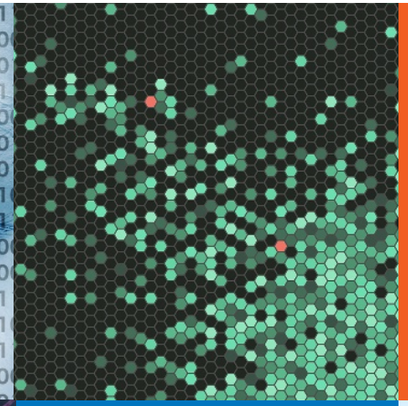
Mary Ann Liebert, Inc. publishers

AN E-BOOK SERIES

Publishers of *Big Data*



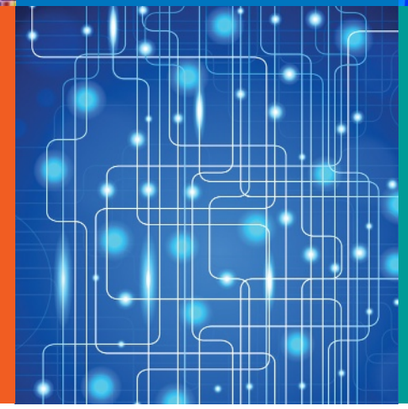
DATA SCIENCE

BIG DATA

PREDICTIVE MODELING

HARNESSING THE POWER OF BIG DATA THROUGH EDUCATION AND DATA-DRIVEN DECISION MAKING

OPEN SOURCE



DATA ANALYTICS

Sponsored by

REVOLUTION ANALYTICS

REVOLUTION ANALYTICS

Bigger. Faster. Better.

Revolution R Enterprise is fast, powerful, and scales to fit your needs.

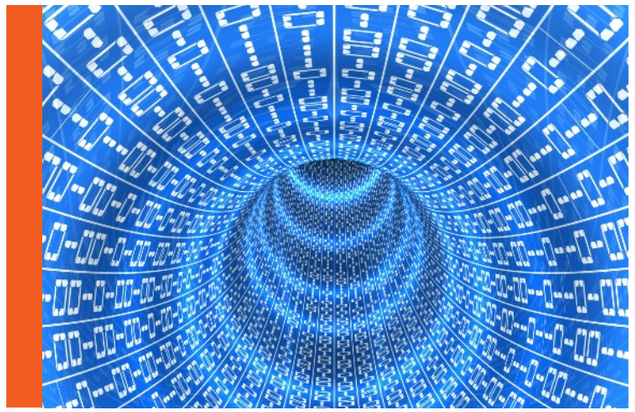
Perform Big Data Analytics at speed **and** at scale. R-based analytics in-Hadoop and in-Database reduce latency and eliminate sampling. RRE is 100% R–get everything you know and love, **plus** innovative Big Data capabilities and management techniques.

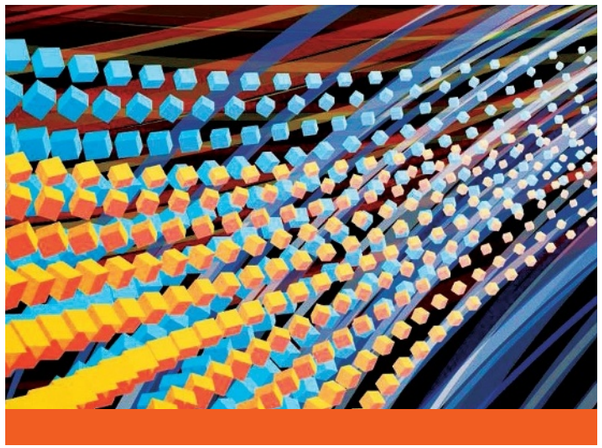
Learn more!

Welcome...

David Smith, Chief Community Officer
Revolution Analytics, @REVODAVID

Contents





Data Science and Its Relationship to Big Data & 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.

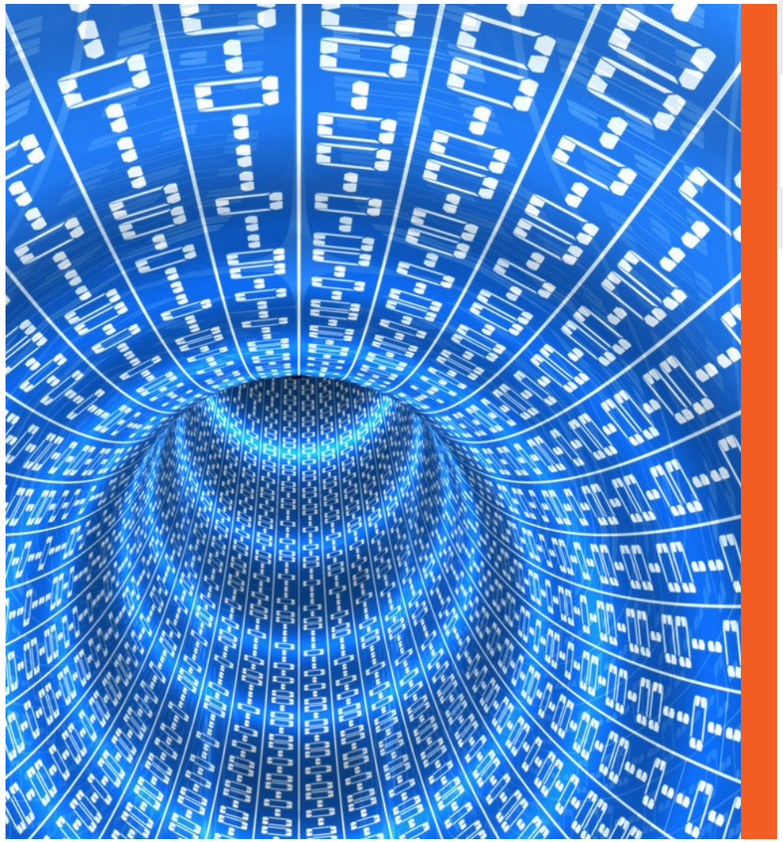
This article is a modification of Chapter 1 of Provost & Fawcett's book,
Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly Media, 2013.

Introduction

With vast amounts of data now available, companies in almost every industry are focused on exploiting data for competitive advantage. The volume and variety of data have far outstripped the capacity of manual analysis, and in some cases have exceeded the capacity of conventional databases. At the same time, computers have become far more powerful, networking is ubiquitous, and algorithms have been developed that can connect datasets to enable broader and deeper analyses than previously possible. The convergence of these phenomena has given rise to the increasingly widespread business application of data science.

Companies across industries have realized that they need to hire more data scientists. Academic institutions are scrambling to put together programs to train data scientists. Publications are touting data science as a hot career choice and even "sexy."[1] However, there is confusion about what exactly is data science, and this confusion could well lead to disillusionment as the concept diffuses into meaningless buzz. In this article, we argue that there are good reasons why it has been hard to pin down what exactly is data science. One reason is that data science is intricately intertwined with other important concepts, like big data and data-driven decision making, which are also growing in importance and attention. Another reason is the natural tendency, in the absence of academic programs to teach one otherwise, to associate what a practitioner actually does with the definition of the practitioner's field; this can result in overlooking the fundamentals of the field.

At the moment, trying to define the boundaries of data science precisely is not of foremost importance. Data-science academic programs are being developed, and in an academic setting we can debate its boundaries. However, in order for data science to serve business effectively, it is important (i) to understand its relationships to these other important and closely related concepts, and (ii) to begin



to understand what are the fundamental principles underlying data science. Once we embrace (ii), we can much better understand and explain exactly what data science has to offer. Furthermore, only once we embrace (ii) should we be comfortable call-

[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.

ing it data *science*.

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

Data Science

At a high level, *data science* is a set of fundamental principles that support and guide the principled extraction of 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.

"Publications are touting data science as a hot career choice and even sexy."

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.

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]

> "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."

Consider a second, more typical business scenario and how it might be treated from a data

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

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

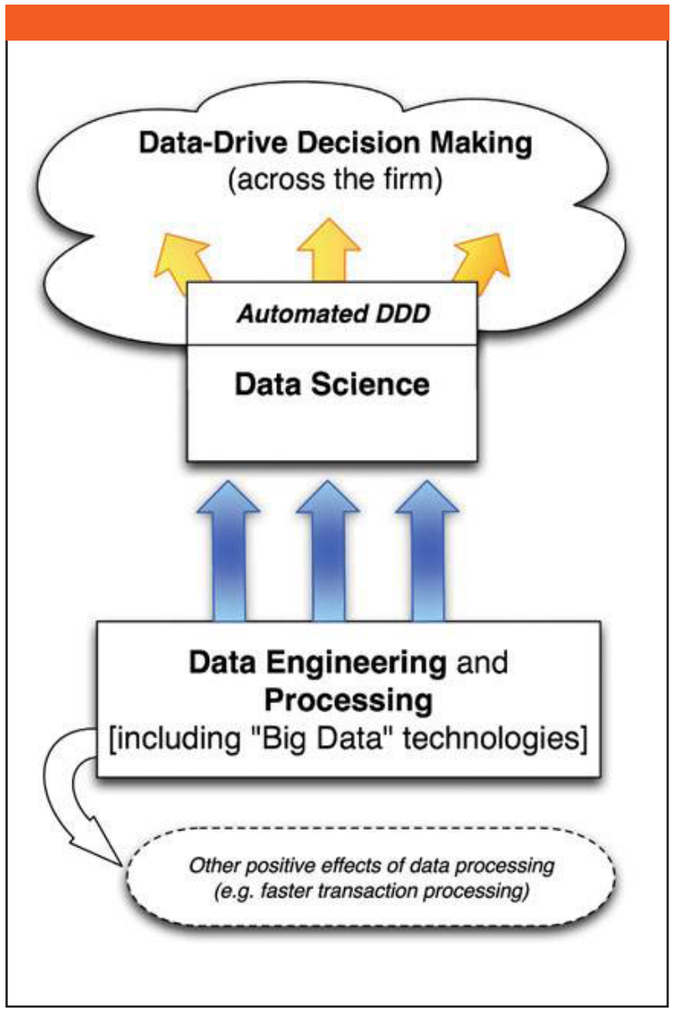


Figure 1. Data science in the context of closely related processes in the organization.

long experience in the field and her eye for what will work. Or, she could base her selection on the analysis of data regarding how consumers react to different ads. She could also use a combination of these approaches. DDD is not an all-or-nothing practice, and different firms engage in DDD to greater or lesser degrees.

The benefits of data-driven decision making have been demonstrated conclusively. Economist Erik Brynjolfsson and his colleagues from MIT and Penn's Wharton School recently conducted a study of how DDD affects firm performance.[3] They developed a measure of DDD that rates firms as to how strongly they use data to make decisions across the company. They show statistically that the more data-driven a firm is, the more productive it is—even controlling for a wide range of possible confounding factors. And the differences are not small: one standard deviation higher on the DDD scale is associated with a 4–6% increase in productivity. DDD also is correlated with higher return on assets, return on equity, asset utilization, and market value, and the relationship seems to be causal.

Our two example case studies illustrate two different sorts of decisions: (1) decisions for which "discoveries" need to be made within data, and (2) decisions that repeat, especially at massive scale, and so decision making can benefit from even small increases in accuracy based on data analysis. The Wal-Mart example above illustrates a type-1 problem. Linda Dillman would like to discover knowledge that will help Wal-Mart prepare for Hurricane Frances's imminent arrival. Our churn example illustrates a type-2 DDD problem. A large telecommunications company may have hundreds of millions of customers, each a candidate for defection. Tens of millions of customers have contracts expiring each month, so each one of them has an increased likelihood of defection in the near future. If we can improve our ability to estimate, for a given customer, how profitable it would be for us to focus on her, we can potentially reap large benefits by applying this ability to the millions of customers in the population. This same logic applies to many of the areas where we have seen the most intense application of data science and data mining: direct marketing, online advertising, credit scoring, financial trading, help-desk management, fraud detection, search ranking, product recommendation, and so on.

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'

"The benefits of data-driven decision making have been demonstrated conclusively."

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.

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. 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.

"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."

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 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.

> "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."

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]

Some Fundamental Concepts of Data Science

There is a set of well-studied, fundamental concepts

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

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.

"Without academic programs defining the field for us, we need to define the field for ourselves."

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*

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

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 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.

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

This article is a modification of Chapter 1 of Provost & Fawcett's book, *Data Science for Business: What you need to know about data mining and data-analytic thinking*, O'Reilly Media, 2013.

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 West 4th Street, 8th Floor
New York, NY 10012

E-mail:

fprovost@stern.nyu.edu

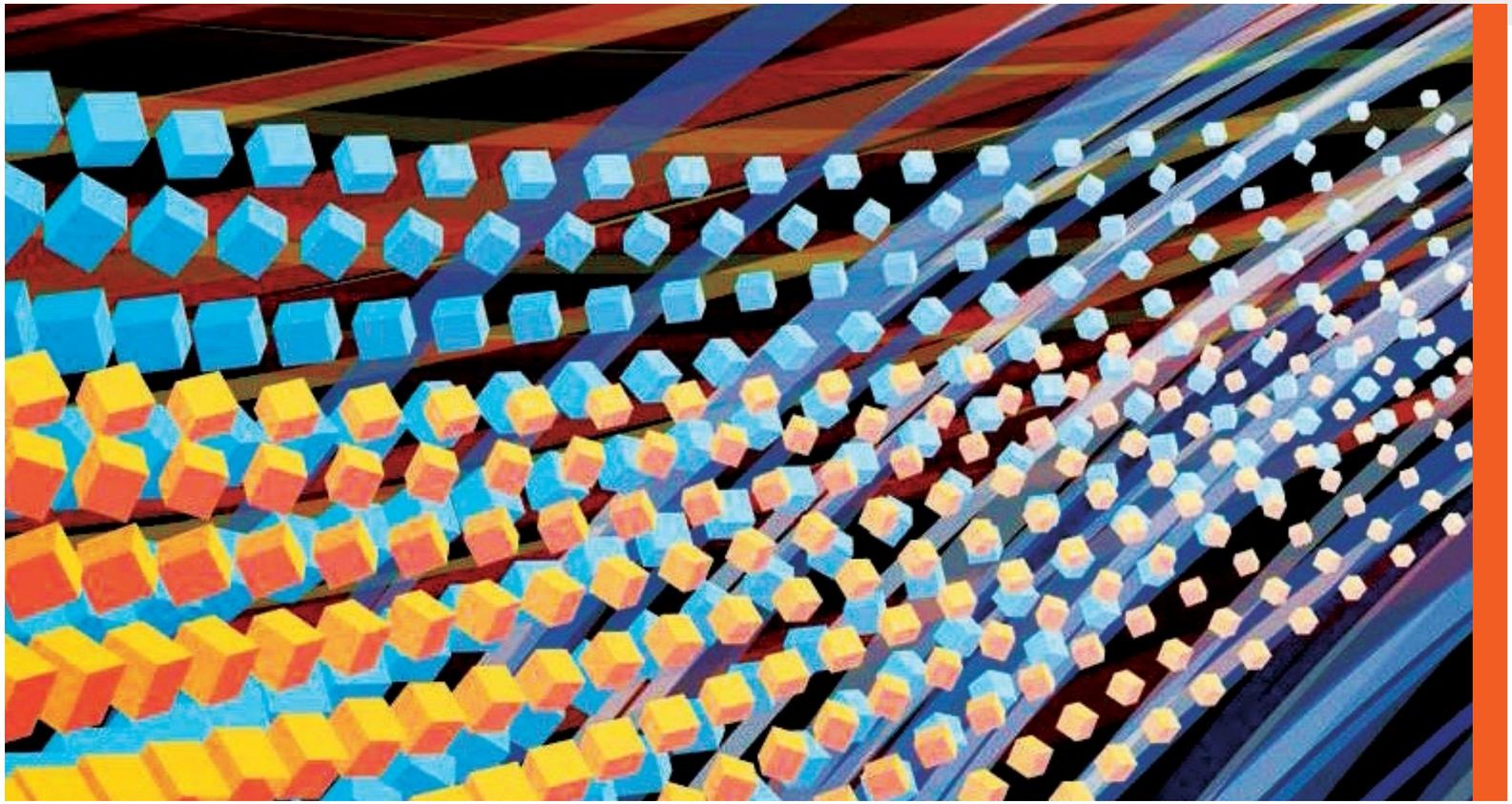
Predictive Modeling With Big Data: Is Bigger Really Better?

Enric Junqué de Fortuny,[†] David Martens,[†] and Foster Provost[‡]

Abstract

With the increasingly widespread collection and processing of "big data," there is natural interest in using thesedata assets to improve decision making. One of the best understood ways to use data to improve decision making is via predictive analytics. An important, open question is: to what extent do larger data actually lead to better predictive models? In this article we empirically demonstrate that when predictive models are built from sparse, fine-grained data—such as data on low-level human behavior—we continue to see marginal increases in predictive performance even to very large scale. The empirical results are based on data drawn from nine different predictive modeling applications, from book reviews to banking transactions. This study provides a clear illustration that larger data indeed can be more valuable assets for predictive analytics. This implies that institutions with larger data assets—plus the skill to take advantage of them—potentially can obtain substantial competitive advantage over institutions without such access or skill. Moreover, the results suggest that it is worthwhile for companies with access to such fine-grained data, in the context of a key predictive task, to gather both more data instances and more possible data features. As an additional contribution, we introduce an implementation of the multivariate Bernoulli Naïve Bayes algorithm that can scale to massive, sparse data.

Introduction



One of the exciting opportunities presented by the proliferation of big data architectures is the ability to conduct predictive analytics based on massive data. However, an important and open question is whether and when massive data actually will improve predictive modeling. Authors have argued for decades for the need to scale up predictive modeling algorithms to massive data.[1,2] However, there is surprisingly scant empirical evidence supporting continued scaling up to modern conceptions of "massive data." Increases in computing power and memory, as well as improved algorithms and algorithmic understanding allow us now to build models from very large data sets. It is striking how quaint the massive datasets of yesteryear look today.

In this paper we will focus on a particular sort of data as being a reason for continued attention to scaling up. To highlight this sort of data, it is worthwhile to draw a contrast. Predictive modeling is based on one or more data *instances* for which we want to predict the value of a *target* variable. Data-driven predictive modeling generally induces a model from *training data*, for which the value of the target (the *label*) is known. These instances typically are described by a vector of *features* from which the predictions will be made. This is the setting we will consider. Now, the most common, tried-and-true

† Applied Data Mining Research Group, Department of Engineering Management, University of Antwerp, Antwerp, Belgium.
‡ Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University, New York, New York.

predictive modeling formulation is to describe these instances with a small-to-moderate number—say, dozens up to hundreds—of features summarizing characteristics of the instances. For lack of a better term, let's call this the "traditional" setting, or more generally "traditional predictive analytics." At the risk of oversimplifying for contrast, a key aspect of the traditional setting is that the dataset is *dense*—each instance has a non-trivial value for each feature (or at least for most of them).

Indeed, for many of the most common business applications of predictive analytics, such as targeted marketing in banking[3] and telecommunications,[4] credit scoring,[5] and attrition management,[6] the data used for predictive analytics are very similar. In these example applications, the features tend to be demographic, geographic, and psychographic characteristics of individuals, as well as statistics summarizing particular behaviors, such as their prior purchase behavior with the firm.[7] Recent work has shown that even with rich descriptive data, traditional predictive analytics may not receive much benefit from increasing data beyond what can currently be handled well on typical modern platforms.[3]

In contrast, big data thinking opens our view to non-traditional data for predictive analytics—datasets in which each data point may incorporate less information, but when taken in aggregate may provide much more.* In this article we focus on one particular sort of data: sparse, fine-grained feature data, such as that derived from the observation of the behaviors of individuals. In the context of behavior data, we can draw a contrast: we specifically do not mean data that are *summaries* of individuals' behaviors, as used traditionally, but data on the actual fine-grained behaviors themselves. Businesses increasingly are taking advantage of such data. For example, data on individuals' visits to massive numbers of specific web pages are used in predictive analytics for targeting online display advertisements.[8–10] Data on individual geographic locations are used for targeting mobile advertisements.[11] Data on the individual merchants with which one transacts are used to target banking advertisements.[3] A key aspect of such datasets is that they are *sparse*: for any given instance, the vast majority of the features have a value of zero or "not present." For example, any given consumer has not transacted with the vast majority of merchants, has not visited the vast majority of geographic locations, has not visited the vast majority of web pages, etc. Predictive modeling based on sparse, finegrained (behavior) data is not a new phenomenon. Individual communications events have been used to improve targeting offers[4] and to predict customer attrition.[6] For two decades, fraud detection has been a popular application of predictive modeling,[12] where fine-grained behavioral

> "Certain telling behaviors may not be observed in sufficient numbers without massive data."

*Predictive modeling of properties of text documents is a shining instance of a nontraditional predictive modeling application that has seen moderately broad application. In fact, predictive modeling with text is quite similar to the sort of predictive modeling that is the focus of this article.

data are also employed. Credit card companies, for example, use specific details from payment transaction data, and telco operators use specific locations and particular numbers called.

Despite research and development specifically geared toward sparse, high-dimensional data[12–14] we do not yet have a large body of empirical results helping us to understand whether and when bigger data indeed are better for improving the generalization performance of predictive modeling for real-world applications. Some hints can be found in the literature. Perlich *et al.*[15] show increasing returns in generalization performance with increasingly larger data across dozens of datasets; however, although large for empirical machine learning studies at the time, the data sizes are not massive by present standards. Their largest dataset, which indeed shows increases to scale for over a million instances, is in fact a sparse (text) classification problem. Moreover, for the sort of linear models often used in business applications (viz., logistic regression), their results largely show the contrary: that massive data by and large do not provide additional improvements in predictive performance. On the other hand, Agarwal *et al.*[2] observe a noticeable drop in the performance of their proposed learning method after subsampling an online advertising dataset, leading to their conclusion that the entire training dataset is needed to achieve optimal performance. Martens and Provost[3] show similar results for predicting customer interest in new financial products. These two datasets are sparse.

Looking from a different angle, certain telling behaviors may not even be observed in sufficient numbers without massive data.[16] If in aggregate such rare-but-important behaviors make up a substantial portion of the data, due to a heavy tail of the behavior distribution, then again massive data may improve predictive modeling. Our experimental results below show that this indeed is the case for the sparse, fine-grained data we study.

This article makes two main contributions. The primary contribution is an empirical demonstration that indeed when predictive models are built from such sparse, fine-grained data, we continue to see-marginal increases in predictive performance even to massive scale. The empirical results are based on data drawn from nine different predictive modeling applications.

The second contribution is a technical contribution: We introduce a version of Naïve Bayes with a multivariate event model that can scale up efficiently to massive, sparse datasets. Specifically, this version of the commonly used multivariate Bernoulli Naïve Bayes only needs to consider the "active" elements of the dataset—those that are present or non-zero—which can be a tiny fraction of the elements in the matrix for massive, sparse data. This means that predictive modelers wanting to work with the very convenient Naïve Bayes algorithm are not forced to use the multinomial event model simply because it is more scalable. This article thereby makes a small but important addition to the cumulative answer to a current open research question[17]: How can we learn predictive models from lots of data?

Note that our use of Naïve Bayes should not be interpreted as a claim that Naïve Bayes is by any means the best modeling technique for these data. Other methods exist that handle large transactional datasets, such as the popular Vowpal Wabbit software

based on scalable stochastic gradient descent and input hashing.[2,18,19] Moreover, results based on Naïve Bayes are conservative. As one would expect theoretically[20] and as shown empirically,[15] nonlinear modeling and less-restrictive linear modeling generally will show continued improvements in predictive performance for much larger datasets than will Naïve Bayes modeling. (However, how to conduct robust, effective nonlinear modeling with massive high-dimensional data is still an open question.) Nevertheless, Naïve Bayes is popular and quite robust. Using it provides a clear and conservative baseline to demonstrate the point of the article. If we see continued improvements when scaling up Naïve Bayes to massive data, we should expect even greater improvements when scaling up more sophisticated induction algorithms.

These results are important because they help provide some solid empirical grounding to the importance of big data for predictive analytics and highlight a particular sort of data in which predictive analytics is likely to benefit from big data. They also add to the observation[3] that firms (or other entities) with massive data assets[21] may indeed have a considerable competitive advantage over firms with smaller data assets.

Sparse, Fine-Grained (Behavior) Data

As discussed above, it is clear from prior work that more data do not necessarily lead to better predictive performance. It has been argued that sampling (reducing the number of instances) or transformation of the data to lower dimensional spaces (reducing the number of features) is beneficial,[22] whereas others have argued that massive data can lead to lower estimation variance and therefore better predictive performance.[3] The bottom line is that, not unexpectedly, the answer depends on the type of data, the distribution of the signal (the information on the target variable) across the features, as well as the signal-to-noise ratio.

Therefore, we will focus on a certain sort of data: sparse, fine-grained data, such as data created by the detailed behavior of individuals. Such data from different domains have similar characteristics that would lead one to expect increasing benefits with very large data, an expectation that does not come with traditional data for predictive modeling. Modern information systems increasingly are recording fine-grained actions of individuals. As we use our telecommunications devices, make financial transactions, surf the Web, send e-mail, tag online photos, "like" postings, and so on, our behaviors are logged.

When data are drawn from human actions, noise rates often are high because of the ceiling imposed by behavioral reliability.[23] Social scientists have long argued that one way to circumvent the poor predictive validity of attitudes and traits is to aggregate data across occasions, situations, and forms of actions.[24] This provides an early suggestion that more (and more varied) data might indeed be useful when modeling human behavior data. The implication for predictive analytics based on data drawn from human behaviors is that by gathering more data over more behaviors or individuals (aggregated by the modeling), one could indeed hope for better predictions.

To show the actual predictive performance on these types of datasets in a realistic setting, we have gathered a collection of datasets representing such sparse, fine-grained feature information. Most of

the datasets incorporate a task to predict some particular "target" characteristic based on features created from human behavior.

The Yahoo Movies dataset,† for which we are predicting the gender of the user, provides data on which movies each user has rated. In the Book-Crossing dataset,[25] the age of a user is predicted (higher or lower than the average age) based on the books rated. In the Ta-Feng dataset,‡ age is predicted based upon which products are purchased. The Dating site describes user rating profiles on an online dating site,[26] for which we predict the gender of the users. The Flickr dataset also describes behavioral data[27] in the form of users "favoriting" pictures, and we predict whether each picture will be highly commented on or not (more or less than the dataset

† http://webscope.sandbox.yahoo.com/

‡ http://aiia.iis.sinica.edu.tw/

Table 1. Sparse Datasets Considered

Dataset	*Active elements*	*Instances*	*Features*	*Sparseness*	*Target variable*
Yahoo Movies	220,831	7,620	11,914	0.99756753	Gender
Book Crossing	680,194	61,308	282,700	0.999960755	Age
Ta-Feng	723,449	31,640	23,718	0.999035964	Age
Dating	17,359,100	135,358	168,790	0.999240205	Gender
URL	277,058,644	2,396,130	3,230,441	0.999964207	Malicious or not
KDD-A	305,613,510	8,407,752	19,306,082	0.999998117	Correctness of answer
KDD-B	566,345,888	19,264,097	28,875,156	0.999998982	Correctness of answer
Flickr	34,645,468	11,195,143	497,470	0.999993779	Comments (few or many)
Bank	20,914,533	1,204,727	3,139,575	0.99999447	Product interest

The table reports the total number of active (non-zero) elements, the number of data instances, the number of features, the resultant sparseness for a traditional data matrix representation (fraction of matrix elements that are not active), and the target variable.

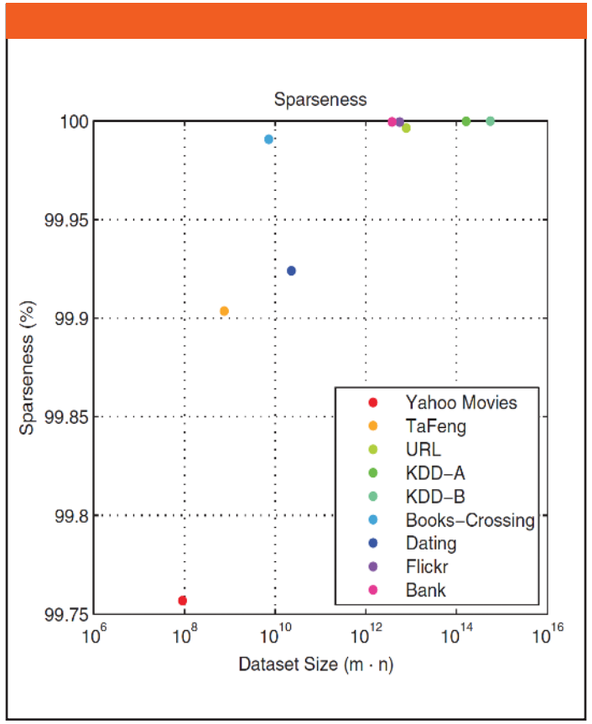


Figure 1. The sparseness is the percent of the entries that are not active (i.e., are zero) in the feature-vector matrix for each dataset. The horizontal axis represents the total size of the feature-vector matrix and the vertical axis shows the sparsensess. The sparseness increases (the density decreases) as the dataset size increases. Note the scale of the vertical axis.

average). Nonbehavioral sparse data are captured in the URL dataset,where for each URL a large set of lexical and host-based features have been defined to predict whether a URL is malicious or not.[28] In the KDD Cup datasets we are predicting the performance of a student on a test, where a large number of binary features have been constructed based on other students answers.[29] Finally, one year of fine-grained payment transaction data is provided in the banking dataset, from consumers to merchants or other persons, to predict interest in a pension fund product (the payment receivers being the features).[3] Size characteristics of these datasets are listed in Table 1. The total number of elements in the traditional instance x feature matrix is the product of n and m (cf. Fig. 1); the "active" elements are those that actually have a value in the data.

As shown in Table 1 and Figure 1, these datasets are quite sparse, and the sparseness increases as the datasets grow bigger. For behavioral data there is a clear explanation as to why the sparseness increases as dataset size increases. People only have a limited amount of "behavioral capital." That is, any person only can take a limited number of actions. Most importantly here, when the total number of possible actions is huge, any individual only can take a very small fraction of all the possible actions. Increasing the dataset size may increase the total number of behaviors observed across all individuals, but increasing the dataset size will not increase an individual's behavioral capital. Concretely, there are only so many movies one can see (Yahoo Movies dataset), products one can buy (Bank dataset), or pictures one can favorite (Flickr dataset). The actual amount depends on the application context, but in every case the amount of behavioral capital is finite and is more or less constant. Therefore, the density of the actions in a traditional dataset formulation, such as the user-byaction matrix, will tend to decrease as the dataset size increases. This sparseness provides a key to efficient predictive modeling from massive data. It also provides a key reason why massive data are crucial to effective predictive modeling in these domains.

Predictive Techniques for Analyzing Fine-Grained Behavior Data

For the experiments we report, we use the Naïve Bayes classifier to learn models and to make predictions. Despite its simplicity, Naïve Bayes has been shown to have surpris-

ingly good performance on sparse datasets. A very attractive property of Naïve Bayes is that it is extremely fast to run on a large, sparse dataset when it is formulated well. The main speedup stems from the fact that Naïve Bayes completely ignores interfeature dependencies by assuming that the within-class covariances of the features are zero. Combining this "naïve" assumption with Bayes' rule for conditional probability produces the following formulation:

$$P(C=c|x_1,\ldots,x_m)=\frac{P(C=c)\cdot P(x_1,\ldots,x_m|C=c)}{P(x_1,\ldots,x_m)}$$
$$\propto P(C=c)\cdot\prod_{j=1}^{m}P(X_j=x_j|C=c),$$

There are a several variants of Naïve Bayes in common use, which differ in their calculation of the conditional probabilities. Each of these variants assumes a different *event model*, which postulates how the features from the data were generated.[30] The two most popular choices are the multivariate and the multinomial event models. The reader is referred to the appendix and to the literature[30] for details, but the main idea behind the *multinomial* event model is that each input sample results from a series of independent draws from the collection of all features. The main advantage of this model is that it requires only computing over those features that have a count that is nonzero. For a sparse dataset, this results in a huge reduction in run time since all the zeros can be ignored. Although results have shown this multinomial Naïve Bayes to be superior to the multivariate event model (described next) for text mining, the multinomial model makes the generative assumption that the features are drawn with replacement, which is not well specified for most binary feature data (although the multinomial model is used commonly in practice for binary data). The multinomial model further makes the implicit assumption that the number of draws in the generative model is independent of the class, which roughly implies in our setting that each instance contains on average the same number of active features (behaviors) per class, which might not be true for the datasets that we are considering. Therefore, it is helpful to consider the multivariate model as well, even though the traditional formulation is not efficient for massive datasets.

"When data are drawn from human actions, noise rates often are high because of the ceiling imposed by behavioral reliability."

The *multivariate* event model takes the point-of-view that the features are generated according to an independent Bernoulli process with probability parameters θ_{ij} for class C (see the appendix and the literature[30] for technical details). Unfortunately, the usual multivariate formulation does not have the same attractive sparsity properties as the multinomial formulation. In the Appendix we present an alternative, equivalent formulation that circumvents the need to consider the massive number of inactive elements for binary classification problems containing only Boolean-valued features (which is the case

in all our datasets). The implementations can be obtained from the authors.**

Bigger Is Better (Volume)

We now can ask: *when using sparse fine-grained (behavior) data as features, do we indeed see increasing predictive performance as the training dataset continues to grow to massive size?* We present the results of experiments examining this question on our datasets. The results are shown in Figure 2.

To determine whether bigger volumes indeed lead to better performance, we built learning curves for each dataset.[15] Learning curves are an important tool for model analytics, as they help an organization decide how to invest in its data assets.[21] In particular, is it worth undertaking the investment to collect, curate, and model from larger and larger datasets? Or will a sample of the data be sufficient for building predictive models.

Specifically, our learning curves are built for each dataset via a series of experiments that simulated different data sizes by sampling (uniformly at random) increasingly larger amounts of data from the original datasets, training a predictive model for each of these datasets, and estimating its generalization performance using standard predictive modeling holdout evaluation (viz., computing averaged areas under the receiver operating characteristic curve via five-fold cross-validation[21]).

As Figure 2 shows, for most of the datasets the performance keeps improving even when we sample more than millions of individuals for training the models. One should note, however, that the curves do seem to show some diminishing returns to scale. This is typical for the relationship between the amount of data and predictive performance, especially for linear models.[15] The marginal increase in

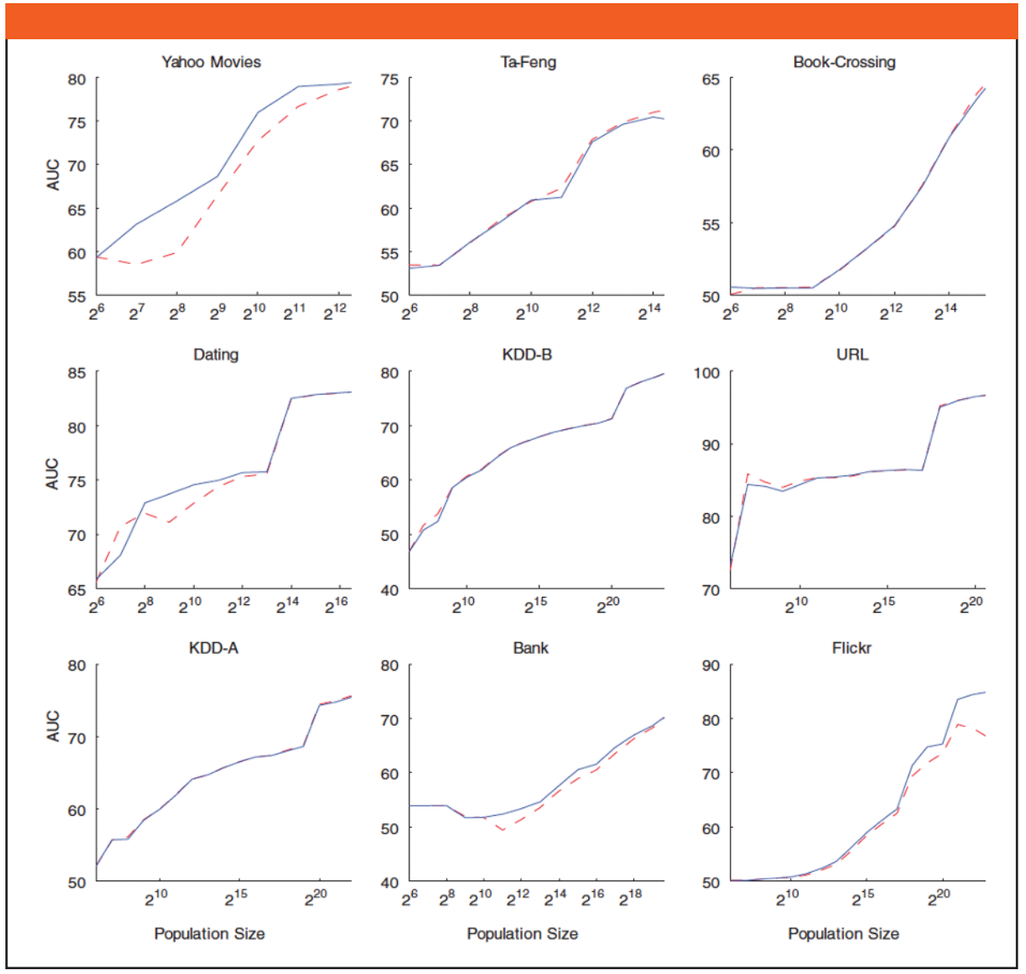


Figure 2. Learning curves for the multivariate (solid, blue) and multinomial (dashed, red) Naïve Bayes variants. For all of the datasets, adding more data (x-axis) leads to continued improvement of predictive power (area under the curve [AUC]).

**www.applieddatamining.com

generalization accuracy decreases with more data for several reasons. First, there simply is a maximum possible predictive performance (the "Bayes rate") due to the inherent randomness in the data and the fact that accuracy can never be better than perfect. Second, modeling (especially linear modeling) tends to find the larger generalities first; modeling with larger datasets usually helps to work out nuances, "small disjuncts," and other nonlinearities that are difficult or impossible to capture from smaller datasets.[1,15]

What is surprising here is that, even for this simple, linear model, a ceiling on generalization accuracy is not reached even after sampling a very large number of data points. We would expect more sophisticated modeling procedures (esp., non-linear methods) to extend the advantages with larger data.[15,20]

Interestingly, the multinomial Naïve Bayes did not perform better than the multivariate model for these datasets, contrary to results reported for text mining.[30] In fact, for these datasets, the multivariate event model sometimes outperformed the multinomial event model. Recall that the multinomial model makes the additional assumption that the numbers of observations made per subject are independent of the subjects' classes. Therefore, our introduction of an efficient big data implementation of multivariate Naïve Bayes is not just of theoretical interest. These results show that it should be included as an alternative in practice.

Fine-Grained Behavior Data Have Massive Feature Sets

As discussed briefly above, fine-grained behavior data often have massive numbers of possible features that predictive modeling may incorporate. In these domains, there simply are a very large number of different possible behaviors. When examining the importance of data size, separate for the moment from the question of predictive modeling, it is interesting to examine whether we need to observe large numbers of individuals even simply to see all the different possible behaviors. The notion of limited behavioral capital would suggest so.

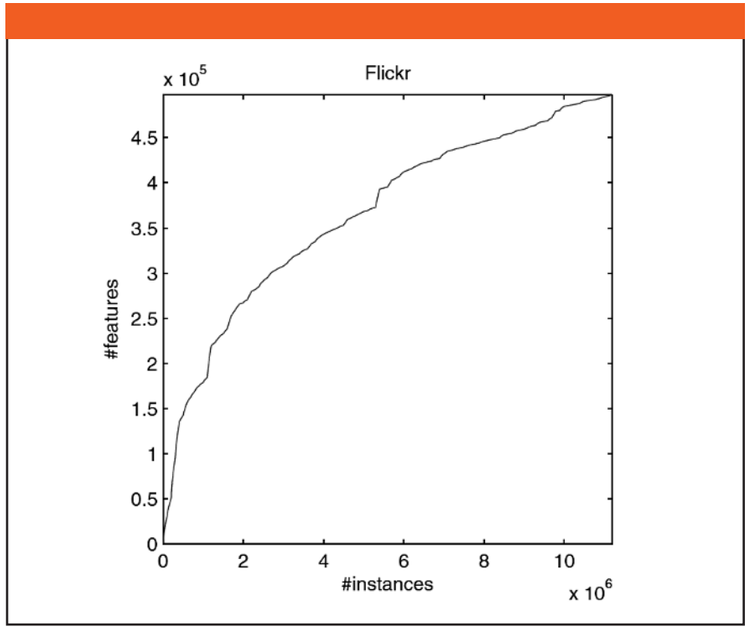


Figure 3. The number of features (behaviors) that have been observed (i.e., are present for at least one data instance), as the number of data instances is increased.

Figure 3 examines this empirically for the Flickr dataset (the other datasets show similar patterns). Figure 3 plots the number of behaviors observed as more and more instances are added to the data (uniformly at random). As we expect, with few data instances, given the very sparse data only a portion of the

behaviors will be revealed. (Realize of course that the behaviors we see here are not the total set of behaviors, but only those revealed by the individuals in our data.)

Bigger Is Better (Variety)

The experiments reported above demonstrate the improvement in predictive modeling accuracy from increasing the number of instances (individuals) observed. An alternative perspective on increasing data size is to consider a fixed population of individuals and increase the number of different possible data items (features, signals, behaviors) about each of them that are collected, stored, or utilized.

We would like to believe that indeed, more variety in the features ought to lead to richer models with better predictive ability. Intuitively it might make sense that adding more features to our total set of possible features would increase predictive performance, but this seems contrary to the long tradition of (and vast amount of research on) "feature selection" for predictive modeling. Therefore, we should think carefully about what are the deeper motivations for feature selection and whether we believe that they indeed apply in our context.

Feature selection is employed for at least two interrelated reasons. First, technically, larger feature sets lead to higher variance in predictive modeling[31] and relatedly increase overfitting,[21] and thereby can increase prediction error. Second, in many applications there are only a few important features among a large number of irrelevant features. The large number of irrelevant features simply increases variance and the opportunity to overfit, without the balancing opportunity of learning better models (presuming that one can actually select the right subset). Domains with this structure (many irrelevant variables) are the motivation behind the development of many feature selection techniques, both explicit (e.g., filter and wrapper methods[22]) and implicit (e.g., via L1 regularization[31]).

"Is it worth undertaking the investment to collect, curate, and model from larger and larger datasets?"

This line of reasoning does not provide such clear motivation in domains with a different "relevance structure" to the features. If most of the features provide a small—but nonzero—amount of additional information about the target prediction, then selecting a small set of features is not ideal. (If one has only a small number of instances, feature selection may be practically necessary, but that is a different story.) When presented with many low-signal but possibly relevant features, when there also are many instances, predictive modeling may benefit from a large number of features.††

We can argue intuitively that fine-grained be-

†† One might observe that instead of feature selection, when faced with such problems, one should engage in more sophisticated dimensionality reduction, such as via methods for matrix factorization. This may indeed allow the construction of predictive models with fewer direct features. However, in the context of our current discussion, this begs the question—one must process all the relevant features in the dimensionality reduction, and one must process all the features to create the values along the reduced dimensions in order to apply the resultant model(s).

haviors ought to provide such underlying relevance structures: If the target prediction is related to the same internal factors that drive the fine-grained behaviors, each should provide a little more information toward predicting the target concept. For example, if we are predicting the likelihood that an individual will respond positively to an advertisement from a particular brand or an offer for a particular product, that individual's taste is a key internal factor. If we have created features based on other signals of consumer taste, such as the web pages or the merchants that an individual frequents, each of these features may have a small amount of additional information and in total may give better predictions than simply choosing a small set of the best signals.

Our goal in this article is not to determine how best to model these domains, as there are a tremendous number of factors to consider, nor how best to perform feature selection. Nevertheless, as a follow-up analysis to our main results, let us consider how easy it is in our set of domains to approximate the true underlying knowledge based on a substantially smaller dataset via feature selection. If we can do this, it would mean that most of the features actually do not contain any meaningful information (or at least none that our methods can find). As a first, simple analysis, we conducted an experiment in which we repeatedly sampled increasing numbers of features (uniformly at random) from the entire feature set and learned a predictive model from the reduced representation. The result (for the Flickr dataset) is shown in Figure 4 (upper graph), where the predictive performance is shown for different numbers of features (over multiple repetitions of the experiment for stability in the results). The graph clearly shows that including random subsets of the features results in a disastrous performance drop.

"We would like to believe that indeed, more variety in the features ought to lead to richer models with better predictive ability."

The foregoing result, selecting features at random, provides a lower bound. One cannot simply use a random subset of the features. This shows that for these datasets, the behavior data do not contain a large amount of overlapping information that is predictive of the target. However, it still may be the case that there is a small set of informative behaviors, and that those are all we need to build accurate predictive models.

As alluded to above, there exist many ways to perform intelligent feature or "input" selection. Our purpose is not to find the best feature selection methods for these data, or to perform a large comparative study of feature selection. (Moreover, many of the classic feature selection techniques do not work on data of this magnitude in a reasonable amount of time.)

Ideally we would like to complement the lower bound provided by the random selection above with an approximate upper bound. Specifically, if we knew a priori which small set of features mattered (e.g., if they were provided by an omniscient oracle), how

close would be the predictive accuracy to that obtained using all the features? Unfortunately, conducting such a study is far beyond the scope of this article. Optimal feature selection itself is a big data problem in such domains; further, using Naïve Bayes we would not be able to determine a true upper bound on what would be possible using more sophisticated, non-linear modeling (cf., the striking differences in the learning curves for the linear and non-linear models shown by Perlich *et al.*[15])

Nonetheless, it is useful to examine the results of a moderately intelligent selection of feature subsets, tailored to the modeling we perform. Specifically, given that we have learned Naïve Bayes models from all the data already, we can approximate the "best" feature set for each dataset by choosing the features which have the largest a posteriori marginal correlations with the target—in our case, the a posteriori marginal probabilities in each Naïve Bayes model. Let's call this Naïve Bayes feature selection.

As can be seen from the resulting graph on the same dataset (Flickr in Fig. 4, lower graph), the results are much better with Naïve Bayes feature selection. Nevertheless, despite this improved performance, this approximate upper bound still is substantially below the performance with the full feature set, up until the reduced feature set is itself massive (in this case, 100K features instead of 400K features).

Returning to the main contribution of the article, the results in Figure 4 reiterate that we cannot easily achieve top-notch predictive performance without massive numbers of instances. Thus, even when reducing the feature set somewhat, bigger still is better.

Final Words

Our main contribution is to provide empirical support that when mining sparse,

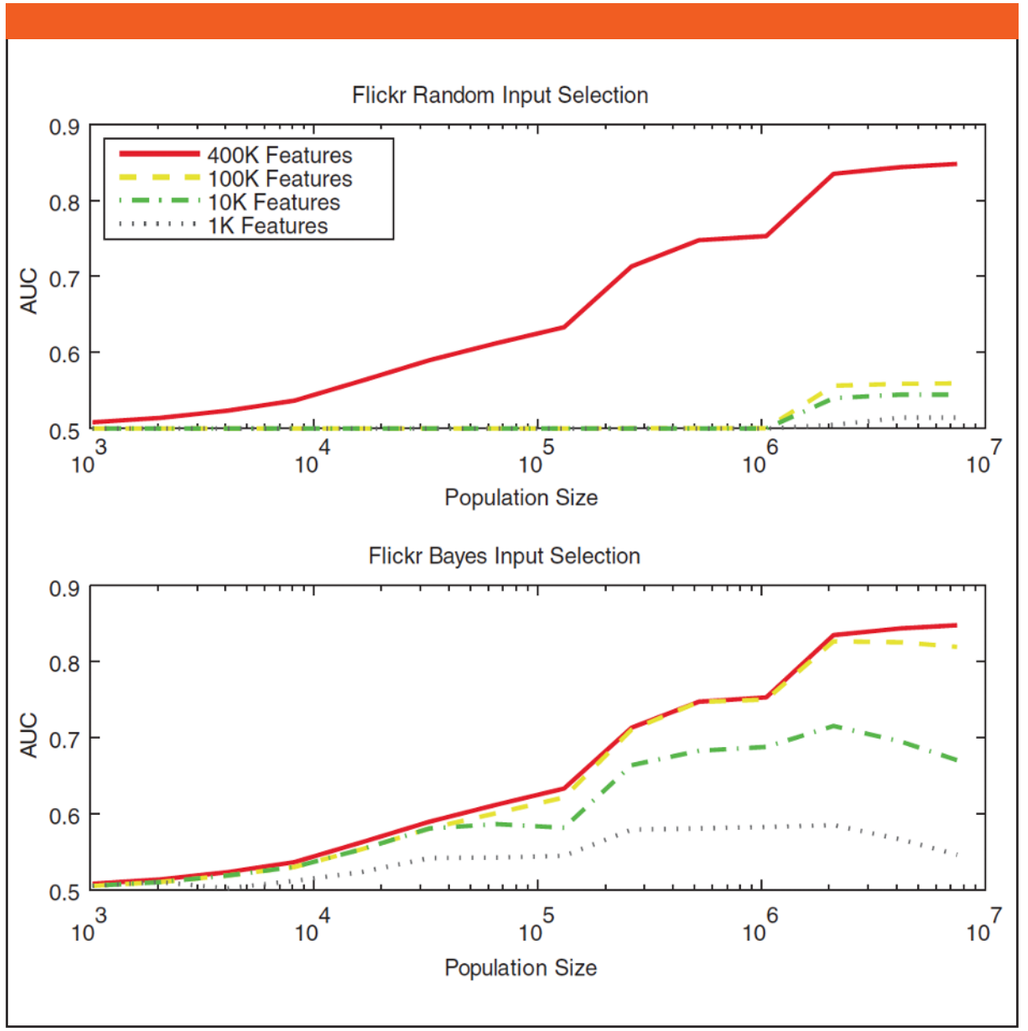


Figure 4. Learning curves with increasing numbers of features: sampled randomly (top) and using a posteriori information (bottom).

fine-grained data, we do indeed continue to see increasing predictive performance from more data as the datasets become massive. This holds even for simple, linear models; the increases with larger data should be even greater as technical and theoretical advances allow us to build effective non-linear models from massive, sparse, noisy data. As a second, technical contribution, we provide an implementation of multivariate Naïve Bayes that can mine massive, sparse data extremely efficiently.

These results are not only of interest to pundits preaching about the value of big data, but also have important implications for businesses and other institutions faced with competition. We have shown that predictive modeling with large, transactional data can be made substantially more accurate by increasing the data size to a massive scale. This provides one of the clearest illustrations (of which we are aware) that larger data assets are indeed potentially more valuable[21] when it comes to predictive analytics. This implies that institutions with larger data assets—plus the ability to take advantage of them—potentially can obtain substantial competitive advantage over institutions without access to so much data, or without the ability to handle it.

Compare our results with the results of the study conducted by economist Prasanna Tambe of NYU's Stern School, which examined the extent to which *big data* technologies seem to help firms.[32] As described by Provost and Fawcett,[21] Tambe finds that, after controlling for various possible confounding factors, using big data technologies is associated 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.

Our results suggest that it is worthwhile for companies with access to such fine-grained data, in the context of a key predictive task, to gather data both on more data instances and on more possible features. For example, for an online advertising company that uses web page visitation data to target online ads, it is worthwhile to invest in additional data to increase the number of features—to see even more web pages a user visits. Such an online advertising company may also decide to diversify the sort of behavioral signals it collects, for example, by acquiring a mobile targeting company and collecting behavior on location visitations in addition to web page visitations.[33] It is worthwhile for two or more smaller banks to consider pooling their customer transaction data for better targeting, credit scoring, or fraud detection (and dealing with the obvious privacy concerns of such an endeavor), if each is faced with a large competitor. Similarly, it may be beneficial for two telco operators to pool their location visitation data for improved fraud detection. Another example implication is that a social network site with more users and more "likes" or "checkins" has a significant competitive advantage when it comes to using predictive modeling to target ads.

It is important not to interpret these results as indicating that all one needs to do is to apply Naïve Bayes to sparse, finegrained data and one will

be competitive in predictive analytics. These results are a demonstration that for predictive analytics with sparse, fine-grained data, even with simple models it indeed can be valuable to scale up to millions of instances and millions of features. Importantly, we have not shown that the benefit of increasing data size is unbounded. Nonetheless, looking forward, increases in value may continue for data that are orders of magnitude larger, especially as our ability to build non-linear models with such data improves. On the other hand, the evidence provided in this article notwithstanding, it is important for organizations to conduct their own analyses of whether and how the value of their data assets scales to massive data,[21] for example by building problem- and model-specific learning curves.

It also is important to keep in mind that achieving business value from massive, sparse, fine-grained data can be quite complicated (see, e.g., Perlich *et al.*[9]). Developing a worldclass data science team is essential, and even the best data scientists in the world do not yet know how best to build complex predictive models from this sort of massive data.[17,19]

Acknowledgments

Our sincere thanks to Brian Dalessandro, Tom Fawcett, and John Langford for their helpful comments on a prior draft manuscript; these improved the article substantially. Our thanks also to Brian, Tom, John, Josh Attenberg, Jessica Clark, Claudia Perlich, and Marija Stankova for many discussions of predictive analytics from sparse, massive-dimensional data. Foster Provost thanks NEC for a Faculty Fellowship.

Author Disclosure Statement

F.P. is the author of the book, *Data Science for Business*.

References

1. Provost F, Kolluri V. A survey of methods for scaling up inductive algorithms. *Data Mining Knowledge Discov* 1999; 3:131–169.
2. Agarwal A, Chapelle O, Dudik M, Langford J. A reliable effective terascale linear learning system. 2011; ar-Xiv: 1110.4198.
3. Martens D, Provost F. Pseudo-social network targeting from consumer transaction data. Working paper CeDER-11-05. Stern School of Business, New York University, 2011.
4. Hill S, Provost F, Volinsky C. Network-based marketing: Identifying likely adopters via consumer networks. *Stat Sci* 2006; 22:256–276.
5. Martens D, Baesens B, Van Gestel T, Vanthienen J. Comprehensible credit scoring models using rule extraction from support vector machines. *Eur J Oper Res* 2007; 183:1466–1476.
6. Verbeke W, Martens D, Baesens B. Social network analysis for customer churn prediction. *Appl Soft Comput* 2013 [in press].
7. Fader PS, Hardie BGS, Ka Lok L. RFM and CLV: Using iso-value curves for customer base analysis. *J Mark Res* 2005; 42:415–430.
8. Provost F, Dalessandro B, Hook R, Zhang X, Murray A. Audience selection for on-line brand advertising: privacyfriendly social network targeting. In: *KDD'09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining. New York: ACM, 2009, pp. 707–716.

9. Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning* 2013. Published online; to appear in print. DOI 10.1007/ s10994-013–5375-2.
10. Stitelman O, Perlich C, Dalessandro B, Hook R, Raeder T, Provost F. Using co-visitation networks for detecting large scale online display advertising exchange fraud. *KDD'13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2013:1240–1248.
11. Provost F. Geo-social targeting for privacy-friendly mobile advertising: Position paper. Working paper CeDER- 11-06A. Stern School of Business, New York University, 2011.
12. Fawcett T, Provost F. Adaptive fraud detection. *Data Mining Knowledge Discov* 1997; 1:291–316.
13. Perlich, C, Provost F. Distribution-based aggregation for relational learning from identifier attributes. *Machine Learning* 2006; 62(1/2):65–105.
14. Weinberger K, Dasgupta A, Langford J, Smola A, Attenberg J. Feature hashing for large scale multi-task learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1113–1120.
15. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: A learning-curve analysis. *JMLR* 2003; 4:211–255.
16. Provost F, Aronis J. Scaling up inductive learning with massive parallelism. *Machine Learning* 1996; 23:33–46.
17. Langford J, Ortega R. Machine learning and algorithms: Agile development. *Commun ACM* 2012; 55:10–11.
18. Langford J, Li L, Strehl A. 2007. Vowpal Wabbit. http:// hunch.net/*vw/
19. Dalessandro B. Bring the noise: Embracing randomness is the key to scaling up machine learning algorithms. *Big Data J* 2013; 1:105–109.
20. Ng A, Jordan M. On discriminative vs. generative classifiers: A comparison of logistic regression and Naïve Bayes. *Adv Neural Inf Process Syst* 2002; 14:841.
21. Provost F, Fawcett T. *Data Science for Business—What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013.
22. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell* 1997; 97:273–324.
23. Ajzen I. The theory of planned behavior. *Theor Cogn Self Regul* 1991; 50:179–211.
24. Fishbein M, Ajzen I. Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychol Rev* 1974; 81:59–74.
25. Ziegler C-N, McNee S, Konstan JA, Lausen G. Improving recommendation lists through topic diversification. In: *Proceedings of the 14th International Conference on World Wide Web*. New York: ACM, 2005, pp. 22–32.
26. Brozovsky L, Petricek V. *Recommender sytem for online dating service*. 2007; http://arXiv:cs/0703042v1.
27. Cha M, Mislove A, Gummadi KP. A measurement-driven analysis of information propagation in the Flickr social network. In: *WWW'09 Proceedings of the 18th International Conference on World Wide Web*, 2009.

28. Ma J, Saul LK, Savage S, Voelker GM. Identifying suspicious URLs: An application of large-scale online learning. In: *ICML'09 Proceedings of the 26th Annual International Conference on Machine Learning,* 2009, pp. 681–688.

29. Yu HF, Lo HY, Hsieh HP. Feature engineering and classifier ensemble for KDD cup 2010. *JMLR Workshop and Conference Proceedings,* 2011.

30. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. *AAAI Workshop on Learning for Text Categorization,* 1998.

31. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer, 2009.

32. Tambe, P. Big Data Investment, Skills, and Firm Value (May 8, 2013). *Management Science*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=2294077 or http:// dx.doi.org/10.2139/ssrn.2294077

33. Gray T. 2013. Dstillery is Picasso in the dark art of digital advertising. *Fast Company,* September 16, 2013. www.fastcompany.com/3017495/dstillery-is-picasso-in-thedark- art-of-digital-advertising. (Last accessed on October 1, 2013).

34. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli 2004*; 10:989–1010.

Address correspondence to:

Enric Junqué de Fortuny
Applied Data Mining Research Group
Department of Engineering Management
Faculty of Applied Economics
University of Antwerp
Prinsstraat 13
2000 Antwerp
Belgium

E-mail:

enric.junquedefortuny@uantwerpen.be

Appendix: Multivariate Bernoulli Naïve Bayes for Big Data

In the case of classification of high-dimensional sparse data, we can achieve computational time reductions for Naïve Bayes by reformulating the calculation all of the probabilities in terms of only the *active* (non-zero) elements of the data matrix. Specifically, let us denote the *i*-th instance in the dataset as x_i and the *j*-th feature of the *i*-th instance as $x_{i'j}$. An element in this data matrix is said to be inactive if its value is equal to zero $(x_{i'j} = 0)$ and active otherwise. For the sparsedata problems (with binary features) on which this paper focuses, the average number of active elements per instance $|x|$ is much smaller than the total number of features *m*. Thus, it makes sense to use a sparse formulation.

The Multivariate Bernoulli event model assumes that the data are generated according to a Bernoulli process. This is in fact a compound assumption consisting of the following three components:

1. The problem is binary classification.
2. The data are generated i.i.d.

3. The features can only take on a distinct number of discrete levels (e.g. 0 and 1). We will assume they are binary.

These assumptions enable very fast calculation of the empirical probabilities encountered when computing the class probability of an instance. If the data were generated by independent Bernoulli processes with probability parameters θ_j for class *C*, we ought to see that:

$$P(\mathbf{x}_i|C=c) = \prod_{j=1}^{m} P(X_j = x_{i,j}|C=c)^{x_{i,j}} \cdot (1 - P(X_j = x_{i,j}|C=c))^{(1-x_{i,j})}$$
$$= \prod_{j=1}^{m} (\theta_j)^{x_{i,j}} (1-\theta_j)^{(1-x_{i,j})}$$

We can then determine the log likelihood of the combination of all of the probabilities gathered in probability vector θ, given all of the training instances *x* as:

$$L(\theta) = \log P(\mathbf{x}|C=c)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j} \log \theta_j + (1-x_{i,j}) \log(1-\theta_j)$$

Where $L(\theta)$ represents the log likelihood and $x_{i,j}$ again represents the value of feature *j* of training instance *i*. Maximizing the log likelihood yields:

$$\hat{\theta}_{X_j=1|C=c} = P(X_j = 1|C=c;\theta_c)$$
$$= \frac{|X_j = 1 \wedge C = c|}{|C=c|}$$

For computational reasons, this is usually reformulated as:

$$\hat{\theta}_{X_j=1|C=c} = \frac{1 + |X_j = 1 \wedge C = c|}{2 + |C=c|}$$

This corresponds to assuming a Laplacian prior on the probability in a Bayesian setting. Similarly, we can derive that:

$$\hat{\theta}_c = P(C=c) = \frac{|C=c|}{n}$$

This leads to the most common formulation of the Naïve Bayes classifier.

The Bernoulli event model for sparse data: For the Bernoulli event model, a sparse formulation is also possible by expanding the summands in the calculation of the log likelihood of an input vector into two parts. This is possible due to the assumption that a feature value X_j can only take on two values (either one or zero). The log conditional probability of the instance x_i to belong to class *C*, given its values for the binary features then becomes:

$$\log P(\mathbf{x}_i|C) = \log\left(\prod_{j=1}^{m} P(X_j = x_{i,j}|C)\right)$$
$$= \sum_{j=1}^{m} \log\left(P(X_j = x_{i,j}|C)\right)$$
$$= \sum_{j|x_{i,j}=1} \log\left(P(X_j = 1|C)\right) + \sum_{j|x_{i,j}=0} \log\left(P(X_j = 0|C)\right)$$
$$= \sum_{j|x_{i,j}=1} \log\left(P(X_j = 1|C)\right) + \sum_{j|x_{i,j}=0} \log\left[P(C) - P(X_j = 1|C)\right]$$

Revealing the following log likelihood for an instance xi:

$$\log P(C|\mathbf{x}) \propto \log P(C) + \sum_{j|x_{i,j}=1} \log\left(P(X_j = 1|C)\right) + \sum_{j|x_{i,j}=0} \log\left[P(C) - P(X_j = 1|C)\right]$$

With:

$$\sum_{j|x_{i,j}=0} \log\left[P(C) - P(X_j = 1|C)\right] = \sum_{j=0}^{m} \log\left[P(C) - P(X_j = 1|C)\right] - \sum_{j|x_{i,j}=1} \left[\log P(C) - P(X_j = 1|C)\right]$$

For a dataset with $\overline{m}$ active elements per instance, this formulation of log likelihood only needs $O(\overline{m} \cdot n)$ amortized time to finish for n instances (as opposed to $O(m \cdot n)$ for the original formulation) under the assumption that m is of the same order as *n*. This leads to a reduction in computational time proportional to $p = m/\overline{m}$, which increases as the sparseness of the dataset increases.

Note that a separate problem with computing and using Naïve Bayes estimates in practice is that often the exponentials to be calculated to recover the probabilities are so small that they cannot be represented by a computer using double precision. One example of avoiding the exponentials is to compare the ratio of the probabilities instead of the actual probabilities,[34] computing a score such as:

$$S(\mathbf{x}) = \frac{P(C = 1|\mathbf{x})}{P(C = 0|\mathbf{x})} \propto \log P(C = 0|\mathbf{x}) - \log P(C = 1|\mathbf{x})$$

This value can be interpreted as being proportional to Kononenko's information gain of choosing one class over the other. In summary, this implementation of multivariate Bernoulli Naïve Bayes under the binary assumptions can be calculated very fast as they only require one pass over each of the *active* elements of the instances for training as well as in use.

Educating the Next Generation of Data Scientists

Moderator: Edd Dumbill[1], Founding Editor, Big Data
Participants: Elizabeth D. Liddy,[2] Jeffrey Stanton,[2] Kate Mueller,[2] and Shelly Farnham [3]

Big Data *approached the School of Information Studies (iSchool) at Syracuse University to discuss the issue of big data education through an expert roundtable discussion. The participants discuss big data from all sides—from wrestling with the definition and the science behind big data, to how the iSchool is cultivating a full data lifecycle perspective with its on-campus and online program. The importance of industrial collaboration is also emphasized, in order to meet the growing demand for big data analysis. Read on for more, including a link to Syracuse iSchool's free introductory e-book on data science.*

[1] O'Reilly Media, Sebastopol, California.
[2] The School of Information Studies, Syracuse University, Syracuse, New York.
[3] Department of Research, Microsoft, Redmond, Washington.

Edd Dumbill: Let's start with the biggest question about big data that a lot of people are asking, which is: What is it? We have heard a lot this year, but from your point-of-view running this course, how do you break down big data?

Elizabeth D. Liddy: I personally think of big data as a very large collection set of either textual or numeric data and, increasingly, image data, which can either be collected for a common purpose or it could come from very diverse sources.The clear goal is to reuse it.

To look at another main point: How is it we learn from it? We now know that everything we do is monitored or tracked in some way, so not just the formal collection of data, but that data from all the social media resources, from buying habits—everything is data of some type right now. The point is, no matter if you are in industry, a government agency, a hospital, or a school—there are things we can learn from data. My goal with big data is to learn something to improve the next version of whatever it is we are doing.

Edd Dumbill: Jeff, I wonder if you can give an overview of what you mean by data science?

Jeff Stanton: I think about data science as three really large areas. The first one that people identify most closely with data science, particularly in

industry discussions, is the analytics component. I keep visualization in that same bucket. Analytics and visualization are ways of taking data that you may have and that you may have cleaned and turning it into something that is actionable—turning it into conclusions.

Analytics is just one area though. There are two other big areas that figure in, and one is the infrastructure. If you imagine data that is used in industry, it often begins with point-of-sale systems that are collecting the data and goes all the way through long-term archival storage, but there is a lot of technology in there. And there is a lot of different areas where people can specialize.

A third is data curation, and that may be a little bit of the library influence. Librarians have had this curation area under control for quite a long time. That is over the long-term making the data available for reuse, for repurposing. A key component of that is having good metadata.

Edd Dumbill: Kate, as a student approaching this course, what was your grasp of data science and its importance?

Kate Mueller: I would probably make similar divisions as Jeff, but coming at it from a standpoint as a student and seeing the other students in the program, I would say that each of us specializes in one of those three broad areas.

My perception of big data from the student side is that there is not a great deal of consensus about what it actually is. It is a very popular buzzword. It gets thrown around a lot. How people who are creating and using the tools use the term is sometimes very different from how people in business who are trying to figure out what big data is use the term. I think a lot of the business side is a little bit more about master data management than it is necessarily big data in the more technical sense.

> "So instead of thinking about it as a social scientist studying a few people in the lab, you are talking about actually analyzing patterns across an entire society."

Edd Dumbill: Shelly, coming from your viewpoint on the industry side, are these labels useful and does the breakdown that Jeff discussed make sense to you when you are looking for new hires?

Shelly Farnham: Yes. Two things I would add are thinking about what big data is in its more practical standpoint: Data that is so large you cannot analyze it on your own machine, and so you need the tools and infrastructure in order to do that kind of large-scale analysis. That is not something that you can usually find in desktop applications. You usually need a Hadoop server somewhere, or Cosmos, that sort of a thing.

Also, in thinking about the nature of the problems that you can solve, when you are thinking about big data, the scale is very different. So instead of thinking about it as a social scientist studying a

few people in the lab, you are talking about actually analyzing patterns across an entire society, which then affects what kind of issues or problems you can address.

Edd Dumbill: Given this overview of data science, where do you think the biggest challenges lie in educating students? What are the main skill sets they need and the biggest demand from industry?

Elizabeth D. Liddy: People come to data science either from the collection and curation side—and this may be someone who has worked in a science lab who has been responsible for keeping track of the data for a number of researchers—or the business side, which is, what are the business analytics? What is it we are trying to learn from it? What are the algorithms? What is the language we are going to use?

A focus that we have had is putting these pieces together, because it is a true life cycle, and what I find from the students is they come to our Certificate of Advanced Study (CAS) in data science to get in-depth knowledge on both of those sides of it. That is the goal of our CAS.

Jeff Stanton: From a teaching perspective, as a faculty member I can teach someone how to do a t-test in 10 minutes, and I can teach them how to write a Python program in half an hour, but what I cannot teach them very easily is the domain knowledge. In other words, in a given area, if you are from healthcare, what you need to know in order to be effective at analysis is very different than if you are in retail. That underlying domain knowledge, to be able to have a student come up to speed on that is very hard.

For that, we really need a partnership with people in industry because there is a component of the educational process that really depends on our industry partners to help our students understand the domains where their skills will be applied.

Edd Dumbill: That makes a lot of sense. Do you find, incidentally, that you have many students coming from the physical sciences? One of the things that is often said is the first data scientists were all high-energy physicists.?

Jeff Stanton: Well, we in fact had a different label for our program a few years ago. We called it e-science. At that time, a substantial number of folks in the program were coming in not necessarily from high-energy physics but from the sciences in general. What we have seen across the board is that as the industry interests in data science across all sectors—education and government and so forth—have grown, so has the interest level of students. Students are coming from all over now, rather than just from the sciences.

Kate Mueller: Yes, I would agree with that. I am a "data geek," so for me it is a natural fit, but there are a lot of fellow students coming at it more with a library background or more of a business background, so less emphasis on the tools themselves but more emphasis on trying to figure out what kind of insight you are trying to derive and what kind of

problems you are trying to solve. That is definitely a big shift just from the more strict science side of it.

Shelly Farnham: I was just thinking through these categories—analytics, infrastructure, and curation—and the direction of this conversation and perhaps another category I might add is design. How big data informs design, how can you leverage data in the user experience, both for end-user consumers but also research scientists.

"I do not know how you teach someone to love to learn, but being self-motivated is integral to this field."

Elizabeth D. Liddy: That actually fits very well into one of the courses that is an elective for students. It is information design, which we now have been teaching for several semesters and is, like you say, very important.

Jeff Stanton: The user experience—is a very good point.

Elizabeth D. Liddy: Yes. And I wanted to add that another huge group of students who are interested in this are cohorts of individuals we see coming to us from either a particular corporation or from a division of the government who say, "You have a 15-credit, online available CAS. What we would like to do is have a whole cohort of our employees come take the course as a cohort and work with our own data within that section of the course."

Edd Dumbill: Let me bring in the idea of the toolkit of a data scientist, because what we are observing now is that this is a field that is exploding in terms of the tools that can be used, especially around Hadoop. This leads to the question: How can an institution actually prepare students in an environment like this where all the tools are moving so quickly?

Elizabeth D. Liddy: One possible way is partnership with those who are creating the tools. Shelly was just here and gave a wonderful talk, and a lot of it is sharing the tools with the students and ensuring they are involved in next versions. That is one way for the students in our instruction to keep up with the latest and the best tools.

Shelly Farnham: I do not know how you teach someone to love to learn, but being self-motivated is integral to this field. Once you have the core concepts, to be able to be really excited about, and continue to seek out, new information is something that I look for, for example, when we are recruiting people. Are you the kind of person who does not just learn what is in the school, but you are tracking the blogs and you are tracking the industry? Are you excited to download something and try it out just because you got a recommendation from a friend? Are you that kind of person who is constantly engaged and able to learn independently?

Jeff Stanton: That is really key for me. If you have curiosity and you have learned one thing, like you have learned Hadoop, then you can move on to other

things. You get the concepts and you can apply it. There is no way educationally that we can keep up with the pace of generating new tools, but somebody who loves to learn and knows how to learn, they can continue to do it on their own.

Kate Mueller: One of the interesting things for me is that I come from a database development background, and the difference in looking at relational databases versus the field of big data is that if you are looking at relational database jobs, a lot of what you are talking about is: do you have familiarity with SQL? If you know that, you have the right kind of mindset and sense of logic to learn any specific detailed system.

I think the same kind of principle applies to the field of big data, except that there is not one agreed upon technology that is a standard. So what the curriculum here focuses on is teaching that kind of analytical, engaging mindset so that you can pick up any tool that you need to pick up, rather than sending you down the path of: this is *the* tool you are going to use for this, since it is such a budding and constantly changing field.

I think the emphasis has to be on teaching and encouraging students to think outside the classroom, as Shelly mentioned, but also be curious and willing to try things and fail in the effort of learning something new that has just kind of hit the scene and figuring out what it is capable of doing.

Edd Dumbill: Certainly that entrepreneurial and exploratory element seems to be one of the big keys for data scientists. Thinking about the students that you are preparing, I know, Liz, you mentioned that sometimes companies come to you. We have all heard through the McKinsey reports and various other stories that there is a shortage of data scientists. In your interaction with industry, do you believe that they are feeling that shortage right now? If so, how do you think, as a school, you can help address this shortage?

Elizabeth D. Liddy: I think they are. I was speaking at an event recently and talked to representatives of three different large corporations, all of whom were very interested in pursuing the opportunity to have cohorts of theirs take the online CAS. The way they are preparing is they know they have clever people. They have people who have an interest in this. What they are going to do is update their skills, so you may know one part of it, but in order to do the task and the full life cycle that we want done, you need some additional skills. That is what I am hearing.

Jeff Stanton: The demand side is very strong. I was talking to another faculty member recently who was on the phone with a large manufacturing company from the Midwest who was saying, "We need about 80 or 90 people with analytic skills. How soon can we come and recruit your graduates?"

Edd Dumbill: And what did you say to that?

Jeff Stanton: We had to tell the truth, which is, we have got about 30 people in our program right now, with the expectation of having a fair number more pretty soon. But it is going to be a while before we can supply 80 or 90 people to one company. We are a ways off from that.

Edd Dumbill: Given the number of intakes that you would need for that then, do you think there is work to do to publicize data science as a career path for students?

Jeff Stanton: I think there is definitely work to do, but there are some really good signs that we are starting to turn the corner. Kate, earlier in the conversation, said that she was a data geek. Just based on what I know of the research, even five years ago people would not label themselves as geeks because it was stigmatizing. The tide has turned and we are beginning to see a groundswell certainly among undergraduates who are okay identifying themselves as data geeks.

We have to work with that pipeline of students and in some cases need to bring them up to speed on quantitative skills and some of the technology skills, but I am optimistic that that is going to happen. I think if the interest in industry sustains data science, we will have people that want to do it because they see the potential there.

Elizabeth D. Liddy: Another area of emphasis here in our school is entrepreneurship. So at the moment, there are 34 student-startup companies. Most of them are kind of geeky. They like hands-on, intense time to solve a problem together. Not all may end up being so interested on the business side of entrepreneurship, but they sure like that hands-on viewpoint: "This is data, let me have an impact. Let me do something with it."

"It is an exciting time where there is so much data available to play with, to mash up and merge in new and interesting ways."

Jeff Stanton: The demand for our programming courses a few years ago was not very strong. Then once we started really focusing on entrepreneurship and people began to try starting their own companies and coming up with their own ideas, they realized that they really needed those skills in order to be able to create. It ties in beautifully with the Maker Movement that we have heard so much about—data science is really about building things that can lead to better decision-making.

Shelly Farnham: It is an exciting time where there is so much data available to play with, to mash up and merge in new and interesting ways. I think just that ease of availability is one of the really unique aspects of our time.

Jeff Stanton: Do you want to talk about your project a little bit, Shelly?

Shelly Farnham: Sure, I work in FSE Labs, which is Future Social Experiences, and we explore projects trying to think 3 to 5 to 10 years in the future. One of our projects is social, SO.CL, which is really focused on helping people leverage the Internet and finding and searching and sharing with each other.

One of the projects that we have been doing is

an experimental project where we really want to use it as a vehicle for developing relationships with the research community as a whole. And so we instrumented the whole system, all the public behavior within the system, and made it available for researchers to play with and do research around.

"So as more of these data-science educational programs come online, I think we have a bit of an educational job to do with talking to hiring managers."

Jeff Stanton: That is a great example of the increase in the amount of interesting data that is available. I know for my students, one of the things that really got them interested in playing with data was the availability of the API for Twitter, so you can mine Twitter data with a couple of commands, and people love that. That is really powerful.

Edd Dumbill: The human element in data is definitely very strong and has brought a lot of people into it. Next, I'd like to look at the students you are turning out. I wonder if you could give us a bit of a flavor of the kinds of jobs they are going into, and the kind of companies that are asking for them?

Jeff Stanton: Maybe we could start by asking Kate to see what kind of job she would like?

Kate Mueller: That is the million-dollar question. Trying to pick an area to focus on has probably been the biggest challenge for me personally. I probably fall more into either infrastructure or design if we are talking about big-data buckets. I enjoy analytics, but I think first and foremost it is about the database person in me. I think you have to have a method to get the data that you need to analyze. You need to have a way to present that in some fashion that is actionable and is accessible.

That said, when I look at jobs, if I am not searching "data scientists," it is very hard to tell what kind of job really fits into my interests. If it does not say "analytics" or "data science," it becomes a lot harder. There seems to be a very sharp distinction in that most of the jobs using those titles require a very high level of familiarity with very specific tools.

As a student, it is challenging to know what I should be searching for other than looking for more general IT or database jobs that are actually big-data design problems called something else.

I am in my second year, so I am very actively job hunting right now. I am hoping I get good answers from this conversation because it has been a challenge for me.

Shelly Farnham: One of the challenges is that data science is not agnostic of domain. For example, when we are looking for people, interns or full-time people on our team, we definitely look for people who have experience analyzing data, but they also should be deeply engaged with the topic of social media. I know in the health industry and dealing with health data, that is a very different problem. So it would not make sense to hire somebody who spent the past

four years studying health data. I think that the domain knowledge is a very important aspect of what we are looking for.

Elizabeth D. Liddy: When I think about the field, representatives from companies who come to interview and are looking to hire our students, one of the largest areas is in the financial industry as well as in the consulting fields. If they are consulting for a number of other large companies, there could be great variety there, but many of the companies who are going to them for consulting are saying, "Do you have data scientists on your payroll?"

Another area that has come looking is universities—either the CIO office, or, in at least one university I know of, the library has assumed the responsibility for managing all the research data for that university. Now of course there are all the requirements from NIH, NEH, and NSF, that if you write a proposal and you are a big university doing lots of research, you have to have data-management plans for every one of those projects.

So if you are the central place at that university, whether it is the CIO or the library, there are some who have told us that they will be in line to hire everyone we graduate. That is not a very specific response, but I think it is an indication.

Jeff Stanton: I have a little bone to pick with hiring managers though. I did a little study of the advertisements like in one of the big online places where they advertise jobs and people can upload their resumes. And a lot of the jobs, which looked like data-science positions, required a computer science degree. To me that is sort of misaligned. Yes, a lot of computer science graduates are good with languages. They might be great Python programmers, for instance. But that is not the whole package. So as more of these data-science educational programs come online, I think we have a bit of an educational job to do with talking to hiring managers, talking to HR people to help them understand where the talent can come from, not just computer science.

Edd Dumbill: Absolutely. So what are the most striking things that are lacking that computer science does not provide?

Jeff Stanton: Well, a lot of times computer scientists do not have any background in statistics or visualization. Many times they are not really up on the key elements of the infrastructure, unless they were computer engineering majors, and they definitely do not get training in curation, in the data management/curation side of it.

Edd Dumbill: What kind of background are your students coming from? We have talked about science already, but do you see computer science and math predominate at all?

Jeff Stanton: We do not. There are some people that are coming more from the library science side of things. We get a lot of English majors. There are a lot of people who have an interest in data who do not come from that highly mathematical or computer science-oriented background, and so that is part of the feature that we have built into our program, the ability to help those folks ramp up to speed on some of the quantitative skills.

Kate Mueller: In full disclosure, I should say I am a former English major.

Elizabeth D. Liddy: I am a former English major too.

Kate Mueller: What has prompted me to come back into the field is that I really like solving problems. Even as an English major, I always read for structure. Structure was what was interesting to me. I was a terrible writer because plot was irrelevant. But what I find is that that sense of analysis and enjoying the puzzle and enjoying trying to solve it translates so much better, not in terms of job prospects, but just in terms of day-to-day interest in this field than it ever did on the English side of it.

I also have the added bonus of having worked in the industry doing database work, and so I come to it with a fairly strong understanding of how data works itself and having done very small-scale level analytics, but being very steeped in the broadcasting industry. To say that only computer science people make good big-data experts or good data analysts is missing the boat in terms of people who see the structure underlying things and solve problems even if they do not have really high-level technical skills.

Jeff Stanton: Kate mentioned one really important thing that I will emphasize. We did a panel at a recent conference where we talked about the skills that data scientists need. One of the key things that came out in almost everybody's discussion was communication. A lot of times you have mounds of data and you need to be able to translate that, in effect, for consumption by managers and other people and to be able to communicate clearly, to be able to tell a story with the data. That is really a critical skill.

Edd Dumbill: That brings up something that Kate mentioned earlier about how different people have different sorts of primary skills, and that would be the data palette you mentioned. As you see people go into industry, do you see data scientists working solo or with complementary skills coming together to be a data team? Do you see both or does the team philosophy predominate?

Shelly Farnham: I would say as a data scientist you end up collaborating with people who are not. Certainly in my industry, I might be working with a designer or a developer or an IT person.

And, coming from social media, I wanted to mention a couple of other areas. We see a lot of people with degrees in human–computer interaction, who also have these data analytics skills, or there is a new term that has been emerging lately, computational social science.

Elizabeth D. Liddy: One other aspect of this I might add, and I like what Shelly says here about the computational social scientists. A lot of the data that I have ever worked with is textual data. What is amazing is the huge amounts of textual data that is available for analysis, interpretation. A lot of the work I did in the past was for the intelligence agencies, and there is just tons of data, way too much for anyone to be able to read and get a picture either of an organization, or of an individual. But by doing the natural language processing of it, particularly the sentiment analysis, the emotional side, you are able to provide insight and answers very quickly that would have

taken analysts days and weeks of reading in order to come up with a similar picture.

Edd Dumbill: Right. Communicating this can be interesting though. You mentioned that a lot of financial companies are looking to hire. That is of course because they know they have a lot of data. But the kind of things you are talking about may be obvious to the intelligence community right now, but when you are talking to organizations, what about expectations for employing data scientists, and what kind of expectations do you give them that they will be able to do if they have a data science team?

Jeff Stanton: At the entry level, for the most part, they are looking for people who can work with particular tools, as Kate mentioned before. So they need someone who can hit the ground and be productive pretty quickly. But that actually contrasts with a more experienced or higher level employee, someone who might have already been employed in the area before and has gotten data-science skills added onto their skill set.

What I found there is that employers are looking for generalists who can create bridges between the different areas. Often, you find the hardware people and the analytics people do not speak the same language. So how do you actually get a project to work?

> "You are able to provide insight and answers very quickly that would have taken analysts days and weeks of reading in order to come up with a similar picture?"

One of the things that we have tried to do is to allow our students to have specializations in the curriculum, and project management for instance is one of those specializations. A good project manager who has skills in data science and other areas is a very valuable asset because that is the kind of person that you need to oversee the level of complexity that is coming out of these projects.

Edd Dumbill: One of the followup questions I have is around the development of a career path. With these kind of generalist and projectmanagement skills you are mentioning, it sounds like you almost have the expectation that it might be the data science people that end up climbing up the ladder to pretty strong management positions in the end?

Jeff Stanton: That is definitely true. There is, in some sense, a parallel to the evolution of the CIO. I do not know how many years we have to go back before we started to really see that C-level position be very common and popular in companies. But Chief Data Officer is actually a position that I have seen named at a few places. But I do not think it is very common yet. If we check back in five years or seven years, I bet we will see a number of places that have Chief Data Officers.

Edd Dumbill: A lot of people who may be reading this may not be data scientists yet, but have a peripheral interest from some other career angle. If people have been working in IT already or they have been working in business intelligence or finance or library science,

etc., are there any tips? How would you brief people to say this is how you get into big data?

Jeff Stanton: My advice would be to, first of all, have a direction where you want to go. I think we heard from Kate before that from such a rich palate of opportunities, you need to make a few preliminary decisions about what industry or sector or problem area that you would like to be in. As we discussed before, that will guide the extent to which you need to work on your domain skills.

If you have been working in financial services but you have a hankering to do genomics, you have got a little bit of work to do to come up to speed there. Then the other thing is just to do an inventory of the kinds of skills that you would need in that position.

As I mentioned before, you get some people that are great in writing programs in a certain language, but maybe they do not have much of a background in statistics or visualization. So they know that they would need to pick out a program to go where they would be able to polish those skills.

Shelly Farnham: I would add that it is kind of a similar problem with computer science, that there is this notion that it can be very dry and involve a lot of math. But actually if you look a lot at what people are doing with data analytics and big data, it is a lot of fun.

So highlighting and emphasizing the fun aspects—like for me my work is to create social applications and think about the future of social experiences. That is a lot of fun. Conveying that effectively can be a little challenging. I think a lot of people are scared of the word data.

> "If we check back in five or seven years, I bet we will see a number of places that have chief data officers."

Jeff Stanton: We have a remedy for that, by the way.

Edd Dumbill: What is that?

Jeff Stanton: We published a free e-book, an electronic textbook that people can download for free either on their iPad or as a PDF, available at http://ischool.syr.edu/ datascience It is like a gentle introduction to data science. It is pretty fun and it is light. It is really aimed at people who do not know or who may be fearful of that dry, mathematical world.

Edd Dumbill: This is a great wrap-up note in many ways, because we have discussed that you can apply data science to any domain and have a lot of fun doing it as well. So as we come to the end of this, could I ask for a few closing remarks from the group?

Elizabeth D. Liddy: It is a new area, and I think what is going to happen is probably every week or so we are going to hear about another field that has recognized, if they have not already, how important understanding and being able to utilize large datasets is to their field.

Jeff Stanton: I would close by saying that if you are old enough to remember the energy crisis of the 1970s, I think we have a modern data crisis, where

there is way too much data and not nearly enough people to help us sort through it. If we do not get a handle on that, we are wasting a precious resource. Just like energy is a precious resource, data is a precious resource if we simply collect it and then let it sit around unused.

Kate Mueller: One of the things that keeps me the most interested in the field is because data is so prevalent, there is a huge potential for that data to serve a larger purpose. So you see things like follow-ups to the human genome project and the encode project. You see things like more advanced disaster response systems that incorporate social media aspects. You see smart energy grid development in which renewable energy sources can be implemented, integrated, and monitored in real-time and you can be much more responsive to outages and things like that.

That, to me, is what is the most interesting. It is not necessarily that we have all of these tools, but the end purposes that are now possible because of the quantity of data that we have, because of the sophistication of the tools that we have. That is what keeps the field so engaging and interesting and, frankly, why I hope we get a lot more students who start jumping into it, because there is so much possibility here; it is so young, but there is so much to work with.

Acknowledgment

This Roundtable Discussion was funded by the School of Information Studies at Syracuse University.

Address correspondence to:

Edd Dumbill
E-mail: edumbill@acm.org.

This article has been cited by:

1. Meredith A. Barrett, Olivier Humblet, Robert A. Hiatt, Nancy E. Adler. 2013. Big Data and Disease Prevention:
From Quantified Self to Quantified Communities. Big Data 1:3, 168-175. [Abstract] [Full Text HTML]
[Full Text PDF] [Full Text PDF with Links]

Delivering Value from Big Data with Revolution R Enterprise and Hadoop

Bill Jacobs, Director of Product Marketing
Thomas W. Dinsmore, Director of Product Management

Abstract

Businesses are rapidly investing in Hadoop to manage analytic data stores due to its flexibility, scalability and relatively low cost. However, Hadoop's native tooling for advanced analytics is immature; this makes it difficult for analysts to use without significant additional training, and limits the ability of businesses to drive value from Big Data.

Revolution R Enterprise leverages open source R, the standard platform for modern analytics. R has a thriving community of more than two million users who are trained and ready to deliver results.

Revolution R Enterprise runs in Hadoop. It enables users to perform advanced analysis on Big Data while working exclusively in the familiar R environment. The software transparently distributes work across the nodes and distributed data of Hadoop without complex programming.

By leveraging Revolution R Enterprise in Hadoop, organizations tap a large and growing community of analysts and all of the capabilities in R with a true cross-platform and open standards architecture.

Learn more!

Read the entire white paper on "Delivering Value from Big Data with Revolution R Enterprise and Hadoop"