DOI:10.1145/2500499

Big data promises automated actionable knowledge creation and predictive models for use by both humans and computers.

BY VASANT DHAR

Data Science and Prediction

USE OF THE term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

The term "science" implies knowledge gained through systematic study. In one definition, it is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions.[11] Data science might therefore imply a focus involving data and, by extension, statistics, or the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference. Why then do we need a new term like data science when we have had statistics for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term.

The short answer is data science is different from statistics and other existing disciplines in several important ways. To start, the raw material, the "data" part of data science, is increasingly heterogeneous and unstructured—text, images, video—often emanating from networks with complex relationships between their entities. Figure 1 outlines the relative expected volumes of unstructured and structured data from 2008 to 2015 worldwide, projecting a difference of almost 200 petabytes (PB) in 2015 compared to a difference of 50PB in 2012. Analysis, including the combination of the two types of data, requires integration, interpretation, and sense making that is increasingly derived through tools from computer science, linguistics, econometrics, sociology, and other disciplines. The proliferation of markup languages and tags is designed to let computers interpret data automatically, making them active agents in the process of decision making. Unlike early markup languages (such as HTML) that emphasized the display of information for human consumption, most data generated by humans and computers today is for consumption by computers; that is, computers increasingly do background work for each other and make decisions automatically. This scalability in decision making has become possible because of big data that serves as the raw material for the creation of new knowledge; Watson, IBM's "Jeopardy!" champion, is a prime illustration of an emerging machine intelligence fueled by data and state-of-the-art analytics.

» key insights

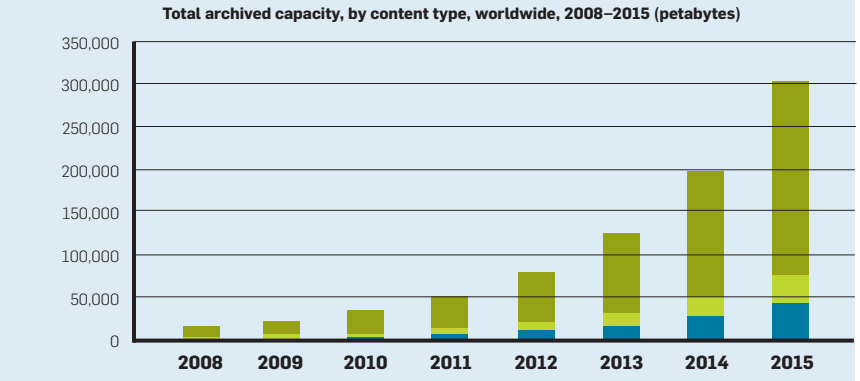
- **Data science is the study of the generalizable extraction of knowledge from data.**
- **A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to explain the past.**
- **A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.**

PHOTO ILLUSTRATION BY BARRY DOWNARD


ansition to the deficit period
Repeated recessions
Surplus
Deficit
January
February
March
April
600
500
400
300
200
100
0

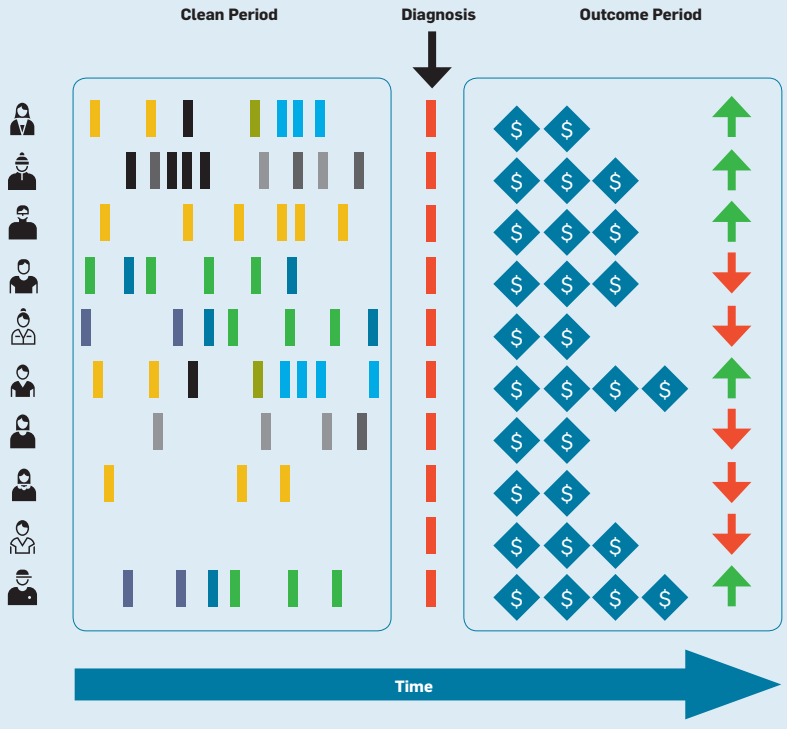
From an engineering perspective, scale matters in that it renders the traditional database models somewhat inadequate for knowledge discovery. Traditional database methods are not suited for knowledge discovery because they are optimized for fast access and summarization of data, given what the user wants to ask, or a query, not discovery of patterns in massive swaths of data when users lack a well-formulated query. Unlike database querying, which asks "What data satisfies this pattern (query)?" discovery asks "What patterns satisfy this data?" Specifically, our concern is finding interesting and robust patterns that satisfy the data, where "interesting" is usually something unexpected and actionable and "robust" is a pattern expected to occur in the future.

Figure 1. Projected growth of unstructured and structured data.



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

Figure 2. Health-care-use database snippet.



What makes an insight actionable? Other than domain-specific reasons, it is its predictive power; the return distribution associated with an action can be reliably estimated from past data and therefore acted upon with a high degree of confidence.

The emphasis on prediction is particularly strong in the machine learning and knowledge discovery in databases, or KDD, communities. Unless a learned model is predictive, it is generally regarded with skepticism, a position mirroring the view expressed by the 20th-century Austro-British philosopher Karl Popper as a primary criterion for evaluating a theory and for scientific progress in general.[24] Popper argued that theories that sought only to explain a phenomenon were weak, whereas those that made "bold predictions" that stand the test of time despite being readily falsifiable should be taken more seriously. In his well-known 1963 treatise on this subject, *Conjectures and Refutations*, Popper characterized Albert Einstein's theory of relativity as a "good" one since it made bold predictions that could be falsified; all attempts at falsification of the theory have indeed failed. In contrast, Popper argued that theories of psychoanalyst pioneers Sigmund Freud and Alfred Adler could be "bent" to accommodate virtually polar opposite scenarios and are weak in that they are virtually unfalsifiable.[a] The emphasis on predictive accuracy implicitly favors "simple" theories over more complex theories in that the accuracy of sparser models tends to be more robust on future data.[4,20] The requirement on predictive accuracy on observations that

a Popper used opposite cases of a man who pushes a child into water with the intention of drowning the child and that of a man who sacrifices his life in an attempt to save the child. In Adler's view, the first man suffered from feelings of inferiority (producing perhaps the need to prove to himself that he dared to commit the crime), and so did the second man (whose need was to prove to himself that he dared to rescue the child at the expense of his own life).

will occur in the future is a key consideration in data science.

In the rest of this article, I cover the implications of data science from a business and research standpoint, first for skills, or what people in industry need to know and why. How should educators think about designing programs to deliver the skills most efficiently and enjoyably? And what kinds of decision-making skills will be required in the era of big data and how will they differ from the past when data was less plentiful?

The second part of my answer to defining big-data skills is aimed at research. How can scientists exploit the abundance of data and massive computational power to their advantage in scientific inquiry? How does this new line of thinking complement traditional methods of scientific inquiry? And how can it augment the way data scientists think about discovery and innovation?

It is not uncommon for two experts in the social sciences to propose opposite relationships among the variables and offer diametrically opposite predictions based on the same sets of facts.

Implications

A 2011 McKinsey industry report[19] said the volume of data worldwide is growing at a rate of approximately 50% per year, or a roughly 40-fold increase since 2001. Hundreds of billions of messages are transmitted through social media daily and millions of videos uploaded daily across the Internet. As storage becomes almost free, most of it is stored because businesses generally associate a positive option value with data; that is, since it may turn out to be useful in ways not yet foreseen, why not just keep it? (One indicator of how inexpensive storage is today is the fact that it is possible to store the world's entire stock of music on a $500 device.)

Using large amounts of data for decision making became practical in the 1980s. The field of data mining burgeoned in the early 1990s as relational database technology matured and business processes were increasingly automated. Early books on data mining[6,7,17] from the 1990s described how various methods from machine learning could be applied to a variety of business problems. A corresponding explosion involved software tools geared toward leveraging transactional and behavioral data for purposes of explanation and prediction.

An important lesson learned in the 1990s is that machine learning "works" in the sense that these methods detect subtle structure in data relatively easily without having to make strong assumptions about linearity, monotonicity, or parameters of distributions. The downside of these methods is they also pick up the noise in data,[31] often with no way to distinguish between signal and noise, a point I return to shortly.

Despite their drawbacks, a lot can be said for methods that do not force us to make assumptions about the nature of the relationship between variables before we begin our inquiry. This is not trivial. Most of us are trained to believe theory must originate in the human mind based on prior theory, with data then gathered to demonstrate the validity of the theory. Machine learning turns this process around. Given a large trove of data, the computer taunts us by saying, "If only you knew what question to ask me, I would give you some very interesting answers based on the data." Such a capability is powerful since we often do not know what question to ask. For example, consider a health-care database of individuals who have been using the health-care system for many years, where among them a group has been diagnosed with Type 2 diabetes, and some subset of this group has developed complications. It could be very useful to know whether there are any patterns to the complications and whether the probability of complications can be predicted and therefore acted upon. However, it is difficult to know what specific query, if any, might reveal such patterns.

To make this scenario more concrete, consider the data emanating from a health-care system that essentially consists of "transactions," or points of contact over time between a patient and the system. Records include services rendered by health-care providers or medication dispensed on a particular date; notes and observations could also be part of the record. Figure 2 outlines what the raw data would look like for 10 individuals where the data is separated into a "clean period" (history prior to diagnosis), a red bar ("diagnosis"), and the "outcome period" (costs and other

outcomes, including complications). Each colored bar in the clean period represents a medication, showing the first individual was on seven different medications prior to diagnosis, the second on nine, the third on six, and so on. The sixth and tenth individuals were the costliest to treat and developed complications, as did the first three, represented by the upward-pointing green arrows.

Extracting interesting patterns is nontrivial, even from a tiny temporal database like this. Are complications associated with the yellow meds or with the gray meds? The yellows in the absence of the blues? Or is it more than three yellows or three blues? The list goes on. Even more significant, perhaps if we created "useful" features or aggregations from the raw data, could physicians, insurers, or policy makers predict likely complications for individuals or for groups of people?

Feature construction is an important creative step in knowledge discovery. The raw data across individuals typically needs to be aggregated into some sort of canonical form before useful patterns can be discovered; for example, suppose we could count the number of prescriptions an individual is on without regard to the specifics of each prescription as one approximation of the "health status" of the individual prior to diagnosis. Such a feature ignores the "severity" or other characteristics of the individual medications, but such aggregation is nonetheless typical of feature engineering.

Suppose, too, a "complications database" would be synthesized from the data, possibly including demographic information (such as patient age and medical history); it could also include health status based on a count of current medications; see Figure 3, in which a learning algorithm, designated by the right-facing blue arrow, could be applied to discover the pattern on the right. The pattern represents an abstraction of the data, or the type of question we should ask the database, if only we knew what to ask. Other data transformations and aggregations could yield other medically insightful patterns.

What makes the pattern on the right side of Figure 3 interesting? Suppose the overall complication rate in the population is 5%; that is, a random sample of the database includes, on average, 5% complications. In this scenario, the snippet on the right side of Figure 3 could be very interesting since its complication rate is many times greater than the average. The critical question is whether this is a pattern that is robust and hence predictive, likely to hold up in other cases in the future. The issue of determining robustness has been addressed extensively in the machine learning literature and is a key consideration for data scientists.[23]

A new powerful method is available for theory development not previously practical due to the paucity of data.

If Figure 3 is representative of the larger database, the box on the right tells us the interesting question to ask the database: "What is the incidence of complications in Type 2 diabetes for people over age 36 who are on six or more medications?" In terms of actionability, such a pattern might suggest being extra vigilant about people with such a profile who do not currently have a complication in light of their high susceptibility to complications.

The general point is that when data is large and multidimensional, it is practically impossible for us to know a priori that a query (such as the one here concerning patterns in diabetes complications) is a good one, or one that provides a potentially interesting and actionable insight. Suitably designed machine learning algorithms help find such patterns for us. To be useful both practically and scientifically, the patterns must be predictive. The emphasis on predictability typically favors Occam's razor, or succinctness, since simpler models are more likely to hold up on future observations than more complex ones, all else being equal;[4] for example, consider the diabetes complication pattern here:

Age > 36 and #Medication > 6 → Complication_rate=100%

A simpler competing model might ignore age altogether, stating simply that people on six or more medications tend to develop complications. The reliability of such a model would be more apparent when applied to future data; for example, does simplicity lead to greater future predictive accuracy in terms of fewer false positives and false negatives? If it does, it is favored. The

practice of "out of sample" and "out of time" testing is used by data scientists to assess the robustness of patterns from a predictive standpoint.

When predictive accuracy is a primary objective in domains involving massive amounts of data, the computer tends to play a significant role in model building and decision making. The computer itself can build predictive models through an intelligent "generate and test" process, with the end result an assembled model that is the decision maker; that is, it automates Popper's criterion of predictive accuracy for evaluating models at a scale in ways not feasible before.

If we consider one of these patterns—that people with "poor health status" (proxied by number of medications) have high rates of complications—can we say poor health status "causes" complications? If so, perhaps we can intervene and influence the outcome by possibly controlling the number of medications. The answer is: it depends. It could be the case that the real cause is not in our observed set of variables. If we assume we have observed all relevant variables that could be causing complications, algorithms are available for extracting causal structure from data,[21] depending how the data was generated. Specifically, we still need a clear understanding of the "story" behind the data in order to know whether the possibility of causation can and should be entertained, even in principle. In our example of patients over age 36 with Type 2 diabetes, for instance, was it the case that the people on seven or more medications were "inherently sicker" and would have developed complications anyway? If so, it might be incorrect to conclude that large numbers of medications cause complications. If, on the other hand, the observational data followed a "natural experiment" where treatments were assigned randomly to comparable individuals and enough data is available for calculating the relevant conditional probabilities, it might be feasible to extract a causal model that could be used for intervention. This issue of extracting a causal model from data is addressed in the following sections; for a more complete treatment on causal models, see Pearl,[21] Sloman,[29] and Spirtes et al.[30]

Skills

Machine learning skills are fast becoming necessary for data scientists as companies navigate the data deluge and try to build automated decision systems that hinge on predictive accuracy.[25] A basic course in machine learning is necessary in today's marketplace. In addition, knowledge of text processing and "text mining" is becoming essential in light of the explosion of text and other unstructured data in health-care systems, social networks, and other forums. Knowledge about markup languages like XML and its derivatives is also essential, as content becomes tagged and hence able to be interpreted automatically by computers.

Data scientists' knowledge about machine learning must build on more basic skills that fall into three broad classes: The first is statistics, especially Bayesian statistics, which requires a working knowledge of probability, distributions, hypothesis testing, and multivariate analysis. It can be acquired in a two- or three-course sequence. Multivariate analysis often overlaps with econometrics, which is concerned with fitting robust statistical models to economic data. Unlike machine learning methods, which make no or few assumptions about the functional form of relationships among variables, multivariate analysis and econometrics by and large focus on estimating parameters of linear models where the relationship between the dependent and independent variables is expressed as a linear equality.

The second class of skills comes from computer science and pertains to how data is internally represented and manipulated by computers. This involves a sequence of courses on data structures, algorithms, and systems, including distributed computing, databases, parallel computing, and fault-tolerant computing. Together with scripting languages (such as Python and Perl), systems skills are the fundamental building blocks required for dealing with reasonable-size datasets. For handling very large datasets, however, standard database systems built on the relational data model have severe limitations. The recent move toward cloud computing and non-relational structures for dealing with enormous datasets in a robust manner signals a new set of required skills for data scientists.

The third class of skills requires knowledge about correlation and causation and is at the heart of virtually any modeling exercise involving data. While observational data generally limits us to correlations, we can get lucky. Sometimes plentiful data might represent natural randomized trials and the possibility of calculating conditional probabilities reliably, enabling discovery of causal structure.[22] Building causal models is desirable in domains where one has reasonable confidence as to the completeness of the formulated model and its stability, or whether the causal model "generating" the observed data is stable. At the very least, a data scientist should have a clear idea of the distinction between correlation and causality and the ability to assess which models are feasible, desirable, and practical in different settings.

The final skill set is the least standardized and somewhat elusive and to

Figure 3. Extracting interesting patterns in health outcomes from health-care system use.

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
Complication = Yes (100% confidence)

some extent a craft but also a key differentiator to be an effective data scientist—the ability to formulate problems in a way that results in effective solutions. Herbert Simon, the 20th-century American economist who coined the term "artificial intelligence" demonstrated that many seemingly different problems are often "isomorphic," or have the identical underlying structure. He demonstrated that many recursive problems could be expressed as the standard Towers of Hanoi problem, or involving identical initial and goal states and operators. His larger point was it is easy to solve seemingly difficult problems if represented creatively with isomorphism in mind.[28]

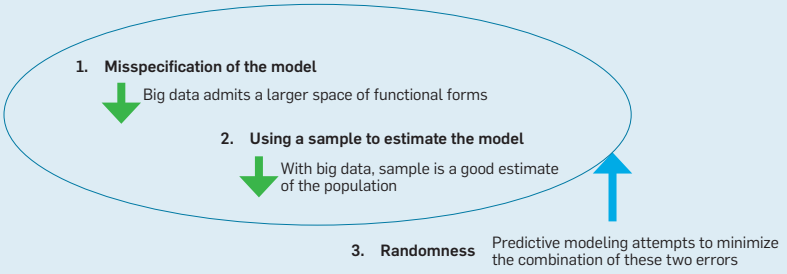
In a broader sense, formulation expertise involves the ability to see commonalities across very different problems; for example, many problems have "unbalanced target classes" usually denoting the dependent variable is interesting only sometimes (such as when people develop diabetes complications or respond to marketing offers or promotions). These are the cases of interest we would like to predict. Such problems are a challenge for models that, in Popperian terms, must go out on a limb to make predictions that are likely to be wrong unless the model is extremely good at discriminating among the classes. Experienced data scientists are familiar with these problems and know how to formulate them in a way that gives a system a chance to make correct predictions under conditions where the priors are stacked heavily against it.

Problem-formulation skills represent core skills for data scientists over the next decade. The term "computational thinking" coined by Papert[21] and elaborated by Wing[32] is similar in spirit to the skills described here. There is considerable activity in universities to train students in problem-formulation skills and provide electives structured around the core that are more suited to specific disciplines.

The data science revolution also poses serious organizational challenges as to how organizations manage their data scientists. Besides recognizing and nurturing the appropriate skill sets, it requires a shift in managers' mind-sets toward data-driven decision making to replace or augment intuition and past practices. A famous quote by 20th-century American statistician W. Edwards Demming—"In God we trust, everyone else please bring data"—has come to characterize the new orientation, from intuition-based decision making to fact-based decision making.

From a decision-making standpoint, we are moving into an era of big data where for many types of problems computers are inherently better decision makers than humans, where "better" could be defined in terms of cost, accuracy, and scalability. This shift has already happened in the world of data-intensive finance where computers make the majority of investment decisions, often in fractions of a second, as new information becomes available. The same holds in areas of online advertising where millions of auctions are conducted in milliseconds every day, air traffic control, routing of package delivery, and many types of planning tasks that require scale, speed, and accuracy simultaneously, a trend likely to accelerate in the next few years.

Figure 4. Sources of error in predictive models and their mitigation.



Knowledge Discovery

Former editor of *Wired* magazine Chris Anderson[1] drew on the quote by British-born statistician George Box that "All models are wrong, but some are useful," arguing, with the huge amounts of data available today, we do not need to settle for wrong models or any models at all. Anderson said prediction is of paramount importance to businesses, and data can be used to let such models emerge through machine learning algorithms, largely unaided by humans, pointing to companies like Google as symbolizing the triumph of machine learning over top-down theory development. Google's language translator does not "understand" language, nor do its algorithms know the contents of webpages. IBM's Watson does not "understand" the questions it is asked or use deep causal knowledge to generate questions to the answers it is given. There are dozens of lesser-known companies that likewise are able to predict the odds of someone responding to a display ad without a solid theory but rather based on gobs of data about the behavior of individuals and the similarities and differences in that behavior.

Anderson's 2008 article launched a vigorous debate in academic circles. How can one have science and predictive models without first articulating a theory?

The observation by Dhar and Chou[5] that "patterns emerge before reasons for them become apparent" tends to resonate universally among professionals, particularly in financial markets, marketing, health care, and fields that study human behavior. If this is true, Box's observation becomes relevant: If a problem is nonstationary and a model is only an approximation anyway, why not build the best predictive model based on data available until that time and just update it periodically? Why bother developing a detailed causal model if it is poor at prediction and, more important, likely to get worse over time due to "concept drift"?

Some scientists would say there is no theory without causality, that all observational data, except total chaos, must be generated from a causal model. In the earlier health-care example involving medical complications in patients with Type 2 diabetes,

this seems obvious; some underlying mechanism must have been responsible for the observed outcomes. But we may not have observed or been capable of observing the causal picture. Even if we observed the right variables we would need to know how the observational data was generated before we can in principle draw causal connections. If the observations represent a natural experiment (such as physicians using a new drug vs. other physicians using an old one for comparable individuals), the data might reveal causality. On the other hand, if the new drug is prescribed primarily for "sicker" individuals, it would represent a specific kind of bias in the data.

Anderson's point has particular relevance in the health, social, and earth sciences in the era of big data since these areas are generally characterized by a lack of solid theory but where we now see huge amounts of data that can serve as grist for theory building[3,12,13] or understanding large-scale social behavior and attitudes and how they can be altered.[14] Contrast physics and social sciences at opposite ends of the spectrum in terms of the predictive power of their theories. In physics, a theory is expected to be "complete" in the sense a relationship among certain variables is intended to explain the phenomenon completely, with no exceptions. Such a model is expected to make perfect predictions—subject to measurement error but not to error due to omitted variables or unintended consequences. In such domains, the explanatory and predictive models are synonymous. The behavior of a space shuttle is, for example, explained completely by the causal model describing the physical forces acting on it. This model can also be used to predict what will happen if any input changes. It is not sufficient to have a model 95% sure of outcomes and leave the rest to chance. Engineering follows science.

In contrast, the social sciences are generally characterized by incomplete models intended to be partial approximations of reality, often based on assumptions of human behavior known to be simplistic. A model correct 95% of the time in this world would be considered quite good. Ironically, however, the emphasis in social science theory development is often on proposing theories that embody causality without serious consideration of their predictive power. When such a theory claims "A causes B," data is gathered to confirm whether the relationship is indeed causal. But its predictive accuracy could be poor because the theory is incomplete. Indeed, it is not uncommon for two experts in the social sciences to propose opposite relationships among the variables and offer diametrically opposite predictions based on the same sets of facts; for example, economists routinely disagree on both theory and prediction, and error rates of forecasts tend to be high.

Big data makes it feasible for a machine to ask and validate interesting questions humans might not consider.

How could big data put these domains on firmer ground? In the "hard" sciences, where models can be assumed, for practical purposes, to be complete, there exists the possibility of extracting causal models from large amounts of data. In other fields, large amounts of data can result in accurate predictive models, even though no causal insights are immediately apparent. As long as their prediction errors are small, they could still point us in the right direction for theory development. As an example of being pointed in the right direction, a health-care research scientist recently remarked on an observed pattern of coronary failure being preceded months earlier by a serious infection. One of his conjectures was infections might have caused inflamed arteries and loosened plaque that subsequently caused coronary failure. There could be other explanations, but if the observed pattern is predictive, it might be worthy of publication and deeper inquiry. The questions such a case raise for gatekeepers of science is whether to more strongly consider the Popperian test of predictive accuracy on future data and favor simple accurate predictive models as potential components of future theory instead of requiring a causal model up front tested by the data.

What makes predictive models accurate? Conversely, where do errors come from?

Hastie et al.[10] said errors in prediction come from three sources: The first is misspecification of a model, so, for example, a linear model that attempts to fit a nonlinear phenom-

enon could generate an error simply because the linear model imposes an inappropriate bias on the problem. The second is the samples used for estimating parameters; the smaller the samples, the greater the bias in the model's estimates. And the third is randomness, even when the model is specified perfectly.

Big data allows data scientists to significantly reduce the first two types of error (see Figure 4). Large amounts of data allow us to consider models that make fewer assumptions about functional form than linear or logistic regressions simply because there is a lot more data to test such models and compute reliable error bounds.[27] Big data also eliminates the second type of error, as sample estimates become reasonable proxies for the population.

The theoretical limitation of observational data of the sort in these examples, regardless of how big it is, is that the data is generally "passive," representing what actually happened in contrast to the multitude of things that could have happened had circumstances been different. In health care, it is like having observed the use of the health-care system passively and now having the chance of understand it in retrospect and extract predictive patterns from it. Unless we are fortunate enough that the data provided us the right experiments naturally, it does not tell us what could have happened if some other treatment had been administered to a specific patient or to an identical patient; that is, it does not represent a clean controlled randomized experiment where the researcher is able to establish controls and measure the differential effect of treatments on matched pairs.

Interestingly, however, the Internet era is fertile ground for conducting inexpensive large-scale randomized experiments on social behavior; Kohavi et al.[15] provide a number of examples. A 2012 controlled experiment by Aral and Walker[2] on the adoption of video games asked whether it was "influence" or "homophily" that affected choice uncovered profiles of people who are influential and susceptible. Results include patterns (such as "older men are more influential than younger men" and "people of the same age group have more influence on each other than from other age groups"). While specific to games, these results suggest influence is nuanced, certainly more so than existing theories like Malcolm Gladwell's concept of "super influencers"[8] and myriad other popular theories. Big data provides a basis for testing them.

Predictive modeling and machine learning are increasingly central to the business models of Internet-based data-driven businesses.

One of the most far-reaching modern applications of big data is in politics, as exemplified by the Democratic National Committee heavy investment in data and analytics prior to President Barack Obama's winning 2012 campaign, debunking widely held beliefs (such as voters in the "middle" are most critical to outcomes, when in fact issues that resonate with some segments of solidly partisan voters can sway them[14]). In the campaign, the DNC crafted predictive models on the basis of results from large-scale experiments used to manipulate attitudes. The campaign predicted at the level of individual voters how each eligible voter would vote, as well as how to "turn someone into the type of person it wanted you to be."[14]

Social science theory building is also likely to get a good boost from big data and machine learning. Never before have social scientists been able to observe human behavior at the degree of granularity and variability seen today with increasing amounts of human interaction and economic activity mediated by the Internet. While the inductive method has limitations, the sheer volume of data being generated makes induction not only feasible but productive. That is not to say the traditional scientific method is "dead," as claimed by Anderson.[1] On the contrary, it continues to serve us well. However, a new powerful method is available for theory development not previously practical due to the paucity of data. That era of limited data and its associated assumptions is largely over.

Conclusion

Hypothesis-driven research and approaches to theory development have served us well. But a lot of data is emanating around us where these traditional approaches to identifying structure do not scale well or take advantage of observations that would not occur under controlled circum-

stances; for example, in health care, controlled experiments have helped identify many causes of disease but may not reflect the actual complexities of health.[3,18] Indeed, some estimates claim clinical trials exclude as much as 80% of the situations in which a drug might be prescribed, as when a patient is on multiple medications.[3] In situations where we are able to design randomized trials, big data makes it feasible to uncover the causal models generating the data.

As shown earlier in the diabetes-related health-care example, big data makes it feasible for a machine to ask and validate interesting questions humans might not consider. This capability is indeed the foundation for building predictive modeling, which is key to actionable business decision making. For many data-starved areas of inquiry, especially health care and the social, ecological, and earth sciences, data provides an unprecedented opportunity for knowledge discovery and theory development. Never before have these areas had data of the variety and scale available today.

This emerging landscape calls for the integrative skill set identified here as essential for emerging data scientists. Academic programs in computer science, engineering, and business management teach a subset of these skills but have yet to teach the integration of skills needed to function as a data scientist or to manage data scientists productively. Universities are scrambling to address the lacunae and provide a more integrated skill set covering basic skills in computer science, statistics, causal modeling, problem isomorphs and formulation, and computational thinking.

Predictive modeling and machine learning are increasingly central to the business models of Internet-based data-driven businesses. An early success, Paypal, was able to capture and dominate consumer-to-consumer payments due to its ability to predict the distribution of losses for each transaction and act accordingly. This data-driven ability was in sharp contrast to the prevailing practice of treating transactions identically from a risk standpoint. Predictive modeling is also at the heart of Google's search engine and several other products. But the first machine that could arguably be considered to pass the Turing test and create new insights in the course of problem solving is IBM's Watson, which makes extensive use of learning and prediction in its problem-solving process. In a game like "Jeopardy!," where understanding the question itself is often nontrivial and the domain open-ended and nonstationary, it is not practical to be successful through an extensive enumeration of possibilities or top-down theory building. The solution is to endow a computer with the ability to train itself automatically based on large numbers of examples. Watson also demonstrated the power of machine learning is greatly amplified through the availability of high-quality human-curated data, as in Wikipedia. This trend—combining human knowledge with machine learning—also appears to be on the rise. Google's recent foray in the Knowledge Graph[16] is intended to enable the system to understand the entities corresponding to the torrent of strings it processes continuously. Google wants to understand "things," not just "strings."[26]

Organizations and managers face significant challenges in adapting to the new world of data. It is suddenly possible to test many of their established intuitions, experiment cheaply and accurately, and base decisions on data. This opportunity requires a fundamental shift in organizational culture, one seen in organizations that have embraced the emerging world of data for decision making. C

References

1. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired 16*, 7 (June 23, 2008).
2. Aral, S. and Walker, D. Identifying influential and susceptible members of social networks. *Science 337*, 6092 (June 21, 2012).
3. Buchan, I., Winn, J., and Bishop, C. *A Unified Modeling Approach to Data-Intensive Healthcare. The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, WA, 2009.
4. Dhar, V. Prediction in financial markets: The case for small disjuncts. *ACM Transactions on Intelligent Systems and Technologies 2*, 3 (Apr. 2011).
5. Dhar, V. and Chou, D. A comparison of nonlinear models for financial prediction. *IEEE Transactions on Neural Networks 12*, 4 (June 2001), 907–921.
6. Dhar, V. and Stein, R. *Seven Methods for Transforming Corporate Data Into Business Intelligence.* Prentice-Hall, Englewood Cliffs, NJ, 1997.
7. Frawley, W. and Piatetsky-Shapiro, G., Eds. *Knowledge Discovery in Databases.* AAAI/MIT Press, Cambridge, MA, 1991.
8. Gladwell, M. *The Tipping Point: How Little Things Can Make a Big Difference.* Little Brown, New York, 2000.
9. Goel, S., Watts, D., and Goldstein, D. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (2012), 623–638.
10. Hastie, T., Tibsharani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2009.
11. Heilbron, J.L., Ed. *The Oxford Companion to the History of Modern Science.* Oxford University Press, New York, 2003.
12. Hey, T., Tansley, S., and Tolle, K., Eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, WA, 2009.
13. Hunt, J., Baldochi, D., and van Ingen, C. *Redefining Ecological Science Using Data. The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, WA, 2009.
14. Issenberg, S. A more perfect union: How President Obama's campaign used big data to rally individual voters. *MIT Technology Review* (Dec. 2012).
15. Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery 18* (2009), 140–181.
16. Lin, T., Patrick, P., Gamon, M., Kannan, A., and Fuxman, A. Active objects: Actions for entity-centric search. In *Proceedings of the 21st International Conference on the World Wide Web* (Lyon, France). ACM Press, New York, 2012.
17. Linoff, G. and Berry, M. *Data Mining Techniques: For Marketing, Sales, and Customer Support.* John Wiley & Sons, Inc., New York, 1997.
18. Maguire, J. and Dhar, V. Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed Type 2 diabetics: Data-driven predictive analytics in healthcare. *Health Systems 2* (2013), 73–92.
19. McKinsey Global Institute. *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* Technical Report, June 2011.
20. Meinshausen, N. Relaxed lasso. *Computational Statistics & Data Analysis 52*, 1 (Sept. 15, 2007), 374–393.
21. Papert, S. An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning 1*, 1 (1996), 95–123.
22. Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, U.K., 2000.
23. Perlich, C., Provost, F., and Simonoff, J. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research 4*, 12 (2003), 211–255.
24. Popper, K. *Conjectures and Refutations.* Routledge, London, 1963.
25. Provost, F. and Fawcett, T. *Data Science for Business.* O'Reilly Media, New York, 2013.
26. Roush, W. Google gets a second brain, changing everything about search. *Xconomy* (Dec. 12, 2012); http://www.xconomy.com /san-francisco/2012/12/12/google-gets-a-second-brain-changing-everything-about-search/?single_page=true
27. Shmueli, G. To explain or to predict? *Statistical Science 25*, 3 (Aug. 2010), 289–310.
28. Simon, H.A. and Hayes, J.R. The understanding process: Problem isomorphs. *Cognitive Psychology 8*, 2 (Apr. 1976), 165–190.
29. Sloman, S. *Causal Models.* Oxford University Press, Oxford, U.K. 2005.
30. Spirtes, P., Scheines, R., and Glymour, C. *Causation, Prediction and Search.* Springer, New York, 1993.
31. Tukey, J.W. *Exploratory Data Analysis.* Addison-Wesley, Boston, 1977.
32. Wing, J. Computational thinking. *Commun. ACM 49*, 3 (Mar. 2006), 33–35.

Vasant Dhar (vdhar@stern.nyu.edu) is a professor and co-director of the Center for Business Analytics at the Stern School of Business at New York University, New York.