correlation" often is used loosely to mean data items that provide information about other data items—specifically, known quantities that reduce our uncertainty about unknown quantities. In our churn example, a quantity of interest is the likelihood that a particular customer will leave after her contract expires. Before the contract expires, this would be an unknown quantity. However, there may be known data items (usage, service history, how many friends

*And they do; see http://data-science-for-biz.com.

have canceled contracts) that correlate with our quantity of interest. This fundamental concept underlies a vast number of techniques for statistical analysis, predictive modeling, and other data mining.

Fundamental concept: *Entities that are similar with respect to known features or attributes often are similar with respect to unknown features or attributes.* Computing similarity is one of the main tools of data science. There are many ways to compute similarity and more are invented each year.

Fundamental concept: *If you look too hard at a set of data, you will find something—but it might not generalize beyond the data you're observing.* This is referred to as "overfitting" a dataset. Techniques for mining data can be very powerful, and the need to detect and avoid overfitting is one of the most important concepts to grasp when applying data-mining tools to real problems. The concept of overfitting and its avoidance permeates data science processes, algorithms, and evaluation methods.

Fundamental concept: *To draw causal conclusions, one must pay very close attention to the presence of confounding factors, possibly unseen ones.* Often, it is not enough simply to uncover correlations in data; we may want to use our models to guide decisions on how to influence the behavior producing the data. For our churn problem, we want to intervene and *cause* customer retention. All methods for drawing causal conclusions—from interpreting the coefficients of regression models to randomized controlled experiments—incorporate assumptions regarding the presence or absence of confounding factors. In applying such methods, it is important to understand their assumptions clearly in order to understand the scope of any causal claims.

Chemistry Is Not About Test Tubes: Data Science vs. the Work of the Data Scientist

Two additional, related complications combine to make it more difficult to reach a common understanding of just what is data science and how it fits with other related concepts.

First is the dearth of academic programs focusing on data science. Without academic programs defining the field for us, we need to define the field for ourselves. However, each of us sees the field from a different perspective and thereby forms a different conception. The dearth of academic programs is largely due to the inertia associated with academia and the concomitant effort involved in creating new academic programs—especially ones that span traditional disciplines. Universities clearly see the need for such programs, and it is only a matter of time before this first complication will be resolved. For example, in New York City alone, two top universities are creating degree programs in data science. Columbia University is in the process of creating a master's degree program within its new Institute for Data Sciences and Engineering (and has founded a center focusing on the foundations of data science), and NYU will commence a master's degree program in data science in fall 2013.

"WITHOUT ACADEMIC PROGRAMS DEFINING THE FIELD FOR US, WE NEED TO DEFINE THE FIELD FOR OURSELVES."

The second complication builds on confusion caused by the first. Workers tend to associate with their field the tasks they spend considerable time on or those they find challenging or rewarding. This is in contrast to the tasks that *differentiate* the field from other fields. Forsythe described this phenomenon in an ethnographic study of practitioners in artificial intelligence (AI):

> The AI specialists I describe view their professional work as science (and in some cases engineering)…The scientists' work and the approach they take to it make sense in relation to a particular view of the world that is taken for granted in the laboratory…Wondering what it means to "do AI," I have asked many practitioners to describe their own work. Their answers invariably focus on one or more of the following: problem solving, writing code, and building systems.[8]

Forsythe goes on to explain that the AI practitioners focus on these three activities even when it is clear that they spend much time doing other things (even less related specifically to AI). Importantly, *none* of these three tasks differentiates AI from other scientific and engineering fields. Clearly just being very good at these three things does not an AI scientist make. And as Forsythe points out, technically the latter two are not even necessary, as the lab director, a famous AI Scientist, had not written code or built systems for years. Nonetheless, these are the tasks the AI scientists saw as defining their work—they apparently did not explicitly consider the notion of what makes doing AI different from doing other tasks that involve problem solving, writing code, and system building. (This is possibly due to the fact that in AI, there were academic distinctions to call on.)

Taken together, these two complications cause particular confusion in data science, because there are few academic distinctions to fall back on, and moreover, due to the state of the art in data processing, data scientists tend to spend a majority of their problem-solving time on data preparation and processing. The goal of such preparation is either to

subsequently apply data-science methods or to understand the results. However, that does not change the fact that the day-to-day work of a data scientist—especially an entry-level one—may be largely data processing. This is directly analogous to an entry-level chemist spending the majority of her time doing technical lab work. If this were all she were trained to do, she likely would not be rightly called a chemist but rather a lab technician. Important for being a chemist is that this work is in support of the application of the science of chemistry, and hopefully the eventual advancement to jobs involving more chemistry and less technical work. Similarly for data science: a chief scientist in a data-science-oriented company will do much less data processing and more data-analytics design and interpretation.

At the time of this writing, discussions of data science inevitably mention not just the analytical skills but the popular tools used in such analysis. For example, it is common to see job advertisements mentioning data-mining techniques (random forests, support vector machines), specific application areas (recommendation systems, ad placement optimization), alongside popular software tools for processing big data (SQL, Hadoop, MongoDB). This is natural. The particular concerns of data science in business are fairly new, and businesses are still working to figure out how best to address them. Continuing our analogy, the state of data science may be likened to that of chemistry in the mid-19th century, when theories and general principles were being formulated and the field was largely experimental. Every good chemist had to be a competent lab technician. Similarly, it is hard to imagine a working data scientist who is not proficient with certain sorts of software tools. A firm may be well served by requiring that their data scientists have skills to access, prepare, and process data using tools the firm has adopted.

Nevertheless, we emphasize that there is an important reason to focus here on the general principles of data science. In ten years' time, the predominant technologies will likely have changed or advanced enough that today's choices would seem quaint. On the other hand, the general principles of data science are not so differerent than they were 20 years ago and likely will change little over the coming decades.

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.

Author Disclosure Statement

F.P. and T.F. are authors of the forthcoming book, *Data Science for Business*.

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 Capellá 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. Working paper, CEDER-11-05, Stern School of Business, 2011. Available at SSRN: http://ssrn.com/abstract=1934670.

Address correspondence to:

F. Provost
Department of Information, Operations, and Management Sciences
Leonard N. Stern School of Business
New York University
44 W. 4th Street, 8th Floor
New York, NY 10012

E-mail: fprovost@stern.nyu.edu