Big data: are we making a big mistake?

Economist, journalist and broadcaster **Tim Harford** delivered the 2014 *Significance* lecture at the Royal Statistical Society International Conference. In this article, republished from the *Financial Times*, Harford warns us not to forget the statistical lessons of the past as we rush to embrace the big data future

Five years ago, a team of researchers from Google announced a remarkable achievement in one of the world's top scientific journals, *Nature*. Without needing the results of a single medical check-up, they were nevertheless able to track the spread of influenza across the US. What's more, they could do it more quickly than the Centers for Disease Control and Prevention (CDC). Google's tracking had only a day's delay, compared with the week or more it took for the CDC to assemble a picture based on reports from doctors' surgeries. Google was faster because it was tracking the outbreak by finding a correlation between what people searched for online and whether they had flu symptoms.

Not only was "Google Flu Trends" quick, accurate and cheap, it was theory-free. Google's engineers didn't bother to develop a hypothesis about what search terms – "flu symptoms" or "pharmacies near me" – might be correlated with the spread of the disease itself. The Google team just took their top 50 million search terms and let the algorithms do the work.

The success of Google Flu Trends became emblematic of the hot new trend in business, technology and science: "Big Data". What, excited journalists asked, can science learn from Google?

As with so many buzzwords, "big data" is a vague term, often thrown around by people with something to sell. Some emphasise the sheer scale of the data sets that now exist – the Large Hadron Collider's computers, for example, store 15 petabytes a year of data, equivalent to about 15,000 years' worth of your favourite music.

But the "big data" that interests many companies is what we might call "found data", the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast. Google Flu Trends was built on found data and it's this sort of data that interests me here. Such data sets can be even bigger than the LHC data – Facebook's is – but just as noteworthy is the fact that they are cheap to collect relative to their size, they are a messy collage of data points collected for disparate purposes and they can be updated in real time. As our communication, leisure and commerce have moved to the internet and the internet has moved into our phones, our cars and even our glasses, life can be recorded and quantified in a way that would have been hard to imagine just a decade ago.

Cheerleaders for big data have made four exciting claims, each one reflected in the success of Google Flu Trends: that data analysis produces uncannily accurate results; that every single data point can be captured, making old statistical sampling techniques obsolete; that it is passé to fret about what

causes what, because statistical correlation tells us what we need to know; and that scientific or statistical models aren't needed because, to quote "The End of Theory", a provocative essay published in *Wired* in 2008, "with enough data, the numbers speak for themselves".

Unfortunately, these four articles of faith are at best optimistic oversimplifications. At worst, according to David Spiegelhalter, Winton Professor of the Public Understanding of Risk at Cambridge University, they can be "complete bollocks. Absolute nonsense."

The data exhaust

Found data underpin the new internet economy as companies such as Google, Facebook and Amazon seek new ways to understand our lives through our data exhaust. Since Edward Snowden's leaks about the scale and scope of US electronic surveillance it has become apparent that security services are just as fascinated with what they might learn from our data exhaust, too.

Consultants urge the data-naive to wise up to the potential of big data. A recent report from the McKinsey Global Institute reckoned that the US healthcare system could save $300bn a year – $1,000 per American – through better integration and analysis of the data produced by everything from clinical trials to health insurance transactions to smart running shoes.

But while big data promise much to scientists, entrepreneurs and governments, they are doomed to disappoint us if we ignore some very familiar statistical lessons. "There are a lot of small data problems that occur in big data," says Spiegelhalter. "They don't disappear because you've got lots of the stuff. They get worse."

Four years after the original *Nature* paper was published, *Nature News* had sad tidings to convey: the latest flu outbreak had claimed an unexpected victim: Google Flu Trends. After reliably providing a swift and accurate account of flu outbreaks for several winters, the theory-free, data-rich model had lost its nose for where flu was going. Google's model pointed to a severe outbreak but when the slow-and-steady data from the CDC arrived, they showed that Google's estimates of the spread of flu-like illnesses were overstated by almost a factor of two.

The problem was that Google did not know – could not begin to know – what linked the search terms with the spread of flu. Google's engineers weren't trying to figure out what caused what. They were merely finding statistical patterns in the data. They cared about correlation rather than causation. This is common in big data analysis. Figuring out what causes what is hard (impossible, some say). Figuring out what is correlated with what is much cheaper and easier. That is why, according to Viktor Mayer-Schönberger and Kenneth Cukier's book, *Big Data*, "causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning".

But a theory-free analysis of mere correlations is inevitably fragile. If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down. One explanation of the Flu Trends failure is that the news was full of scary stories about flu in December 2012 and that these stories provoked internet searches

by people who were healthy. Another possible explanation is that Google's own search algorithm moved the goalposts when it began automatically suggesting diagnoses when people entered medical symptoms.

Google Flu Trends will bounce back, recalibrated with fresh data – and rightly so. There are many reasons to be excited about the broader opportunities offered to us by the ease with which we can gather and analyse vast data sets. But unless we learn the lessons of this episode, we will find ourselves repeating it.

Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. The data are bigger, faster and cheaper these days – but we must not pretend that the traps have all been made safe. They have not.

In 1936, the Republican Alfred Landon stood for election against President Franklin Delano Roosevelt. The respected magazine, *The Literary Digest*, shouldered the responsibility of forecasting the result. It conducted a postal opinion poll of astonishing ambition, with the aim of reaching 10 million people, a quarter of the electorate. The deluge of mailed-in replies can hardly be imagined but the *Digest* seemed to be relishing the scale of the task. In late August it reported, "Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totalled."

After tabulating an astonishing 2.4 million returns as they flowed in over two months, *The Literary Digest* announced its conclusions: Landon would win by a convincing 55 per cent to 41 per cent, with a few voters favouring a third candidate.

The election delivered a very different result: Roosevelt crushed Landon by 61 per cent to 37 per cent. To add to *The Literary Digest's* agony, a far smaller survey conducted by the opinion poll pioneer George Gallup came much closer to the final vote, forecasting a comfortable victory for Roosevelt. Mr Gallup understood something that *The Literary Digest* did not. When it comes to data, size isn't everything.

Opinion polls are based on samples of the voting population at large. This means that opinion pollsters need to deal with two issues: sample error and sample bias. Sample error reflects the risk that, purely by chance, a randomly chosen sample of opinions does not reflect the true views of the population. The "margin of error" reported in opinion polls reflects this risk and the larger the sample, the smaller the margin of error. A thousand interviews is a large enough sample for many purposes and Mr Gallup is reported to have conducted 3,000 interviews.

But if 3,000 interviews were good, why weren't 2.4 million far better? The answer is that sampling error has a far more dangerous friend: sampling bias. Sampling error is when a randomly chosen sample doesn't reflect the underlying population purely by chance; sampling bias is when the sample isn't randomly chosen at all. George Gallup took pains to find an unbiased sample because he knew that was far more important than finding a big one.

Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. We must not pretend that the traps have all been made safe

The Literary Digest, in its quest for a bigger data set, fumbled the question of a biased sample. It mailed out forms to people on a list it had compiled from automobile registrations and telephone directories – a sample that, at least in 1936, was disproportionately prosperous. To compound the problem, Landon supporters turned out to be more likely to mail back their answers. The combination of those two biases was enough to doom *The Literary Digest*'s poll. For each person George Gallup's pollsters interviewed, *The Literary Digest* received 800 responses. All that gave them for their pains was a very precise estimate of the wrong answer.

History repeating?

The big data craze threatens to be *The Literary Digest* all over again. Because found data sets are so messy, it can be hard to figure out what biases lurk inside them – and because they are so large, some analysts seem to have decided the sampling problem isn't worth worrying about. It is.

Professor Viktor Mayer-Schönberger of Oxford's Internet Institute, co-author of *Big Data*, told me that his favoured definition of a big data set is one where "*N* = All" – where we no longer have to sample, but we have the entire background population. Returning officers do not estimate an election result with a representative tally: they count the votes – all the votes. And when "*N* = All" there is indeed no issue of sampling bias because the sample includes everyone.

But is "*N* = All" really a good description of most of the found data sets we are considering? Probably not. "I would challenge the notion that one could ever have all the data," says Patrick Wolfe, a computer scientist and professor of statistics at University College London.

An example is Twitter. It is in principle possible to record and analyse every message on Twitter and use it to draw conclusions about the public mood. (In practice, most researchers use a subset of that vast "fire hose" of data.) But while we can look at all the tweets, Twitter users are not representative of the population as a whole. (According to the Pew Research Internet Project, in 2013, US-based Twitter users were disproportionately young, urban or suburban, and black.)

There must always be a question about who and what is missing, especially with a messy pile of found data. Kaiser Fung, a data analyst and author of *Numbersense*, warns against simply assuming we have everything that matters. "*N* = All is often an assumption rather than a fact about the data," he says.

Consider Boston's Street Bump smartphone app, which uses a phone's accelerometer to detect potholes without the need for city workers to patrol the streets. As citizens of Boston download the app and drive around, their phones automatically notify City Hall of the need to repair the road surface. Solving the technical challenges involved has produced, rather beautifully, an informative data exhaust that addresses a problem in a way that would have been inconceivable a few years ago. The City of Boston proudly proclaims that the "data provides the City with real-time information it uses to fix problems and plan long term investments."

Yet what Street Bump really produces, left to its own devices, is a map of potholes

that systematically favours young, affluent areas where more people own smartphones. Street Bump offers us "N = All" in the sense that every bump from every enabled phone can be recorded. That is not the same thing as recording every pothole. As Microsoft researcher Kate Crawford points out, found data contain systematic biases and it takes careful thought to spot and correct for those biases. Big data sets can seem comprehensive but the "N = All" is often a seductive illusion.

Who cares about causation or sampling bias, though, when there is money to be made? Corporations around the world must be salivating as they contemplate the uncanny success of the US discount department store Target, as famously reported by Charles Duhigg in *The New York Times* in 2012. Duhigg explained that Target has collected so much data on its customers, and is so skilled at analysing that data, that its insight into consumers can seem like magic.

Duhigg's killer anecdote was of the man who stormed into a Target near Minneapolis and complained to the manager that the company was sending coupons for baby clothes and maternity wear to his teenage daughter. The manager apologised profusely and later called to apologise again – only to be told that the teenager was indeed pregnant. Her father hadn't realised. Target, after analysing her purchases of unscented wipes and magnesium supplements, had.

Found data contain systematic biases and it takes careful thought to spot and correct for those biases. "*N* = All" is often a seductive illusion

Statistical sorcery? There is a more mundane explanation. "There's a huge false positive issue," says Kaiser Fung, who has spent years developing similar approaches for retailers and advertisers. What Fung means is that we didn't get to hear the countless stories about all the women who received coupons for babywear but who weren't pregnant.

Hearing the anecdote, it's easy to assume that Target's algorithms are infallible – that everybody receiving coupons for onesies and wet wipes is pregnant. This is vanishingly unlikely. Indeed, it could be that pregnant women receive such offers merely because everybody on Target's mailing list receives such offers. We should not buy the idea that Target employs mind-readers before considering how many misses attend each hit.

In Charles Duhigg's account, Target mixes in random offers, such as coupons for wine glasses, because pregnant customers would feel spooked if they realised how intimately the company's computers understood them.

Fung has another explanation: Target mixes up its offers not because it would be weird to send an all-baby coupon-book to a woman who was pregnant but because the company knows that many of those coupon books will be sent to women who aren't pregnant after all.

None of this suggests that such data analysis is worthless: it may be highly profitable. Even a modest increase in the accuracy of targeted special offers would be a prize worth winning. But profitability should not be conflated with omniscience.

The multiple-comparisons problem

In 2005, John Ioannidis, an epidemiologist, published a research paper with the self-explanatory title, "Why Most Published Research Findings Are False". The paper became famous as a provocative diagnosis of a serious issue. One of the key ideas behind Ioannidis's work is what statisticians call the "multiple-comparisons problem".

It is routine, when examining a pattern in data, to ask whether such a pattern might have emerged by chance. If it is unlikely that the observed pattern could have emerged at random, we call that pattern "statistically significant".

The multiple-comparisons problem arises when a researcher looks at many possible patterns. Consider a randomised trial in which vitamins are given to some primary schoolchildren and placebos are given to others. Do the vitamins work? That all depends on what we mean by "work". The researchers could look at the children's height, weight, prevalence of tooth decay, classroom behaviour, test scores, even (after waiting) prison record or earnings at the age of 25. Then there are combinations to check: do the vitamins have an effect on the poorer kids, the richer kids, the boys, the girls? Test enough different correlations and fluke results will drown out the real discoveries.

There are various ways to deal with this but the problem is more serious in large data sets, because there are vastly more possible comparisons than there are data points to compare. Without careful analysis, the ratio of genuine patterns to spurious patterns – of signal to noise – quickly tends to zero.

Worse still, one of the antidotes to the multiple-comparisons problem is transparency, allowing other researchers to figure out how many hypotheses were tested and how many contrary results are languishing in desk drawers because they just didn't seem interesting enough to publish. Yet found data sets are rarely transparent. Amazon and Google, Facebook and Twitter, Target and Tesco – these companies aren't about to share their data with you or anyone else.

How can statisticians rise to the big data challenge?

At the conclusion of his 2014 *Significance* lecture, Tim Harford was asked for his view on what statisticians need to do to help users of data avoid falling into the big data traps.

"One of the things we have to do is demonstrate examples where mistakes have been made, and explain how, with the appropriate statistical tools, preparation, wisdom and insight, those mistakes would not have been made," he said.

Proving the value of statistics would also come from interdisciplinary working; from statisticians "teaming up with computer scientists, astronomers, the bioinformatics people – anybody else who is working with these large data sets – and showing them that statistics has a tremendous amount to offer".

He concluded: "Statistics has never been cooler; it's never been more useful. It just seems to me to be a wonderful time to be a statistician."

Brian Tarran

New, large, cheap data sets and powerful analytical tools will pay dividends – nobody doubts that. And there are a few cases in which analysis of very large data sets has worked miracles. David Spiegelhalter of Cambridge points to Google Translate, which operates by statistically analysing hundreds of millions of documents that have been translated by humans and looking for patterns it can copy. This is an example of what computer scientists call "machine learning", and it can deliver astonishing results with no preprogrammed grammatical rules. Google Translate is as close to a theory-free, data-driven algorithmic black box as we have – and it is, says Spiegelhalter, "an amazing achievement". That achievement is built on the clever processing of enormous data sets.

But big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on, and figuring out how we might intervene to change a system for the better.

"We have a new resource here," says Professor David Hand of Imperial College London. "But nobody wants 'data'. What they want are the answers."

To use big data to produce such answers will require large strides in statistical methods.

"It's the wild west right now," says Patrick Wolfe of UCL. "People who are clever and driven will twist and turn and use every tool to get sense out of these data sets, and that's cool. But we're flying a little bit blind at the moment."

Big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on

Statisticians are scrambling to develop new methods to seize the opportunity of big data. Such new methods are essential but they will work by building on the old statistical lessons, not by ignoring them.

Recall big data's four articles of faith. Uncanny accuracy is easy to overrate if we simply ignore false positives, as with Target's pregnancy predictor. The claim that causation has been "knocked off its pedestal" is fine if we are making predictions in a stable environment but not if the world is changing (as with Flu Trends) or if we ourselves hope to change it. The promise that "N = All", and therefore that sampling bias does not matter, is simply not true in most cases that count. As for the idea that "with enough data, the numbers speak for themselves" – that seems hopelessly naive in data sets where spurious patterns vastly outnumber genuine discoveries.

"Big data" has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers – without making the same old statistical mistakes on a grander scale than ever.