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 datapoints 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: 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.”

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.”

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. Unless we learn the lessons of this episode, we will find ourselves repeating it. 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.

Mr Gallup understood something that The Literary Digest did not. When it comes to data, size isn’t everything. 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. 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. Gallup took pains to find an unbiased sample because he knew that was far more important than finding a big one. 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.

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.

when “N = All” there is indeed no issue of sampling bias because the sample includes everyone. 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.

There must always be a question about who and what is missing, especially with a messy pile of found data.

*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?

“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.

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. 🡪 ei olekaan varmaa, eivät viitsi lähettää niin radikaalisti.

Profitability should not be conflated with omniscience(=kaikkitietävyys).

“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.

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. 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.

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 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. Nobody wants ‘data’. What they want are the answers. To use big data to produce such answers will require large strides in statistical methods. “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.”

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.

**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.