MITSloan
Management Review

WINTER 2017
ISSUE

Why Big Data Isn't Enough

There is a growing belief that sophisticated algorithms can explore huge databases and find relationships independent of any preconceived hypotheses. But in businesses that involve scientific research and technological innovation, the authors argue, this approach is misguided and potentially risky.

Sen Chai
Willy Shih

Vol. 58, No. 2 | Reprint #58227 | http://mitsmr.com/2faldUh

Why Big Data Isn't Enough

There is a growing belief that sophisticated algorithms can explore huge databases and find relationships independent of any preconceived hypotheses. But in businesses that involve scientific research and technological innovation, the authors argue, this approach is misguided and potentially risky.

BY SEN CHAI AND WILLY SHIH

AS "BIG DATA" becomes increasingly integrated into many aspects of our lives, we are hearing more calls for revolutionary changes in how researchers work. To save time in understanding the behavior of complex systems or in predicting outcomes, some analysts say it should now be possible to let the data "tell the story," rather than having to develop a hypothesis and go through painstaking steps to prove it. The success of companies such as Google Inc. and Facebook Inc., which have transformed the advertising and social media worlds by applying data mining and mathematics, has led many to believe that traditional methodologies based on models and theories may no longer be necessary. Among young professionals (and many of the MBA students we see), there is almost a blind faith that sophisticated algorithms can be used to explore huge databases and find interesting relationships independent of any theories or prior beliefs. The assumption is: The bigger the data, the more powerful the findings.

As appealing as this viewpoint may be, we think it's misguided — and could be risky for companies. For example, what if the data appears to support a new drug design or a new scientific approach when there actually isn't a causal relationship? Although we acknowledge that data mining has enabled tremendous advances in business intelligence and in the understanding of consumer behavior — think of how Amazon.com Inc. figures out what you might want to buy or how content recommendation engines such as those used by Netflix Inc. work — it's important for executives who oversee technical disciplines to be thoughtful about how they apply this approach to their areas.

Recently, we looked at several fields where massive amounts of data are available and collected: drug discovery and pharmaceutical research; genomics and species improvement; weather forecasting; the design of complex products like gas turbines; and speech recognition. (See "About the Research," p. 58.) In each setting, we asked a series of broad questions, including the following: How do data-driven

?

THE LEADING QUESTION

How should big data alter the way scientific research is conducted?

FINDINGS

- **While some have argued that big data obviates the need for hypotheses in research, this belief has pitfalls.**
- **Researchers need to pay close attention to issues such as biases in data collection and spurious correlations.**
- **Rather than replacing traditional research methods, data-intensive approaches should be treated as complementary.**

ABOUT THE RESEARCH

The idea for this article arose from a series of research interviews with leaders at a large gene-sequencing company. Several of the individuals, all of whom had been classically trained in various branches of the life sciences, asserted that data-driven research was the new paradigm for science. This view is consistent with the arguments contained in the book "The Fourth Paradigm: Data-Intensive Scientific Discovery,"[i] a collection of essays that examine how different fields of scientific endeavor are being transformed by data. At the same time, we found that working scientists and engineers at a number of other companies were questioning the degree to which managers were embracing data-driven research as a replacement for traditional methods. This prompted us to spend time with researchers in fields that were in the throes of applying large-scale data collection and analysis. We conducted semi-structured interviews in several segments of the life sciences, weather modeling, industrial engineering, and engineering analysis fields to understand how data-intensive research methods complemented traditional methods. Our hope is that our analysis provides executives in science- and engineering-intensive businesses with a more nuanced view of what data-driven methods can and cannot do.

research approaches fit with traditional research methods? In what ways could data-driven research extend the current understanding of scientific and engineering problems? And what cautions do managers need to exercise about the limitations and the proper use of statistical inference?

Pursuing Data-Driven Research

Based on what we found, we have developed some guidelines for using big data effectively: how to extract meaning from open-ended searches; how to determine appropriate sample sizes; and how to avoid systematic biases. We have also identified several opportunities in which the use of large datasets can complement traditional hypothesis generation and testing, and have reaffirmed the importance of theory-based models.

1. Beware of spurious correlations in open-ended searches. Using large datasets to support the testing of theories and development of novel insights is not new; this approach has been used for many years in fields such as drug discovery, genomics, and weather forecasting. What is new is open-ended searching for relationships and correlations without having a clear goal in mind[1] — something the director of a large Asian genomics organization based in China described to us as "letting the data speak."

A pitfall in studying large datasets with billions of observational data points is that large deviations are often more attributable to the noise than to the signal itself; searches of large datasets inevitably turn up coincidental patterns that have no predictive power. A 2010 paper by economists Carmen Reinhart and Kenneth Rogoff, for example, reported a correlation between a country's economic growth and its debt-to-GDP ratio: When public debt exceeded 90% of GDP, the economists found, countries experienced slower economic growth than countries with lower levels of debt.[2] However, other researchers subsequently argued that the supposed correlation was meaningless — indeed, later research found that high debt did not necessarily cause slower economic growth.[3] This turned out to be another reminder that correlation is different from causation.

2. Be conscious of sample sizes and sample variation when mining for correlations. In traditional statistical analysis, it's common to collect many data points while varying a small number of independent variables. The number of data points typically far exceeds the number of variables. However, with open-ended searches, the raw data is likely to be spread across a wide range of new (but not necessarily relevant) variables.[4] It's important to recognize that the number of data points required for statistically significant results needs to increase as the number of variables grows. Otherwise, there will be a greater risk of false correlations. In the area of precision medicine, where researchers might be looking at the gene sequences of a few hundred subjects to study a rare disease, the number of variables (in this case, genes) might be in the thousands. Adding new data sources such as online repositories or aggregated databases increases the likelihood that the number of independent variables will exceed the number of observations.

As the number of dimensions expands, the need for sufficient variation in the sample size is also important. Statistically, low variation can lead to biased estimates and limit predictive power — especially at the tails of the distribution. With the drastic growth in dimensionality, researchers must be mindful of both sample size and sample variation.

3. Beware of systematic biases in data collection. An additional caution involves the potential for systematic biases in measurement, which can lead to spurious results. With very large datasets drawn from multiple pools of data, variations in experimental or measurement conditions can produce data that is not comparable. For example, the data may have been collected at different times, or with different technologies, or aggregated from multiple sources using different collection conditions. Researchers in genomics have long understood the importance of recognizing such biases, and they use standardized normalization techniques to remove the distortions.[5] But with open-ended searches, researchers need to understand potential measurement biases and pay close attention to how experiments are designed.

The Role of Hypothesis Generation and Models

Large scientific datasets have been around for years. Pharmaceutical researchers, for example, have long used technologies such as high-throughput screening and combinatorial chemistry to synthesize large

numbers of compounds. But simply making new chemicals robotically and looking at data apparently doesn't lead to new drugs.[6] Rather, effective drug discovery comes from having a theoretical framework for structure-activity patterns, or models of drug absorption and metabolism that enable the selection of targets and drug candidates, and inform the design of new experiments.

Although fields such as drug discovery and genomics build upon explicit models, sometimes the models can be *implicit*. Even the Google Flu Trends service, which attempted to predict the spread of influenza based purely on a big data analysis of internet searches rather than traditional epidemiological studies, relied on an implicit model of how the disease spreads based on location and physical proximity.[7] While a new technique appears to be relying solely on data for its conclusions, usually there is a theoretical underpinning that we might not recognize or simply take for granted.

One area we examined in which large datasets and machine learning appear to have made tremendous advances over model-based approaches is speech recognition. Traditional speech recognition systems use probabilistic models for things like determining how a speaker's voice varies or measuring acoustic characteristics of the speaking environment.[8] A computer program optimizes parameters by having the speaker read a sequence of words to "train" the system, and algorithms then match the incoming sounds to words.

Google's data-driven approach leverages the recordings of millions of users talking to Google's voice search or Android's voice input service, which are then fed into machine learning systems. Does this mean that there is no longer a need for models? Quite the contrary — these machine learning systems invariably incorporate a model that includes knowledge of the structure of the data to train the system. Often researchers use what are known as generative models: The first layer of the machine learning algorithm trains the next layer, applying what it knows.[9] So progress comes from more than just having and using the data.

Having a priori hypotheses helps researchers spot and exploit natural experiments. In a modern factory that collects large-scale data, it's possible to exploit natural variation in process conditions without actually comparing different approaches. The director of manufacturing technology at a large pharmaceutical company pointed out to us that it is often impractical to do experiments on actual production batches of drugs: The manufacturers will not allow it. But by using established theories and models, managers can design data collection strategies that exploit the natural variation in the data being acquired — without having to explicitly design and run separate experiments.

Opportunities to Improve Models

Cautions aside, we believe that combining data-driven research with traditional approaches provides managers with opportunities to refine or develop new and more powerful models. Unexpected correlations that arise from mining large data can strengthen existing models or even establish new ones.

Strengthening Existing Models Weather modeling and forecasting offer a good example. Modern meteorology is based on dividing the Earth's atmosphere into a three-dimensional grid of interconnected cells and then mathematically modeling how the conditions in each cell evolve over time. Each meteorological prediction needs to be divided into two stages. First, during the data assimilation stage, data collected from radar and satellites are used as the initial inputs into a physics model that can be calculated within each cell. For points in the atmosphere that can't be measured directly, the models infer temperature, humidity, wind speed, and other factors from light wavelengths, infrared readings, and thermal radiances acquired by

By using established theories and models, managers can design data collection strategies that exploit the natural variation in the data being acquired — without having to explicitly design and run separate experiments.

radar pulses and satellite cameras. Then, in the forecasting stage, the computer simulation model looks ahead through tiny time steps and produces an output for each cell as a function of time. The outputs are aggregated to develop forecasts for the upcoming days for a particular location.

Increasing the scale and scope of data observations has led to the discovery of unexpected phenomena that have helped to improve underlying physics models. One such example can be found in the Madden-Julian Oscillation (MJO), which is the largest element of the intraseasonal variability in the tropical atmosphere over the Indian and Pacific oceans.[10] Although meteorologists had previously thought that MJO events (which manifest themselves in clusters of thunderstorms) were limited to latitudes within 10 degrees of the equator, researchers have found through data mining that MJO events can have significant impacts on weather phenomena outside of the tropics. "MJOs may start in the southwest Pacific, but [they] can have an influence on the weather in Boston 30 days later," noted Peter Neilley, senior vice president of The Weather Co.'s global forecasting services unit.[11] Once researchers discovered the connection, they were able to modify the model.

Creating New Models Numerical modeling of physical systems has been used in a wide range of disciplines, including aerodynamics, seismology, materials science, and medicine, and it has played an important role in the design of industrial products, civil engineering structures, medical devices, and more. What began as the application of basic principles of physics and continuum mechanics has evolved to include sophisticated numerical methods based on the idea that large complex objects can be broken down and modeled as sets of individual elements. The simple equations for the individual elements are then assembled into a larger system that models the entire problem.[12] Today, advanced software running on high-performance computing systems enables engineers to build increasingly sophisticated simulation models before building prototypes, and then to test them with an eye toward continual refinement.

Even with today's inexpensive computing resources, however, there are limits to how big and how complex simulations models can get — you still can't model everything. Data-driven research might provide an opportunity to guide the evolution of simulation models in a more efficient way. The new PW1000G jet engines from Pratt & Whitney, a division of United Technologies Corp. that designs, manufactures, and services aircraft engines, are a good example. While the designers incorporated a vast array of sensors into the engine design to enable predictive maintenance and to anticipate when a critical part might fail, the data could also be used to improve theory and advance simulation models. Expanded dimensionality will likely include more external parameters such as atmospheric conditions or new measures that have not yet been considered. In the context of weather forecasting, for example, improvements in observation will provide better inputs to the physics and forecasting models, which will lead to greater forecast accuracy.

We think such synergy will be possible in other areas as well, especially in areas like genomics and the life sciences, where so many dimensions are yet unknown. Rather than big data and modeling working separately, we see them complementing each other: The models will help define what data to collect, and the collected data will be used to refine models and improve their designs.

A New Paradigm?

How will new scientific knowledge be produced? Empirical observation and experimentation are often referred to as the first paradigm of scientific research and the foundation upon which natural sciences got started. But experiments in and of themselves can't tell researchers why things occur — that required the development of theoretical models (the second paradigm). As scientists tried to expand their scope to large-scale systems, models were frequently seen as inadequate. Paul Dirac, the Nobel Prize-winning theoretical physicist, observed in 1929 that one could use models to calculate the behavior of a few atoms at a time.[13] But at the time of his remark, the calculations required for something that one could actually see in a test tube were far beyond what was even imaginable. Even today, although theoretical models are able to predict behavior at a microscopic level, solving the equations for real systems beyond a few hundred or a few thousand atoms is difficult. Achieving the next level required the ability to conduct computer simulations and modeling — the third paradigm.

We think of data-intensive methods as *supplemental* to existing methods — a way to expand dimensionality, discover potentially new relationships, and refine theory.

The idea of a fourth paradigm based on data-intensive scientific research has been credited to Jim Gray, an influential software pioneer and Microsoft Corp. researcher.[14] Gray argued that prior beliefs weren't necessary and that results were fully and solely driven by what was found in the collected data. Our interviews with managers of science- and technology-based companies, however, made the case that research should not be solely data-driven. Rather than it being a fourth paradigm, we think of data-intensive methods as *supplemental* to existing methods — a way to expand dimensionality, discover potentially new relationships, and refine theory. Clearly, data-intensive methods are important complements to experimentation, theoretical models, computer modeling, and simulation because they take us into a realm beyond what such methods are capable of today. Researchers just need to be careful about how they use them.

Sen Chai *is an assistant professor of management at ESSEC Business School in Cergy-Pontoise, France.* ***Willy Shih*** *is the Robert and Jane Cizik Professor of Management Practice in Business Administration at Harvard Business School in Boston. Comment on this article at http://sloanreview.mit.edu/x/58227, or contact the authors at smrfeedback@mit.edu.*

REFERENCES

1. D. Simchi-Levi, "OM Research: From Problem-Driven to Data-Driven Research," Manufacturing & Service Operations Management 16, no. 1 (February 2014): 2-10.

2. C.M. Reinhart and K.S. Rogoff, "Growth in a Time of Debt," American Economic Review 100, no. 2 (May 2010): 573–578.

3. T. Herndon, M. Ash, and R. Pollin, "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff," Cambridge Journal of Economics 38, no. 2 (March 2014): 257–279.

4. A. Hero and B. Rajaratnam, "Large Scale Correlation Mining for Biomolecular Network Discovery," technical report no. 2015-02, Stanford Department of Statistics, Stanford, California, January 2015; and C. Rudin, D. Dunson, R. Irizarry et al., "Discovery With Data: Leveraging Statistics With Computer Science to Transform Science and Society," white paper, American Statistical Association, Alexandria, Virginia, June 2014.

5. J. Aleksic, S.H. Carl, and M. Frye, "Beyond Library Size: A Field Guide to NGS Normalization," June 19, 2014, http://dx.doi.org/10.1101/006403.

6. J.G. Lombardino and J.A. Lowe 3rd, "The Role of the Medicinal Chemist in Drug Discovery — Then and Now," Nature Reviews Drug Discovery 3, no. 10 (October 2004): 853-862.

7. Google discontinued this program after a failure that missed the peak of the 2013 flu season by 140%; see, for example, D. Lazar and R. Kennedy, "What We Can Learn From the Epic Failure of Google Flu Trends," October 1, 2015, www.wired.com.

8. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE 77, no. 2 (February 1989): 257-286.

9. For a more extensive discussion of this topic, see, for example, G. Hinton, L. Deng, D. Yu et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Processing Magazine 29, no. 6 (November 2012): 82-97; X.-W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access 2 (May 2014): 514-525; and L. Deng and N. Jaitly, "Deep Discriminative and Generative Models for Pattern Recognition," chap. 1.2 in "Handbook of Pattern Recognition and Computer Vision," 5th ed., ed. C.H. Chen (Singapore: World Scientific Publishing, 2016).

10. C. Zhang, "Madden-Julian Oscillation," Reviews of Geophysics 43, no. 2 (June 2005): 1-36.

11. P. Neilley, interview with authors, Aug. 17, 2015.

12. P.E. Grafton and D.R. Stome, "Analysis of Axisymmetrical Shells by the Direct Stiffness Method," AIAA Journal 1, no. 10 (1963): 2342-2347; and M.J. Turner, R.W. Clough, H.C. Martin, and L.J. Topp, "Stiffness and Deflection Analysis of Complex Structures," Journal of the Aeronautical Sciences 23, no. 9 (September 1956): 805-823.

13. P.A.M. Dirac, "Quantum Mechanics of Many — Electron Systems," Proceedings of the Royal Society A 123, no. 792 (April 6, 1929): 714-33.

14. T. Hey, S. Tansley, and K.M. Tolle, eds., "The Fourth Paradigm: Data-Intensive Scientific Discovery" (Microsoft Research, Redmond, Washington, 2009).

i. Hey, Tansley, and Tolle, "The Fourth Paradigm."

Reprint 58227.

MITSloan Management Review

PDFs ■ Reprints ■ Permission to Copy ■ Back Issues