Little Data
By Jerry W. Thomas

You are not feeling well, so you visit your friendly family doctor. He puts you in a new, electronic scanner and generates 28 trillion measurements of your temperature all over the surface of your body. He then saves all of these big-data measurements and, using advanced statistical algorithms and supercomputers, announces that your temperature is 98.6 degrees Fahrenheit. What a relief! Big data to the rescue.

The Bandwagon

As the "big data bandwagon" picks up momentum, consultants, professors, conference organizers, authors, magazines, blogs, software firms, pundits, crooks, private equity firms, and computer hardware manufacturers clamor to get aboard. Rarely has a bandwagon attracted so much attention or so many passengers.

The basic premises of big data appear to be that:

- More data are always better than less data.

- Volume, variety, and velocity of data create new sources of potential knowledge and prescience.

- With big data, all questions can be answered; the "why" will finally be revealed to the human race, and the future can be accurately predicted.

Is big data an accurate picture of the future, or is it simply a mirage shimmering in the distant desert heat? Is it the pathway to ultimate truth, or is it only a bandwagon of exaggerated promises and illusory dreams?

As the "big-data bandwagon" picks up momentum, consultants, professors, conference organizers, authors, magazines, blogs, software firms, pundits, crooks, private equity firms, and computer hardware manufacturers clamor to get aboard.

Decision Analyst
strategic research ■ analytics ■ modeling ■ optimization

1.817.640.6166 or 1.800. ANALYSIS • www.decisionanalyst.com

The truth is the solution to marketing and business problems—and the identification of strategic opportunities—often lies in the realm of little data, not big data. You don't have to boil the ocean to determine its salt content. You don't have to eat the whole steer to know it's tough. Most times a doctor only needs to take your temperature with a $20 thermometer, not a $10 million scanning machine. The great opportunity is not more data faster, but better data—and better analytics.

The Limits of Data

The preponderance of business data— indeed, all data—in the world is historical data, or "tracking" data, such as financial data, sales data, customer behavioral data, weather data, and inventory data. Virtually all data tend to be backward-looking, analogous to looking in the rearview mirror to steer a car forward.

No matter how current or instantaneous data are (i.e., the velocity) or the sheer amount of data, the backward-looking bias is an omnipresent limitation. We might see trends in that data that give us an inkling of the near-term future, and we might be able to find out what has driven a firm's success in the past, but most historical data are of limited value in predicting the future.

Another limitation of most business data is *noise*. All data consist of two main components:

- *Noise*—Random variations, static, aberrations, errors, missing values, and distortions.

- *Signal*—Relevant, valid information in a recognizable, reliable pattern.

The human race has spent the better part of the last two million years trying to separate the signal from the noise, and for the vast majority of that time our ancestors conjured up myths and magic to explain their experiences and observations. It is a pattern that the perceptive can still see at work in corporate America and Washington, D.C. The scientific thinking and scientific systems that have evolved over the past 5,000 years represent a revolution, a great leap forward, in our ability to separate the signal from the noise.

No matter how current or instantaneous data are (i.e., the velocity) or the sheer amount of data, the backward-looking bias is an omnipresent limitation.

Inflationary distortion is another major limitation of business data. Inflationary and deflationary forces vary by product category, by industries, and by time, and rarely does anyone have the measures or means to correct for these distortions. We can think of the changing value of the dollar and variations in currency exchange rates as other types of pervasive noise.

Data You Can Trust

Often, without thinking, we tend to see all data as equal, but rarely is this true. The corporate world is awash in data. It streams in from all directions 24 hours a day, and the data deluge continues to worsen. Most large companies today have at least 100 to 200 times as much data in their collective databases as they had 30 years ago. Has all this data helped them make better decisions? Would (or could) anyone argue that today's corporations are making better decisions than they did 30 years ago?

I would argue that most major corporations are making poorer decisions now, compared to 30 years ago, despite the rising tide of data. In fact, the growing flood of data is part of the problem. More data often means more confusion. Which data are correct? What data can be trusted?

Here's a point of view on the trustworthiness of various types of data, ranked from most trustworthy to least:

1. *Experimental Data.* Carefully designed and carefully controlled experiments, conducted by objective third parties who are experts in such experiments, yield the most trustworthy data. Before-after and side-by-side controls are employed, along with sophisticated statistical analyses, to separate the signal from the noise.

2. *Survey Research Data.* Scientific research studies, conducted by experienced professionals who are objective third parties, yield trustworthy data. Often this data is experimental in nature. Research design, normative data, mathematical modeling, stimulus controls, statistical controls, historical experience, quality-assurance standards, etc., tend to

More data often means more confusion. Which data are correct? What data can be trusted?

make this data very precise. Noise tends to be minimal.

3. **Marketing-Mix Modeling Data.** The creation of an analytical database, the cleansing and normalizing of that data, and the use of multivariate statistics and modeling to isolate and neutralize some of the noise tend to make marketing-mix modeling data better than actual sales data.

The signal in marketing-mix modeling data is more stable, more reliable, and more measurable. This type of data can be valuable in helping companies understand what variables are driving their businesses (is it media advertising, or the number of salesmen, or pricing differentials?), but it generally takes multiple years of data to get maximum value out of marketing-mix modeling.

4. **Media-Mix Modeling Data.** This is the same concept as marketing-mix modeling, just applied to a different set of variables. The same general rules apply. An analytic database, data cleansing, modeling, and statistics allow the noise in the data to be minimized, so that the effects of various media can be isolated. Again, if combined with controlled experiments, the data and analyses are much more explanatory.

5. **Sales Data.** Sales data are pretty good, but not perfect, measures of actual sales. But sales are not reliable and valid measures of advertising effectiveness, optimal media spending, product quality, service productivity, competitive activities, etc.

Sales data can only be trusted so far. The noise often drowns out the signal. The economy, competitive activity, the weather, inflation, the vacation cycle, news events, political events, aberrations in inventories and distribution, pricing disturbances, etc., create false echoes and distorted illusions. Sales data are not good measures of cause and effect. Sales are reasonably good measures of what happened, but not why it happened or what forces caused it to happen.

6. **Communities or Advisory Panel Data.** Many large companies have bought into systems that allow them to frequently talk to and survey a small group of target consumers over and over again.

Surveys among this group are conducted by various folks in the corporation on a daily or weekly basis. The cost per survey or measurement is relatively low—if the quality of outcomes is not taken into account. Such communities are not truly representative, not randomly

chosen, and seldom ever validated. Over time conditioning and learning undermine the representativeness of the community, assuming it existed at the outset.

7. *Eye-Tracking Data.* With steady improvements in measurement equipment and software, the direction the human eye is pointing can be determined with a high degree of accuracy—less than one degree of error in a controlled environment with high-quality equipment. This can provide useful diagnostic information to help understand why a package, website, or advertisement is failing to attract attention or failing to register certain messages or images.

8. *Biometric or Physiological Measurements.* Galvanic skin response, eye pupil dilation, heart rate, EEG (brainwave) measurements, facial emotions recognition, etc., are very interesting and exciting, and they may one day open portals into the human soul, but for the present these measures are largely speculative and unproven. Many anecdotes are cited as proof, but real experimental proof is lacking. Physiological measures became very popular during the 1970s, but they faded over time as validation of backers' claims and assertions proved elusive. Some of

these measures are reasonably good at tracking arousal, but there's no precise way to know if the arousal is positive or negative without bringing in survey or qualitative research.

9. *Social-Media Data.* Social-media data are very popular in corporate America. The data are comparatively inexpensive, often massive, and real-time (day by day, hour by hour). Many new software tools and systems make analyses of the data relatively easy.

Social-media data are, perhaps, most valuable as an early-warning system—of something going wrong, of a competitive initiative, or of an unexpected aberration. Social-media data, however, must always be viewed with suspicion and skepticism, for several reasons:

Physiological measures became very popular during the 1970s, but they faded over time as validation of backers' claims and assertions proved elusive.

Social-media data are, perhaps, most valuable as an early-warning system—of something going wrong, of a competitive initiative, or of an unexpected aberration.

- Many product categories and brands are scarcely ever mentioned in social media, making sample sizes too small for data reliability.

- Social-media comments are influenced by the news cycle, special events, media advertising, promotions, publicity, movies, competitive activity, and television shows (i.e., there is a lot of noise in the data).

- Social-media data are subject to manipulation. You may think you are following an important trend in the data, only to learn later it was a clever ruse to confuse by a competitor. Increasingly, corporations and other organizations are striving to create social-media content and manage social- media comments, so the "research'' value of the data is rapidly diminishing.

- As social-media comments are identified and collected via web scraping, we almost never know the exact source, the context, the stimulus, or the history that underlie a comment. These unknowns make interpretation risky, indeed. That's why social-media data must be viewed with trepid spirit and jaundiced eye.

10. *Your Spouse's Opinions (After the Honeymoon).* Your spouse will often tell you the unvarnished truth (and you should kiss her or him for this act of honesty and bravery, not call the divorce lawyer). However, your spouse tends to share some of your own biases, can have hidden agendas at times, and is not likely to be representative of your firm's target market. Listen politely, nod approval and acceptance, but don't put too much weight on your spouse's opinions.

11. *Your Own Personal Opinions.* Rarely is your personal opinion trustworthy, despite the fact that emotionally you trust your own inner feelings and intuitions over all external data. Your experiences, genetics, hopes, and biases cloud perceptions of reality. You are rarely objective and are often driven by intense emotions. You are a member of the human race.

Objectivity

Social and political pressures within a large organization tend to undermine the objectivity required for determining the trustworthiness of business data and accurately determining its meaning and import. The trend toward low-cost, do-it-yourself surveys by the inexperienced in corporate America is creating a growing

number of bad decisions and big mistakes, based upon survey data of questionable objectivity, accuracy, and validity.

The normal controls and safeguards practiced by independent research companies are rarely followed by do-it-yourself research functions in larger companies (and these internal groups often use copyrighted questionnaires "borrowed" from research companies—which expose large companies to unknown legal risks). The objectivity of professional, independent, experienced third parties can greatly improve the quality of data and its interpretation. *(Confession: The author might benefit financially if corporate America acted on the assertions in this paragraph).*

Little Data

Corporate decision-makers often would be better served if they relied on tried-and-true tools and systems from the world of little data, rather than illusions from big data. Sampling theory teaches that if the sample is random, one can measure the behavior or mood of the whole by talking to very few people.

A sample of 1,500 is sufficient to predict who will win a presidential election. A sample of 200 to 300 respondents is generally sufficient to predict how much the whole

population will like a new product or service. A sample of 200 users can test a new peanut butter in-home for a week, and from this it can be precisely determined if the product is optimal and what its market share will be once introduced. These are examples of little data.

Survey research is relatively inexpensive, yet very accurate, because professional researchers know the source, stimulus, context, and history—and have tried-and-true measuring instruments, normative data, quality assurance, and controls.

Marketing research can be designed to be forward-looking and predictive, rather than backward-looking. Experienced researchers

Social and political pressures within a large organization tend to undermine the objectivity required for determining the trustworthiness of business data and accurately determining its meaning and import.

can create alternative futures and measure the relative appeal of the differing visions of the future. These professional researchers can predict the sales volume of new products within narrow tolerances, based on survey research. They can optimize the formulation of a new product via product testing. They can accurately predict the effectiveness of new commercials long before they air. They can measure the size and composition of an industry or category with amazing precision, based solely on scientific sampling and surveys.

They can use qualitative research methods (ethnography, depth interviews, focus groups, and online forums) to discover unmet needs and hidden dreams that can become templates for new product development (long before a trend in anyone's tracking data suggests an opportunity). These researchers can optimize products, services, ads, and store designs to maximize sales, all based on choice-modeling experiments within surveys.

All of this research is based on little data. The data are derived from random sampling, carefully controlled experiments, and/or scientific surveys. The sample and sampling error are known; the stimulus is known; the questions are known; the context is understood; and the meaning of the answers is known. Despite the marketing hoopla and gurus touting big data, little data often provides a more accurate basis for sound corporate decision-making.

About the Author

Jerry W. Thomas is the President/CEO of Decision Analyst. The author may be reached by email at **jthomas@decisionanalyst.com** or by phone at *1-800-262-5974* or *1-817-640-6166*.

Decision Analyst is a global marketing research and analytical consulting firm. The company specializes in advertising testing, strategy research, new products research, and advanced modeling for marketing-decision optimization.

Decision Analyst

strategic research ∎ analytics ∎ modeling ∎ optimization

604 Avenue H East ∎ Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800. ANALYSIS ∎ www.decisionanalyst.com