The New York Times | http://nyti.ms/1kgErs2

THE OPINION PAGES | OP-ED CONTRIBUTORS | NYT NOW

Eight (No, Nine!) Problems With Big Data

By GARY MARCUS and ERNEST DAVIS APRIL 6, 2014

BIG data is suddenly everywhere. Everyone seems to be collecting it, analyzing it, making money from it and celebrating (or fearing) its powers. Whether we're talking about analyzing zillions of Google search queries to predict flu outbreaks, or zillions of phone records to detect signs of terrorist activity, or zillions of airline stats to find the best time to buy plane tickets, big data is on the case. By combining the power of modern computing with the plentiful data of the digital era, it promises to solve virtually any problem — crime, public health, the evolution of grammar, the perils of dating — just by crunching the numbers.

Or so its champions allege. “In the next two decades,” the journalist Patrick Tucker writes in the latest big data manifesto, “The Naked Future,” “we will be able to predict huge areas of the future with far greater accuracy than ever before in human history, including events long thought to be beyond the realm of human inference.” Statistical correlations have never sounded so good.

Is big data really all it's cracked up to be? There is no doubt that big data is a valuable tool that has already had a critical impact in certain areas. For instance, almost every successful artificial intelligence computer program in the last 20 years, from Google's search engine to the I.B.M. “Jeopardy!” champion Watson, has involved the substantial crunching of large bodies of data. But precisely because of its newfound popularity and growing use, we need to be levelheaded about what big data can — and can't — do.

The first thing to note is that although big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, it never tells us which correlations are meaningful. A big data analysis might reveal, for instance, that from 2006 to 2011 the United States murder rate was well correlated with the market share of Internet Explorer: Both went down sharply. But it’s hard to imagine there is any causal relationship between the two. Likewise, from 1998 to 2007 the number of new cases of autism diagnosed was extremely well correlated with sales of organic food (both went up sharply), but identifying the correlation won’t by itself tell us whether diet has anything to do with autism.

Second, big data can work well as an adjunct to scientific inquiry but rarely succeeds as a wholesale replacement. Molecular biologists, for example, would very much like to be able to infer the three-dimensional structure of proteins from their underlying DNA sequence, and scientists working on the problem use big data as one tool among many. But no scientist thinks you can solve this problem by crunching data alone, no matter how powerful the statistical analysis; you will always need to start with an analysis that relies on an understanding of physics and biochemistry.

Third, many tools that are based on big data can be easily gamed. For example, big data programs for grading student essays often rely on measures like sentence length and word sophistication, which are found to correlate well with the scores given by human graders. But once students figure out how such a program works, they start writing long sentences and using obscure words, rather than learning how to actually formulate and write clear, coherent text. Even Google’s celebrated search engine, rightly seen as a big data success story, is not immune to “Google bombing” and “spamdexing,” wily techniques for artificially elevating website search placement.

Fourth, even when the results of a big data analysis aren’t intentionally gamed, they often turn out to be less robust than they initially seem. Consider Google Flu Trends, once the poster child for big data. In 2009, Google reported — to considerable fanfare — that by analyzing flu-related search queries, it had been able to detect the spread of the flu as accurately and more quickly than the Centers for Disease Control and Prevention. A few years later, though, Google Flu Trends began to falter; for the last two years it has made more bad predictions than good ones.

As a recent article in the journal Science explained, one major contributing cause of the failures of Google Flu Trends may have been that the Google search engine itself constantly changes, such that patterns in data collected at one time do not necessarily apply to data collected at another time. As the statistician Kaiser Fung has noted, collections of big data that rely on web hits often merge data that was collected in different ways and with different purposes — sometimes to ill effect. It can be risky to draw conclusions from data sets of this kind.

A fifth concern might be called the echo-chamber effect, which also stems from the fact that much of big data comes from the web. Whenever the source of information for a big data analysis is itself a product of big data, opportunities for vicious cycles abound. Consider translation programs like Google Translate, which draw on many pairs of parallel texts from different languages — for example, the same Wikipedia entry in two different languages — to discern the patterns of translation between those languages. This is a perfectly reasonable strategy, except for the fact that with some of the less common languages, many of the Wikipedia articles themselves may have been written using Google Translate. In those cases, any initial errors in Google Translate infect Wikipedia, which is fed back into Google Translate, reinforcing the error.

A sixth worry is the risk of too many correlations. If you look 100 times for correlations between two variables, you risk finding, purely by chance, about five bogus correlations that appear statistically significant — even though there is no actual meaningful connection between the variables. Absent careful supervision, the magnitudes of big data can greatly amplify such errors.

Seventh, big data is prone to giving scientific-sounding solutions to hopelessly imprecise questions. In the past few months, for instance, there have been two separate attempts to rank people in terms of their “historical importance” or “cultural contributions,” based on data drawn from Wikipedia. One is the book “Who’s Bigger? Where Historical Figures Really Rank,” by the computer scientist Steven Skiena and the engineer Charles Ward. The other is an M.I.T. Media Lab project called Pantheon.

Both efforts get many things right — Jesus, Lincoln and Shakespeare were surely important people — but both also make some egregious errors. “Who’s Bigger?” claims that Francis Scott Key was the 19th most important poet in history; Pantheon has claimed that Nostradamus was the 20th most important writer in history, well ahead of Jane Austen (78th) and George Eliot (380th). Worse, both projects suggest a misleading degree of scientific precision with evaluations that are inherently vague, or even meaningless. Big data can reduce anything to a single number, but you shouldn’t be fooled by the appearance of exactitude.

FINALLY, big data is at its best when analyzing things that are extremely common, but often falls short when analyzing things that are less common. For instance, programs that use big data to deal with text, such as search engines and translation programs, often rely heavily on something called trigrams: sequences of three words in a row (like “in a row”). Reliable statistical information can be compiled about common trigrams, precisely because they appear frequently. But no existing body of data will ever be large enough to include all the trigrams that people might use, because of the continuing inventiveness of language.

To select an example more or less at random, a book review that the actor Rob Lowe recently wrote for this newspaper contained nine trigrams such as “dumbed-down escapist fare” that had never before appeared anywhere in all the petabytes of text indexed by Google. To witness the limitations that big data can have with novelty, Google-translate “dumbed-down escapist fare” into German and then back into English: out comes the incoherent “scaled-flight fare.” That is a long way from what Mr. Lowe intended — and from big data’s aspirations for translation.

Wait, we almost forgot one last problem: the hype. Champions of big data promote it as a revolutionary advance. But even the examples that people give of the successes of big data, like Google Flu Trends, though useful, are small potatoes in the larger scheme of things. They are far less important than the great innovations of the 19th and 20th centuries, like antibiotics, automobiles and the airplane.

Big data is here to stay, as it should be. But let’s be realistic: It’s an important resource for anyone analyzing data, not a silver bullet.

Gary Marcus is a professor of psychology at New York University and an editor of the forthcoming book “The Future of the Brain.” Ernest Davis is a professor of computer science at New York University.

A version of this op-ed appears in print on April 7, 2014, on page A23 of the New York edition with the headline: Eight (No, Nine!) Problems With Big Data.