Deloitte Review
ISSUE 12 | 2013

Complimentary article reprint

Too Big to Ignore

When does big data provide big value?

BY JAMES GUSZCZA, DAVID STEIER, JOHN LUCKER, VIVEKANAND GOPALKRISHNAN, AND HARVEY LEWIS > ILLUSTRATION BY JON KRAUSE

Deloitte.

About Deloitte
Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee, and its network of member firms, each of which is a legally separate and independent entity. Please see www.deloitte.com/about for a detailed description of the legal structure of Deloitte Touche Tohmatsu Limited and its member firms. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

Copyright 2012 Deloitte Development LLC. All rights reserved.
Member of Deloitte Touche Tohmatsu Limited

Too Big to Ignore

When does big data provide big value?

BY JAMES GUSZCZA, DAVID STEIER, JOHN LUCKER, VIVEKANAND GOPALKRISHNAN, AND HARVEY LEWIS
> ILLUSTRATION BY JON KRAUSE

"Beware of false knowledge; it is more dangerous than ignorance."

—George Bernard Shaw

THE SIGNAL AND THE NOISE

By now, we have all heard the claims. Data is the new oil. Big data is different.[1] Big data is a management revolution.[2] Big data is the next frontier for innovation, competition, and productivity.[3] Big data makes the scientific method obsolete.[4]

On the one hand, it is easy so see why big data has generated so much excitement both in the business press and in the larger culture. New business models, scientific breakthroughs, and profound societal transformations are all observable effects of big data.

Familiar examples abound. Google Translate exploits word associations in massive databases of free-form text to yield a tool that can, if you wish, instantly translate Icelandic to Indonesian. Social, political, economic, and professional relationships—and the societal and marketing mores surrounding them—are rapidly evolving along with the evolution of social networking and social media technologies. Political campaigns increasingly harness the power of social networks and social media to raise funds, motivate constituencies, and get out the vote. Companies use detailed databases about their customers' behavior and lifestyles not only to better target them, but to create such innovative "data products" as playlists, newsfeeds, next-best offers, and recommendation engines for items ranging from airplane tickets to romantic partners. The raw material for all of these innovations—and surely many more to come—is large amounts of data.

So the topic is undoubtedly important. Yet at the same time, much of the language that has come to surround big data conveys a muddled conception of what data, "big" or otherwise, means to the majority of organizations pursuing analytics strategies. Big data is indeed a signature issue of our time. But it is also shrouded in hyperbole and confusion, which can be a breeding ground for strategic errors. In short, big data is a big deal, but it is time to separate the signal from the noise.

V IS FOR …

Of course business applications of data analysis and predictive modeling go back decades.[5] For example, credit scoring dates back to the ENIAC-era late 1950s, and actuaries have long analyzed industrial-strength data to price insurance contracts and more recently to set aside appropriate loss reserves as well as guide underwriting and claim adjustment decisions. And in the decade since the appearance of Michael Lewis's *Moneyball*, statistical approaches to improved business decision making have spread to realms as disparate as improving patient safety, making better hiring decisions, and warding off lapses in child support payments. What then is new about big data?

The term is somewhat hard to pin down in part because it is commonly used in at least two distinct senses. In everyday discussions, "big data" is increasingly used as shorthand for the granular and varied data sources that go into the sorts of projects described above.[6] Here "big" is used in the sense of "as rich and detailed as practical given the business context." Such data is "big" relative to the small,

"clean," easily accessible datasets that can easily be manipulated in spreadsheets and have traditionally been fodder for mainstream academic statistical research. While colloquial, this is actually a useful conception of "big data" for reasons that will become clear.

More officially, "big data" denotes data sources whose very size creates problems for standard data management and analysis tools. Examples include the data continuously emanating in vast quantities from digital sensors, audio and video recording devices, mobile computing devices, Internet searches, social networking and media technologies, and so on. Such examples motivate Doug Laney's widely accepted "three V's" characterization of big data:[7]

- **Volume:** Here, "big" is often taken to mean multiple terabyte- or petabyte-class data, motivating the use of such highly parallel next-generation data management technologies as MapReduce, Hadoop, and NoSQL.[8]
- **Variety:** Big data goes beyond numbers in databases, a.k.a. "structured data." Such "unstructured" data sources as Internet search log files, tweets, call center transcripts, telecom messages, email, and data from sensor networks, video, and photographs can equally well be considered data. The multi-structured nature of big data in part accounts for its large volume and often high degree of "messiness" and "noisiness."
- **Velocity:** Because much of it emanates from sensors, web search logs, real-time feeds, or mobile devices, big data is often generated continuously and at a rapid clip.

Big data is said to be different and revolutionary because its size, scope, and fluidity are so great as to enable it to "speak to us," often in real time, in ways not seen before. This point of view is clearly articulated in Chris Anderson's influential essay, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete":[9]

There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

The thinking is that in the pre-big data era, statistical science was necessary to make up for the inherent limitations of incomplete data samples. Statisticians and scientists were forced to cleanse, hypothesize, sample, model, and analyze data to arrive at contingent lessons that are at least consistent with, and at best strongly

supported by, limited data. Today, so the thinking goes, organizations increasingly have access to something approaching a complete sample. Therefore the process of "learning from data" becomes akin more to a problem in algorithm and architecture design than one of learning from and quantifying uncertain knowledge using statistical science.

The mode of thought represented by Anderson's essay can be seductive. But for reasons we will explore, most organizations would do well to resist this particular seduction.

IS BIG DIFFERENT?

Fitzgerald: "The rich are different from you and me."
Hemmingway: "Yes, they have more money."

Google Translate is a case in point of how big data can be dramatically different. In their widely cited article, "The Unreasonable Effectiveness of Data,"[10] Google researchers Alon Halevy, Peter Norvig, and Fernando Pereira describe an approach to translation that hinges on mining the messy patterns from enormous collections of translations that exist "in the wild." The approach is notable in that it bypasses traditional statistical methodology involving laboriously cleansing, sampling, exploring, and modeling data. The very size and completeness of the data now available employ word associations to do the work that linguistic rules and complicated models did in previous, less effective, approaches.

Halevy, Norvig, and Pereira write:

> **Invariably, simple models and a lot of data trump more elaborate models based on less data. … Currently, statistical translation models consist mostly of large memorized phrase tables that give candidate mappings between specific source- and target-language phrases.**

Analogous discussions could be made, for example, about recommendation engines, online marketing experiments performed in rapid succession, and the use of Internet search data to predict flu outbreaks or "nowcast" economic trends. Its near-completeness and real-time nature can make big data an "unreasonably effective" (to use Halevy, Norvig, and Pereira's phrase) force for innovative data products.

In short, why ask why? In many applications, we only care about a good enough next answer, decision, or recommendation. And this is what big data, processed by the right computer algorithms, gives us. Or does it?

HOW DATA SCIENCE IS A SCIENCE

"The whole of science is nothing more than a refinement of everyday thinking."

—Albert Einstein

As important as Google Translate and any number of other big data products are, cursory thinking about their significance can muddle important issues ranging from the epistemological to the economic to the strategic.

First, a fundamental distinction should not be forgotten: Data is not the same thing as information. In itself, data is nothing more than an inert raw material: uninterpreted symbols on a page or in a database. Furthermore, the size of a dataset is often a poor proxy for the amount of useful information it contains. A single Eminem video accounts for more of the world's exabytes of data than do the complete works of Einstein. A typical collection of holiday snapshots from the Galapagos Islands will occupy more disk space than *On the Origin of Species*.

Second, it is deeply misguided to view the skills needed to convert raw data into usable information in software engineering terms. Contrary to the view articulated by Chris Anderson, these skills are inherently scientific in nature. Indeed, they are increasingly labeled by the helpful umbrella term "data science."[11] Data science should be viewed as a synthesis of statistical science, computer science, and domain knowledge appropriate to the problem at hand.[12] The term originated from a farsighted group of statisticians who understood that as data continues to grow in volume and availability, the ability to interact with, visually explore, and generally compute with data becomes an inescapable part of doing serious statistics.[13]

Business projects with data science at their core—a.k.a. business analytics projects—have three primary phases: design, analysis, and execution. Critical thinking, scientific judgment, creativity, and pragmatism are inherent to each of them. It is, generally speaking, unrealistic to expect that any of these phases could be automated or outsourced to computer algorithms processing big data.

Strategy and design: Perhaps a data scientist's most important skill is in understanding (and often helping to articulate) an organization's questions, problems, or strategic challenges and then translating them into the design of one or more data analysis projects. John Tukey's famous slogan captures the need for such a strategy-led approach: "Better to have an approximate answer to the right question than a precise answer to the wrong question."

To illustrate, consider a hypothetical chief medical officer of a hospital group seeking to improve patient outcomes. Should the focus of data analysis be on preventing medical errors? Identifying physicians at high risk of being sued for

malpractice? Identifying patients with potentially high medical utilization? Identifying patients likely to fall off their treatments? Predictive approaches to guiding diagnosis and treatment decisions? Even after the various alternatives have been articulated and prioritized, there remains the design challenge of outlining a sensible series of data analysis and/or predictive modeling steps. Indeed the decision of which issues to tackle and in what order might be partially informed by the opinion of a seasoned data scientist regarding their relative feasibility.

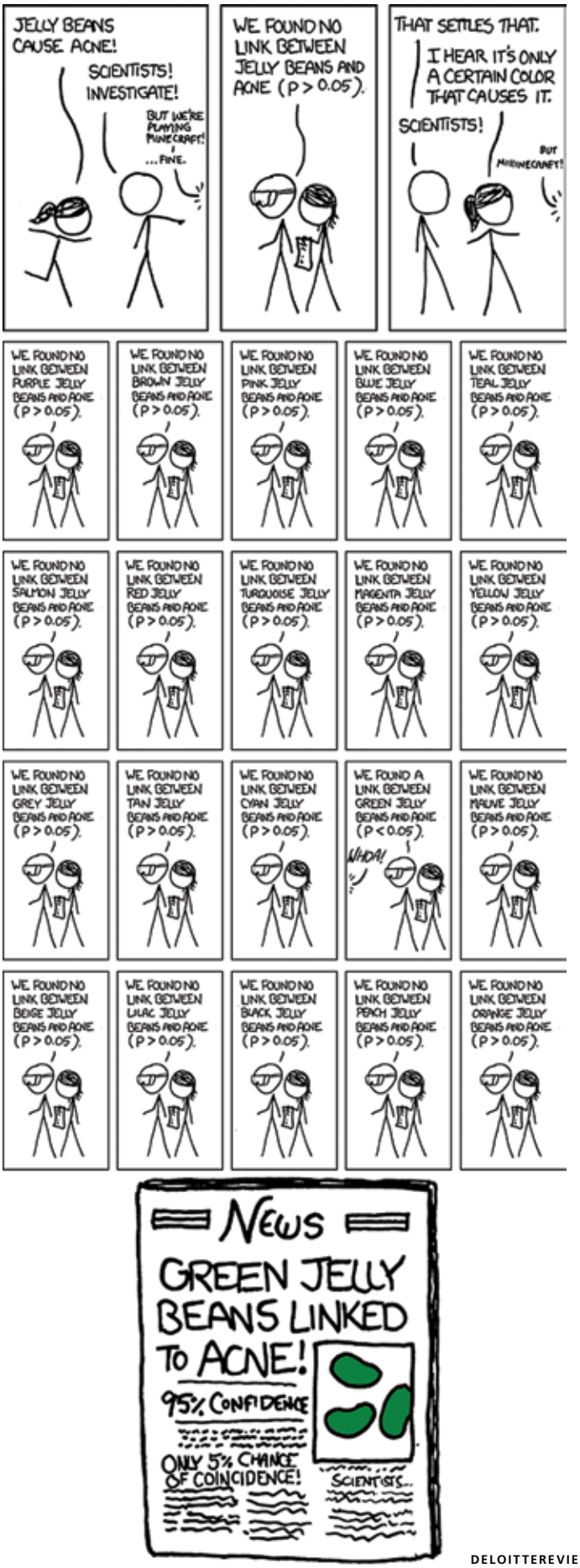
"Better to have an approximate answer to the right question than a precise answer to the wrong question."

Analytical process: The technical aspect of an analytics project itself comprises three distinct phases: data scrubbing and preparation, data analysis, and validation. Each requires judgment, generally making it nonautomatic in character.

The need for statistical programming (computing with data) is rightly emphasized because data comes in such messy and disparate forms: scanned documents, web logs, electronic medical records, billing records and other financial transactions, geospatial coordinates, audio/video streams, and increasingly free-form unstructured text and beyond. But it would be mistaken to conceive of this process purely as a programming challenge.

Analogous to Halevy, Norvig, and Pereira's observation that more data trumps more elaborate models, we have repeatedly found that the creation of innovative data "features" (or "synthetic variables") from raw data elements does more to enhance the effectiveness of analytics projects than does elaborate methodology. Examples of creating explanatory or predictive power where none existed before include calculating the distance between two addresses; calculating social network centrality using one or more shared associations; using behavioral data in a novel way; or creating index variables to serve as proxies for latent traits not directly observable.[14] In short, characterizing the feature creation aspect of data analysis as a programming exercise would be akin to characterizing creative writing as word processing.[15]

After data scrubbing comes data exploration, analysis, and validation. Each of these steps should be viewed as an exercise in creative scientific investigation, aided with the tools of data visualization, statistics, and machine learning. Many datasets used in analytics projects contain considerable amounts of random noise, spurious correlations, ambiguity, and redundancy alongside the abiding "signal" that we wish to capture in a model or analysis.



Reprinted by permission from xkcd.com under Creative Commons license.

A particularly diabolical and underappreciated problem is known in the vernacular as "multiple comparisons." Imagine 100 people in a room flipping fair coins. Because so many coins are being repeatedly flipped, it is likely that at least one of the coins will—by chance—produce so many heads or so many tails that the coin will be found to be biased with a high degree of "statistical significance."[16] This is despite the fact that the coin thus selected is in reality no different from the other coins and is likely to produce heads in roughly half of the future flips.

A venerable motto of computer science is GIGO: "garbage in, garbage out." Perhaps an analogous motto for the nascent field of data science ought to be NIINO: "no insight in, none out."

In fields such as medicine, drug testing, and psychology, essentially this phenomenon is known as the "file drawer problem": A purely mechanical approach of filing away nonsignificant (or sometimes unwanted) results and publicizing "significant" results systematically yields spurious findings that will likely not hold up over time.[17] The xkcd cartoon reproduced here conveys the issue humorously and succinctly.[18]

While this is a problem for any large-scale analysis involving many possible data dimensions and relationships, it is particularly acute in the age of big data and brute-force algorithms for extracting interesting relationships. Furthermore, as will be discussed below, the cost of such false positives should be taken into account.

The problem of multiple comparisons is but one of many reasons why data science should not, in general, be viewed as the brute-force harvesting of patterns from stores of big data. For others, see the inset, "Putting the Science in Data Science."

Business execution: As the statistician George Box wrote, "All models are wrong, but some are useful." A major implication of this statement is that it is generally advisable to blend model indications with common sense and expert judgment to arrive at a decision. This happens often even in *bona fide* big data applications. For example, if Jim uses Google Translate to write to a francophone friend he will hopefully know enough to correct various mistakes and change the second person singular from the formal *vous* to the familiar *tu*. Similarly, the recommendations of book or movie collaborative filtering algorithms are often useful but other times should be taken with a grain of salt.[19]

In these everyday examples, the stakes are low, and it is easy to blend model indications (translations or recommendations) with judgment founded on experience (our prior beliefs about the correct translation or what we will enjoy). In more high-stakes business analytics applications, such as using models to make medical diagnoses, screening potential employees, security rating, fraud investigation, or complex loan or insurance underwriting decisions, the blending of expert judgment with model indications should be viewed as a risk management issue of strategic importance.[20] It is important that the data scientist guiding the analytical process play a role in communicating the assumptions and limitations of the analysis or model so that these issues are effectively addressed.[21] This is yet another reason why data science should not be framed in purely programming or engineering terms and why big data should not be characterized as an automatic source of reliable predictions regardless of context.

Amassing repositories of big data and purchasing software, therefore, is not sufficient for business analytics. Data scientists—a.k.a. people—are essential to the process. It should also be kept in mind that in many situations, the appropriate data, at least to start, will not be "big in the 3V sense," and much can be done with open-source statistical computing software. Again, domain knowledge and scientific judgment are important factors in such decisions.

A venerable motto of computer science is GIGO: "garbage in, garbage out." Perhaps an analogous motto for the nascent field of data science ought to be NIINO: "no insight in, none out."

WHAT HAPPENS IN VAGUENESS STAYS IN VAGUENESS

It is natural to find the discussions of big data in the business press and blogosphere bewildering. The field is inherently multidisciplinary, and terms such as "big data," "business analytics," and "data science" mean different things to different people.[22]

Confusion also may arise for a more fundamental reason: Concepts relating to statistical uncertainty simply do not come naturally to the human mind. The same body of psychological research that underpins behavioral economics also suggests that we are very poor natural statisticians. We are naturally prone to find spurious information in data where none exists, latch on to causal narratives that are unsupported by sketchy statistical evidence, ignore population base rates when estimating probabilities for individual cases, be overconfident in our judgments, and generally be "fooled by randomness."[23] There is little wonder that misleading narratives about big data have multiplied.

PUTTING THE SCIENCE IN DATA SCIENCE

(OR, THOSE WHO IGNORE STATISTICS ARE CONDEMNED TO REINVENT IT.)[24]

Not only is information different from data in general, there is no automatic, purely algorithmic way to extract the right islands of information from oceans of raw data. Generally speaking, this process requires a combination of domain knowledge, creativity, critical thinking, an understanding of statistical reasoning, and the ability to visualize and program with data. Theory and sound causal understanding remain important checks on the innate human tendency to be "fooled by randomness." Far from making the scientific method obsolete, increasing volumes of data are making data science a core strategic capability of many organizations. Here are some reasons why.

Data contains too few patterns:

- In many situations, one can fit many models to, and draw a variety of conclusions from, the same data. Examples are all around and include forecasting the profitability of a cohort of insurance policies, estimating Value at Risk (VaR) of a portfolio of securities, evaluating the effectiveness of an ad campaign or human resources policy, analyzing a financial time series, and predicting the outcome of a political election. While more data is often helpful, it is misleading to characterize these activities as "data driven." Rather they are driven by the creative and judgmental application of scientific principles of data analysis. Data—"big" or otherwise—is an input into the process, not the driving force of the process.
- Human creativity and domain knowledge are often necessary to create synthetic data features. Examples include body mass index, measures of social network centrality, distances between relevant physical addresses, composite measures of employee performance, and proxy variables for unobservable latent traits. Once again data is a raw material, not a source of automatic insight.

Data contains too many patterns:

- Datasets contain a mixture of "signal" and "noise," and many big data sources have low signal-to-noise ratios, perhaps earning a fourth "V": "vagueness." While statistical and machine learning techniques continue to improve in their ability to separate signal from noise, they are often best viewed as tools that facilitate an inherently judgmental process. Examples include selecting which variables and variable interactions should be considered for inclusion in a model, which data points should be excluded or down-weighted for various reasons, and whether various apparently linear or nonlinear relationships among variables are real or spurious.
- The problem of "multiple comparisons": It often happens that the more associations you test, the more apparently significant patterns you will detect, even when nothing is actually happening. This is generally regarded as a basic fact of statistical life and becomes more challenging as data sets become larger and messier.

Big datasets are unnecessarily big:

- Summary statistics can be sufficient for the task at hand. An elementary example: Suppose a coin has been repeatedly tossed. Assuming a binomial process, two numbers (the total number of tosses and the number of heads) contain as much information about the probability of heads on the next toss as a complete history of the previous tosses. More bytes do not always translate to more information.
- Scores and composite indices (such as credit scores, social media sentiment scores, or lifestyle clusters) can reduce hundreds of data elements to a handful of numbers.
- Often a small, carefully chosen sample of data contains as much usable information as a large "messy" dataset.

Big datasets are too small:

- Another broadly accepted fact of statistical life is "the curse of dimensionality." In many situations, even the largest conceivable data is "sparse" because of the large number of dimensions involved and/or a rare "outcome" variable (such as fraud or infrequent purchases). Imagine, for example a marketing researcher confronting a dataset containing few or no purchases for many of the hundreds of different products offered; an actuary confronted with setting professional liability insurance rates for a multitude of specialty/geography combinations; or a geneticist analyzing many thousands of gene combinations for a relatively small patient population.
- Biased sampling frames: In 1936 the *Literary Digest* conducted a poll that received over 2 million responses predicting that Alfred Landon would prevail over Franklin Roosevelt by a double-digit margin. In fact, Roosevelt won by a landslide. The *Literary Digest* erred in conducting its poll by telephone, which at the time was disproportionately used by the wealthy. In this sense the huge sample was still too "small." Analogously today, petabytes of social media data aren't guaranteed to accurately reflect the membership or sentiments of the population of interest.[25]

The patterns in the data are trivial:

- While "garbage in, garbage out" is a well-known danger in data science, a less frequently noted GIGO is "gigabytes in, generalities out." Often, an automatic approach to data analysis results in generalities that are obvious, nonactionable, or both. For example, a first attempt at a recommendation engine might suggest the "obvious," most popular movies or songs to most people; similarly a novice analysis of insurance claim severity might uncover obvious facts such as back injuries costing more than fractured arms. Human insight is required to design data analysis approaches that go beyond the obvious.
- On the other hand, "obviousness" is sometimes a good thing. Some of the most valuable models insightfully combine a number of fairly obvious variables. The value of such models is due less to incorporating surprise nuggets of insight (although those are always nice) than to outperforming the human mind's ability to quantify their relative importance and interactions. Also important is that such models help enforce consistency on groups of cognitively bounded, and perhaps distracted and tired, human professionals.[26]

The patterns in the data are diabolically misleading:

- A famous example of "Simpson's Paradox" occurred when the University of California, Berkeley was sued for gender bias in its graduate school admissions decisions: In 1973 Berkeley admitted 44 percent of its male graduate school applicants but only 35 percent of its female applicants. However when the data was broken down by department, the apparent bias disappeared. Why? Women tended to apply to departments (such as humanities) with lower admission rates—doh! Similarly naive analyses can spuriously suggest that higher prices lead to higher demand, or that marketing campaigns can lead to lower sales.
- Studies find that highly intelligent women tend to marry men who are less intelligent than they are.[27] Top-performing companies tend to do worse over time. These facts do not require sociological or management theoretic explanations. They are instances of "regression to the mean." The concept was first identified by Francis Galton, a cousin of Charles Darwin and the inventor of regression analysis, when he studied such phenomena as tall parents having shorter offspring (and vice versa). Misunderstanding of this phenomenon can lead to such bad business decisions as paying for past performance of sports players who have experienced winning streaks, or movie genres and franchises whose moment has passed.

Nevertheless, avoiding misconceptions about big data should be regarded as a prerequisite for avoiding analytics projects with negative ROI. First, data should not be confused with useful information. Deriving insight from data, as discussed above, should generally not be viewed as a form of programming or software implementation but as a type of scientific investigation requiring the judgmental evaluation of ambiguous data.

It is equally important to pay attention to the economic and strategic aspects of big data. While the potential benefits of big data get a lot of attention, less attention is given to the costs. Many of these costs are straightforward: It costs money to acquire, store, back up, secure, integrate, manage, audit, document, and make available any data source.[28] And while inexpensive multi-terabyte drives are available at the local electronics store, they are not characteristic of the enterprise hardware many organizations use to manage big data.

More subtle economic points can be equally relevant. First, economic decisions should be made at the margins. Therefore, a big data project should not be evaluated in isolation but in terms of how much insight or predictive power it is likely to add over and above a less costly analytics project.

Second, big data projects often carry opportunity costs. Many organizations have a menu of potential analytics projects with limited resources to execute them. For a large retailer wishing to make real-time next-best offers or an Internet company aiming to enjoy the "winner take all" effect of network externalities, big data naturally rises to the top of the list of priorities. For many other organizations, a more likely path to analytics success is paved with a sequence of smaller, well-targeted projects, with the benefits of one helping fund the next. For such organizations, a focus on big data could be an expensive distraction.

Finally, the human capital and organizational costs required to work with, analyze, and act upon big data should not be underestimated. Data science skills are clearly in demand and, therefore, can be difficult and expensive to acquire. While this is subject to change as supply ramps up to meet demand, a deeper and more abiding point was made over 40 years ago by the artificial intelligence pioneer and management theorist Herbert Simon. Simon wrote that:

> **In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: It consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.[29]**

To be sure, there is ample evidence that when properly analyzed and acted upon, data helps enable organizations to make decisions more accurately,

consistently, and economically.[30] But this does not imply the cost-effectiveness of big data technology in any particular application.

Moneyball is often held up as a prime example of the sort of transformative innovation enabled by data-driven decision making. In fact, variations on "Big Data: *Moneyball* for Business" are common names for articles about the topic. But Billy Beane's achievement did not result from using terabyte- or petabyte-class data. As author Michael Lewis recounts, it resulted from an inspired use of the *right* data (for example on-base percentage rather than batting average) to address the right business opportunity (an inefficient market for talent due to a widespread culture of intuition-driven decision making). Much the same could be said about the use of statistical analysis and predictive models to help guide medical decisions; loan or insurance underwriting decisions; hiring, admissions, and resource allocation decisions; and so on.

This last point sheds light on the "data is the new oil" metaphor. The metaphor is helpful in driving home the point that data should be treated as a valuable asset that, analogous to oil, can be refined to help power insights and better decisions. But unlike oil, data is not an undifferentiated quantity. More volumes and varieties of data are not necessarily better. Indeed, if not pursued strategically, they can lead to missed opportunities, expensive distractions, and what Simon called "a poverty of attention." Rather than let the data tail wag the strategic dog, it is crucial to begin with a plan within which the organization can judge which data is the *right* data and which analysis is the right analysis.

TWO CHEERS FOR BIG DATA

Is big data a big deal? Yes. Instances of innovative big data-driven products, business models, and scientific breakthroughs are already common and are likely to multiply in the future.

But for a leader seeking an appropriate business analytics strategy specific to his or her organization, the answer to this question is less straightforward. There is indeed abundant evidence that organizations—large and small, public and private—benefit from employing the analysis of data to guide strategic, operational, and tactical decisions. But this does not always entail processing terabyte- or petabyte-class data on Hadoop clusters.

Above all, business strategy should guide the choice of data and technology, not vice versa. Laying out a clear vision and analytical roadmap helps avoid being sidetracked by narratives in which the phrase "big data" is defined one way and used in another; data is conflated with usable information; the judgment-infused process of data science is mischaracterized as mining patterns and associations from raw data; and the economics of big data are downplayed.

Properly harnessed, the right data can indeed be an organization's new oil. But it is important not to lose sight of two fundamental points. First, analytics initiatives ultimately do not begin with data; they begin with clearly articulated problems to be addressed and opportunities to be pursued. Second, more data does not guarantee better decisions. But the right data—properly analyzed and acted upon—often does. Organizations that lose sight of these principles risk experiencing big data not as the new oil, but as the new turmoil. **DR**

James Guszcza *is the National Predictive Analytics lead for Deloitte Consulting LLP's Actuarial, Risk, and Analytics practice.*

David Steier *is a director with Deloitte Consulting LLP and leads the Deloitte Analytics Solutions Group.*

John Lucker *is a principal with Deloitte Consulting LLP and Global Advanced Analytics & Modeling Market Offering leader and a US leader of Deloitte Analytics.*

Vivekanand Gopalkrishnan *is a director with Deloitte & Touche Financial Advisory Services Pte Ltd Singapore and leads Deloitte Analytics Institute Asia.*

Harvey Lewis *is a director with Deloitte UK and leads the research program for Deloitte Analytics in the UK.*

Endnotes

1. Thomas H. Davenport, Paul Barth, and Randy Bean. "How Big Data is Different," *MIT Sloan Management Review*, July 30, 2012 <http://sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different/>.
2. Andrew McAfee and Erik Brynjolfsson. "Big Data: The Management Revolution", *Harvard Business Review*, October 2012 <http://hbr.org/2012/10/big-data-the-management-revolution/ar/1>.
3. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute. <http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation>.
4. Chris Anderson. "The End of Theory: Big Data Makes the Scientific Method Obsolete," *Wired*, June 2008. <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory>.
5. Or perhaps centuries. The article "Beyond the Numbers: Analytics as a Strategic Capability" pointed out that actuarial science, which dates back 250 years, might reasonably be considered an early instance of what is today called business analytics. <http://www.deloitte.com/view/en_US/us/Insights/Browse-by-Content-Type/deloitte-review/24cda920f718d210VgnVCM2000001b56f00aRCRD.htm>.
6. For example, see "The New Boss: Big Data," *The Wall Street Journal*, September 20, 2012. This article is about the use of predictive models to improve hiring decisions. Here, "big data" is used in the increasingly common colloquial sense of "supporting a business analytics application," not the sense suggesting the involvement of terabyte- or petabyte-class data.
7. See Doug Laney's recent blog post for a link to the original research note as well as observations about the conception's widespread adoption: <http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/>.
8. A petabyte, which equals 1 million gigabytes and 1,000 terabytes, is approximately the storage space consumed by the movie *Avatar*, one-twentieth of the storage space consumed by the photographs comprising Google Street View, and the amount of data generated by the CERN Large Hadron Collider (LHC) experiments each second. See: <http://cacm.acm.org/news/110048-cern-experiments-generating-one-petabyte-of-data-every-second/fulltext>.
9. See <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory>.
10. IEEE *Intelligent Systems*, March/April 2009. <http://www.csee.wvu.edu/~gidoretto/courses/2011-fall-cp/reading/TheUnreasonable%20EffectivenessofData_IEEE_IS2009.pdf>.

(Continued)

11. Unfortunately, like the term "big data," "data science" has been garbled in the popular press to the extent that prominent analysts have come to express frustration with it. In the blog associated with her Columbia University "Introduction to Data Science" course, Rachel Shutt wrote, "I don't call myself a 'data scientist.' I call myself a statistician. I refuse to be called a data scientist because as it's currently used, it's a meaningless, arbitrary marketing term." <http://columbiadatascience.com/2012/10/04/next-gen-data-scientists/>. Similarly, the mathematician/blogger Cathy O'Neil declared, "There are far too many posers out there in the land of data scientists, and it's getting to the point where I'm starting to regret throwing my hat into that ring." <http://mathbabe.org/2012/07/31/statisticians-arent-the-problem-for-data-science-the-real-problem-is-too-many-posers/>. We sympathize with Shutt's and O'Neil's sentiments and attempt to use the term in a way that is consistent with their writings.
12. In a blog entry, Drew Conway memorably used a Venn diagram to define data science. According to Conway's characterization, data science is the intersection of three circles representing substantive expertise, math and statistics knowledge, and hacking skills. Conway's excellent definition could perhaps be improved with the addition of the abilities to identify and frame problems as well as communicate effectively. <http://www.drewconway.com/zia/?p=2378>.
13. The phrase "data science" in the sense used here originated in a prescient 2001 essay by the Bell Labs statistician and data visualization pioneer William Cleveland: "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics" <http://cm.bell-labs.com/cm/ms/departments/sia/doc/datascience.pdf>. Among other things, Cleveland called for "computing with data" and the evaluation of statistical computing tools to be added to traditional university curricula. It is notable that Cleveland was a collaborator of the other legendary Bell Labs statisticians John Tukey and John Chambers. Tukey is known as the father of exploratory data analysis [EDA] and Chambers was the inventor of the S (as in "statistics") statistical computing language. The original S language, which implemented many of Tukey's and Cleveland's data exploration and data visualization ideas, is the basis for the open-source R statistical computing environment. Today, R is the lingua franca of applied statistics and has played no small role in bringing Cleveland's vision of "applied statistics as data science" to life. Ironically, although Tukey himself never used computers, he coined the terms "bit" and "software," and Chambers discussed him as an inspiration for the development of S. See "John Tukey and 'Software'" by John Chambers: <http://cm.bell-labs.com/cm/ms/departments/sia/tukey/tributes.html>.
14. For example, one of the most successful innovations with regard to personal insurance applications of the 1990s was the adoption of credit scoring to select and price personal auto and homeowner policies. Not only is credit score predictive, it is one of the most powerful rating variables in situations where its use is allowed. This is likely because credit behavior is a proxy for non-directly measurable traits that also manifest themselves in risky behavior. It would be misleading to characterize such innovations as mechanical processes of putting numbers into an algorithm and letting computers find the relevant patterns. Rather, deciding to test this novel application of financial stability information was a minor innovative breakthrough.
15. Which in turn brings to mind Truman Capote's famous takedown of Jack Kerouac: "That's not writing, that's typing."
16. So-called p-values are often used as measures of "significance" which can be thought of as a type of "surprise." A result with a very low p-value (such as 18 heads in 20 tosses) means that we would be very surprised to get such an extreme (or more extreme) result were the coin in fact fair. The problem is that tossing a large number of fair coins—analogous to performing a large number of medical trials or examining a large number of correlations in a big data base—is virtually guaranteed to produce a number of such "significant" results purely due to random chance.
17. The epidemiologist John Ioannidis made essentially this point about medical research in a widely cited 2005 article entitled, "Why Most Published Research Findings Are False." <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/>. Ioannidis's point is illustrated by the recent *New England Journal of Medicine* article "Chocolate Consumption, Cognitive Function, and Nobel Laureates" by Franz Messerli. The article reports "a close, significant linear correlation ($r=0.791$, $p<0.0001$) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries." The article speculated on the effects of certain antioxidants found in chocolate on improved thinking ability. <http://www.nejm.org/doi/full/10.1056/NEJMon1211064>. The article, which was intended as a Nobel prize-season parody, was seriously reported in major media outlets, for example: <http://www.forbes.com/sites/larryhusten/2012/10/10/chocolate-and-nobel-prizes-linked-in-study/>. Note that while the article is a parody, the highly "statistically significant" correlation is an actual fact of the dataset studied by Dr. Messerli. This illustrates that studying multiple correlations is bound to lead to results that are "significant" in a certain technical sense, but in substantive terms are either misleading or simply meaningless. It is also interesting to note that the international team of physicists that announced the discovery of the Higgs boson earlier this year, mindful of the problem of multiple comparisons, used an extremely low "five-sigma" p-value (< 0.000001) threshold. See for example: <http://normaldeviate.wordpress.com/2012/07/11/the-higgs-boson-and-the-p-value-police/>.
18. <http://www.xkcd.com/882/>.
19. Amusing examples are recounted in the *Wall Street Journal* article, "If TiVo Thinks You Are Gay, Here's How to Set It Straight". <http://online.wsj.com/article/SB10382619368723569 08.html>.
20. Paul Meehl, the quantitative psychologist who pioneered the "Actuarial versus Clinical Prediction" school of research, identified a major reason why models can be wrong and dubbed it "the broken leg problem." Suppose a model predicts with high accuracy someone's attendance at a movie theater every Friday. The model might fit the historical data perfectly, yet provide the wrong prediction if the person suffered a broken leg earlier in the week. Someone who has both this detailed knowledge and also understands the inherent limitations of models will know enough to ignore (or at least down-weight) the model indication in this case. See the 1989 Science survey article "Clinical versus Actuarial Judgment" by Dawes, Faust, and Meehl. <http://apsychoserver.psych.arizona.edu/JJBAReprints/PSYC621/Dawes_Faust_Meehl_Clinical_vs_actuarial_assessments_1989.pdf>.

21. This was a major theme of our previous *Deloitte Review,* Issue 10, article, "A Delicate Balance."
22. Even the academic statistics community itself has undergone considerable change in recent decades. Some comments from a recent blog post of the Carnegie-Mellon statistician Cosma Shalizi vividly convey a sense both of this change, as well as the heterogeneity of university-level statistical training: "Suppose you were exposed to that subject as a sub-cabalistic ritual of manipulating sums of squares and magical tables according to rules justified (if at all) only by a transparently false origin myth—that is to say, you had to endure what is still an all-too-common sort of intro. stats. class—or, perhaps worse, a 'research methods' class whose content had fossilized before you were born. Suppose you then looked at the genuinely impressive things done by the best of those who call themselves 'data scientists.' Well then no wonder you think 'This is something new and wonderful;' and I would not blame you in the least for not connecting it with statistics. Perhaps you might find some faint resemblance, but it would be like comparing a child's toy wagon to a Ducati. Modern statistics is not like that, and has not been for decades... the skills of a 'data scientist' are those of a modern statistician." <http://masi.cscs.lsa.umich.edu/~crshalizi/weblog/925.html>.
23. See *Thinking, Fast and Slow* by Daniel Kahneman. Kahneman reported that his landmark research exploring systematic biases in human cognition and leading to behavioral economics was motivated by his experience teaching statistics to university students in Israel. He found what he was teaching to be very unintuitive. This observation ran counter to a then-prominent theory that humans are natural statisticians. In the language of *Thinking, Fast and Slow*, "System 1" thinking is rapid and biased towards belief and causal narratives rather than skepticism and rational analysis. System 1-style thinking accounts for the bulk of our mental operations and—here's the rub—is very poor at statistical reasoning. "System 2" thinking is slow, effortful, and strives for logical coherence rather than unsupported narrative coherence. The phrase "fooled by randomness" is of course borrowed from the Nassim Nicholas Taleb book of the same name. Regarding Taleb, Kahneman writes, "The trader-philosopher-statistician Nassim Taleb could also be considered a psychologist. … Taleb suggests that we humans constantly fool ourselves by constructing flimsy accounts of the past and believing they are true." See our previous *Deloitte Review* article "A Delicate Balance" for a discussion of how the Kahnenan school's findings relate to some of the organizational biases that can impede analytics projects.
24. This quote is attributed to the Stanford statistician Brad Efron, who is best known as the inventor of the statistical technique known as bootstrapping. Another Efron quote partially inspired this essay: "In some ways I think that scientists have misled themselves into thinking that if you collect enormous amounts of data you are bound to get the right answer … the fallacy that if we could just get it all inside the computer we would get the answer." <http://www-stat.stanford.edu/~ckirby/brad/other/2010Significance.pdf>.
25. An intriguing election-year example concerns differing treatments of the presidential candidates in the traditional and social media. The authors thank our colleague Michael Greene for bringing this example to their attention. <http://www.journalism.org/commentary_backgrounder/how_social_and_traditional_media_differ_their_treatment_conventions_and_beyo>.
26. A dramatic example of "decision fatigue" (a.k.a. "ego depletion") is judges' parole decisions being strongly influenced by time of day, and presumably blood sugar level. <http://www.nytimes.com/2011/08/21/magazine/do-you-suffer-from-decision-fatigue.html>.
27. In *Thinking, Fast and Slow* (Chapter 17), Daniel Kahneman provided this example to illustrate how regression to the mean tends to be misinterpreted in causal terms.
28. It is interesting to speculate whether behavioral economics helps account for an apparent bias toward hoarding data that one often encounters. The tendency of humans to feel losses more keenly than commensurate gains is known in the behavioral economics literature as "loss aversion." As the availability of computer storage power grows exponentially, perhaps there is a tendency to the loss avoidance strategy of storing as much data as possible to ward off the fear of future feelings of loss that would result from having thrown out something useful. This is not to say that organizations should never err on the side of keeping information with no obvious or immediate use; only that the process should be guided by cost/benefit estimates rather than a default strategy of keeping nearly everything.
29. H. A. Simon, (1971), "Designing Organizations for an Information-Rich World," in Martin Greenberger, *Computers, Communication, and the Public Interest*, Baltimore, MD: The Johns Hopkins Press, pp. 40–41.
30. Popular books such as *Moneyball* and Ian Ayres's *Super Crunchers* have offered considerable anecdotal evidence for this, as have our previous *Deloitte Review* articles "Irrational Expectations" and "Beyond the Numbers". <http://www.deloitte.com/view/en_US/us/Insights/Browse-by-Content-Type/deloitte-review/24cda920f718d210VgnVCM2000001b56f00aRCRD.htm>. For corroborating academic evidence, see *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* by Erik Brynjolfsson, Lorin Hitt, and Heekyung Kim. The authors find that "data-driven decision making" was associated with 5–6 percent higher productivity in a sample of 179 publicly traded firms studied. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486>.