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.

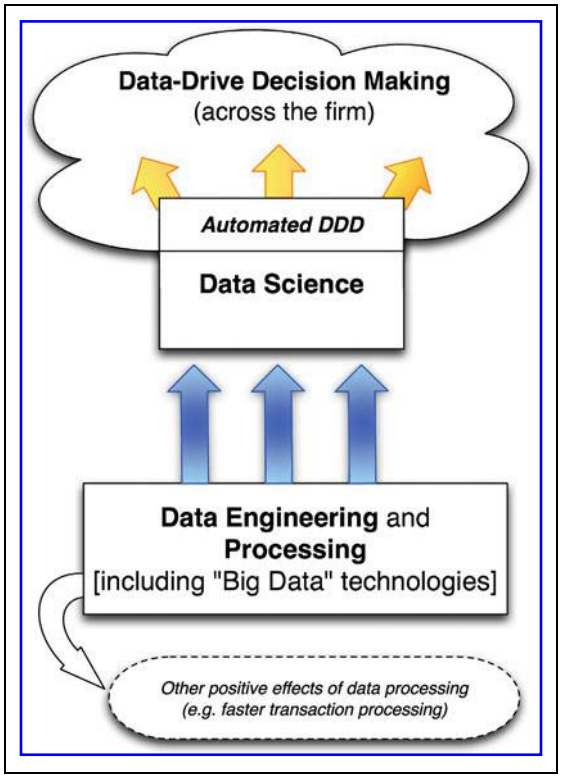

FIG. 1. Data science in the context of closely related processes in the organization.

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

This article has been cited by:

1. Frank Emmert-Streib, Matthias Dehmer. 2019. Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. *Machine Learning and Knowledge Extraction* **1**:1, 521-551. [Crossref]
2. Fábio C. P. Navarro, Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson, Mark Gerstein. 2019. Genomics and data science: an application within an umbrella. *Genome Biology* **20**:1. . [Crossref]
3. Mei Li, Ying Wu, Yi He, Shuai Huang, Anand Nair. 2019. Sparse Inverse Covariance Estimation: A Data Mining Technique to Unravel Holistic Patterns among Business Practices in Firms. *Decision Sciences* **9**. . [Crossref]
4. Roberto Espinosa, Diego García-Saiz, Marta Zorrilla, José Jacobo Zubcoff, Jose-Norberto Mazón. 2019. S3Mining: A model-driven engineering approach for supporting novice data miners in selecting suitable classifiers. *Computer Standards & Interfaces* **65**, 143-158. [Crossref]
5. Mikko Hänninen, Lasse Mitronen, Stephen K. Kwan. 2019. Multi-sided marketplaces and the transformation of retail: A service systems perspective. *Journal of Retailing and Consumer Services* **49**, 380-388. [Crossref]
6. Jennifer Patterson, George Foxcroft. 2019. Gilt Management for Fertility and Longevity. *Animals* **9**:7, 434. [Crossref]
7. Cristina Orsolin Klingenberg, Marco Antônio Viana Borges, José Antônio Valle Antunes Jr. 2019. Industry 4.0 as a data-driven paradigm: a systematic literature review on technologies. *Journal of Manufacturing Technology Management* **11**. . [Crossref]
8. Somayeh Dodge. 2019. A Data Science Framework for Movement. *Geographical Analysis* **92**. . [Crossref]
9. Hui Cai, Yanmin Zhu, Jie Li, Jiadi Yu. 2019. Double Auction for a Data Trading Market with Preferences and Conflicts of Interest. *The Computer Journal* **1**. . [Crossref]
10. Hamed M. Zolbanin, Dursun Delen, Durand Crosby, David Wright. 2019. A Predictive Analytics-Based Decision Support System for Drug Courts. *Information Systems Frontiers* **28**. . [Crossref]
11. Shuyang Li, Guo Chao Peng, Fei Xing. 2019. Barriers of embedding big data solutions in smart factories: insights from SAP consultants. *Industrial Management & Data Systems* **119**:5, 1147-1164. [Crossref]
12. Andrea Urbinati, Marcel Bogers, Vittorio Chiesa, Federico Frattini. 2019. Creating and capturing value from Big Data: A multiple-case study analysis of provider companies. *Technovation* **84-85**, 21-36. [Crossref]
13. Frank Emmert-Streib, Matthias Dehmer. 2019. Large-Scale Simultaneous Inference with Hypothesis Testing: Multiple Testing Procedures in Practice. *Machine Learning and Knowledge Extraction* **1**:2, 653-683. [Crossref]
14. Luis Bote-Curiel, Sergio Muñoz-Romero, Alicia Gerrero-Curieses, José Luis Rojo-Álvarez. 2019. Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners. *Applied Sciences* **9**:11, 2331. [Crossref]
15. Emmanouil Perakakis, George Mastorakis, Ioannis Kopanakis. 2019. Social Media Monitoring: An Innovative Intelligent Approach. *Designs* **3**:2, 24. [Crossref]
16. Sofia Aparicio, Joao Tiago Aparicio, Carlos J. Costa. Data Science and AI: Trends Analysis 1-6. [Crossref]
17. Daniel J. Power, Dale Cyphert, Roberta M. Roth. 2019. Analytics, bias, and evidence: the quest for rational decision making. *Journal of Decision Systems* **5**, 1-18. [Crossref]
18. Vedika Gupta, Vivek Kumar Singh, Udayan Ghose, Pankaj Mukhija. 2019. A quantitative and text-based characterization of big data research. *Journal of Intelligent & Fuzzy Systems* **36**:5, 4659-4675. [Crossref]
19. Hsia-Ching Chang, Chen-Ya Wang, Suliman Hawamdeh. 2019. Emerging trends in data analytics and knowledge management job market: extending KSA framework. *Journal of Knowledge Management* **23**:4, 664-686. [Crossref]
20. Diego Buenaño-Fernández, David Gil, Sergio Luján-Mora. 2019. Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability* **11**:10, 2833. [Crossref]
21. William Villegas-Ch, Xavier Palacios-Pacheco, Sergio Luján-Mora. 2019. Application of a Smart City Model to a Traditional University Campus with a Big Data Architecture: A Sustainable Smart Campus. *Sustainability* **11**:10, 2857. [Crossref]
22. Alex Singleton, Daniel Arribas-Bel. 2019. Geographic Data Science. *Geographical Analysis* **37**. . [Crossref]
23. Srikar Velichety, Sudha Ram, Jesse Bockstedt. 2019. Quality Assessment of Peer-Produced Content in Knowledge Repositories using Development and Coordination Activities. *Journal of Management Information Systems* **36**:2, 478-512. [Crossref]
24. Ao Huang, Yanzhu Huo, Juan Yang, Guangqiang Li. 2019. Computational Simulation and Prediction on Electrical Conductivity of Oxide-Based Melts by Big Data Mining. *Materials* **12**:7, 1059. [Crossref]
25. Patrick Mikalef, Maria Boura, George Lekakos, John Krogstie. 2019. Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment. *British Journal of Management* **30**:2, 272-298. [Crossref]

26. Pervaiz Akhtar, Jędrzej George Frynas, Kamel Mellahi, Subhan Ullah. 2019. Big Data-Savvy Teams' Skills, Big Data-Driven Actions and Business Performance. *British Journal of Management* **30**:2, 252-271. [Crossref]
27. Patrick Mikalef, John Krogstie. Investigating the Data Science Skill Gap: An Empirical Analysis 1275-1284. [Crossref]
28. Daniel Badura, Michael Schulz. 2019. Kleine Barrieren für große Analysen – Eine Untersuchung der Eignung aktueller Plattformen für Self-Service Data Mining. *HMD Praxis der Wirtschaftsinformatik* . [Crossref]
29. Shah Jahan Miah, Huy Quan Vu, John G. Gammack. 2019. A Location Analytics Method for the Utilisation of Geotagged Photos in Travel Marketing Decision-Making. *Journal of Information & Knowledge Management* **18**:01, 1950004. [Crossref]
30. Shadi A. Aljawarneh. 2019. Reviewing and exploring innovative ubiquitous learning tools in higher education. *Journal of Computing in Higher Education* **33**. . [Crossref]
31. Bing Wang, Chao Wu, Lang Huang, Liangguo Kang. 2019. Using data-driven safety decision-making to realize smart safety management in the era of big data: A theoretical perspective on basic questions and their answers. *Journal of Cleaner Production* **210**, 1595-1604. [Crossref]
32. Michael Gaies, Jeffrey Anderson, Alaina Kipps, Angela Lorts, Nicolas Madsen, Bradley Marino, John M. Costello, David Brown, Jeffrey P. Jacobs, David Kasnic, Stacey Lihn, Carole Lannon, Peter Margolis, Gail D. Pearson, Jonathan Kaltman, John R. Charpie, Andrew N. Redington, Sara K. Pasquali. 2019. Cardiac Networks United: an integrated paediatric and congenital cardiovascular research and improvement network. *Cardiology in the Young* **29**:2, 111-118. [Crossref]
33. Sonali Vyas, Sai Sathya Jain, Isha Choudhary, Aryaman Chaudhary. Study on Use of AI and Big Data for Commercial System 737-739. [Crossref]
34. Francesco Caputo, Valentina Cillo, Elena Candelo, Yipeng Liu. 2019. Innovating through digital revolution. *Management Decision* **37**. . [Crossref]
35. Iván García-Magariño, Carlos Medrano, Jorge Delgado. 2019. Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction methods. *Neural Computing and Applications* **3**. . [Crossref]
36. Celia Satiko Ishikiriyama, Carlos Francisco Simoes Gomes. Big Data: A Global Overview 35-50. [Crossref]
37. Brian J. Galli. 2019. Role of Big Data in Continuous Improvement Environments. *International Journal of Applied Logistics* **9**:1, 53. [Crossref]
38. Iztok Fister, Iztok Fister, Dušan Fister. Visualization of Sports Activities Created by Wearable Mobile Devices 223-246. [Crossref]
39. Miftachul Huda, Ulfatmi, Muhammad Ja'far Luthfi, Kamarul Azmi Jasmi, Bushrah Basiron, Mohd Ismail Mustari, Ajmain Safar, Wan Hassan Wan Embong, Ahmad Marzuki Mohamad, Ahmad KIlani Mohamed. Adaptive Online Learning Technology 163-195. [Crossref]
40. Ben Kei Daniel. 2019. Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology* **50**:1, 101-113. [Crossref]
41. Sumira Jan, Parvaiz Ahmad. Ecological Metabolomics: Challenges and Perspectives 293-378. [Crossref]
42. Yi-Cheng Tsai, Mu-En Wu, Jia-Hao Syu, Chin-Laung Lei, Chung-Shu Wu, Jan-Ming Ho, Chuan-Ju Wang. 2019. Assessing the Profitability of Timely Opening Range Breakout on Index Futures Markets. *IEEE Access* 1-1. [Crossref]
43. Mikel Canizo, Angel Conde, Santiago Charramendieta, Raul Minon, Raul G. Cid-Fuentes, Enrique Onieva. 2019. Implementation of a Large-Scale Platform for Cyber-Physical System Real-Time Monitoring. *IEEE Access* **7**, 52455-52466. [Crossref]
44. James Ming Chen. 2019. Models for Predicting Business Bankruptcies and Their Application to Banking and to Financial Regulation. *SSRN Electronic Journal* . [Crossref]
45. Khaled Salah Mohamed. IoT Cloud Computing, Storage, and Data Analytics 71-91. [Crossref]
46. Agata Mardosz-Grabowska. Big Data Myth 223-237. [Crossref]
47. Richard Berntsson Svensson, Robert Feldt, Richard Torkar. The Unfulfilled Potential of Data-Driven Decision Making in Agile Software Development 69-85. [Crossref]
48. Jens Prufer, Patricia Prufer. 2019. Data Science for Entrepreneurship Research: Studying Demand Dynamics for Entrepreneurial Skills in the Netherlands. *SSRN Electronic Journal* . [Crossref]
49. Thilo Stadelmann, Martin Braschler, Kurt Stockinger. Introduction to Applied Data Science 3-16. [Crossref]
50. Ray Cooksey, Gael McDonald. What Data Gathering Strategies Should I Use? 555-687. [Crossref]
51. Stefaan G. Verhulst, Zeynep Engin, Jon Crowcroft. 2019. Data & Policy : A new venue to study and explore policy–data interaction. *Data & Policy* **1**. . [Crossref]
52. Trevor Bogani Shihundla, Khumbulani Mpofu, Olukorede Tijani Adenuga. 2019. Integrating Product-Service Systems into the manufacturing industry: Industry 4.0 perspectives. *Procedia CIRP* **83**, 8-13. [Crossref]

53. Alex Young Pedersen, Francesco Caviglia, Tom Gislev, Anders Hjortskov Larsen. Learning in Hybrid Protopublic Spaces: Framework and Exemplars 89-110. [Crossref]
54. John R. Walkup, Roger A. Key, Patrick R. M. Talbot, Michael A. Walkup. 2019. Data-driven decision making in an introductory physics lab. *American Journal of Physics* **87**:8, 654. [Crossref]
55. Pier Francesco De Maria, Leonardo Tomazeli Duarte, Álvaro de Oliveira D'Antona, Cristiano Torezzan. Digital Humanities and Big Microdata: New Approaches for Demographic Research 217-231. [Crossref]
56. Lorenzo Ardito, Veronica Scuotto, Manlio Del Giudice, Antonio Messeni Petruzzelli. 2018. A bibliometric analysis of research on Big Data analytics for business and management. *Management Decision* **90**. . [Crossref]
57. Vian Ahmed, Zeeshan Aziz, Algan Tezel, Zainab Riaz. 2018. Challenges and drivers for data mining in the AEC sector. *Engineering, Construction and Architectural Management* **25**:11, 1436-1453. [Crossref]
58. Paulo Rita, Nicole Rita, Cristina Oliveira. 2018. Data science for hospitality and tourism. *Worldwide Hospitality and Tourism Themes* **10**:6, 717-725. [Crossref]
59. JingMing Zhang, ShuZhen Zhu, Wei Yan, ZhiPeng Li. 2018. The construction and simulation of internet financial product diffusion model based on complex network and consumer decision-making mechanism. *Information Systems and e-Business Management* **130**. . [Crossref]
60. Changtian Ying, Changyan Ying, Chen Ban. 2018. A performance optimization strategy based on degree of parallelism and allocation fitness. *EURASIP Journal on Wireless Communications and Networking* **2018**:1. . [Crossref]
61. Virginia Ortiz-Repiso, Jane Greenberg, Javier Calzada-Prado. 2018. A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science* **44**:6, 768-784. [Crossref]
62. Aleksander Fabijan, Pavel Dmitriev, Colin McFarland, Lukas Vermeer, Helena Holmström Olsson, Jan Bosch. 2018. Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. *Journal of Software: Evolution and Process* **30**:12, e2113. [Crossref]
63. Frank Emmert-Streib, Matthias Dehmer. 2018. Defining Data Science by a Data-Driven Quantification of the Community. *Machine Learning and Knowledge Extraction* **1**:1, 235-251. [Crossref]
64. Trang VoPham, Jaime E. Hart, Francine Laden, Yao-Yi Chiang. 2018. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environmental Health* **17**:1. . [Crossref]
65. Zhuangli Hu, Tong He, Yihui Zeng, Xiangyuan Luo, Jiawen Wang, Sheng Huang, Jianming Liang, Qinzhang Sun, Hengbin Xu, Bin Lin. 2018. Fast image recognition of transmission tower based on big data. *Protection and Control of Modern Power Systems* **3**:1. . [Crossref]
66. Gabriele Santoro, Fabio Fiano, Bernardo Bertoldi, Francesco Ciampi. 2018. Big data for business management in the retail industry. *Management Decision* **6**. . [Crossref]
67. Kwangha Lee. 2018. Critical Care Research Using "Big Data": A Reality in the Near Future. *Acute and Critical Care* **33**:4, 269-270. [Crossref]
68. Muhammad Anwar, Sher Zaman Khan, Syed Zulfiqar Ali Shah. 2018. Big Data Capabilities and Firm's Performance: A Mediating Role of Competitive Advantage. *Journal of Information & Knowledge Management* **26**, 1850045. [Crossref]
69. Bas van Raaij, Arco van de Ven. 2018. Betere financiële prognoses in business cases door de toepassing van big data. *Maandblad Voor Accountancy en Bedrijfseconomie* **92**:9/10, 255-263. [Crossref]
70. Shaofu Du, Wenzhi Tang, Jiajia Zhao, Tengfei Nie. 2018. Sell to whom? Firm's green production in competition facing market segmentation. *Annals of Operations Research* **270**:1-2, 125-154. [Crossref]
71. L. Nelson Sanchez-Pinto, Yuan Luo, Matthew M. Churpek. 2018. Big Data and Data Science in Critical Care. *Chest* **154**:5, 1239-1248. [Crossref]
72. P BastinThiyagaraj, A Aloysius. 2018. Distance Based Measurement Approach for Truth Discovery by Resolving the Conflicts in Big Data. *Journal of Physics: Conference Series* **1142**, 012013. [Crossref]
73. Lin Wang. 2018. Twinning data science with information science in schools of library and information science. *Journal of Documentation* **74**:6, 1243-1257. [Crossref]
74. Adrian Gardiner, Cheryl Aasheim, Paige Rutner, Susan Williams. 2018. Skill Requirements in Big Data: A Content Analysis of Job Advertisements. *Journal of Computer Information Systems* **58**:4, 374-384. [Crossref]
75. Forrest J. Bowlick, Dawn J. Wright. 2018. Digital Data-Centric Geography: Implications for Geography's Frontier. *The Professional Geographer* **70**:4, 687-694. [Crossref]

76. Junius Gunaratne, Lior Zalmanson, Oded Nov. 2018. The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems* **35**:4, 1092-1120. [Crossref]
77. Abhilash Acharya, Sanjay Kumar Singh, Vijay Pereira, Poonam Singh. 2018. Big data, knowledge co-creation and decision making in fashion industry. *International Journal of Information Management* **42**, 90-101. [Crossref]
78. Dorota Kamrowska-Zaluska, Hanna Obracht-Prondzyńska. 2018. The Use of Big Data in Regenerative Planning. *Sustainability* **10**:10, 3668. [Crossref]
79. Kyoung-Yun Kim, Fahim Ahmed. 2018. Semantic weldability prediction with RSW quality dataset and knowledge construction. *Advanced Engineering Informatics* **38**, 41-53. [Crossref]
80. Gabriela Viale Pereira, Gregor Eibl, Constantinos Stylianou, Gilberto Martínez, Haris Neophytou, Peter Parycek. 2018. The Role of Smart Technologies to Support Citizen Engagement and Decision Making. *International Journal of Electronic Government Research* **14**:4, 1-17. [Crossref]
81. Changtian Ying, Jiong Yu, JingSha He. 2018. Towards fault tolerance optimization based on checkpoints of in-memory framework spark. *Journal of Ambient Intelligence and Humanized Computing* **275**. . [Crossref]
82. Luca Dezi, Gabriele Santoro, Heger Gabteni, Anna Claudia Pellicelli. 2018. The role of big data in shaping ambidextrous business process management. *Business Process Management Journal* **24**:5, 1163-1175. [Crossref]
83. Andrea De Mauro, Marco Greco, Michele Grimaldi, Paavo Ritala. 2018. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management* **54**:5, 807-817. [Crossref]
84. Sophie Cockcroft, Mark Russell. 2018. Big Data Opportunities for Accounting and Finance Practice and Research. *Australian Accounting Review* **28**:3, 323-333. [Crossref]
85. Jürgen Bajorath. 2018. Foundations of data-driven medicinal chemistry. *Future Science OA* **4**:8, FSO320. [Crossref]
86. Stefan Thalmann. 2018. Data driven decision support. *it - Information Technology* **60**:4, 179-181. [Crossref]
87. Milla Ratia, Jussi Myllärniemi, Nina Helander. 2018. The new era of business intelligence. *Meditari Accountancy Research* **26**:3, 531-546. [Crossref]
88. David Contreras, Maria Salamó. 2018. Data-driven decision making in critique-based recommenders: from a critique to social media data. *Journal of Intelligent Information Systems* **310**. . [Crossref]
89. Patrick Mikalef, Ilias O. Pappas, John Krogstie, Michail Giannakos. 2018. Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and e-Business Management* **16**:3, 547-578. [Crossref]
90. Sean Robert Valentine, David Hollingworth, Patrick Schultz. 2018. Data-based ethical decision making, lateral relations, and organizational commitment. *Employee Relations* **19**. . [Crossref]
91. Ramon Wenzel, Niels Van Quaquebeke. 2018. The Double-Edged Sword of Big Data in Organizational and Management Research. *Organizational Research Methods* **21**:3, 548-591. [Crossref]
92. Jeeyae Choi, Jeungok Choi, Hee-Tae Jung. 2018. Applying Machine-Learning Techniques to Build Self-reported Depression Prediction Models. *CIN: Computers, Informatics, Nursing* **36**:7, 317-321. [Crossref]
93. Jürgen Bajorath. 2018. Data analytics and deep learning in medicinal chemistry. *Future Medicinal Chemistry* **10**:13, 1541-1543. [Crossref]
94. . References 311-314. [Crossref]
95. Anastasia Griva, Cleopatra Bardaki, Katerina Pramatari, Dimitris Papakiriakopoulos. 2018. Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications* **100**, 1-16. [Crossref]
96. Sharon Hewner, Suzanne S. Sullivan, Guan Yu. 2018. Reducing Emergency Room Visits and In-Hospitalizations by Implementing Best Practice for Transitional Care Using Innovative Technology and Big Data. *Worldviews on Evidence-Based Nursing* **15**:3, 170-177. [Crossref]
97. Basanta-Val Pablo, Sánchez-Fernández Luis. 2018. Big-BOE: Fusing Spanish Official Gazette with Big Data Technology. *Big Data* **6**:2, 124-138. [Abstract] [Full Text] [PDF] [PDF Plus]
98. Catherine Chen, Haoqiang Jiang. 2018. Important Skills for Data Scientists in China: Two Delphi Studies. *Journal of Computer Information Systems* **37**, 1-10. [Crossref]
99. Shirish Jeble, Rameshwar Dubey, Stephen J. Childe, Thanos Papadopoulos, David Roubaud, Anand Prakash. 2018. Impact of big data and predictive analytics capability on supply chain sustainability. *The International Journal of Logistics Management* **29**:2, 513-538. [Crossref]
100. Jenny Farmer, Donald Jacobs. 2018. High throughput nonparametric probability density estimation. *PLOS ONE* **13**:5, e0196937. [Crossref]

101. Aleksej Heinze, Marie Griffiths, Alex Fenton, Gordon Fletcher. 2018. Knowledge exchange partnership leads to digital transformation at Hydro-X Water Treatment, Ltd. *Global Business and Organizational Excellence* **37**:4, 6-13. [Crossref]
102. Kristen L. Fessele. 2018. The Rise of Big Data in Oncology. *Seminars in Oncology Nursing* **34**:2, 168-176. [Crossref]
103. Yin Long, Zhi Chen, Jun Fang, Chintha Tellambura. 2018. Data-Driven-Based Analog Beam Selection for Hybrid Beamforming Under mm-Wave Channels. *IEEE Journal of Selected Topics in Signal Processing* **12**:2, 340-352. [Crossref]
104. Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* **48**:2, 204-231. [Crossref]
105. Patrick Mikalef, Michail N. Giannakos, Ilias O. Pappas, John Krogstie. The human side of big data: Understanding the skills of the data scientist in education and industry 503-512. [Crossref]
106. Ehi E. Aimiuwu. 2018. Using Social Media to Target Customers for Green Technology Use. *International Journal of Virtual Communities and Social Networking* **10**:2, 41-61. [Crossref]
107. Marynia Kolak. 2018. Ian Foster, Rayid Ghani, Ron S Jarmin, et al. (eds), Big data and social science: A practical guide to methods and tools. *Environment and Planning B: Urban Analytics and City Science* **45**:2, 388-389. [Crossref]
108. Qingqi Long. 2018. Data-driven decision making for supply chain networks with agent-based computational experiment. *Knowledge-Based Systems* **141**, 55-66. [Crossref]
109. G. Rejikumar, A. Aswathy Asokan, V. Raja Sreedharan. 2018. Impact of data-driven decision-making in Lean Six Sigma: an empirical analysis. *Total Quality Management & Business Excellence* **11**, 1-18. [Crossref]
110. Cecilia Fredriksson. 2018. Big data creating new knowledge as support in decision-making: practical examples of big data use and consequences of using big data as decision support. *Journal of Decision Systems* **27**:1, 1-18. [Crossref]
111. Konstantinos Vassakis, Emmanuel Petrakis, Ioannis Kopanakis. Big Data Analytics: Applications, Prospects and Challenges 3-20. [Crossref]
112. Alessandro Ancarani, Carmela Di Mauro. Successful digital transformations need a focus on the individual 11-26. [Crossref]
113. # #. 2018. Research and Design of Hydrological Big Data Sharing Platform. *Journal of Water Resources Research* **07**:01, 10-18. [Crossref]
114. Weisheng Lu, Yi Peng, Fan Xue, Ke Chen, Yuhan Niu, Xi Chen. The Fusion of GIS and Building Information Modeling for Big Data Analytics in Managing Development Sites 345-359. [Crossref]
115. Ettore Bolisani, Constantin Bratianu. Strategic Performance and Knowledge Measurement 175-198. [Crossref]
116. Kamaluddeen Usman Danyaro, M. S. Liew. A Proposed Methodology for Integrating Oil and Gas Data Using Semantic Big Data Technology 30-38. [Crossref]
117. X. Chen, W. S. Lu. Scenarios for Applying Big Data in Boosting Construction: A Review 1299-1306. [Crossref]
118. Kyungtae Kim, Sungjoo Lee. 2018. How Can Big Data Complement Expert Analysis? A Value Chain Case Study. *Sustainability* **10**:3, 709. [Crossref]
119. John Storm Pedersen, Adrian Wilkinson. 2018. The digital society and provision of welfare services. *International Journal of Sociology and Social Policy* **38**:3/4, 194. [Crossref]
120. Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach, Constantine E. Kontokosta. 2018. Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City. *Journal of Technology in Human Services* **36**:1, 89. [Crossref]
121. Alon Friedman. 2018. Measuring the promise of Big Data syllabi. *Technology, Pedagogy and Education* **27**:2, 135. [Crossref]
122. Stephen Jia Wang, Patrick Moriarty. The Potential for Big data for Urban Sustainability 45-63. [Crossref]
123. W. Velasquez, A. Munoz-Arcentales, T. M. Chalen, Joaquin Salvachua. Survival analysis of people with cardiac problems in a simulated earthquake environment 702-706. [Crossref]
124. Riccardo Fini, Monica Bartolini, Stefano Benigni, Paolo Ciancarini, Angelo Di Iorio, Alan Johnson, Marcello Maria Mariani, Silvio Peroni, Francesco Poggi, Einar Rasmussen, Riccardo Silvi, Maurizio Sobrero, Laura Toschi. Collaborative Practices and Multidisciplinary Research: The Dialogue Between Entrepreneurship, Management, and Data Science 129-152. [Crossref]
125. Rafael de Oliveira Werneck, Waldir Rodrigues de Almeida, Bernardo Vecchia Stein, Daniel Vatanabe Pazinato, Pedro Ribeiro Mendes Júnior, Otávio Augusto Bizetto Penatti, Anderson Rocha, Ricardo da Silva Torres. 2018. Kuaa: A unified framework for design, deployment, execution, and recommendation of machine learning experiments. *Future Generation Computer Systems* **78**, 59-76. [Crossref]
126. Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani, Mohammed Ali Al-Garadi. 2018. Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges. *IEEE Access* 1-1. [Crossref]

127. Longbing Cao. What Is Data Science 29-58. [Crossref]
128. Małgorzata Przybyła-Kasperek. Application of the Pairwise Comparison Matrices into a Dispersed Decision-Making System With Pawlak's Conflict Model 392-404. [Crossref]
129. Arnold Picot, Yvonne Berchtold, Rahild Neuburger. Big Data aus ökonomischer Sicht: Potenziale und Handlungsbedarf 309-416. [Crossref]
130. Russell Tatenda Munodawafa, Satirenjit Kaur Johl. 2018. Eco-Innovation and Industry 4.0: A Big Data Usage conceptual model. *SHS Web of Conferences* **56**, 05003. [Crossref]
131. Houcine Dammak, Mickaël Gardoni. Improving the Innovation Process by Harnessing the Usage of Content Management Tools Coupled with Visualization Tools 642-655. [Crossref]
132. Luca Urciuoli. 2018. An algorithm for improved ETAs estimations and potential impacts on supply chain decision making. *Procedia Manufacturing* **25**, 185-193. [Crossref]
133. Carlos Costa, Maribel Yasmina Santos. 2017. The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *International Journal of Information Management* **37**:6, 726-734. [Crossref]
134. Alberto Fernández, Sara del Río, Abdullah Bawakid, Francisco Herrera. 2017. Fuzzy rule based classification systems for big data with MapReduce: granularity analysis. *Advances in Data Analysis and Classification* **11**:4, 711-730. [Crossref]
135. Ming-Chi Liu, Yueh-Min Huang. 2017. The use of data science for education: The case of social-emotional learning. *Smart Learning Environments* **4**:1. . [Crossref]
136. Thomas M. Kreuzer, Martina Wilde, Birgit Terhorst, Bodo Damm. 2017. A landslide inventory system as a base for automated process and risk analyses. *Earth Science Informatics* **10**:4, 507-515. [Crossref]
137. Mario Jose Divan. Data-driven decision making 50-56. [Crossref]
138. Jun Li, Ming Lu, Guowei Dou, Shanyong Wang. 2017. Big data application framework and its feasibility analysis in library. *Information Discovery and Delivery* **45**:4, 161-168. [Crossref]
139. Murray E. Jennex. 2017. Big Data, the Internet of Things, and the Revised Knowledge Pyramid. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* **48**:4, 69-79. [Crossref]
140. Dominikus Kleindienst. 2017. The data quality improvement plan: deciding on choice and sequence of data quality improvements. *Electronic Markets* **27**:4, 387-398. [Crossref]
141. Ivana Semanjski, Sidharta Gautama, Rein Ahas, Frank Witlox. 2017. Spatial context mining approach for transport mode recognition from mobile sensed big data. *Computers, Environment and Urban Systems* **66**, 38-52. [Crossref]
142. Kasper Welbers, Wouter Van Atteveldt, Kenneth Benoit. 2017. Text Analysis in R. *Communication Methods and Measures* **11**:4, 245-265. [Crossref]
143. Sarah Cheah, Shenghui Wang. 2017. Big data-driven business model innovation by traditional industries in the Chinese economy. *Journal of Chinese Economic and Foreign Trade Studies* **10**:3, 229-251. [Crossref]
144. Ju Yeon Lee, Joo Seong Yoon, Bo-Hyun Kim. 2017. A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. *International Journal of Precision Engineering and Manufacturing* **18**:10, 1353-1361. [Crossref]
145. S. Vijayakumar Bharathi. 2017. Prioritizing and Ranking the Big Data Information Security Risk Spectrum. *Global Journal of Flexible Systems Management* **18**:3, 183-201. [Crossref]
146. Sarah Giest. 2017. Big data for policymaking: fad or fasttrack?. *Policy Sciences* **50**:3, 367-382. [Crossref]
147. Jie Sheng, Joseph Amankwah-Amoah, Xiaojun Wang. 2017. A multidisciplinary perspective of big data in management research. *International Journal of Production Economics* **191**, 97-112. [Crossref]
148. Il-Yeol Song, Yongjun Zhu. 2017. Big Data and Data Science: Opportunities and Challenges of iSchools. *Journal of Data and Information Science* **2**:3, 1-18. [Crossref]
149. Marko Bohanec, Marko Robnik-Šikonja, Mirjana Kljajić Borštnar. 2017. Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management & Data Systems* **117**:7, 1389-1406. [Crossref]
150. Dominik Krimpmann, Anna Stühmeier. 2017. Big Data and Analytics. *International Journal of Service Science, Management, Engineering, and Technology* **8**:3, 79-92. [Crossref]
151. Kevin Daniel André Carillo. 2017. Let's stop trying to be "sexy" – preparing managers for the (big) data-driven business era. *Business Process Management Journal* **23**:3, 598-622. [Crossref]

152. Matthias Murawski, Markus Bick. 2017. Digital competences of the workforce – a research topic?. *Business Process Management Journal* **23**:3, 721-734. [Crossref]
153. Mikel Canizo, Enrique Onieva, Angel Conde, Santiago Charramendieta, Salvador Trujillo. Real-time predictive maintenance for wind turbines using Big Data frameworks 70-77. [Crossref]
154. Samuel T. McAbee, Ronald S. Landis, Maura I. Burke. 2017. Inductive reasoning: The promise of big data. *Human Resource Management Review* **27**:2, 277-290. [Crossref]
155. Diana Heredia Vizcaino, Wilson Nieto Bernal. Proposal for a methodology for the adoption of the big data 1-6. [Crossref]
156. Ling Cen, Dymitr Ruta. A Map-Based Gender Prediction Model for Big E-Commerce Data 1025-1029. [Crossref]
157. Ruben Costa, Paulo Figueiras, Ricardo Jardim-Goncalves, Jose Ramos-Filho, Celson Lima. Semantic enrichment of product data supported by machine learning techniques 1472-1479. [Crossref]
158. Hansol Lee, Eunkyung Kweon, Minkyun Kim, Sangmi Chai. 2017. Does Implementation of Big Data Analytics Improve Firms' Market Value? Investors' Reaction in Stock Market. *Sustainability* **9**:6, 978. [Crossref]
159. Guillaume Coqueret. 2017. Approximate NORTA simulations for virtual sample generation. *Expert Systems with Applications* **73**, 69-81. [Crossref]
160. Allen D. Allen. 2017. Algorithms that extract knowledge from fuzzy big data: Conserving traditional science1. *Journal of Intelligent & Fuzzy Systems* **32**:5, 3689-3694. [Crossref]
161. Saša Baškarada, Andy Koronios. 2017. Unicorn data scientist: the rarest of breeds. *Program* **51**:1, 65-74. [Crossref]
162. Marko Bohanec, Mirjana Kljajić Borštnar, Marko Robnik-Šikonja. 2017. Explaining machine learning models in sales predictions. *Expert Systems with Applications* **71**, 416-428. [Crossref]
163. Andoni Elola, Javier Del Ser, Miren Nekane Bilbao, Cristina Perfecto, Enrique Alexandre, Sancho Salcedo-Sanz. 2017. Hybridizing Cartesian Genetic Programming and Harmony Search for adaptive feature construction in supervised learning problems. *Applied Soft Computing* **52**, 760-770. [Crossref]
164. Ali Intezari, Simone Gressel. 2017. Information and reformation in KM systems: big data and strategic decision-making. *Journal of Knowledge Management* **21**:1, 71-91. [Crossref]
165. Elias G. Carayannis, Evangelos Grigoroudis, Manlio Del Giudice, Maria Rosaria Della Peruta, Stavros Sindakis. 2017. An exploration of contemporary organizational artifacts and routines in a sustainable excellence context. *Journal of Knowledge Management* **21**:1, 35-56. [Crossref]
166. Dorota Jelonek. 2017. Big Data Analytics in the Management of Business. *MATEC Web of Conferences* **125**, 04021. [Crossref]
167. Samantha Hautea, Sayamindu Dasgupta, Benjamin Mako Hill. Youth Perspectives on Critical Data Literacies 919-930. [Crossref]
168. Carlos Costa, Maribel Yasmina Santos. A Conceptual Model for the Professional Profile of a Data Scientist 453-463. [Crossref]
169. Md Tarique Hasan Khan, Fahim Ahmed, Kyoung-Yun Kim. 2017. Weldability Knowledge Visualization of Resistance Spot Welded Assembly Design. *Procedia Manufacturing* **11**, 1609-1616. [Crossref]
170. Abeer Ahmed Abdullah AL-Hakimi. Big Data Skills Required for Successful Application Implementation in the Banking Sector 381-392. [Crossref]
171. Jaime Campos, Pankaj Sharma, Unai Gorostegui Gabiria, Erkki Jantunen, David Baglee. 2017. A Big Data Analytical Architecture for the Asset Management. *Procedia CIRP* **64**, 369-374. [Crossref]
172. Jiang Zeyu, Yu Shuiping, Zhou Mingduan, Chen Yongqiang, Liu Yi. 2017. Model Study for Intelligent Transportation System with Big Data. *Procedia Computer Science* **107**, 418-426. [Crossref]
173. Wardah Zainal Abidin, Nur Amie Ismail, Nurazean Maarop, Rose Alinda Alias. Skills Sets Towards Becoming Effective Data Scientists 97-106. [Crossref]
174. Chhaya S. Dule, H. A. Girijamma. Guaging the Effectivity of Existing Security Measures for Big Data in Cloud Environment 209-219. [Crossref]
175. P. Y. Zhao, Y. M. Shi. Predicting the likelihood of purchase by big data 040002. [Crossref]
176. Joohee Choi, Yla Tausczik. Characteristics of Collaboration in the Emerging Practice of Open Data Analysis 835-846. [Crossref]
177. I. Tikito, N. Souissi. Data Collect Requirements Model 1-7. [Crossref]
178. Lynda R. Hardy, Philip E. Bourne. Data Science: Transformation of Research and Scholarship 183-209. [Crossref]
179. Eduan Kotzé. Augmenting a Data Warehousing Curriculum with Emerging Big Data Technologies 128-143. [Crossref]
180. Pan Xiang. Decision Support Assistant Management in Intelligent Logistics System 254-258. [Crossref]

181. Rudolph Pienaar, Ata Turk, Jorge Bernal-Rusiel, Nicolas Rannou, Daniel Haehn, P. Ellen Grant, Orran Krieger. CHIPS – A Service for Collecting, Organizing, Processing, and Sharing Medical Image Data in the Cloud 29-35. [Crossref]
182. Fahim Ahmed, Kyoung-Yun Kim. 2017. Data-driven Weld Nugget Width Prediction with Decision Tree Algorithm. *Procedia Manufacturing* **10**, 1009-1019. [Crossref]
183. Fernando Rosas, Jen-Hao Hsiao, Kwang-Cheng Chen. 2017. A Technological Perspective on Information Cascades via Social Learning. *IEEE Access* **5**, 22605-22633. [Crossref]
184. Muhammad Fahim, Thar Baker. Knowledge-Based Decision Support Systems for Personalized u-lifecare Big Data Services 187-203. [Crossref]
185. Eduardo Nigri, Ognjen Arandjelovic. Light Curve Analysis From Kepler Spacecraft Collected Data 93-98. [Crossref]
186. Andrzej Szwabe, Pawel Misiorek, Michal Ciesielczyk. Logistic Regression Setup for RTB CTR Estimation 61-70. [Crossref]
187. Youakim Badr, Soumya Banerjee. Developing Modified Classifier for Big Data Paradigm: An Approach Through Bio-Inspired Soft Computing 109-122. [Crossref]
188. Florin Gheorghe Filip, Constantin-Bălă Zamfirescu, Cristian Ciurea. Essential Enabling Technologies 121-176. [Crossref]
189. Hiroshi Mamiya, Arash Shaban-Nejad, David L. Buckeridge. Online Public Health Intelligence: Ethical Considerations at the Big Data Era 129-148. [Crossref]
190. Andrzej Szwabe, Paweł Misiorek, Michał Ciesielczyk. Evaluation of Tensor-Based Algorithms for Real-Time Bidding Optimization 160-169. [Crossref]
191. Megan K. Sutherland, Meghan E. Cook. Data-Driven Smart Cities 471-476. [Crossref]
192. Stavros Sindakis. Applying Data Analytics for Innovation and Sustainable Enterprise Excellence 271-275. [Crossref]
193. Mansi Khurana, Deepak Kumar. 140. [Crossref]
194. Xiaohong Liu, Junfei Chu, Pengzhen Yin, Jiasen Sun. 2017. DEA cross-efficiency evaluation considering undesirable output and ranking priority: a case study of eco-efficiency analysis of coal-fired power plants. *Journal of Cleaner Production* **142**, 877-885. [Crossref]
195. Marko Bohanec, Marko Robnik-Šikonja, Mirjana Kljajić Borštnar. 2017. Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. *Organizacija* **50**:3. . [Crossref]
196. Lourdes S. Martinez. Data Science 1-4. [Crossref]
197. Elena Falletti. 2017. Could Wearable Technology Transform the Traditional Concept of Habeas Corpus?. *SSRN Electronic Journal* . [Crossref]
198. Andrzej Szwabe, Pawel Misiorek, Michal Ciesielczyk. Tensor-Based Modeling of Temporal Features for Big Data CTR Estimation 16-27. [Crossref]
199. Leonardo M. Millefiori, Dimitrios Zissis, Luca Cazzanti, Gianfranco Arcieri. Scalable and Distributed Sea Port Operational Areas Estimation from AIS Data 374-381. [Crossref]
200. Babak Yadranjiaghdam, Nathan Pool, Nasseh Tabrizi. A Survey on Real-Time Big Data Analytics: Applications and Tools 404-409. [Crossref]
201. Yan Li, Manoj Thomas, Kweku-Muata Osei-Bryson, Jason Levy. 2016. Problem Formulation in Knowledge Discovery via Data Analytics (KDDA) for Environmental Risk Management. *International Journal of Environmental Research and Public Health* **13**:12, 1245. [Crossref]
202. Yan Li, Manoj A. Thomas, Kweku-Muata Osei-Bryson. 2016. A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems* **91**, 1-12. [Crossref]
203. Shirley Coleman, Rainer Göb, Giuseppe Manco, Antonio Pievatolo, Xavier Tort-Martorell, Marco Seabra Reis. 2016. How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International* **32**:6, 2151-2164. [Crossref]
204. Kayode Ayankoya, Andre P. Calitz, Jean H. Greyling. 2016. Real-Time Grain Commodities Price Predictions in South Africa: A Big Data and Neural Networks Approach. *Agrekon* **55**:4, 483-508. [Crossref]
205. Ajaya K. Swain. 2016. Mining big data to support decision making in healthcare. *Journal of Information Technology Case and Application Research* **18**:3, 141-154. [Crossref]
206. Satyanarayana V Nandury, Beneyaz A Begum. Strategies to handle big data for traffic management in smart cities 356-364. [Crossref]
207. Xiangzheng Deng. 2016. Urgent need for a data sharing platform to promote ecological research in china. *Ecosystem Health and Sustainability* **2**:9, e01241. [Crossref]

208. Shanliang Yang, Mei Yang, Song Wang, Kedi Huang. 2016. Adaptive immune genetic algorithm for weapon system portfolio optimization in military big data environment. *Cluster Computing* **19**:3, 1359-1372. [Crossref]
209. Russell Newman, Victor Chang, Robert John Walters, Gary Brian Wills. 2016. Model and experimental development for Business Data Science. *International Journal of Information Management* **36**:4, 607-617. [Crossref]
210. Lisbeth Rodríguez-Mazahua, Cristian-Aarón Rodríguez-Enríquez, José Luis Sánchez-Cervantes, Jair Cervantes, Jorge Luis García-Alcaraz, Giner Alor-Hernández. 2016. A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing* **72**:8, 3073-3113. [Crossref]
211. Il-Yeol Song, Yongjun Zhu. 2016. Big data and data science: what should we teach?. *Expert Systems* **33**:4, 364-373. [Crossref]
212. Rong Tang, Watinee Sae-Lim. 2016. Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information* **32**:3, 269-290. [Crossref]
213. Hao Peifeng, Cui Yuzhe, Song Jingping, Hu Zhaomu. Smart wardrobe system based on Android platform 279-285. [Crossref]
214. Jianwei Yin, Yan Tang, Wei Lo, Zhaohui Wu. From Big Data to Great Services 165-172. [Crossref]
215. Carolina Lagos, Sebastian Gutierrez, Felisa Cordova, Guillermo Fuertes, Raul Carrasco. Data analysis methods related to energetic consumption in copper mining — A Test case in Chile 244-249. [Crossref]
216. Zhenning Xu, Gary L. Frankwick, Edward Ramirez. 2016. Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research* **69**:5, 1562-1566. [Crossref]
217. Myongho Yi. 2016. A Study on the Curriculums of Data Science. *Journal of the Korean BIBLIA Society for library and Information Science* **27**:1, 263-290. [Crossref]
218. Thomas Hart, Lei Xie. 2016. Providing data science support for systems pharmacology and its implications to drug discovery. *Expert Opinion on Drug Discovery* **11**:3, 241-256. [Crossref]
219. Meiyu Fan, Jian Sun, Bin Zhou, Min Chen. 2016. The Smart Health Initiative in China: The Case of Wuhan, Hubei Province. *Journal of Medical Systems* **40**:3. . [Crossref]
220. 2016. A Project-Based Case Study of Data Science Education. *Data Science Journal* **15**. . [Crossref]
221. David M. Steinberg. 2016. Industrial statistics: The challenges and the research. *Quality Engineering* **28**:1, 45-59. [Crossref]
222. Gerald Schermann, Dominik Schöni, Philipp Leitner, Harald C. Gall. Bifrost 1-14. [Crossref]
223. Go Muan Sang, Lai Xu, Paul de Vrieze. A reference architecture for big data systems 370-375. [Crossref]
224. Harald Foidl, Michael Felderer. Data Science Challenges to Improve Quality Assurance of Internet of Things Applications 707-726. [Crossref]
225. Ivana Semanjski, Rik Bellens, Sidharta Gautama, Frank Witlox. 2016. Integrating Big Data into a Sustainable Mobility Policy 2.0 Planning Support System. *Sustainability* **8**:11, 1142. [Crossref]
226. Saurabh Brajesh. Big Data Analytics in Retail Supply Chain 269-289. [Crossref]
227. Jenna K. Simandl, Andrew J. Graettinger, Randy K. Smith, Steven Jones, Timothy E. Barnett. 2016. Making Use of Big Data to Evaluate the Effectiveness of Selective Law Enforcement in Reducing Crashes. *Transportation Research Record: Journal of the Transportation Research Board* **2584**:1, 8-15. [Crossref]
228. Siddhartha Duggirala. Big Data Architecture 315-344. [Crossref]
229. Saurabh Brajesh. Big Data Analytics in Retail Supply Chain 1473-1494. [Crossref]
230. Kurt Stockinger, Thilo Stadelmann, Andreas Ruckstuhl. Data Scientist als Beruf 59-81. [Crossref]
231. Yuriko Yano, Yukari Shirota. SVD and Text Mining Integrated Approach to Measure Effects of Disasters on Japanese Economics 20-29. [Crossref]
232. Christos Alexakos, Konstantinos Arvanitis, Andreas Papalambrou, Thomas Amorgianiotis, George Raptis, Nikolaos Zervos. ERMIS: Extracting Knowledge from Unstructured Big Data for Supporting Business Decision Making 611-622. [Crossref]
233. Tien-Chi Huang, Chieh Hsu, Zih-Jin Ciou. 2015. Systematic Methodology for Excavating Sleeping Beauty Publications and Their Princes from Medical and Biological Engineering Studies. *Journal of Medical and Biological Engineering* **35**:6, 749-758. [Crossref]
234. Pedro Quelhas Brito, Carlos Soares, Sérgio Almeida, Ana Monte, Michel Byvoet. 2015. Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing* **36**, 93-100. [Crossref]
235. Hamid Afshari, Qingjin Peng. 2015. Modeling and quantifying uncertainty in the product design phase for effects of user preference changes. *Industrial Management & Data Systems* **115**:9, 1637-1665. [Crossref]

236. Banage T.G.S. Kumara, Incheon Paik, Jia Zhang, T.H.A.S. Siriweera, Koswatte R.C. Koswatte. Ontology-Based Workflow Generation for Intelligent Big Data Analytics 495-502. [Crossref]
237. Rameshwar Dubey, Angappa Gunasekaran. 2015. Education and training for successful career in Big Data and Business Analytics. *Industrial and Commercial Training* **47**:4, 174-181. [Crossref]
238. Amy Sliva, Joe Gorman, Christopher Bowman, Martin Voshell. Dual node decision wheels: an architecture for interconnected information fusion and decision making 94640F. [Crossref]
239. Seth C. Lewis. 2015. Journalism In An Era Of Big Data. *Digital Journalism* **3**:3, 321-330. [Crossref]
240. Giulio Aliberti, Alessandro Colantonio, Roberto Di Pietro, Riccardo Mariani. 2015. EXPEDITE: EXPress closED ITemset Enumeration. *Expert Systems with Applications* **42**:8, 3933-3944. [Crossref]
241. Ashok Kumar Wahi, Yajulu Medury, Rajnish Kumar Misra. 2015. Big Data. *International Journal of Service Science, Management, Engineering, and Technology* **6**:2, 1-17. [Crossref]
242. Li Kung, Hsiao-Fan Wang. A recommender system for the optimal combination of energy resources with cost-benefit analysis 1-10. [Crossref]
243. Jussi Ronkainen, Antti Iivari. 2015. Designing a Data Management Pipeline for Pervasive Sensor Communication Systems. *Procedia Computer Science* **56**, 183-188. [Crossref]
244. Yi Shen. 2015. Strategic planning for a data-driven, shared-access research enterprise: virginia tech research data assessment and landscape study. *Proceedings of the Association for Information Science and Technology* **52**:1, 1-4. [Crossref]
245. Laura Azzimonti, Marzia A. Cremona, Andrea Ghiglietti, Francesca Ieva, Alessandra Menafoglio, Alessia Pini, Paolo Zanini. BarCamp: Technology Foresight and Statistics for the Future 53-67. [Crossref]
246. Vincenzo Morabito. Managing Change for Big Data Driven Innovation 125-153. [Crossref]
247. Vincenzo Morabito. Big Data and Analytics for Competitive Advantage 3-22. [Crossref]
248. Conrad Boton, Gilles Halin, Sylvain Kubicki, Daniel Forgues. Challenges of Big Data in the Age of Building Information Modeling: A High-Level Conceptual Pipeline 48-56. [Crossref]
249. Jiuyong Li, Sarowar A. Sattar, Muzammil M. Baig, Jixue Liu, Raymond Heatherly, Qiang Tang, Bradley Malin. Methods to Mitigate Risk of Composition Attack in Independent Data Publications 179-200. [Crossref]
250. Shiming Yang, Mary Njoku, Colin F Mackenzie. 2014. 'Big data' approaches to trauma outcome prediction and autonomous resuscitation. *British Journal of Hospital Medicine* **75**:11, 637-641. [Crossref]
251. Stefan Debortoli, Oliver Müller, Jan vom Brocke. 2014. Comparing Business Intelligence and Big Data Skills. *Business & Information Systems Engineering* **6**:5, 289-300. [Crossref]
252. Stefan Debortoli, Oliver Müller, Jan vom Brocke. 2014. Vergleich von Kompetenzanforderungen an Business-Intelligence- und Big-Data-Spezialisten. *WIRTSCHAFTSINFORMATIK* **56**:5, 315-328. [Crossref]
253. Alberto Fernández, Sara del Río, Victoria López, Abdullah Bawakid, María J. del Jesus, José M. Benítez, Francisco Herrera. 2014. Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **4**:5, 380-409. [Crossref]
254. A.H.M. Sarowar Sattar, Jiuyong Li, Jixue Liu, Raymond Heatherly, Bradley Malin. 2014. A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowledge-Based Systems* **67**, 361-372. [Crossref]
255. Provost Foster. 2014. ACM SIGKDD 2014 to be Held August 24–27 in Manhattan. *Big Data* **2**:2, 71-72. [Citation] [Full Text] [PDF] [PDF Plus]
256. Prasanna Tambe. 2014. Big Data Investment, Skills, and Firm Value. *Management Science* **60**:6, 1452-1469. [Crossref]
257. Jay Lee, Hung-An Kao, Shanhu Yang. 2014. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. *Procedia CIRP* **16**, 3-8. [Crossref]
258. Siddhartha Duggirala. Big Data Architecture 129-156. [Crossref]
259. Dylan Walker, Lev Muchnik. 2014. Design of Randomized Experiments in Networks. *SSRN Electronic Journal* . [Crossref]
260. Shujaat Hussain, Byeong Ho Kang, Sungyoung Lee. A Wearable Device-Based Personalized Big Data Analysis Model 236-242. [Crossref]
261. Tiago Cunha, Carlos Soares, Eduarda Mendes Rodrigues. TweeProfiles: Detection of Spatio-temporal Patterns on Twitter 123-136. [Crossref]
262. Matthew A. Waller, Stanley E. Fawcett. 2013. Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics* **34**:2, 77-84. [Crossref]

263. Prasanna Tambe. 2013. Big Data Investment, Skills, and Firm Value. *SSRN Electronic Journal* . [Crossref]

264. Xiuli He, Satyajit Saravane, Qiannong Gu. Supply Chain Analytics 2364-2375. [Crossref]

265. Zhecheng Zhu, Heng Bee Hoon, Kiok-Liang Teow. Interactive Data Visualization Techniques Applied to Healthcare Decision Making 46-59. [Crossref]

266. Zhecheng Zhu, Heng Bee Hoon, Kiok-Liang Teow. Interactive Data Visualization Techniques Applied to Healthcare Decision Making 1157-1171. [Crossref]

267. Murray E. Jennex. Using a Revised Knowledge Pyramid to Redefine Knowledge Management Strategy 1-18. [Crossref]

268. Pratiyush Guleria, Manu Sood. Big Data Analytics 77-108. [Crossref]