Systems Research and Behavioral Science
Syst. Res. **32**, 28–49 (2015)
Published online 20 September 2013 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/sres.2219

■ *Research Paper*

The False Promise of Big Data: Can Data Mining Replace Hypothesis-Driven Learning in the Identification of Predictive Performance Metrics?

David Apgar*

Social Impact, Inc., Arlington, VA, USA

This paper argues US manufacturers still fail to identify metrics that predict performance results despite two decades of intensive investment in data-mining applications because indicators with the power to predict complex results must have high information content as well as a high impact on those results. But data mining cannot substitute for experimental hypothesis testing in the search for predictive metrics with high information content—not even in the aggregate—because the low-information metrics it provides require improbably complex theories to explain complex results. So theories should always be simple but predictive factors may need to be complex. This means the widespread belief that data mining can help managers find prescriptions for success is a fantasy. Instead of trying to substitute data mining for experimental hypothesis testing, managers confronted with complex results should lay out specific strategies, test them, adapt them—and repeat the process. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords predictive performance metrics; complexity; information content; hypothesis-driven learning; data mining

INTRODUCTION

Despite decades of intensive investment in data analysis applications, US firms still have trouble identifying metrics that can predict and explain performance results. Research on complexity and predictive metrics is beginning to show why. Data mining, or the search for patterns among variables captured by data applications, cannot replace a form of experimental hypothesis testing here called hypothesis-driven learning in the identification of complex predictive metrics. Confronted with complex results, managers should lay out hypotheses in the form of highly specific strategies, test them in practice, adapt them—and then repeat the process.

This paper rejects as a fantasy the widespread belief that data mining can identify prescriptions for success. The reason is that it cannot generate metrics with the potential to predict, explain or control complex patterns of results. It may be useful in testing the predictive power of metrics once managers propose them, as in Karl Popper's conjectures and refutations

*Correspondence to: David Apgar, 2804 Dumbarton Street NW, Washington, DC 20007, USA.
E-mail: d_apgar@msn.com

Received 15 March 2012
Accepted 10 August 2013

model of scientific discovery (Popper, 2002). Indeed, others have suggested this approach as a planning model for the US Army (Dewar *et al.*, 1993) and start-ups (McGrath and MacMillan, 1995). But data mining cannot replace experimental hypothesis testing or simply hypothesis-driven learning.

Hypothesis-driven learning for the purposes of this paper is defined as recursive experimental hypothesis testing. In the context of business these hypotheses are generally strategies. Experimental hypothesis testing is taken in the Bayesian sense (pace Popper) of updating the probabilities assigned to competing hypotheses or strategies in the light of new evidence or results based on how much more or less likely each hypothesis makes that evidence. Recursive hypothesis testing is Bayesian in the sense that each new set of results updates the probabilities resulting from the last update to the extent possible while always allowing for the introduction of novel hypotheses or strategies. In sum, hypothesis-driven learning refers here to the repeated use of results in testing business hypotheses to find one with high predictive power.

The first section of the paper illustrates the argument by contrasting HP's recent efforts to derive strategy from data with the tradition of hypothesis-driven learning at Toyota. The second section shows what is at stake by assessing the weak impact of decades of intensive investment in data analysis on the capacity of US manufacturers to explain performance results.

The main argument lies in the third and fourth sections. The third section links ideas about requisite variety, mutual information and probabilistic causation to show why metrics with the power to predict complex results must have high information content as well as a high impact on those results. The fourth section proposes an approach to proving Pareto's conjecture that only a few metrics are vital to most problems (Pareto, 1896). Data mining cannot substitute for hypothesis-driven learning in the search for predictive metrics with high information content—not even in the aggregate—because it relies on linking multiple low-information metrics in theories that turn out to be improbably complex.

A final section illustrates the thesis by contrasting WellPoint's reliance on data-driven requirements with Nestlé's continual negotiated experiments to cope with complex results.

DIVERGENT APPROACHES TO SIMILAR CHALLENGES AT HP AND TOYOTA

This section contrasts HP's efforts to derive strategy from data with Toyota's tradition of hypothesis-driven learning. The strategic challenges are similar, but Toyota is faring better.

HP's recent succession of chief executive officers (CEOs) has made the broad outlines of its strategic debates a matter of public record. The story starts with Carly Fiorina, appointed CEO in 1999 as the firm debated whether to split the businesses serving companies from the ones making printers and other consumer devices. Fiorina's answer was a bold move sideways—the audacious 2002 purchase of Compaq to shore up HP's consumer business—before deciding the larger question.

Industry analysts credited Fiorina with a well-executed merger, but the results were lacklustre. Whereas the combined PC shares of Compaq and HP were 21% against Dell's 12% in 2000, the 2004 share of Dell was 18% against 16% for the merged firm (Gaitonde, 2005). Moreover, HP's deeper dive into PCs shed little light on the value of synergies between its profitable printer business and growing enterprise business.

Mark Hurd took over as CEO from 2005 to 2010 and his HP bio summarizes his strategic direction for the firm. He concentrated on enterprise data centre architecture and services, mobile technologies, and the move from analogue to digital imaging and printing. In other words, except for a nod at the evolution from personal to mobile computing, he stayed the course.

Analysts remember Hurd fondly for the steady 7.5% per year growth in revenue he oversaw and his relentless focus on cost control and execution. As did Fiorina, he managed a major acquisition—data services firm Electronic Data Systems (EDS)—without really inflecting strategy. But he consolidated HP data centres from 85 to 6; he pruned software applications from 6000 to 1500, and he fattened operating margins from 4% in 2005 to 8.8% in 2008 (Lashinsky, 2009).

A self-described 'rack-and-stack' guy, Hurd was so focused on numbers that he reportedly disliked graphics in presentations. After HP's board ousted him over inappropriate expenses

DOI: 10.1002/sres.2219

and allegations about sexual advances, he crossed Route 101 to help run Oracle, one of the three largest enterprise data management companies in the world.

And then HP hired a strategist from SAP, another of the three largest enterprise data managers in the world. Once he took the reins, Leo Apotheker announced decisively that HP would let go of its PC business and buy British enterprise software star Autonomy.

Pascal Gobry wrote in August of 2011 that HP was 'ditching doomed…businesses and focusing on where…it has a competitive advantage by being a big company based in Silicon Valley with a tradition of R&D…' (Gobry, 2011). Whether or not the strategy was right, the important thing, Gobry argued, was to execute it and find out. But the share price sagged as PC sentimentalists protested, and the board replaced Apotheker after just 10 months with E-Bay's Meg Whitman.

At her September 2011 press conference, Whitman said the important thing was to rebuild confidence by executing well because HP's strategy was sound. And yet the firm made no move to sell its PC business or double down on enterprise software. In fact, the only clear move related to strategy was the departure of long-time chief strategist Shane Robison and the decision not to replace him. There was a lot of market analysis while Autonomy floundered.

HP's board may be afraid to jump into the deep end of the enterprise software swimming pool. But it could not offer a more heartfelt testimonial than its experiment-shy stewardship of HP's decade of management changes to the hope that some kind of advance in analysis will somehow free the modern enterprise from the messy business of trial and error.

Toyota looks out at a landscape similar in many ways to those of HP. It has led its industry. Changes in automobile power sources are in some ways as momentous as changes in mobile and enterprise connectivity. Doubts about Toyota's quality standards—similar to those of HP—grew as the firm consolidated its leadership position. And both firms have distinctive cultures—the Toyota Way of continuous improvement and the egalitarian, decentralized HP Way.

Moreover, both firms face a growing rift between the two parts of their business. For HP, the needs of consumers focused on mobility, and businesses focused on data capture and recall, are diverging. For Toyota, rural drivers looking to improve on the performance of a pickup, and urban drivers looking for lower emissions, are at a proverbial crossroads. And if HP has spread its resources across both sides of a growing customer divide, so has Toyota.

With gas prices scraping $1.50 per gallon and all of its competitors building bigger and taller sport utility vehicles (SUVs), Toyota introduced the hybrid Prius to the US market in 2001. That looked smart when oil prices doubled from 2003 to 2006. Toyota had foreseen peak oil, the end of cheap fuel and a rush to environmentalism on US roads, crooned the critics and the case writers.

That ignores the fact, however, that Toyota added two mid-sized to large-sized lines of SUVs—the Highlander (fourth largest) and the Sequoia (second largest)—to the three it already had in the US market in 2001. And it did not stop there. It added the FJ Cruiser (sixth largest) in 2007 and the Venza (fifth largest) in 2009. There has been no effort to reform the American driver or even to predict the American driver's preferences about car sizes.

What has emerged, instead, is a distinctly barbell-shaped line-up of vehicles. With five lines of hybrid vehicles—including the Highlander Hybrid SUV—Toyota now has twice as many models of SUVs and hybrids as sedan cars in the US market. That line-up, moreover, reflects constant tinkering with configuration, size, price point and power source. Both the SUV lines and the hybrid lines are evolving as quickly as they can in their separate tracks. The appearance of a hybrid adapted for city driving—the Prius C—extended the trend.

In sum, both firms kept straddling divided businesses. And yet the record suggests Toyota went about it in a very different way. Nearly every year of the past decade brought major new hypothesis-driven experiments in hybrids, SUVs or both. In keeping with the continuous improvement on the factory floor for which it is famous, Toyota tends to explore the ends of the range of any line of vehicles before filling in the

DOI: 10.1002/sres.2219

middle—as if it wanted to understand the full range of engineering challenges each vehicle line will generate. At no point did Toyota management debate grandiose ideas about abandoning hybrids (e.g. when the Prius had a slow start) or SUVs (even when oil prices peaked before the financial crisis) without testing them. The firm probed the specifics of the two sectors with new offerings every year.

And Toyota is the one that seems to have come up with an answer. If there is an overarching bet that the company is playing out, it is that US driving habits will keep polarizing. The markets, at any rate, have substantially more confidence that Toyota has determined what will drive results in its sector than HP. Zacks Investment Research reported a consensus 2012 5-year earnings growth forecast for HP of 8% per year compared with 15% for the computer sector—or 7% below average. The forecast for Toyota was 23% per year compared with 19% for the automobile industry—or 4% above average.

US MANUFACTURING FIRMS STILL CHALLENGED TO IDENTIFY PREDICTIVE METRICS

US manufacturers have not overcome the challenge of identifying metrics that can predict or explain performance results despite two decades of intensive investment in data analysis. This section starts with evidence from the scarcity of sustained high performance over the past 40 years that US companies have not adapted well to changes in their business environment. Research on the proportion of US business stalls due to controllable factors shows that the root of this performance sustainability problem is a failure of causal analysis—confusion about the drivers of organizational performance. The disparity between pay and labour productivity over the past decade suggests this failure persists despite the prior decade's widespread investment in large-scale data systems for performance improvement.

In a study of over 6700 companies from 1974 to 1997, Robert Wiggins and Tim Ruefli found that very few companies sustain superior performance (2002, 2005). Defining superior performance relative to each company's sector competitors at levels ensuring statistical significance, they determined that just one out of 20 firms in their enormous sample sustained superior results for at least a decade. Just one out of 200 companies (32 in their sample) managed it for 20 years. And only 3M, American Home Products, and Eli Lilly kept it up for 50 years. Perhaps more important, the authors found that the average duration of periods of sustained superior performance declined towards the end of the sample period. This means that the amount of performance churning accelerated through the late 1990s.

Citing the result in his survey of evolutionary economics *The Origin of Wealth*, Eric Beinhocker takes it as evidence of what he, following evolutionary biologists, calls a Red Queen race in business competition. As the Red Queen warned Lewis Carroll's Alice, it takes all the running you can do to stay in place. Much as the species studied by those biologists, Beinhocker suggests, firms are 'locked in a never-ending co-evolutionary arms race with each other' (2006: 332). In fact, one can go further—as Beinhocker does later in his book—and argue that companies are not adapting to the dynamism that the complexity of their competitive interactions creates in their business environments.

In *Stall Points*, Matt Olson and Derek van Bever anatomize this adaptive failure. Although more than three quarters of large companies stall seriously and never recover, they report, the reasons for 87% of these stalls are controllable (2008: 35). Each of the four principal controllable reasons that they identify, furthermore, turns out to be a kind of failure in diagnosing the causes of change in organizational performance. Because these four reasons account for 55% of all stalls, the failure to understand what drives performance change appears to be a major reason firms do not sustain strong performance.

Updating a major strategic research project of the Corporate Executive Board inspired by an internal study and request from HP, Olson and van Bever looked at the more than 500 companies that have ranked at some point in the top half of the Fortune 100 over the 50 years to 2006. 76% suffered stalls defined as points after which

DOI: 10.1002/sres.2219

10-year growth averaged at least 4% per year less than average growth for the preceding 10 years. Thus defined, stalls are serious—nine-tenths of the companies that suffered them lost at least 50% of their market capitalization.

The authors find that premium position captivity is the largest reason for stalls. As does the innovator with a dilemma in Clay Christensen's book, these firms focus resources too long on a proven technology or innovation (Christensen, 1997). They correspond to organizations in Stafford Beer's Viable System Model (VSM) whose strategic planning process (system five) favours current operations (system three) over new ones (system four) (Beer, 1972). They tend to deprive internal technology experiments of resources until a competitor concentrating on an upstart technology develops it to the point of creating an insuperable competitive advantage. The failure here is one of causal analysis—misjudging the value of product or service attributes. For example, most laptop manufacturers failed to understand how increasing connectivity would affect the trade-off consumers make between computing power and mobility.

The steady stream of innovations that companies bring to the market might seem to belie this first reason for stalls. A long-term series of studies by Michael Hannon and John Freeman on the organizational ecology of markets suggests, however, that most firms bring new ideas to a market by entering it (1984a, 1984b). Once they have entered the market, they tend to be much less creative, often executing their original game plan past its period of prime relevance (Hannan and Freeman, 1977; Hannan and Carroll, 1992). Hannan and his collaborators give the example of the shift from transistors to very large-scale integration chips in the 1980s. Among the top 10 semiconductor firms of the transistor era, only Texas Instruments and Motorola adapted to chip technology and sustained their competitive edge. The leaders in the earlier period—Hughes, Philco and Transitron—did not adapt, whereas new entrants such as Hitachi, Intel and Philips revolutionized computer hardware. This pattern of innovative entry and lack of later adaptability suggests that while companies rarely test their causal assumptions, markets always do. The start-ups with well-adapted ideas survive until their performance drivers evolve, and we tend not to see the rest.

The second largest category of stalls is innovation management breakdown, including inconsistent funding, disorganized research and development (R&D), inattention to time and simplicity, and conflicts with core firm technologies. The common failure of causal analysis underlying stalls in this category is misjudging the value of past innovation. The third category, premature core abandonment, mirrors the first. Stalls here reflect confusion about the drivers of customer satisfaction and overestimate the potential to increase it through new features and services. Companies heading for this kind of stall look elsewhere and miss nearby opportunities. Volvo's 1970s expansion into oil exploration, food products, biotech and leisure products is an example.

The fourth category, talent bench shortfall, explicitly turns on failures of causal analysis regarding the contribution to success of specific skills, experience, kinds of talent and key individuals. Accounting for 9% of stalls, these stalls raise questions about firms' ability to determine the drivers of their employees' productivity. As do the other leading kinds of stalls, they reflect a failure to identify what drives success.

The long time-frame of *Stall Points* shows that the problem of a poor understanding of what drives success has been around for many years. But this is the problem that information systems for the large-scale analysis of performance data—variously called business analytics or just big data—specifically set out to address over the past decade. The *Stall Points* authors nevertheless see no drop in the threats to growth they have identified for the coming generation of corporate giants, which they dub the class of 2015. Indeed, they believe premium position captivity, and the failures of causal analysis underlying it, are growing.

There is already some evidence from the US labour market that they may be right. During the decade from 2001 through 2010, the corporate use of data analysis tools exploded. In general, expenditure grew disproportionately in the 1990s, whereas usage widened and deepened in

DOI: 10.1002/sres.2219

the 2000s. Most analyses show US business expenditure on IT rising faster than GDP every year of the earlier decade (e.g. Goldman Sachs, IT spending survey, 2 Nov 2009). The pace falls off in the later decade, when the Census Bureau's information and communication technology (ICT) survey finds US manufacturer spending on IT rising to $36bn a year (or 0.2% of GDP) before dropping 20% in 2009, as it usually does in recessions.

In the years from 2001 to 2010, however, businesses put large-scale data analysis to work more intensively in performance planning. This is when many terms for the strategic analysis of enterprise data became widespread. Business analytics, business intelligence and enterprise data visualization all became mainstream categories of IT tools during the decade.

More powerful computing tools no doubt contributed to this most recent flowering of business empiricism. The 1998 appearance of the first online analytic processing (OLAP) server, Microsoft Analysis Services, provides a useful marker. But the Sarbanes–Oxley audit reforms following the Enron scandal of 2001, and particularly their focus on the adequacy of controls for financial reporting processes, may have been equally important. Following Sarbanes–Oxley, businesses systematized their controls for businesses processes ranging far beyond accounting. And that provided more coherent structures for a wide range of business data that OLAP could chew over.

The decade is notable here for the dog that did not bark—wage growth failed to keep up with inflation. Between 2001 and 2010, unit labour costs in manufacturing grew 0.4% per year, whereas the consumer price index grew 2.4% per year in the US [Bureau of Labor Statistics (BLS) data]. It is true that hourly manufacturing compensation rose 3.7% per year over the decade. But output per hour grew 3.3% per year in manufacturing, or 2% per year more than the 1.3% per year real increase in hourly compensation. For all of the decade's new uses of data analysis, labour compensation growth continued to lag labour productivity growth.

Granted, other factors beyond labour affect labour productivity, not least technology capital investment. The data given earlier suggest we should attribute as much as 2% per year of the growth in labour productivity to these other factors. That same 2% per year impact of other factors on labour productivity growth would arguably also explain the 2% gap between inflation and growth in manufacturing unit labour costs. A starting point for much of the research over decades of labour productivity analysis, however, is the expectation that the cost of labour will reflect its marginal productivity in a competitive market. For that to be true between 2001 and 2010 in US manufacturing, the 1.3% per year rise in real hourly compensation would have to have reflected a similar rise in marginal output per hour—that is, the rise in output per hour keeping other factors constant. If so, the rise in *marginal* output per hour was 2% per year less than the 3.3% rise in *average* output per hour in US manufacturing over the decade. But why would technology affect *average* more than *marginal* labour productivity?

A number of studies of total factor productivity try to explain the part of the growth in labour productivity that hourly compensation does not capture (Jorgenson *et al.*, 2008). Perhaps the workforce captured 39% of the growth in US manufacturing labour productivity over the last decade, for example—taking the 1.3% real annual increase in labour compensation as a proportion of the 3.3% annual growth in hourly labour productivity—because the value of skills in using new technology was 39% of the total value of the technology.

The problem is that no allocation of the value of new technology to contributions from new skills and new investment explains the sustained difference between marginal and average productivity growth rates. If anything, one would expect new workers to capture more of their productivity growth as they internalized process innovations and technology advances. 2% per year is a lot, at any rate, to take out of the pockets of manufacturing workers for an extended period. Over a decade, it amounts to nearly 22% of foregone wage increases.

Critics of US labour arrangements such as Hedrick Smith simply conclude that the market is unfair. Even if US employers are exercising

DOI: 10.1002/sres.2219

unfair power in the labour market, however—by cooperating to ration hiring, for example, or through wage discrimination—it is not clear why average productivity would grow faster than marginal productivity. Moreover, it is unclear why employment of a price-repressed resource would not go up. An ongoing inability of manufacturing firms to evaluate employee contributions is a much stronger explanation because it severs the link between marginal productivity and wages. Indeed, it would be surprising if this failure were limited to firms that stall.

In sum, a failure of causal analysis—and in particular a failure to understand the drivers of performance—seems to lie at the root of performance sustainability problems that US companies have suffered for years. Despite widespread investment in IT systems to analyze performance in the 1990s and strides in applying those systems in the past decade, furthermore, the problems are not going away.

The rest of the paper looks closely at how increasing complexity in performance results intensifies organizations' need for hypothesis-driven learning. It suggests part of the reason US firms still have trouble identifying explanatory performance metrics is their overreliance on the data analysis systems they procured to generate them.

HIGH INFORMATION CONTENT AS WELL AS IMPACT NEEDED TO PREDICT COMPLEX RESULTS

The next two sections are the heart of the argument. This section shows why only predictive metrics with high information content as well as high impact can explain complex results. The next section argues data mining cannot substitute for hypothesis-driven learning in the search for predictive metrics with high information content, because it relies on low-information metrics. And it can link them to form higher information aggregate metrics only in theories that turn out to be improbably complex. Managers therefore cannot hope to understand complex results by substituting data mining for the ongoing experiments of hypothesis-driven learning.

The story so far is that US manufacturing executives are failing to explore the right kind of causal assumptions about performance results, and this manifests itself as a failure to find predictive performance metrics. Predictive metrics here are indicators whose values, when measured, determine or narrow the range of possible performance results of interest to a management team. If the result is a measure of profit for a company that sells software to improve factory operations, for example, a predictive metric might be the frequency of conversations with prospective customers about their operating problems.

Assumptions about cause and effect are really hypotheses. In the last example, for instance, the company's executives may believe that asking customers about unsolved problems will help make inroads in the process automation market and that customers will be candid about them. To test a causal assumption, naturally, it is necessary to measure outcomes for both the effect and the proposed cause. Fortunately, measures of the results that matter to most management teams are widely tracked—measures such as profit, return on assets, return on equity, return on invested capital and economic profit. So the challenge is finding indicators that evaluate the causes those teams think matter most.

This is why many so many consultants use metrics as a shorthand way of talking about assumptions regarding performance drivers. This paper will also use metrics to discuss the predictive power of performance drivers and the adequacy of the assumptions underlying them. But the reason is not just convenience—metrics have two characteristics that turn out to be the measurable counterparts of what make performance assumptions and hypotheses useful.

One of these characteristics is the impact of indicator outcomes on the results of interest to managers—the higher the impact, the greater the indicator's predictive power. In the case of our software company, for example, an indicator for the frequency of problem-solving conversations with customers would have high impact on profit if changes in profit always accompany changes in the indicator. Indicators that have a high impact on a result thus have a high correlation with it (for a more precise statement, please see Appendix).

DOI: 10.1002/sres.2219

High impact metrics, in turn, reflect assumptions about causes that bear strongly on results. If the outcomes of a customer-conversation metric have a high impact on profit, for example, the software company's management team has probably found a strong driver of profitability in the form of problem-seeking conversations with prospects.

Measures of the impact of a cause on a result are related to several well-known statistics. The regression coefficients of independent variables in statistical analysis, for example, directly capture the sensitivity of changes in a result to changes in other variables. The correlation coefficient between two metrics, furthermore, compares the impact of each on the other (without presuming knowledge of which is the cause and which is the effect) with the variability of the two variables. Nonparametric impact measures (independent of any particular model) that capture the indeterminacy of a driver or causal variable given information about the result it is thought to affect are increasingly widespread (see Appendix for more detail).

The impact of metrics reflecting performance drivers lends itself to a useful ranking technique borrowed from risk management. The technique can help managers focus on what potentially matters most to a result, and it makes the concept of impact more concrete. It involves two steps. For any identified driver of a result—whether it is an operating variable or a risk factor—first project expected and worst-case outcomes. Then, project the impact on the result of the scenario producing the worst-case outcome of the driver. Let us call this the worst-case impact (WCI) of the driver. For example, if the expected and worst-case defect rates for a factory are 1% and 4%, and the worst-case scenario reduces the factory's profit from $100 000 to $90 000, then the WCI of the default rate is $10 000.

Worst-case impacts are useful because they allow the direct comparison of different drivers' potential effects on a result. By the same token, they provide an impact ranking of all of the identified drivers of a result and their associated metrics. It is possible to simplify most scorecards by estimating WCIs for each driver and keeping just the ones with the highest WCI. (The worst-case scenarios for each identified driver must be equally remote.)

The WCIs are closely related to value-at-risk calculations for risk factors. Similar to value-at-risk, WCIs combine the concepts of volatility and exposure—the WCI of a factor or driver rises with both its volatility and the sensitivity of results to its changes. So, for example, cooking times have a big impact on enchiladas—not because they burn easily but because the times vary so much from one recipe to another. The kind of oil matters for the opposite reason—not because vegetable and olive oil are wildly different but because enchiladas are exquisitely sensitive to the oil (vegetable, for once, being better).

Moreover, WCIs provide a natural framework for incorporating an analysis of risk factors into performance assessment because there is no reason to treat performance levers under management's control any differently from risk factors outside of its control. The importance of a performance driver depends on its WCI—not whether it is controllable.

The impact on results of metrics reflecting the operating variables, risk factors and other parameters assumed to affect an organization is fairly easy to understand —even if managers do not always use it to simplify their performance scorecards and measurement systems. The big question is what else metrics need to have predictive power.

The answer is that understanding or predicting a complex pattern of results requires metrics with high information content as well as high impact. It has emerged repeatedly over the past century in statistics, Bayesian probability, genetics, logic, model theory, cybernetics, statistical mechanics, thermodynamics and decision theory. And yet it still surprises.

The English statistician Ronald Fisher made something of a crusade out of the related argument that correlation does not imply causation. He even claimed to prove that apple imports raise divorce rates in England (because they were correlated) to rebut a paper by British epidemiologists Doll and Hill on the correlation between smoking and lung cancer (Doll and Hill, 1950). Although he was wrong about cancer, there can be little question about his larger point, because even when there is a causal relation between

DOI: 10.1002/sres.2219

two variables, correlation does not show which way it points. Moreover, there are plenty of examples of spurious regularities.

The correlation between a drop in a column of mercury and storms is an example of a spurious regularity. Here a storm's cause may be a drop in barometric pressure, which also causes the change in barometer readings. It would be wrong, of course, to infer from the correlation that barometer readings cause storms. Spurious regularities are so problematic for any attempt to understand causality in terms of things that can be observed that they provoked Hans Reichenbach to define causes as regularities between two variables persisting even when you account for every other variable that might be a common cause of them. He also stipulated the Common Cause Principle that some common cause explains any correlation between two well-defined variables neither of which causes the other (Reichenbach, 1956).

Reichenbach's definition has been of great value to the understanding of causal hypotheses even though it does not tell you how to find common causes—also called confounding factors because they confound conclusions about causation among other variables if they are not taken into account. Indeed, both Fischer's point about correlation and causation and Reichenbach's concern about hidden common causes raise the larger question here: what does a causal variable have to have beyond an impact on or correlation with the result it affects? If not correlation then what does imply causation?

The English cybernetics pioneer Ross Ashby provided an answer in his 1956 Law of Requisite Variety (Ashby, 1956). According to Ashby's principle, any control or regulatory system such as a thermostat must have as many alternative states or settings as the external disturbances (such as extreme weather), it is designed to offset. If a bank branch experiences three widely different numbers of customers waiting in line each week, for example, then it should probably have three different staffing plans for the appropriate hours to keep waiting time constant. The same goes for measuring tools: if a thermometer designed to work up to 100°F bursts on a hot afternoon in Austin, its failure is one of requisite variety—the thermometer could not reflect as many possible states as the climate it measured.

Ashby's law implies that control systems need to be able to match the complexity of their environment. By extension, a predictive system needs to match the complexity of what it tries to predict. So when it tries to predict results that are complex in the sense of having high information content, a predictive system also needs to have high information content.

Before going further, it may help to contrast indicators with high and low information content in intuitive terms. Indicators with high information content compared with a result that one wants to understand or predict are similar to recipes, whereas those with low information content are similar to lists of ingredients. If your palate is good, you can probably guess the ingredients in a dish once you have tasted it. But knowing the list of ingredients is not enough to know how to prepare the dish. If you know a recipe, however, you can probably prepare the dish. No matter how good your palate, just tasting the dish will probably not help you guess the recipe.

Implementation metrics for highly specific strategies are good examples of indicators with high information content. As in the case of recipes, knowing you implemented a proven strategy will let you forecast a successful result with confidence. But knowing that someone hit a target is not enough to tell you what specific strategy she followed.

Requirements are a good example of indicators with low information content relative to the result. As in the case of ingredients, just knowing you hit a target may let you ascertain the value of an indicator for one of its requirements. Knowing that you have met a requirement for hitting a target, however, is not enough to reassure you that you hit it. A requirement for our software company to hit its profit target, for example, is completion of the latest version of its software—but completing that version is no guarantee of any particular level of profits.

Recipes have much higher information content than lists of ingredients just as good strategies have much higher information content than requirements for hitting a goal. And root causes, as we shall see, have more information content

DOI: 10.1002/sres.2219

than later causes. That certainly makes recipes—and well-specified strategies—more useful than lists of ingredients and requirements for knowing what to do. But recipes and well-specified strategies offer an additional advantage in understanding or predicting an uncertain result.

The advantage is that you always know when to modify a specific strategy or recipe. The failure to hit a goal after properly executing it shows the recipe or strategy was wrong. If you find a highly specific strategy or recipe that starts to work, in contrast, it will probably help you predict or control results until environmental changes undermine it.

Linking Ashby's law of requisite variety to Bayesian probability, the American physicist E. T. Jaynes proposed the related principle of maximum entropy (Jaynes, 1957). Suppose a set of data reflects an unknown underlying probability distribution of possible outcomes. Of all the possible probability distributions that could explain the data, the best one to infer until more data are available—the best prior—is the one whose outcomes have the highest information content, or are most surprising or seemingly random on average.

Say, for instance, you have collected a data set large enough to give you confidence in the mean value of the underlying probability distribution it reflects, but not much else. This might be the case if you have repeatedly measured the time it takes to complete a process in a factory. And suppose you want to infer an underlying probability distribution to make some predictions. The distribution you infer should be the one whose outcomes have the highest information content—are most surprising on average—and the mean you have measured.

One can speak about the information content of a distribution just as well as its outcomes. So, for example, a distribution that is perfectly flat across all possible outcomes has higher information content than others with the same number of possible outcomes. This is because the average surprise or improbability of its outcomes turns out to be higher than the average surprise or improbability of the outcomes of less even distributions.

The rationale behind Jaynes' principle is modesty about what we know. The probability distributions thought to underlie a data set that have the highest information content—the flattest ones—are also the ones that presume the least knowledge about what gave rise to them. They are the closest to being random. And they are also the ones that reflect the fewest prior causes or forces that might have shaped them. To infer that a data set (or knowledge about the data) arises from a probability distribution with low information content—that is, from an uneven one—is to infer the effect of prior causes for which the data set provides no evidence by itself.

The principle of maximum entropy helps explain why predictive metrics should have high information content. By searching for explanatory probability distributions with high information content, once again, Jaynes made sure he did not presume the effect of factors for which the results he wanted to explain provided no evidence. But that is similar to asking which chapter of a novel you would consider most likely to be true if you knew some events in the middle of the book really happened to someone. If you had to choose one chapter, it would make sense to select the one that happened first because the other chapters all limit the possibilities still open in the earliest one.

The flip side of the chapter in a novel with the most open future possibilities is the great amount you learn by filling in the chapter's future. You learn about the possibilities open at the end of each chapter of a novel by reading the rest of the novel. What you learn by reading from the end of any chapter to the end of the novel is similar to the outcome of a metaphorical indicator that picks out an actual future from the possibilities still left open by that chapter.

Now ask which of those indicators is most useful in understanding how a novel turns out. You would choose the indicator reflecting what one learns reading from the end of the earliest chapter to the end of the book—this has the highest information content of the indicators reflecting what you learn about the possibilities left open after each chapter.

It is in just this way that indicators with high information content resemble indicators for root causes. Root causes—such as the earliest chapter of a novel—precede other causes that determine

DOI: 10.1002/sres.2219

a set of results. But that means they generally have higher information content than indicators for successive causes. This is actually an implication of Ashby's law of requisite variety. To see why, suppose you measure two indicators A and B and you know that the outcome of A completely determines the outcome of B. A might be income, for example, and B might be income tax. There is no way A could have less information content than B and still determine B. Income might have less information content than tax if there were two levels of tax payment corresponding to a single level of income—but then it would not make sense to say that income determined income tax.

So Jaynes' search for probability distributions with the least baggage in terms of hidden assumptions is similar to the search for root causes in one tantalizing respect. Both push to the beginning of causal chains. Without aiming directly at it, Jaynes provided an answer to our question about what an indicator needs to have, beyond a high impact on results, to help managers understand, predict or control those results. It needs to have high information content, as does the outcome of a root cause indicator or variable.

Management theorists have occasionally emphasized the dependence of predictive power on information content. A good example is the section on Potential Problem Analysis in Charles Kepner's and Ben Tregoe's classic *The New Rational Manager* (Kepner and Tregoe, 1997). Using weather as an example, the authors focus on the problem-solving power of specificity in identifying risk factors, writing: 'In Potential Problem Analysis, the purposeful identification of specific potential problems leads to specific actions.' The general risk of bad weather, for example, involves two possible states of affairs—good and bad weather. The more specific risk of a windstorm with heavy rain involves at least four—weather with and without strong wind and with and without heavy rain. Anyone suspecting a windstorm would prepare for a windstorm rather than just bad weather.

More recently, Ricardo Hausmann, Dani Rodrik and Andres Velasco set up their growth diagnostic approach to economic development with a search for root causes of growth constraints working backwards through a causal tree (Hausmann *et al*., 2005). The present analysis suggests these root causes have greater potential to explain disappointing growth results, because they have higher information content than proximate causes.

The dependence of predictive power on information content also seems to motivate Russell Ackoff's concept of idealized design. In *Creating the Corporate Future*, Ackoff recommends designing the organization to deal with the future rather than just forecasting the future to know what will happen to the organization (Ackoff, 1981). To accomplish this, Ackoff defines idealized design as the structure with which the designers

> …would replace the existing organization right now, if they were free to replace it with any organization they wanted subject to only two constraints (technological feasibility and operational viability) and one requirement (an ability to learn and adapt rapidly and effectively). (Ackoff, 2001: 7)

The concept underlies the widespread use of mission statements in non-profit and for-profit US organizations today which are supposed to launch the process of self-renewal in idealized design. But the anodyne and indistinct mission statements organizations tend to mail in bear little resemblance to the purposive statements Ackoff recommends. "An organization's mission statement should be unique, not suitable for any other organization" (Ackoff, 1981). In other words, to renew the organization, it is necessary to state a mission as a means to the organization's goals so specific that it is unlike any other. Mission statements are effective, Ackoff is telling us, when they have high information content.

Generalizing Bayes' theorem, the chain rule in modern information theory not only proves the importance of high information content to predictive metrics but provides a way to measure it. Here is a conceptual account. Suppose F stands for the outcome of some predictive factor you are measuring such as the frequency of in-depth customer conversations. And suppose E stands for a business result such as earnings that you want to understand, predict or control. Then, the *mutual information* of F and E measures the amount of information that F contains about E

DOI: 10.1002/sres.2219

and that E contains about F. The *entropy* of F or E measures the information content of the indicator (which, again, reflects the average improbability or surprise of an outcome). And the *conditional entropy* of F given E is the amount of information remaining in F, the factor, given an outcome for E, the result. (You might say it measures the information in F that is not in E.)

The amount of information that F contains about E is a good measure of the degree to which outcomes for the predictive factor F determine earnings or outcomes for E. This, once again, is the mutual information of F and E. A foundational theorem in information theory called the chain rule shows it must equal the information content of F less the amount of information in F that is not in E (Cover and Thomas, 1992: 19). Using the definitions earlier, this is the entropy of F less the conditional entropy of F given E. In other words, these two quantities govern predictive power. (See Appendix for an outline of the proof using Bayes' theorem.)

To summarize the relation between predictive power and information content:

> *Chain rule*: (Mutual info of F and E) = (Entropy of F) less (conditional entropy of F given E)
>
> *Predictive power*: (What F says about E) = (Info content of F) less (info content of F not in E)

What about impact? The subtracted component—the amount of information in F that is not in E (or the conditional entropy of F given E) turns out to be a precise negative measure of the impact of F on E. Think of it as a measure of F's lack of impact on E or E's insensitivity to F. Because the first component was just the information content of F, you can say the power of a factor F to help us understand, predict or control a result E is the information content of F (its entropy) less the lack of impact of F on E (the conditional entropy F given E). Both of these are measurable, because they are functions of the average probability of the outcomes of F and E (more in Appendix). Here is how predictive power, information content and impact relate:

> *Predictive power*: (What F says about E) = (Info content of F) less (lack of impact of F on E)

Explicit consideration of the information content of predictive metrics may help solve some of the classic challenges in inferring causality based on observed probabilities. Judea Pearl's causal calculus goes farthest towards inferring the effects of interventions such as a change in marketing or procurement from observed probabilities and correlations (Pearl, 2000). His calculus has been put to work in Bayesian networks—causal models of business, economic and environmental systems—that help characterize the confounding factors Reichenbach sought to identify. As the results in these systems become more complex, however, it becomes more difficult to find a sufficient set of common causes that includes all the factors that could confound conclusions about true causes. This section shows why, as complex results require a set of explanatory factors that are at least similarly complex.

Here, as a practical aside, is a simple way to have a rough measure of the information content of an indicator. It may be an indicator that we already know to have a high impact on a result we want to understand or control. What we want to know is whether it also has information content at least as high as that of the result (Table 1). This tool, by the way, is drawn from a chapter investigating the specificity of indicators, illustrating the tight relationship between specificity and information content (Apgar, 2008).

Table 1 Information content of a predictive indicator versus the results it helps predict

	Less even distribution of possible indicator outcomes than of possible results	More even distribution of possible indicator outcomes than of possible results
More different possible indicator outcomes than possible results	Comparable	Higher
Fewer different possible indicator outcomes than possible results	Lower	Comparable

DOI: 10.1002/sres.2219

The tool suggests a way to determine which indicators belong on a performance scorecard that is short but predictive. The first step is to look for indicators with high information content relative to the result you want to understand or control. You can do so by assigning the candidate indicators to the appropriate cells in Table 1. The second step would be to rank the ones in the top right cell by WCI. It is overwhelmingly likely that only three to five will have relatively high WCIs. Throw out the rest. And then add the indicator for the result you are trying to predict or control to update your scorecard.

The basic information relationship between a predictive metric and the results it helps predict thus suggests a procedure for prioritizing predictive indicators and even for designing performance scorecards. Those relationships may be even more important for what they say about identifying predictive metrics when the results they are supposed to predict are complex.

The performance management challenge before managers who want to understand and control complex results is to find predictive metrics with high information content as well as a high impact on those results. And plenty of large-scale IT vendors and consultants have a ready answer in generating more, bigger and faster data. The question is whether the answer works.

NO WAY TO IDENTIFY HIGH-INFORMATION METRICS WITHOUT HYPOTHESIS-DRIVEN LEARNING

The last section showed any explanation of complex results requires predictive metrics with high information content as well as a high impact on those results. This section argues data mining cannot substitute for the successive experiments of hypothesis-driven learning in the search for predictive metrics with high information content because data mining relies on widespread, low-information metrics. And it can link low-information metrics to form higher information aggregates only in theories that turn out to be improbably complex. Managers therefore cannot hope to understand complex results by substituting data mining for hypothesis-driven learning.

At first sight, there could be no more natural answer to the manager's dilemma about finding predictive metrics to understand and control complex results than taking advantage of the huge increases in data capture and analysis now at that manager's disposal. Generating more, bigger and faster data is a natural answer for three reasons.

First, data management is cheaper than ever before. Second, the use of statistical analysis in the social sciences has been so fruitful that the use of related methods in management seems promising. And third, one might hope to address any deficiencies in the predictive power of operating and risk indicators currently used to understand and control results by looking at more of them.

This section shows that, regarding the first of these reasons, a misguided use of business analytics to find predictive metrics will make it very expensive. As to the second, management challenges no longer resemble the kinds of social science problems that statistical methods have illuminated. And as to the third, there are strong reasons to doubt quantity can make up for quality when it comes to indicators with the power to predict complex results.

Ronald Fisher, the English statistician discussed earlier who helped build the foundation of modern data analysis, was acutely aware of the limits of statistical inference. He knew, for instance, that explanatory variables need to have more than a high impact on what they are supposed to explain (Fisher, 1924). The problem for Fisher was that while high-impact variables can explain some of the variability or scatter of a result, they may not be able to explain most of it. (In the terms of the last section, we would say they do not have high enough information content.) Indeed, the modern statistician's *F*-test—named for him—compares the variability in a result that a set of predictive indicators explains with what they leave unexplained.

The temptation for the social scientist is to add more explanatory variables or indicators in the hope of minimizing the unexplained variability in the result that is to be explained. Fisher had doubts about that, too. The challenge is that when the number of explanatory indicators approaches the number of observations of the result

DOI: 10.1002/sres.2219

to be explained, there is no longer much reason to expect those indicators to be able to predict future results that have not yet been observed. This is the so-called degrees of freedom problem faced in a simplified form if one tries to estimate a trend from just one observation—any trend fits the data. Thus, the *F*-test asks each explanatory indicator in a regression model to do its fair share of explaining—for fear that even unrelated variables can appear to explain a little of the variability of results at the margin.

One solution available to the social scientist with an embarrassment of potential explanatory indicators is to try to use them to explain past results going back further in time. If some combination of explanatory indicators can account for the variability in a much larger set of past and present results, it is likely they can do so only by virtue of a deep underlying relation between the indicators and what they are supposed to explain. The strategy does not work, however, if the underlying relationship is changing.

This limitation of statistical inference inspired much work in the 1980s on autoregressive methods. The idea was to determine what explanatory variables matter in the future based on an analysis of historical results. By assuming a regular, predictable change in the relationship between explanatory indicators and results, it is possible to solve this problem. But it is unsafe to assume this relationship changes smoothly in most business environments.

By 1974, moreover, methodologists were formulating broader limits to the strategy of increasing the predictive power of a set of indicators by adding more new indicators to the set. There is a problem with adding explanatory indicators to a model even when ample results and degrees of freedom are available. Hirotsugu Akaike discovered a strict trade-off between the complexity and accuracy of any predictive model (Akaike, 1974).

Akaike's information criterion rewards any model for the accuracy with which it explains a set of complex results but penalizes it for each explanatory indicator that it uses to do so. The reason, broadly, is that as one uses more explanatory indicators to account for a set of results, the quality of the explanation is more likely to arise by chance and less likely to reflect a persistent underlying relationship. Akaike's information criterion has attracted much attention since 2002, especially in computational genetics (Burnham and Anderson, 2002).

Given all these limits on statistical inference, then, where have statistical methods successfully used many explanatory variables to shed light on complex results in the social sciences? Two areas, broadly defined, stand out—both differing sharply from the circumstances where most managers need to provide causal explanations.

One involves the comparison of results across similar situations at a point in time. Evaluations of government programs and development projects in health care and education, for example, can often determine the impact of an intervention such as the introduction of new medical services or new curricula. The usual procedure is to compare results in randomly selected locations subject to the intervention with those not subject to the intervention—and to attribute the differences to the intervention.

Even this kind of impact evaluation has its limitations, however. Although it is relatively easy to implement randomized trials for the evaluation of social development programs in education and health care, it is hard or impossible to do so for infrastructure projects, general government policy support, or large-scale agricultural projects.

Similarly, two business sectors are particularly amenable to quantitative analysis with many explanatory metrics: retailing and finance both generate results that can be compared across many similar stores or similar accounts. And both have a range of interventions that are easy to tailor across stores or types of accounts such as product positioning and targeted financial services. Most business sectors, however, generate a series of one-off results rather than large sets of simultaneous comparable results. Or they face decisions involving commitments such as investment in a new factory that simply cannot be meaningfully tailored from one customer to another. It is hard to imagine where managers could generate large enough sets of comparable results to test long lists of potential predictive indicators in technology,

DOI: 10.1002/sres.2219

telecoms, drugs, power, manufacturing, energy and transportation. Their best hope, in fact, is to analyze historical data.

And that brings up the second area where statistical methods using many variables have explained complex results in social sciences—the comparison of results across time, or time series analysis. As mentioned earlier, arbitrarily long time series provide enough data to test large numbers of explanatory variables. This has been useful in economic analysis, but it has worked for economists only where underlying causal relations are fixed or change smoothly. Economic markets in equilibrium fit these conditions, and empirical economics provides good explanations of their behavior. Markets in disequilibrium do not, and the results show it. Robert Lucas applied this critique to macroeconomic forecasting with policy variables that change the relationship of other variables in the model (Lucas, 1976).

The problem for managers is that the drivers of complex results such as profitability are subject to constant change because their firms are locked in Beinhocker's never-ending coevolutionary arms race with one another. Historical analysis of results can show which explanatory indicators have mattered at some point over long periods of time but not what matters at the moment. The outcomes of stable processes such as steps in a manufacturing cycle that can put historical data to good use are mostly relevant to tactical, not strategic, decisions. Areas such as quality control have indeed become highly analytical because the patterns of outcomes tend to be stable. Managers facing more complex results cannot take that kind of stability for granted, however, and therefore, cannot make the same use of historical data.

The success of statistical methods deploying large numbers of explanatory variables in the social sciences, therefore, provides little reason for optimism about the use of data to find the drivers of complex business results. The idea that more data can help managers find new drivers of complex results, nevertheless, has persistent appeal. That appeal needs a closer look.

The challenge for the manager who needs to understand or control results, once again, is to find indicators reflecting operating levers and risk factors that have high information content as well as a high impact on those results. Two indicators will be better than one if together they have higher impact, higher information content, or both (and are no more costly to track). The logic of big data is straightforward: no single indicator needs to have sufficient impact and sufficient information content to explain complex results. Just keep adding indicators that raise the information content or the impact on results of all of the predictive indicators as a group. You can keep going until you approach the number of independent observations of results that your group of indicators needs to explain.

Although the logic is straightforward, it carries a hidden assumption. The assumption is that as you try to explain results better by using more indicators from a data warehouse—be they measures of process quality, customer satisfaction or something else—it is likely that at least some will have a little of the two qualities needed to raise your predictive capacity. That is, at least some will have an impact on results after accounting for the impact of the other indicators on your list. And, at the same time, they will add at least a little information content to those other indicators. The assumption turns out to be a highly unlikely stretch.

To see why the assumption is so fragile, consider what it means for a new indicator to add to the impact on results of other explanatory indicators. It means changes in results must accompany changes in the new indicator if the outcomes of the other indicators hold steady. So there must be some correlation between the new indicator and those results.

Now consider what it means for a new indicator to add to the information content of a set of other explanatory indicators. It means you should not be able to predict outcomes of the new indicator from outcomes of the other indicators. So there must be very low correlation between the new indicator and the other indicators.

These two conditions for new indicators to be useful in understanding or controlling results become harder and harder to meet as you add metrics to the predictive pile. Each additional indicator must correlate decently with results

DOI: 10.1002/sres.2219

and very little with all of the other explanatory indicators. It is similar to seeking locations for new community clinics connected to a local hospital. You want each new clinic to be close to the hospital. (This is similar to correlating with results.) But you do not want to build a new clinic close to any other clinic. (This is similar to avoiding correlation with other explanatory indicators.) Fairly soon, you run out of places to build clinics—and you run out of logical space in which to find new indicators that add to information content and impact at the same time (see Appendix for related approaches).

An equivalent way of putting this is that the addition of low-information factors to any theory meant to explain complex results makes the theory complex. And complex theories are improbable in any sense of the word. The two conditions are related—complex theories are unlikely for the same reason that a computer printing random figures is likelier to write a simple program by chance than a complex one (Kolmogorov, 1963). The strategy of big data amounts to taking complexity out of the predictive factors of a theory and putting it into the theory, itself. But it is an unpromising strategy because complex theories are improbable.

Pearl's work on the causal Markov condition may in time generate a proof of this. He has shown that a model will include enough indicators for a causal analysis of a result if their unexplained variances (i.e. the variances that the other indicators cannot explain) are independent (Pearl, 2000). This condition may be harder to meet as factors multiply.

These considerations help explain the robustness of the assertion Vilfredo Pareto made in his classic work on probability distributions that only a *vital few* factors drive most results (Pareto, 1896). All versions of the 80–20 rule—that 20% of factors explain 80% of the variability in a result—derive from Pareto's vital few. And quality control methods such as Six Sigma explicitly recognize it in searching for only the vital few root causes of a quality problem.

Business analytics and other kinds of data analysis systems for improving performance make plenty of sense as a way to test the effectiveness of indicators in understanding, predicting or controlling results. But the vast majority of users are hoping for something more. They want to find new indicators in the haystack of newly searchable stores of operating and market data that data analysis applications generate. They are looking in the wrong place.

There is one way to avoid the misuse of data analysis described earlier and its reliance on the fragile assumption that it is possible to build up the joint information content and impact on results of a list of predictive indicators even if the individual indicators have low individual information content. And that is to find a few indicators, each with high information content, that all have a large impact on results. The trouble is that high information-content indicators are not likely to be sitting in a data warehouse, no matter how expensive. They are too specific.

Recall that factors and assumptions reflected by indicators with high information content are similar to recipes rather than lists of ingredients. Their specificity makes them hard to find in any repository of widely shared indicators. To take a new example, consider the original microfinance proposal to ensure credit quality by relying on peer pressure from solidarity groups of women providing loan cross-guarantees to one another. The proposal embodies specific assumptions about the effect of solidarity group formation, cross-guarantees and customer knowledge on lending outcomes. It has far more information content than the typical bank strategy of acquiring adjacent banking companies and closing redundant branches.

That information content gives the microfinance formula the potential to account for complex lending outcomes and—if it turns out to be right and is well executed—achieve success. A broader strategy, in contrast, might be right as far as it goes but could never by itself ensure results because one could never be clear what counted as correct execution. The very information content that gives the more specific formula the potential to determine results, however, makes it unlikely to be found on a virtual shelf of widely measured indicators.

There is a reliable way to find factors, drivers, or assumptions with enough information content to predict or determine complex results and that

DOI: 10.1002/sres.2219

is to stipulate specific hypotheses—big, uncertain bets about how things really work. They are similar to searchlights into the space of root causes—lighting up specific ones that we can readily test. They may be wrong, but they generate prescriptions specific enough to lay down markers for rapid learning.

The learning takes place in the course of deliberating on the outcomes of successive experiments to update those hypotheses. As Richard Jeffrey put it in describing how subjective probabilities shed light on causality: 'In decision making it is deliberation, not observation, that changes your probabilities' (Jeffrey, 2004: 103). Hypothesis-driven learning is not a bad term for this, because it recognizes that deliberating on successive experiments will lead you to metrics with increasing predictive power so long as you search through factors with information content on the order of the results you are trying to predict.

There are more specific kinds of learning that can find predictive metrics where data mining cannot. Ross Ashby defined nested feedback loops that allow both learning within a process and learning to improve the process as double-loop learning, for example, and argued it is the characteristic capability of adaptive systems (1956). But more specific definitions would weaken the conclusion that, in the search for complex predictive factors, data mining cannot replace any kind of learning that revises hypotheses over successive experiments. The point is that any process of stipulating and testing specific hypotheses has a potential that data mining lacks to explain complex results through simple theories employing high-information factors.

Such learning may sound similar to a time-consuming ordeal until one considers how broad a swath of experience will be relevant to predictive factors with high information content. For such factors, the tests come fast and furious. Ian Hacking warned against underestimating the power of actual experience to test specific beliefs 26 years before writing his *Introduction to probability and inductive logic* (2001). In *Why Does Language Matter to Philosophy?*, he defended an account of meaning by Donald Davidson that takes definitions to be hypotheses tested by their experiential use in sentences with observable truth conditions. Critics doubted experience could constraint speakers' definitions enough to ensure mutual comprehension. In an echo of Alfred Tarski's concept of logical consequence (1936), Hacking disagreed. Specific theories offer more predictions for experience to test, and Davidson's dictionary full of hypotheses is nothing if not explicit—indeed, it '…is an elaborate structure with a large number of, in themselves, trifling consequences that can be matched against our experience' (Hacking, 1975: 143). Specific strategies similarly make more of our business experience available to test and refine our plans.

Managers who want to understand, predict and control complex results therefore face a stark choice. They can sort through large volumes of indicators with low information content and mediocre predictive power from enterprise planning systems and other big data applications. Or can they can advance hypotheses about what really drives their results, test them as quickly as possible, and repeat the process.

The first option can generate requirements but not recipes for success—avoiding the embarrassment of wrong starts but never reaching clarity about what drives results. The second option will occasionally lead to a mistaken initiative but will evolve a clear picture of the (possibly changing) root causes of results and thus strengthen firm resilience over time.

Indeed, how organizations evolve to negotiate that stark choice may be the most valuable legacy of the VSM of Stafford Beer (Beer, 1972). Beer contrasted the routine arbitration of the competing demands of production and back-office functions with a more evolved function arbitrating the more complex competing claims of routine management and a function responsible for radical innovation and renewal. Beer's model thus aligns management functions with the varying information content or complexity of the challenges firms face. Indeed, the top layer of management is positioned to manage the most complex problems precisely, because it is in the best position to advance hypotheses and experiment. Modern organizational structure, in other words, has evolved to handle complexity. It is modern management—perhaps persuaded by a

DOI: 10.1002/sres.2219

misguided empiricism or false promise of big data—that is not adapting to the rising complexity of business outcomes.

In sum, managers cannot replace hypothesis-driven learning with data mining to identify predictive metrics that can explain complex results, because large-scale data analysis applications are limited to linking simple factors in theories that are improbably complex. Managers confronting complex results, as a consequence, should test specific hypotheses and adapt them until they work. In fact, the misuse of data analysis may help us understand the ongoing failure of many firms to deal with complexity.

DIVERGENT APPROACHES TO SIMILAR CHALLENGES AT WELLPOINT AND NESTLÉ

This section contrasts the widespread use of business analytics at WellPoint with the process of negotiated experimentation at Nestlé. Nestlé's iterative negotiation of highly specific strategy experiments handles complexity better than WellPoint's requirements-based planning.

Of course, companies do not experiment with strategy unless they must. At Nestlé, executives must tinker, because the strategies they are investigating are so tailored and specific that no one could be sure in advance they will work. That kind of specificity is the hallmark of assumptions about operating levers and risk factors with enough information content to have predictive power. The experiments help Nestlé teams find strategies, as we shall see, that also have a high impact on the complex results they need to understand and control.

Since WellPoint Health Networks and Anthem merged in 2004, WellPoint has become one of the top two health insurers in the USA and serves 34 million plan members or policy holders. As an independent licensee of Blue Cross and Blue Shield, it is active in 14 states, including California and New York. It is a notable pioneer in the use of business analytics for its human resources function and for patient care. An early success in human resources was to analyze the reasons why first-year turnover in services jobs varied between 15% and 65% in order to suggest improvements in talent management (Ibarra, 2008).

Among WellPoint's business intelligence solutions for patient care are SAS business analytics algorithms that classify patients by severity of medical risks. These algorithms search diagnostic lists, run rules-based tests and mine data for patterns. Uses include identifying patients who should have home monitoring devices after being discharged from a hospital.

The provider also rates 1500 hospitals under 51 metrics to measure quality of care and predict hospital readmission rates that boost health care costs. Hospitals falling below a certain threshold for quality of care are not eligible for rate increases. The quality of care formula is based on health outcomes, patient safety precautions and patient satisfaction.

On 13 September 2011, *Analytics* magazine reported the first commercial application of IBM's Watson technology in a WellPoint effort to improve the delivery of up-to-date, evidence-based health care. Watson's most notable achievement prior to this was beating two of the top Jeopardy champions. Watson's ability to interpret natural language and process large amounts of data would go to work proposing diagnoses and treatment options to doctors and nurses.

Back in 2010, however, a WellPoint use of business analytics came to light that was far less flattering. Reuters reporter Murray Waas wrote in April of that year that one of WellPoint's algorithms identified policyholders recently diagnosed with breast cancer and triggered fraud investigations of them. Fraud investigations are apparently not limited to fraud and can turn up any of a wide variety of reasons or pretexts for cancelling a patient's coverage. Health and Human Services Secretary Kathleen Sebelius criticized the practice as deplorable. WellPoint CEO Angela Braly responded that the computer algorithm in question did not single out breast cancer, which led critics to wonder whether the insurer was using business intelligence to deny coverage for all pre-existing conditions retroactively (Noah, 2010). WellPoint agreed to stop nullifying policies in force for reasons other than fraud or material misrepresentation 5 months ahead of the deadline set in the Affordable Care Act (ACA).

DOI: 10.1002/sres.2219

Some may dismiss this as evidence of weak US business ethics in general but the real problem appears to be WellPoint's pursuit of too many conflicting requirements subject to measurement. Annual letters to shareholders are rarely the sources of great business insight any more, but WellPoint's 2010 letter is remarkable for its avoidance of hard choices in the strategy it lays out. It calls for increasing share in high-growth domestic businesses such as Medicare and Medicaid, for example, without explicitly cutting back its core national health-benefits accounts business, which will soon be subject to the ACA.

For its national accounts business, moreover, WellPoint identifies four drivers: the size of its medical network, low cost structure, medical management innovation and customer service. This is a classic case of requirements-based planning, in which the items form a list of necessary ingredients of success but provide nothing such as a recipe for it. It is a list of low-information factors with low specificity and low predictive power.

By leaving strategy undefined, moreover, they make it more likely that the firm will drift. No doubt one headquarters unit works to grow the medical network, another to lower its cost structure, a third to innovate in medical management, and a fourth to strengthen customer service. But this leaves open critical questions such as whether the network will grow regardless of cost, whether cost containment is more important than customer service, or whether innovative medical management programs can cut back the network, raise indirect costs, or encourage customer self-help. Ideally, strong execution in each category of requirements for success will ensure success. In the real world, however, requirements-based planning systems encourage different teams to pursue goals that turn out to conflict.

That may be precisely what happened in the case of the WellPoint business analytics algorithm that featured in the political debate leading up to the ACA. It is hard to imagine someone in a boardroom say: 'Let's figure out a way to nullify the policies of women diagnosed with breast cancer.' But it is easy to imagine the team looking for continuous improvement in the application of business intelligence software to broaden its mandate without checking in with the team trying to improve critical-care services for patients with serious problems.

In other words, WellPoint's misstep looks like a symptom of management by checklist where the checklist includes every major widely used variable subject to measurement. The problem is that such checklists can never replace experimental learning about specific and coherent strategies that may or may not work right away—but that will provide the most guidance about what works over time.

Under Peter Brabeck, Nestlé refined its recipes for success every time it missed a target. The cumulative effect was to enable the company to grow 5.4% per year over the 10 years up to the food crisis in 2007—faster than all of its rivals despite its enormous size, and in an industry that grows only at the pace of the world economy.

The company set broad cross-product goals for each of its regions and cross-region goals for each of its product groups. Managers then had leeway to shift goals across products or regions until goals for products within regions, and for regions within products, reflected equal tension. The result was detailed recipes for meeting regional goals across all products and product goals across all regions. Flaws showed up as conspicuous patterns of hits and misses across products and regions (author interview with controller Jean-Daniel Luthi, 22 May 2007).

For example, suppose the Kit Kat and Nescafé global managers each expect to sell $50m of product in Vietnam and $100m in Indonesia. Knowing more about local tastes, the Vietnam manager might negotiate higher Nescafé and lower Kit Kat targets (say, $70 and $30m) if the Indonesia manager will accept lower Nescafé and higher Kit Kat targets (say, $80 and $120m). Whereas total country and product targets remain the same, the specific product targets in each country now reflect equal stretch or ambition with respect to the market assumptions of the local managers.

As a consequence, missed Kit Kat targets in both countries are more likely to tell you something about Kit Kat strategy, and missed targets for both products in Vietnam are more likely to tell you something about Vietnam strategy.

DOI: 10.1002/sres.2219

Because equal-stretch goals have high information content, results provide much guidance about how to revise them.

Nestlé generates a succession of negotiated experiments, each benefiting from lessons learned from prior results. So there is a natural interplay between proposing new success drivers, which is necessary to set up specific product/country goals in the first place, and experimentation, which corrects and refines those proposals as quickly as possible.

Unlike WellPoint's pursuit of a list of broad and potentially conflicting requirements, Nestlé's system makes managers to negotiate conflicts before setting specific goals. Moreover, the very negotiations that resolve conflicts lead to more specific strategies. The specificity of Nestlé's successive recipes stands in sharp contrast to the list of ingredients WellPoint communicates to its shareholders.

CONCLUSION

In retrospect, it should be no surprise that US manufacturers continue to have trouble identifying metrics that can predict and explain performance results despite two decades of intensive IT investment in data analysis applications. And it should be no surprise that the hypothesis-driven experimentalism of Toyota and Nestlé's stream of negotiated experiments produce better results than HP's efforts to derive strategy from data and WellPoint's pursuit of data-driven requirements. Data mining cannot substitute for the hypothesis-driven learning needed to find metrics with the power to explain complex results.

Reliance on the successive experiments of hypothesis-driven learning to find what improves organizational performance may seem retrograde at a time of unprecedented analytic power. Both systems research and information theory suggest, however, that metrics with the power to predict complex results must have high information content as well as a high impact on those results. And data mining cannot substitute for the experiments of hypothesis-driven learning in the search for predictive metrics with high information content—not even in the aggregate—because it relies on low-information metrics requiring theories that are improbably complex.

This means the widespread belief that data mining can help managers find prescriptions for success is a fantasy. Data analysis can test metrics with the potential to predict or control complex results once identified. But it cannot substitute for hypothesis-driven learning.

REFERENCES

Ackoff RL. 1981. *Creating the Corporate Future*. Wiley: New York.

Ackoff RL. 2001. A brief guide to interactive planning and idealized design. http://www.ida.liu.se/~steho/und/htdd01/AckoffGuidetoIdealized-Redesign.pdf [10 March 2013].

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6): 716–723.

Apgar D. 2008. *Relevance: Hitting Your Goals by Knowing What Matters*. Jossey-Bass: San Francisco.

Ashby R. 1956. *An Introduction to Cybernetics*. Chapman & Hall: London.

Beer S. 1972. *Brain of the Firm: A Development in Management Cybernetics*. Herder and Herder: London.

Beinhocker E. 2006. *The Origin of Wealth*. Harvard Business School Press: Boston.

Burnham KP, Anderson DR. 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* 2nd ed. Springer-Verlag: New York.

Christensen C. 1997. *The Innovator's Dilemma*. Harvard Business School Press: Boston.

Cover T, Thomas J. 1992. *Elements of Information Theory.* John Wiley & Sons, Inc: New York.

Dewar J, Builder CH, Hix WM, Levin M. 1993. *Assumption-Based Planning: A Planning Tool for Very Uncertain Times*. Rand Corporation: Santa Monica, CA.

Doll R, Hill AB. 1950. Smoking and carcinoma of the lung. *British Medical Journal* **2**(4682): 739–748.

Fisher R. 1924. On a distribution yielding the error functions of several well known statistics. In *Proceedings of the International Congress of Mathematics*, Toronto, **2**: 805–813.

Gaitonde R. 2005. An analysis of HP's future strategy, post Carly Fiorina. *OS news* 24 Feb. http://www.osnews.com/story/9810/An-analysis-of-HPs-future-strategy-post-Carly-Fiorina [8 Feb 2012].

Gobry P. 2011. Leave HP alone. *Business Insider* 29 Aug. http://articles.businessinsider.com/2011-08-29/tech/30008570_1_leo-apotheker-hp-mark-hurd [8 February 2012].

Hacking I. 1975. *Why Does Language Matter to Philosophy?* Cambridge University Press: Cambridge, UK.

Hacking I. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge University Press: Cambridge, UK.

DOI: 10.1002/sres.2219

Hannan MT, Carroll GR. 1992. *Dynamics of Organizational Populations: Density, Legitimation, and Competition*. Oxford University Press: Oxford.

Hannan MT, Freeman JH. 1977. The population ecology of organizations. *American Journal of Sociology* **83**: 929–964.

Hannan MT, Freeman JH. 1984a. Structural inertia and organizational change. *American Sociological Review* **49**(2): 149–164.

Hannan MT, Freeman JH. 1984b. *Organizational Ecology*. Harvard University Press: Cambridge, MA.

Hausmann R, Rodrik D, Velasco A. 2005. Growth diagnostics. Working paper summarized in: 2006. Getting the diagnostics right. *Finance and Development* **43**(1).

Ibarra D. 2008. HR driving the business forward: HR analytics at WellPoint. *HR.com* 22 April. http://www.hr.com/SITEFORUM?t=/documentManager/sfdoc.file.detail&amp;e=UTF-8&amp;i=1116423256281&amp;l=0&fileID=1242361406876 [14 February 2012].

Jaynes ET. 1957. Information theory and statistical mechanics. *Physics Review* **2**(106): 620–630.

Jeffrey R. 2004. *Subjective Probabilities*. Cambridge University Press: Cambridge, UK.

Jorgenson D, Ho M, Stiroh K. 2008. A retrospective look at U.S. productivity growth resurgence. *The Journal of Economic Perspectives* **22**(1): 3–24.

Kepner C, Tregoe B. 1997. *The New Rational Manager*. Princeton Research Press: Princeton, NJ.

Kolmogorov A. 1963. On tables of random numbers. *Sankhyā Series A* **25**: 369–375.

Lashinsky A. 2009. Mark Hurd's moment. *CNN Money* 3 February. http://money.cnn.com/2009/02/27/news/companies/lashinsky_hurd.fortune [8 February 2012].

Lucas G. 1976. Econometric policy evaluation: a critique. *Carnegie-Rochester Conference Series on Public Policy* **1**(1): 19–46.

McGrath R, MacMillan I. 1995. Discovery-driven planning. *Harvard Business Review* **1995**: 44–54.

Noah T. 2010. Obama vs. WellPoint. *Slate* 10 May. http://www.slate.com/articles/news_and_politics/prescriptions/2010/05/obama_vs_wellpoint.html [14 February 2012].

Olson M, Van Bever D. 2008. *Stall Points*. Yale University Press: New Haven, CT.

Pareto V. 1896. *Cours d'Économie Politique Professé à l'Université de Lausanne*. Université de Lausanne: Lausanne.

Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press: Cambridge, UK.

Popper K. 2002. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Clarendon Press: London.

Reichenbach H. 1956. *The Direction of Time*. UC Press: Berkeley, CA.

Scholten D. 2010. A primer for Conant and Ashby's good-regulator theorem. (Unpublished.)

Tarski A. 1936. Über den Begriff der logischen Folgerung. *Actes du Congrès international de philosophie scientifique, Sorbonne, Paris 1935*, Hermann **VII**: 1–11.

Wiggins RR, Ruefli TW. 2002. Sustained competitive advantage: temporal dynamics and the incidence and persistence of superior economic performance. *Organization Science* **1**: 81–105.

Wiggins RR, Ruefli TW. 2005. Schumpeter's ghost: Is hypercompetition making the best of times shorter? *Strategic Management Journal* **26**: 887–911.

APPENDIX

For those interested in a precise mathematical interpretation of the term impact, it is used here to refer to backscatter or the sensitivity of a result to a cause or driver of that result. Mathematically, it is the entropy of the driver conditioned on the result it drives. The rough measure of a driver's WCI given in the article is related but not equivalent to this. WCIs, however, are useful in practice. The Conanct–Ashby theorem that every good regulator of a system must be a model of the system is equivalent to the conditional entropy requirement—it shows there is a mapping from the states of a system to the states of anything that effectively regulates, explains, controls or predicts it, thus minimizing the entropy of the regulator's states conditioned on the states of what it regulates (Scholten, 2010).

The conditional entropy of a driver given outcomes for a result that it drives can be measured as $R(h,e_i) = \Sigma \;\; p(e_i/h) \times \log_2 p(h/e_i)$ where $p(h/e_i)$ is the probability of an outcome h for the driver given each of the various outcomes e_i for the result (Apgar, 2008: 198). Note, once again, that this is not equivalent to the concept of WCI in the text.

Here is an analogue to the information-theory relations in the text for those more familiar with Bayesian probability. Let f be an outcome of a predictive factor and let e be a result that it helps predict. For the factor to have predictive power, knowledge of f must transform the probabilities we assign to different possible results e. In other words, $p(e/f)/p(e)$ must be very high or very low for all or most possible results e, where $p(e/f)$ is the probability of e given f and $p(e)$ is the prior probability of e. But by Bayes' theorem, that means $p(f/e)/p(f)$ must also be very high or very low for all or most possible results e. This in turn

DOI: 10.1002/sres.2219

means that p(f/e) must be close to 1 or 0 for all or most possible results e, and therefore, there can generally be no changes in f without corresponding changes in e—which is close to saying f must have a high impact on e. It also means that the denominator p(f) must be low for typical outcomes of the factor—which is the same as saying the factor must have high information content.

The argument on hospital clinics bears a resemblance to the so-called Chinese restaurant process giving rise to a limiting case of certain simple Dirichlet distributions that can also be used to determine limits on the number of useful explanatory factors. See: Ferguson T. 1973. Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**(2): 209–230.

DOI: 10.1002/sres.2219