CHAPTER 10

Business Strategy as Applied Social Science

Strategic Competition Versus Natural Competition

It is hard to exaggerate the strength of America's competitive position in the world economy in September 1945. The United States accounted for about one-half of all global manufacturing output, had the most technologically advanced economy in the world with ample supplies of natural resources, and could protect this state of affairs with an invincible military backed by a nuclear monopoly. Most of the rest of the world was either in ruins, preindustrial, or under the control of Communist regimes that smothered economic energies. The primary business challenge was to ramp up production rapidly and efficiently to meet demand. In the 1950s, the popular culture thought of American big business as so dominant that the primary public policy challenge was to restrict its power. But beneath the surface, the world was changing in ways that became obvious only later.

By the 1960s, global production capacity largely had been restored; by the 1970s, Europe and Japan had started to compete effectively again, and oil shocks battered the economies of the developed world. Large business saw cozy oligopolies under threat, and by the late 1970s, the dominant executive psychology moved from complacency to fear of new competitive challenges. Major companies were losing market share, facing pricing pressure from new lower-cost competitors, and being forced to confront new entrants with the ability to compete on product features and technological sophistication. From the perspective of these companies, it was as if a protected ecological niche suddenly had been invaded by all kinds of competitor species.

In 1963 Bruce Henderson, a farsighted purchasing executive who had recently lost his job at Westinghouse, convinced the Boston Safe Deposit & Trust Company to give him one room and a salary to form a consulting firm that within a couple of years was known as the Boston Consulting Group (BCG). He turned out to be one of the most original and influential business thinkers of the twentieth century.

The first major insight that put BCG on the map was the "experience curve": a quantified prediction rule that costs per unit tend to decline at a fixed rate as a company makes more and more of an item. By illustrative example, an auto manufacturer that had built 10,000 units of a specific type might observe that the 1,000th of these cars had cost $10,000 to manufacture, the 2,000th had cost $8,000, the 4,000th had cost $6,500, and the 8,000th had cost $5,100. Each cumulative doubling of production in this example reduces cost per unit by about 20 percent. (For some other company making toasters at a different factory, this unit cost reduction per doubling might be 10 percent or 30 percent or some other number, but the percentage is constant for any given process.) An experience curve of this type could allow the auto company to predict what cost per unit would be in the future, as a result of subsequent doublings. The company could know *today* what it would cost to produce the 100,000th car.

This would be a powerful scientific finding, in the sense of being a useful, reliable, and nonobvious predictive rule. In theory, the company could use this model to price cars at, say, $4,000 per unit now and lose money on the next few thousand, but seize market share from competitors who priced their cars with the aim of making money at current production costs. This would allow the company to race down the experience curve ahead of rivals, and therefore build an insurmountable cost advantage. It would be able to make money for a very long time by pricing above its own cost, but below the costs of competitors who got behind.

Does it work? Yes and no. The effect is real, and can be measured. It has been used to make correct, nonobvious predictions in many cases. One of the most famous is Texas Instruments, which used predictions of future cost reductions—which turned out to be accurate—to very aggressively price early electronic calculators in the 1970s, and rapidly grow sales and market share. On the other hand, as a basis for making strategic decisions, the experience curve is radically incomplete. For example, what if a competitor develops a new production technology that is vastly more efficient? Or what if a new kind of product is introduced that is superior in cost or functionality? Or what if other competitors have access to lower-cost capital? These and many other complexities make a linear application of the experience curve concept, in isolation from a more holistic understanding of a strategic situation, extremely hazardous. The Texas Instruments calculator business, in fact, imploded after several years of amazing growth when other competitors refused to play along.

Awareness of these kinds of complexities led BCG to try to incorporate more and more factors into various strategic frameworks. The most famous of these is the growth-share matrix, which attempts to allow a large, multidivisional corporation to use capital more effectively by categorizing its business units. For example, the framework argues that those units that generate more free cash flow than they can profitably invest back into their line of business ("cash cows") should provide cash to other parts of the business, and those that have opportunities to invest more cash than they can generate internally at returns above their cost of capital and into high potential market share positions that can exploit experience curve advantages ("stars") should use the excess cash the cash cows generate. Behind kindergarten-like imagery of cows and stars, this framework incorporates a sophisticated consideration of profits, capital structure, and market share. Yet even this is a gross simplification of real-world business competition.

How should such tools fit into business decision-making? Near the end of his professional life, Henderson wrote *The Logic of Business Strategy* (1984), a slim but profound book that summed up his deepest thoughts about this topic.

Henderson held a decidedly Darwinian view of business. He argued that for many generations humans have competed with one another as other biological organisms always have, a phenomenon he termed natural competition. In his view, businesses are simply vehicles for human competition in the form of extended networks of partial cooperation.

However, once humans evolved the specific capabilities of our species—imagination, logic, forethought, and the will to consciously commit resources today in return for future advantage—we could move beyond mere biological competition. Strategic competition, in which some competitors think through the chain of competitive responses and counterresponses that would result if they were to take various potential actions, allows them to choose actions to maximize their competitive success. Henderson argued that strategic competition offers immense time compression versus natural competition. In effect, by figuring out where natural competition is headed over many future trial-and-error steps, and jumping there in one big step, the strategic competitor compresses many evolutionary steps into one premeditated leap. The experience curve was, under this view, an early first step, and the growth-share matrix a second step, on the road to an ever more comprehensive model to predict the evolution of natural competition, and take advantage of it.

Henderson characterized natural competition as "evolutionary" and strategic competition as "revolutionary." His distinction between natural and strategic competition is, of course, an example of exploiting precisely the distinction between implicit and explicit knowledge, as defined earlier.

Careful premeditation is the key to making sure this isn't a potentially disastrous leap in the dark. As with all scientific knowledge, the key is to understand causal relationships well enough to create useful, nonobvious, and reliable predictive rules. Henderson was clear about this: "To accomplish this revolution, the preparation must be conservative, careful, precise, and all inclusive. . . . Meticulous staff work must be continued until cause and effect become sufficiently predictable to justify the massive commitment of non-recoverable resources."

His vision of what must be known to compete strategically was incredibly demanding. We don't just need to have partial or fragmentary predictive rules, but we need to understand the entire system in which we are competing, including the ability "to understand competitive interaction as a complete dynamic system that includes interaction of competitors, customers, money, people, and resources," and "to use this understanding to predict the consequences of a given intervention in that system and how that intervention will result in new patterns of stable dynamic equilibrium."

This is far beyond anything comprehended by experience curves and growth-share matrices. It would be extremely useful, but is anything like it possible in the real world? (It also raises the obvious question of why a society that had access to all of this would allow the messy and expensive process of market competition in the first place.) Henderson claimed we were getting close to this capability. Although strategy development was still embryonic, we had the promise of "precision, elegance, and power within a reasonable time period." He didn't define this time period, but he implied that this was something on the order of a generation or so.

We are more than twenty-five years on from this judgment, and I see no danger of our developing the kind of comprehensive knowledge that Henderson said true strategic competition required. In retrospect, his prediction seems hubristic to the point of outlandishness. Why has it proved so difficult?

Macro-Strategy Versus Micro-Strategy

I graduated from college the year Henderson published *The Logic of Business Strategy,* though I knew nothing about it and probably wouldn't have cared if someone had handed me a copy. I had studied science with the intention of becoming an academic in math and physics. As college progressed, however, I had become increasingly fascinated by applying mathematics to predict human behavior. After a year in a PhD program devoted to this topic, I decided that the academic life wasn't for me. Needing to pay the rent, I took a job at AT&T's research laboratories, and despite being in a technical organization, found myself drawn into business debates. I was shocked to discover that I found them fascinating, and so sought out a job in strategy consulting, which I understood in some vague way to specialize in analyzing these kinds of issues.

This was the late 1980s, and strategy consulting had already become an established industry dominated by BCG, Bain (a BCG spin-off), and McKinsey. There was also a well-worn career path: Ivy League degree, followed by two years as an analyst at a strategy consulting firm, followed by an MBA at Harvard, Stanford, or Wharton, and then a return to consulting as an associate, followed by about seven years of long hours to make partner. I fit almost none of that profile.

But a number of years earlier a young consultant at BCG, W. Walker Lewis, had left to create another spin-off strategy consulting firm, Strategic Planning Associates (SPA). The firm was founded with the specific purpose of using more analytically sophisticated and data-intensive computerized analysis than was then typical in the strategy consulting industry. Therefore, one of the senior partners was open to the idea of a person with a somewhat nontraditional, technical background like mine as an experiment.

I started work at SPA in 1987, at age twenty-three, and immediately loved it. It was as if someone had designed the perfect environment for me to pursue what had become my interests. It was far less theoretical than I thought academia to be, but focused on rigorously applying data and analysis to develop strategies to outsmart the accumulated intuitions and experience of huge companies. It played directly to all of my youthful ambitions and vanities.

My first assignment was as the junior member of a team charged with developing a strategy for the leading competitor in a mature industry that made commoditized glass-based products. I was tasked with modeling the economics for every production line in every factory in the United States for the whole industry. In effect, I built an actual, empirical version of the economist's famous supply curve for each product by combining physical principles of chemical engineering with painstakingly collected data about our client's and its competitors' facilities. This meant that we could determine, by product, the level at which each production line in the industry would maximize profits, and could predict the multiple-year effects on prices, and therefore profits, of potential investments in production capacity. Like a card counter in a casino, the client could use proprietary knowledge to take rational actions, while seeming to make the kind of risky bets that everybody else at the table had to make.

This specific situation was almost ideal for deploying Hendersonian strategy: our models were underwritten by physical science; there were only four relevant competitors; these competitors behaved according to similar rationales; technology change was slow; there were few relevant substitute products; and so forth. Therefore, we could use this knowledge to predict competitive response to our client's prospective actions. For a time it appeared that this client could turn dials in its own business and drive industry behavior so as to make an enormous amount of money. It was a heady experience.

But two problems subsequently emerged.

First, not everyone at the client agreed that the actions this strategy dictated were responsible for driving profit improvements. Evaluating competing claims for program effectiveness in a business usually is not simple, because we have no rigorous answer to the fundamental counterfactual question in all program evaluation: But for this action, what would performance have been? Suppose that over the three years after this client began implementing this strategy, annual profits went from $1 billion to $1.5 billion, but over that same period, the economy as a whole started to grow faster, one competitor exited a key market, a new technology from an adjacent industry began making significant inroads into this industry, and the client also replaced the head of sales and instituted a process improvement program in its factories. Which of these potential factors deserve how much, if any, of the credit for the profit improvement? Executives use experience, observation, and data to form intuitive judgments about this, and everybody has an obvious incentive to inflate his own contribution and to denigrate that of others.

As with evaluations of surgery versus evaluations of therapeutics, as a practical matter some programs have effects that are so obvious that this question becomes academic; in such instances, the informed judgment of an experienced professional can be reliable. A classic case is a cost reduction created by increased productivity in a factory with no reasonable prospect of a change in consumer perception. On the other hand, most business changes that affect consumer behavior, and therefore revenue, tend to be far more ambiguous. In the longer run, and from the perspective of the economy as a whole, some companies survive and grow, while others go bankrupt or are acquired. In this way, packages of such judgments, as embodied by entire firms, receive some form of feedback. But of course, this is just natural competition, or evolution, and as such does not allow us to know which specific decisions contributed to success or failure. It is pure implicit knowledge.

The second problem with our strategic plan was that after several years, changes in the competitive environment made the modeling clearly obsolete—much like what happened to Texas Instruments after executing an analogous strategy for its calculator product line based on the experience curve. Innovative technologies came on the market, and figuring out the best strategy in that new environment required forming judgments about how this technology might change over time, how consumer preferences would evolve, and so on. Further, a new competitor entered the market that was part of a larger, integrated enterprise, and was making decisions that violated the economic assumptions of our framework, because they apparently were less concerned with making money in this market than in serving some larger corporate objectives. They refused to play by the rules assumed in our model. But since even profitable investments in new-factory capacity take years to pay out, this meant that the early capacity investments might create less economic profits than we had originally expected.

These two problems—the inability to rigorously evaluate the effectiveness of strategies, and growing deviation between reality and the assumptions of the strategy within the time frame the strategy required to create economic profits—kept cropping up as I did more projects, even in the most successful strategy work. They were really manifestations of one underlying problem: the analytical model of the business was always incomplete. *The model is never the system.*

One reaction to this observation in the wider business community was strategic nihilism: the rejection of the idea that strategy was useful, and the belief that the whole strategy exercise was more or less a scam. In Henderson's terms, this is the argument that analytically derived strategic competition was impractical and that the route to success lay in superior execution of natural competition. This appeared in several guises. Interestingly, the theme that united all of them was that strategy ignored the "human element."

One major strand was that what really mattered was motivating and empowering the people who made up the organization. In other words, strategy is make-believe, and only execution is real. Tom Peters and Robert Waterman's epochal business best seller, *In Search of Excellence* (1982), was the founding text of what became a huge movement to empower employees. The emotional energy behind this movement was a cri de coeur of the middle manager: *I matter! I'm not just some piece on your chessboard.* Though it eventually descended into a kind of snake oil that verged on promising to undo the inexorable grinding away of stable, high-wage middle management jobs caused by globalization and the ceaseless advance of information technology, this critique was premised on a very real insight: one of the most severe blind spots of strategy as it was actually practiced was insufficient recognition of the importance of human agency and motivation.

The other major strand was less an organized movement than daily resistance by executives. Picture a been-there-done-that senior executive who refused to accept an analytically derived strategy because of some plausible objection that resisted quantification and analysis. This kind of objection was a practical version of the observation that the analytical models the strategists used were incomplete, and importantly, that there was no way to even scope the relative impact of many outside-of-the-model effects versus those the model considered. Typically the most compelling of these objections would be linked to arguments about human behavior: potential customer reactions to proposed strategies, potential competitive reactions that violate the assumptions of our economic framework, and potential creative technological or business process innovations.

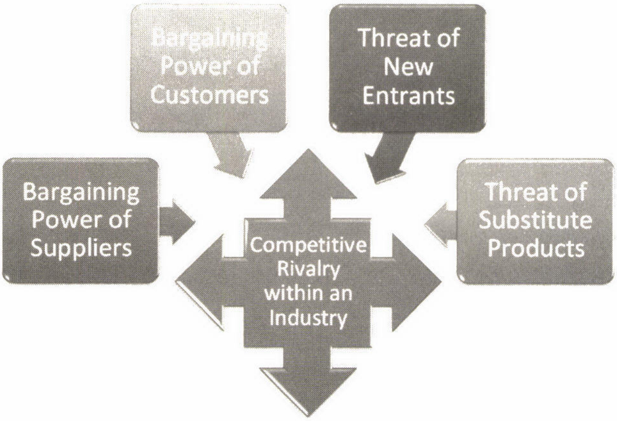
A classic example for American consumer products companies was what we came to call "the Walmart bomb." A senior sales executive often would react to some strategy he didn't support—say, eliminating some products or changing prices—by saying something like, "Sure, that might make us an extra $20 million, but it will put the whole Walmart account at risk, and if we lose them, we go out of business." It's plausible, terrifying, and usually not analyzable. It makes business strategy inherently judgmental. Not coincidentally, like physicians arguing for control of treatment regimens, it retains power in the hands of the relevant executive. But also like those objections, it carries great intellectual weight.

Two alternatives to strategic nihilism were attempted by those who saw that the strategy models were incomplete but wanted to find a way to make strategy work. They went in basically opposite directions.

The first was to build ever more general frameworks that could incorporate things like technological change, human motivation, more complex analyses of competitive intent, and so on. Call this approach "going macro."

Though it continues to the present time, the macro approach probably reached its intellectual apogee with the publication of the massive tomes *Competitive Strategy* (1980) and *Competitive Advantage* (1985) by Harvard Business School professor Michael Porter. In them Porter lays out the most famous and influential of the strategic frameworks in his "five forces" model, which purports to identify the characteristics that make some industries more profitable than others. He combines this with his taxonomy of "generic strategies" to provide a framework to assist corporate executives in creating shareholder value through strategic decision-making.

Here is a standard graphical representation of the five forces model:


Bargaining Power of Customers
Threat of New Entrants
Bargaining Power of Suppliers
Competitive Rivalry within an Industry
Threat of Substitute Products

The Five Forces Model

This is obviously simplified; for example, each of these forces is further subdivided, and book chapters have been written about how to approach and analyze each. Nonetheless, this framework is really just a very detailed and intelligent list of issues, and words about how to think about them, logically grouped into categories, and then visually arranged into a semicircle. It is not a model in the sense of a definitive set of rules that makes falsifiable predictions in the way that the experience curve predicts the cost of the 100,000th car after we have produced only 10,000 of them.

Empirically, as such models or frameworks or whatever we want to call them become more general, they have always become nonfalsifiable methods for organizing our thoughts. This reflects our ignorance. In practice general frameworks like the five forces model serve as (1) a checklist to make sure we don't forget to consider various issues that experience has shown to be potentially important, and (2) narrative description of how to think about them, sometimes incorporating analytical tools such as the experience curve or growth-share matrix for subproblems, as a starting point for the real analytical and intuitive thought process we use to make judgments in an environment of high uncertainty. This is far from useless but is also pretty far from Henderson's vision of analytically derived strategy.

The second approach to reflecting the incompleteness of existing strategy models was to "go micro" by tackling somewhat more bounded problems. That is, rather than asking what tools would be required to do analytical strategy development, we instead ask what we could do with the analytical tools we could actually build.

This was mostly the route we took at SPA. We believed that *our* competitive advantage was skill in inventing and applying creative analytical approaches, so we looked for instances where this could create a lot of economic value. We attacked problems such as establishing where to invest in telecommunications networks, redesigning production processes for manufacturing companies, and matching financial assets and liabilities to increase shareholder value. The whole game was to operate at a more strategic level than either operational businesspeople or analytical technical specialists, but to not become so strategic that nonanalyzable factors would overwhelm the benefits achieved through careful analysis and modeling.

A simplified example was an attempt to improve the economic performance of the glass-based products manufacturer that I referenced earlier. A mathematical method called a linear program (LP) was conventionally used to figure out what combination of products a factory should produce to make the most money. The reason a complex algorithm was useful was that the decision to make any one product on a given line at a given time would affect the economics of all the others (in the language of this book, a factory is holistically integrated). Technical specialists did this work, and it was difficult to create value here, as anybody could learn the math in school, buy commercial software, and apply the technique. It was already commoditized.

But we hypothesized that by adding specific kinds of warehouse space, we could permit new production possibilities in the factory that produced greater profitability. We next hypothesized that by going upstream from the factory, we could integrate decisions about raw materials purchasing that would drive further profit improvements in the factory by changing the trade-offs the LP faced. We iteratively conducted quick analysis to evaluate our hypotheses, improved our understanding, and ultimately built a sufficiently complete analytical model to change decisions for the whole chain from raw materials through manufacturing and on to the warehouses to increase total corporate profit. Like the earlier efforts to model and exploit the industry supply curve for the same company, this made them a lot of money for some period of time.

Guessing about these opportunities, understanding their economic leverage, and developing integrated decision models based on them was not yet standard practice; therefore, it could produce abnormal profits for our client and high wages for us. But because we insisted on creating analytical knowledge, it was not the inherently intuitive guru-like insight of macro-strategy. That meant that eventually it would become commoditized, which is precisely what happened in this case. Across the industrial economy, what were originally manufacturing optimization tools gradually were also used to routinize optimization of functions such as warehouses, raw materials purchasing, and so on. What was innovative yesterday becomes routine tomorrow. So, a successful career in micro-strategy meant constantly inventing new methods. As I'll review in the last part of this book, this is a good example of the overall process by which high-wage jobs are created and then destroyed in the information economy.

Just as we saw with nonexperimental scientific fields, such as parts of evolutionary biology or astrophysics, many of the most successful applications of micro-strategy were tightly linked to experimentally validated physical engineering principles. For example, a key step in improving the economics of the glass products manufacturer was the ability to find relationships between changes in raw materials and changes in yield (some of which previously had been tested in chemical engineering experiments). The manufacturing environment was simple enough that we could find robust, stable statistical relationships that would hold under any reasonable analytical assumptions. This is a simpler problem than smoking-lung cancer, and therefore not nearly as complex as abortion-crime.

I did a lot of this kind of work, but because of my background in predicting human behavior, I ended up trying to build models for pricing, product introductions, and other consumer-oriented decisions. I discovered that I could build analytically sophisticated theories all day long, but it was very difficult to know whether they were correct, because by making slightly different assumptions in the analysis, I could get very different answers for the best predicted course of action. Micro-scaling down from macro-strategy development didn't bound the problem enough to eliminate the same issues that had plagued the original strategists operating at a grander level, because the same root problem was still present: as long as we were trying to predict human behavior, the problem was too complicated for the analytical tools at hand.

I used detailed case studies to illustrate this for social science analyses, and I'll present a similar one that shows these problems for a kind of business analysis I did many times. Unlike the grand themes of politics and morality, this is a down-home example: trying to predict the effect of changing the name of a convenience store.

Will QwikMart Sell More If We Rename It FastMart?

I was once asked a seemingly simple question by a senior executive of a company that operated 10,000 convenience stores, of which 8,000 were named QwikMart, and 2,000 were named FastMart. (I will mostly use retailer examples to describe business analysis of human behavior, because they are so familiar to most readers; and, as in this case, I will anonymize all brands, data, and names to preserve confidentiality.) The executive observed that average annual revenue per store was $1 million in the QwikMart stores and $1.1 million in FastMart stores. She wanted to know whether the company would increase sales by changing the names of all the QwikMart stores to FastMart.

Her question was not easy to answer reliably. We quickly determined that a difference this large was extremely unlikely to have occurred randomly, but we obviously couldn't just assume that the name on the front of the store caused sales to be higher. So, the first logical question to ask was whether there were systematic differences between the QwikMart and FastMart stores other than brand name that might account for the difference in sales. The list of plausible candidate causal sales drivers that could vary on average between the QwikMart and FastMart groups was very long, typically including, as a few practical examples: physical size of the store, how long the store had been open, number of people who lived near the store, average income of people who lived near the store, average number of children per family living near the store, number of nearby competitor locations by brand, relative quality of merchandise at each competitor store, number of parking places, traffic count on the road in front of the store, ease of access from the road, distance to nearest highway, visibility of store and signage, number and quality of other complementary nearby retailers, exact interior store layout, number of open hours per week, number of in-store employees, tenure and background of store manager and employees, mix of employees by skill level, match of employee demographics to customer demographics, amount of shelf space allocated to each department, number of individual products by department, exact position of each product on each shelf, total inventory on hand and inventory mix by department, number of stock-outs by department by day of week and time of day, number of checkout positions or cash registers, deployment of anti-theft technology, cleanliness of the store, quality and maintenance of interior lighting, presence of an ATM in the store, level of TV, radio, print, and other channel of advertising we had done for the market in which the store operated, level of competitive advertising in the same market by channel, relative quality of advertising copy we and competitors had executed for each market, and so on, in practical terms, ad infinitum. It is manifestly an environment of high causal density.

We could do our best, however, to hypothesize as many potential causal factors as we could bring to mind, and then, where practical, collect data on each of these factors for every store. The analytical task in front of us was to determine whether there was a residual difference in sales between the QwikMart and FastMart groups after "holding all other factors equal," then to assert that brand difference must have caused the residual sales difference. Note the parallel between this and the social science models from the last chapter. We would argue a single causal effect could be isolated if all other potential causes are held constant, by analyzing historical data, rather than actually running a field experiment. Henderson's dream of comprehending the entire causal system with sufficient precision to make the decision rationally would be achieved in this one little corner of the world.

As a starting point, one might collect data on, say, 1,000 factors thought to influence sales, then observe that larger stores tend to sell more than smaller stores—in fact, on average, each additional 1,000 square feet of store size is associated with an additional $40,000 in annual sales. (Note the careful and important hedging in the term "associated with," as opposed to "caused by.") We could use this to normalize the sales for each store to reflect its sales versus what would be expected based on its size, and then continue this kind of procedure for each of the factors on our list. We could declare that whatever difference in sales remains after we have adjusted for all factors other than brand is the difference in sales caused by brand. The standard method for doing these adjustments is to create a regression equation of the form:

$$\begin{aligned}
\text{Annual Sales of a Store} = {} & \$40{,}000 \times \text{Store Size} \\
& + \$20{,}000 \times \text{ATM in Store} \\
& \quad \text{(1 if store has ATM, 0 if no ATM)} \\
& + \ldots \\
& + \$50{,}000 \times \text{Store Brand} \\
& \quad \text{(1 if store is FastMart, 0 if store is QwikMart)}
\end{aligned}$$

This is the kind of equation you will see in countless business (and economics, political science, sociology, and other quantitative social science) journals. The conclusion is typically couched as "$50,000 is the estimated impact of store brand after controlling for other factors."

But if I apply the same critical lens to my own work that I applied to the Bartels and Levitt social science regressions, the problems with this assertion can be seen clearly.

First, remember that we can never know we have identified and collected data on all the potential causal drivers of sales. Therefore, we cannot eliminate error that arises because brand is a statistical proxy for a variable either we never considered or for which we could not get data. Adjusting for some but not all of the other potential control variables often does more harm than good in estimating the effect of one particular potential cause of interest, and there is no way to know which without knowing the true list of *all* variables that actually cause sales. This problem goes by many names. I'll refer to it as omitted variable bias.

Second, even among the variables for which we have collected valid data, causal density is higher than it might appear from even a very long list of potential causes of sales, because the various causes typically interact. For example, let's say that having an ATM in the store drives sales in large stores both because it draws incremental customers into the store and puts more money in the pocket of all customers who use it; but in small stores, on balance it reduces sales, because in addition to these effects it also creates so much crowding near the cash register that it discourages customers to an extent that more than outweighs its positive benefits. In other words, sometimes an ATM helps sales, and other times it hurts them. The jargon term for this is an interaction effect.

But in our regression equation, we can have only one coefficient for the variable "ATM in store," which must be either positive or negative. The standard remedy for this problem is to add interaction terms into the equation. We could replace the variable "ATM in store" with two variables: "ATM in store AND store is large" and "ATM in store AND store is small," and therefore create a separate estimate for each variable. Unfortunately, not only are there many such interactions, but also higher-order interaction effects in which interactions themselves interact. For example, an ATM may increase net sales in most large stores, but not in highway rest stops, which tend to be very crowded due to high traffic, and at which speed is more crucial than usual to the customers. So we would need to replace the interaction term "ATM in store AND store is large" with two terms: "ATM in store AND store is large AND store is in highway rest stop" and "ATM in store AND store is large AND store is not in highway rest stop." These interactions-with-interactions can expand indefinitely. In a complex system driven by human behavior, interaction effects are not peripheral issues, but usually are central to the phenomenon under consideration. One thousand potential causal variables could become 10,000 potential causal variables, and entirely overwhelm our dataset of 10,000 stores.

Third, the direction of causality between control variables and the outcome of interest is often unclear. For example, do increases in store size cause increases in sales? There are many intuitive reasons why making a store physically larger could do this: there is more space to display items better; more inventory can be kept in front of consumers; it allows for a more spacious and pleasant shopping experience; it might create more visibility from the street and entice more customers to stop; and so on. On the other hand, higher sales might have led management to expand stores incrementally over time as sales grew faster in them. A third possibility is that there are also lots of ways in which store size could simply be a proxy for other real causes of higher sales: markets in which we do more advertising and so have higher sales may tend to have cheaper land and thus have larger stores; highway rest stop locations are the highest sales stores, and the authority that manages these rest stops might mandate large stores, etc. That is, more size might cause more sales, more sales might cause more size, or some other factor may cause both. Most likely, of course, is that all three are going on at once. But when we "control for" any factor, we implicitly assume that it is a causal agent in a specific direction—that this variable causes the outcome (or at least that it proxies very well for causal agents).

In sum, in real business problems in which we attempt to construct a regression model from historical data to predict some aspect of human behavior, the combination of the three problems I've just described—omitted variable bias, prevalent high-order interaction effects, and variable intercorrelation—presents enormous practical obstacles. Many attempts have been made to circumvent these difficulties.

One is to force structure on the model based on beliefs about causality that are external to the model. As a simple practical example of this, unconstrained application of regression methods might result in a positive coefficient for average price; that is, an indication that higher prices cause higher sales. The model-builder might "know" that higher prices should, all else equal, cause lower sales, and therefore decide not to include this variable. (You might be surprised at how often this kind of thing happens.) But of course such approaches beg the question of how we know our beliefs about causality are correct.

Alternatively, numerous non-regression pattern finding methods have been developed to attempt to build models using different mathematical approaches to the same problem: how to predict the effect of changes in various potential causal factors of an outcome without making structural assumptions. I'll refer to regression and these other non-regression techniques collectively as pattern finding. There is always some hot new pattern-finding algorithm that promises to do this. Most of these approaches have arisen in the new computational environment of large datasets and cheap processing power, and are therefore termed machine learning or data mining techniques. Well-known examples include decision trees, case-based reasoning engines, neural networks, modern implementations of Bayesian statistics, clustering, and support vector machines, as well as various hybrids and extensions of these methods.

I have used such algorithms many times to analyze real business problems. Each tends to have specific application niches in which it demonstrates performance that is better than alternative methods (e.g., neural networks for rapid credit card fraud detection). But none of these can resolve the three core problems of omitted variable bias, interaction effects, and intercorrelation indicated in the example above, because these problems are not at root a result of some unique shortcoming of regression as a method, but are inherent to the phenomenon under study. In fact, though these problems are exacerbated by small samples of, say, thousands of data points, they still apply to the largest datasets. We could have 10 million individual customer records for a very large bank, rather than 10,000 stores, in our database, and though this might (or might not) partially ameliorate the problem of having enough data to specify a large number of interaction effects, it would do nothing about omitted variable bias, and often will not help much at all with the problem of intercorrelation.

A third approach is to add another kind of data: the dimension of time. We could look at data on some stores that have been rebranded from QwikMart to FastMart and see what happened.

The company had previously rebranded one hundred QwikMart stores FastMart, and we could treat this as a natural experiment. If we compared the change in sales in these stores after being rebranded versus before being rebranded, we would entirely eliminate the problem of how to control for differences other than brand between stores, because we would be comparing the same stores at different points in time. The trade-off would be the introduction of a new source of bias: that the rebranded stores might have experienced a change in sales even if we had not rebranded them. If annual sales are down $22,000 (or 2.2 percent) per store on average for these rebranded stores, it might be because the economy entered a major recession. As always, the analyst must answer the question of the counterfactual: But for the rebranding, what would sales have been?

Of course we have the balance of the 9,900 stores in the chain as a potential control group. If the rebranded stores had dropped 2.2 percent in sales, but the rest of the chain was down 3.2 percent, we might attribute the 1 percent difference to the rebranding. But what if all one hundred of the rebranded stores were in Chicago? Then it might seem more sensible to compare them only to other stores in Chicago, since the recession may have been more or less severe in Chicago than nationally. There may be all kinds of other causal changes over this period with disproportionate effects in Chicago versus the rest of the country. But all of the rebranded stores were QwikMart—so shouldn't we compare the change in performance of the rebranded stores to change in performance of only the other QwikMart stores, since various non-brand causal changes may have disproportionately affected QwikMart stores? Or should we compare them only to QwikMart stores in Chicago? Suppose the rebranded stores were larger, on average, than other stores. Should we compare the rebranded stores only to a similar size mix of QwikMart stores in Chicago? And so on. What, in other words, is the appropriate reference class for analysis of this list of one hundred rebranded stores?

The standard methods for addressing this question are to either (1) try to identify a specific subset of stores that are most like the rebranded stores, normally termed matching, or (2) use some pattern-finding method to predict the outcome of *change* in sales after purportedly controlling for all other factors (the typical regression version of such a pattern-finding model on changes is normally termed pooled regression). The first is much like the abortion-crime natural experiment, and the second is much like what Bartels and Levitt did with their regression models. Matching is the more conceptually straightforward, though both methods will have the same underlying weakness: we don't know what factors other than rebranding affected the rebranded store group differently than whatever control group we choose.

Suppose management selected these stores for rebranding because they knew a new competitor was entering this market, or because these stores looked the worst and were therefore believed to be likely to have rapidly deteriorating sales. An infinite number of possible reasons based on future expectations might have introduced significant bias in selecting the stores for rebranding versus the control group. Suppose each store manager had the right to decide whether his or her store was rebranded. All kinds of considerations might have played into, for example, the decision to repaint or the manager's performance in operating the store afterward. With any pattern-finding method, the number of unconsidered or incompletely considered potential sources of bias may be infinitely long.

This bias may be so large that rebranding cannot be considered the primary cause of the effect. In any natural experiment, the first step is to ascertain the bias in selecting the case group versus the control group. But this analysis can be misleading. To take an obvious example, what if an analyst wanted to study the effect of holding higher levels of inventory on later sales, and therefore looked at the natural experiment of the change in sales for stores that did major inventory buildups as compared to the other stores that did not do these buildups over the same time period? The problem, of course, is that inventory buildups or draw-downs are sometimes based on a manager's foreknowledge of local demand changes. Unfortunately, most bias issues inherent in studies of existing data aren't nearly so obvious.

Pooled regression (and similar methods) simply builds a regression equation in which the predicted outcome is "change in store sales" rather than store sales. The terms of the equation attempt to control for all possible causes of change other than rebranding. This is generally superior to building such a model simply on store sales, because the causes of variance *between stores* (e.g., the list of reasons a specific store in downtown Chicago has different sales than a specific store in rural Oregon) are generally far more significant than the causes of variance *within stores over time* (e.g., the list of reasons the change in sales from June to November in the downtown Chicago store is different than the change in sales from June to November in the rural Oregon store). But, as per the discussion of matching, unobserved reasons for exactly such biases in comparing changes within stores over time to changes within other stores over the same time period do exist are often subtle, and typically have larger causal impacts than the causal impact of the program of interest. The same classes of problems observed with straight regression are present with pooled regression, usually less severely, though unfortunately they are still plausibly severe enough to make the method unreliable.

How can we know that even the best of these methods is correct? Various measurements of statistical significance, confidence, and so on cannot tell us, because the open question is whether we have violated the assumptions that go into such models. Having many analysts look at the problem, and seeing whether we reach a consensus can't tell us, because if they are all missing relevant data—which in this kind of situation is always a realistic possibility—they will all reach the same faulty conclusion. For the same reasons, hiring new, smarter analysts, collecting more data, or applying new algorithmic methods cannot tell us whether we're right.

The only generally reliable way to test our theory is the approach that C. S. Peirce, Jerzy Neyman, and R. A. Fisher discovered many decades ago: roughly speaking, pick a random sample of QwikMart stores, rebrand them as FastMart, and compare what happens in them to a control group of stores that we do not rebrand. But of course, if we are going to rely on the experiment as the definitive measurement of the causal effect anyway, then why not just skip all of the analysis and run the experiment?

Well, for one thing, rebranding a test group of, say, fifty stores likely would cost on the order of $1 million. Although it's not free to do research and analysis, it's a lot cheaper than that, and I might find that such models improve my guesses about which theorized programs end up succeeding in experiments. For another, once I've completed the experiment, I will face the problem of how to generalize the results from the specific test stores to predict the effects of rebranding the other 7,950 QwikMarts. In other words, just as we saw with therapeutic RFTs, theory prioritizes some potential experiments over others as users of scarce resources, and also generalizes results from experiments to other untested entities.

A company can earn a lot of money by making experiments a central element of how it makes decisions—specifically as the preferred method for program evaluation. All else equal, an organization in a consumer-focused industry that does this will have a material competitive advantage over those that do not. But experiments must be integrated with other nonexperimental methods of analysis, as well as fully nonanalytical judgments, just as we saw for various scientific fields. I described the manner in which this is done in science as "philosophically unsatisfying," but as we'll see in the next chapter, the way it is accomplished in a real for-profit business makes that look like a discussion on the porch of Plato's Academy.