The Tyranny of Spreadsheets

timharford.com/2021/07/the-tyranny-of-spreadsheets

21 July 2021

21st July, 2021

Early last October my phone rang. On the line was a researcher calling from Today, the BBC's agenda-setting morning radio programme. She told me that something strange had happened, and she hoped I might be able to explain it. Nearly 16,000 positive Covid cases had disappeared completely from the UK's contact tracing system. These were 16,000 people who should have been warned they were infected and a danger to others, 16,000 cases contact tracers should have been running down to figure out where the infected went, who they met and who else might be at risk. None of which was happening.

Why had the cases disappeared? Apparently, Microsoft Excel had run out of numbers.

It was an astonishing story that would, in time, lead me to delve into the history of accountancy, epidemiology and vaccination, discuss file formatting with Microsoft's founder, Bill Gates, and even trace the aftershocks of the collapse of Enron. But above all, it was a story that would teach me about the way we take numbers for granted.

Now, as the UK tentatively reopens against a background of rapidly rising cases, we are hoping that vaccinations will keep us safe. The vaccines have — rightly — been trumpeted as a scientific triumph. Their development and rollout have taken place on a heroic scale.

But back in September and October, when the UK was also reopening against a strikingly similar backdrop of rising cases, we had no vaccine to protect us. Instead, we were trying to defend ourselves with data. And we didn't seem to be nearly as enamoured of data as we now are of vaccines. That is a shame, because when you're relying on numbers to keep you safe, it's important to put some effort into keeping your numbers straight.

*

The origin of Excel can be traced back far further than that of Microsoft. In the late 1300s, the need for a solid system for accounts was evident in the outbursts of one man in particular, an Italian textile merchant named Francesco di Marco Datini. Poor Datini was surrounded by fools.

"You cannot see a crow in a bowlful of milk!" he berated one associate. "You could lose your way from your nose to your mouth!" he chided another.

Iris Origo's vivid book The Merchant of Prato describes Datini's everyday life and explains his problem: keeping track of everything in a complicated world. By the end of the 14th century, merchants such as Datini had progressed from mere travelling salesmen able to keep track of profits by patting their purses. They were now in charge of sophisticated operations.

Datini, for example, ordered wool from the island of Mallorca two years before the sheep had even grown it, a hedge to account for the numerous subcontractors that would process it before it became beautiful rolls of dyed cloth. The supply chain between shepherd and consumer stretched across Barcelona, Pisa, Venice, Valencia, North Africa and back to Mallorca. It took four years between the initial order of wool and the final sale of cloth. No wonder Datini insisted on absolute clarity about where his product was at any moment, not to mention his money.

How did he manage? Spreadsheets. Datini, of course, did not use Excel back in 1396. But he did use its direct predecessor: sheets of paper laid out according to the system of double-entry bookkeeping, otherwise known as bookkeeping alla veneziana. In double-entry bookkeeping, every entry is made twice. (The clue's in the name.) For example, if you spend 100 florins on wool, that is recorded as a credit of 100 florins in your cash account and a debit of 100 florins worth of wool in your assets account. This extra effort of booking everything twice makes it much easier to detect mistakes. If one has been made, the books won't balance.

Double-entry bookkeeping became an essential method for keeping track of who owed what to whom, foreign exchange transactions, profits, losses, everything. It helped Datini and merchants like him ensure nothing was lost, no matter how incompetent their associates.

A century later, the master of double-entry booking was Luca Pacioli. He was a serious mathematician and a friend of Leonardo da Vinci. But he's best known today as the most famous accountant who ever lived. He literally wrote the book on the double-entry method, back in 1494. Pacioli once advised, "If you cannot be a good accountant, you will grope your way forward like a blind man and may meet great losses."

We don't have to accept Pacioli's insensitive simile to understand his point: life is easier when you can see the obstacles and opportunities around you. Good accounts show us clearly what would otherwise be invisible. But if you can't keep your spreadsheets straight, you may meet great losses. (More on that shortly.)

Nearly five hundred years later, in 1978, a student named Dan Bricklin sat in a classroom at Harvard Business School. As he watched his accounting professor filling in rows and columns on the blackboard, an idea dawned on him. Each time the professor made a change he would have to work across and down the grid, erasing and rewriting other numbers to make everything add up. Bricklin knew that this erasing and rewriting was happening every day, millions of times a day, all over the world, as accounting clerks adjusted the entries in what they called spreadsheets: big sheets of paper spread across two pages of an accounting ledger.

Bricklin was a geek and former programmer who immediately thought, “I can do this on a computer.” As Steven Levy described in a classic mid-1980s feature in Harper’s, the rest was history. Bricklin and a friend called their spreadsheet program VisiCalc. It went on sale on October 17 1979. It was a smash hit soon followed by Lotus 1-2-3 and then, in due course, by Excel.

For accountants, digital spreadsheets were revolutionary, replacing hours of painstaking work with a few taps on a keyboard. But some things didn’t change. Accountants still had their professional training and their double-entry system. The rest of us did not, but that did not prevent Excel from becoming ubiquitous. It was, after all, easily accessible and flexible, a tool like a Swiss Army knife for numbers, sitting in your digital back pocket. Any idiot could use it. And goodness, we did.

*

Nobody really knows what happened to the 16,000 positive Covid cases that disappeared from the spreadsheet. Public Health England (PHE), a government agency responsible for the process, still hasn’t published anything very informative on the issue.

“The suggestion that any cases were ‘lost’ is simply incorrect,” they told me. “No cases were missed. There was a delay in referring cases for contact tracing and reporting them in the national figures.”

That delay was typically four or five days, long enough to render the test result almost useless. If I mislaid my passport just before a holiday and then found it after five days staying at home instead, I am not sure I would triumphantly wave it in the air and declare, “The suggestion that my passport was ‘lost’ is simply incorrect.”

For a contact-tracing system, lost for five days is lost. The question is, how were they lost? Somewhere in PHE’s data pipeline, someone had used the wrong Excel file format, XLS rather than the more recent XLSX. And XLS spreadsheets simply don’t have that many rows: 2 to the power of 16, about 64,000. This meant that during some automated process, cases had vanished off the bottom of the spreadsheet, and nobody had noticed.

Everyone could see the funny side of the mishap. The idea of simply running out of space to put the numbers was darkly amusing. The fact that Microsoft was never anyone’s idea of cool simply added to the absurdity. Clippy, the maligned automated assistant from Office 2000, began making the rounds as a meme: “It looks like you’re trying to track a global pandemic. Would you like help?”

A few weeks after the data-loss scandal, I found myself able to ask Bill Gates himself about what had happened. Gates no longer runs Microsoft, and I was interviewing him about vaccines for a BBC programme called How to Vaccinate The World. But the opportunity to have a bit of fun quizzing him about XLS and XLSX was too good to pass up.

I expressed the question in the nerdiest way possible, and Gates's response was so strait-laced I had to smile: "I guess… they overran the 64,000 limit, which is not there in the new format, so…" Well, indeed. Gates then added, "It's good to have people double-check things, and I'm sorry that happened."

Exactly how the outdated XLS format came to be used is unclear. PHE sent me an explanation, but it was rather vague. I didn't understand it, so I showed it to some members of Eusprig, the European Spreadsheet Risks Group. They spend their lives analysing what happens when spreadsheets go rogue. They're my kind of people. But they didn't understand what PHE had told me, either. It was all a little light on detail.

They agreed that the basic problem was that whatever PHE had done wrong, it didn't have the right checks and controls to flag problems. Or as Gates put it, "It's good to have people double-check things."

*

The original paper spreadsheets were designed to help us not lose our way, and one might naturally imagine the digital spreadsheet is not only faster but more accurate. Is it? One clue comes from a wonderful study conducted by Felienne Hermans, a computer scientist. A few years ago, Hermans realised that there was a bountiful source of spreadsheets she could study. That source was Enron, the bankrupt energy company.

After Enron collapsed in 2001 amid an epic accounting scandal, regulators extracted a cache of half a million emails from the company's servers. Those emails are now publicly available and have been studied by researchers trying to understand everything from the evolution of informal written language to the way people use email folders. Hermans was interested in what was attached to some of these emails: spreadsheets.

She started digging through them, not looking for fraud, but for spreadsheets with obvious errors such as missing or circular references. Looking at nearly 10,000 spreadsheets with calculations in them, she found that a quarter had at least one such error. The errors even seemed to multiply. If a spreadsheet had any mistakes at all, on average it contained more than 750.

How can a spreadsheet acquire so many errors? Matt Parker, the author of Humble Pi, a book about mathematical mishaps and their consequences, notes that Excel's own functionality combined with the mistaken assumptions of users will often introduce mistakes.

Type an international phone number into Excel, for example, and the program strips off the leading zeroes, which are redundant in a mathematical integer but not in a phone number. If instead you type in a twenty digit serial number, Excel will decide those 20 digits are a huge quantity and round them off, turning the last few digits into zeroes.

Or say you're a genetics researcher typing in the name of a gene such as "Membrane Associated Ring-CH-Type Finger 1", or March1 for short, or perhaps the Sept1 gene. You can imagine what Excel does next. It turns those gene names into dates. One study estimated that 20 per cent of all genetics papers had errors caused by Excel's autocorrect.

Microsoft's defence is simple enough: the default settings are intended to work in everyday scenarios. Which is the polite way of saying: Guys, Excel wasn't designed for genetics researchers. It was designed for accountants.

But it's understandable that scientists picked up Excel and started to use it. It's powerful, it's flexible. It's ubiquitous. It may not be the right tool, but it's the tool that's right there.

When used by a trained accountant to carry out double-entry bookkeeping, a long-established system with inbuilt error detection, Excel is a perfectly professional tool. But when pressed into service by genetics researchers or contact tracers, it's like using your Swiss Army Knife to fit a kitchen because it's the tool you have closest at hand. Not impossible, but hardly advisable.

And yet when the genetics research community were wrestling with the autocorrecting genes issue, they resigned themselves to the hard truth that they would never wean people off Excel. Instead, the folks in charge — the Hugo Gene Nomenclature Committee — decided to change the names of the genes in question.

The decision is understandable. But it also neatly illustrates the contortions we go through as a result of treating data as an afterthought, just something to slap together on a spreadsheet. That is a shame, because history suggests that well-managed information can be transformative.

*

A few months ago, I asked folks on Twitter if they could recommend some good books about the eradication of smallpox. Most people recommended books about Edward Jenner, who in 1796 was the first to demonstrate an effective smallpox vaccine. That's revealing, because I'd asked about the eradication of smallpox, and smallpox wasn't eradicated in 1796. Not even close.

While eradication would have been impossible without a highly effective vaccine, it also required the highly effective use of information. Or as Datini might have said, it required not losing your way from your nose to your mouth.

Ever since the vaccine for smallpox was demonstrated in 1796, people dreamed of eradicating the disease. But those dreams kept failing to come true. In trying to vaccinate the entire planet, over and over again, the vaccinators never managed to reach quite enough people. In poorer countries, smallpox lingered in rural areas or neglected communities. A generation of babies were born without any immunity and, soon enough, the disease returned.

In the mid-1960s, smallpox was still killing two million people a year. The World Health Organization announced it would redouble its efforts to eradicate the disease and planned to do so by intensifying the mass vaccination campaign. One of those leading these efforts was Bill Foege, an Iowan-born epidemiologist who knew smallpox so well he could detect cases by smell. (Lesion-blistered skin has a distinctive odour.)

Foege would show up in a village in eastern Nigeria, all six foot seven of him, and the elders would put out the word, Come and see the tallest man in the world! And people did. Foege reckons he once vaccinated 11,600 people in a single day. It wasn't enough to quash periodic outbreaks.

Then, late in 1966, Foege received a radio message warning of an outbreak of smallpox in a village about a hundred miles away. He travelled there, found five cases and vaccinated everyone they'd been in contact with. (The smallpox vaccine can still work even if it is given a day or more after people have been exposed to the virus.)

Standard practice then would be to vaccinate everyone for miles around. But Foege's team just didn't have enough doses. Instead, he used radio and the local network of missionaries to spot new cases. Every evening at seven o'clock, they'd switch on the radio and put the word out. Whenever an outbreak was reported, Foege and his team quickly raced to the scene and administered vaccines.

The hope was to create something like a firebreak, keeping the disease from spreading. And it worked. Using this tactic, Foege's team eliminated smallpox from eastern Nigeria within six months. It was 1967, and soon civil war engulfed the country. Despite the chaos and enormous bloodshed of that war, smallpox did not return.

The secret was to worry less about the blanket coverage that was never quite good enough and to worry more about quickly finding exactly where each outbreak had appeared. Eradication was all about information. Up until that point, information had been very patchy. The WHO realised it had been finding only 100,000 or so cases each year against a backdrop of 10 million.

Foege's experience showed that public health workers could beat smallpox if they had the data. The strategy became known as ring vaccination. It's not the same as contact tracing, but it has a lot in common: in both cases you need to rapidly isolate infected people and find their recent contacts.

Ring vaccination worked. In less than a decade, doctors were scrambling to get to an outbreak in India so that they could observe a case of smallpox before the virus went extinct. The last gasp of smallpox in the wild was in Somalia, late in 1977. Ali Maow Maalin, 23 years old, a cook and part-time vaccinator, astonishingly, had not been vaccinated. He developed smallpox symptoms, was vaccinated — along with 91 friends and contacts — and recovered. Maalin devoted his life to the eradication of polio.

The vaccines were important. Essential, in fact. But so was quickly identifying and tracing contacts at risk. Smallpox had survived nearly two centuries of vaccinations — but it could not survive a well-run system that targeted outbreaks and tracked potential cases.

With hindsight, it seems simple. In a way, it was. But of course, keeping track of things is harder than it might first appear. Francesco di Marco Datini could have told you that.

*

One of the striking lessons of the pandemic has been how powerful data can be when handled well — and how much damage is done when the data are fumbled. Almost every question we have asked about this virus requires the deft use of statistics to answer it. Who has it? How does it spread? Who is most at risk? How can we treat it? Without a flow of good data and reliable ways to analyse that data, we haven't a hope of answering such questions.

This is not just a case of having the right boffins solve the right equations. Data do not grow on trees: they must be assembled. An example of this process done right is Recovery (Randomised Evaluations of Covid-19 Therapy). Recovery is a system for running simple but powerful randomised trials of different Covid therapies as an integrated part of the regular treatment of hospital patients with Covid, all over the UK. It was set up at the start of the pandemic in a matter of days by two Oxford academics, Peter Horby and Martin Landray.

Recovery has produced a steady stream of vital findings, notably that the antimalarial drug hydroxychloroquine does not help and the cheap steroid dexamethasone is a lifesaver. (How many lives it has saved is unclear, but it's surely well over a million by now.) It is an example of what can be done when we take seriously not only the data, but the "data infrastructure", the tools and the processes we have to collect, manage and analyse that data.

It is hard to think of a clearer contrast with the misfiring contact-tracing systems in many supposedly sophisticated western democracies. Nature reported late last year that Australia, Washington state and Hawaii were still using phones or faxes to share information about new cases, and that public health professionals from Africa were aghast at the failure of the US system to learn the hard-won lessons of the Ebola outbreak.

There is more to running a good contact-tracing system than data infrastructure. But without good data the task is all but impossible. As with smallpox, success begins with rapidly figuring out where the virus is — and therefore, where it might go next.

Nor has the vaccine rendered contact tracing obsolete. Most people still aren't vaccinated, and some people never will be. One day there will be another pandemic, and another, and another. We can't guarantee vaccines will work every time, and vaccines take time to develop. While we wait, there will always be contact tracing. And good contact tracing, like thousands of other good things we want to achieve, requires investing in serious data infrastructure.

*

Let's say you really want proof that contact tracing works, how would you get it?

Let's also say you're a mad scientist, crazed with power and unhindered by conventional ethics. You'd probably hack into the country's contact-tracing system, then you'd delete some of the positive cases, making sure that some regions lost a lot of cases and some lost very few. This nefarious experiment would allow you to compare what happened in the places where the contact-tracing system was still running smoothly with the places where thousands of cases had gone missing.

If you weren't an evil genius, of course, you wouldn't dream of doing such a thing. Instead, you'd keep an eye out for it happening by accident because somebody bungled the formatting of Excel spreadsheets. Two economists, Thiemo Fetzer and Thomas Graeber, did just that. They decided that no catastrophe should be allowed to occur without trying to learn some lessons. They combed through the evidence from Public Health England's mishap. And by comparing the experiences of different regions, they concluded that the error had led to 125,000 additional infections.

The story about Excel running out of numbers just seemed so bizarre at first. That's why we were sharing Clippy memes, and why I took pleasure in teasing Gates about it. But his response, which seemed po-faced at the time, was right. He wasn't laughing, because he understood that this wasn't a comedy; it was a tragedy.

Fetzer and Graeber have calculated a conservative estimate of the number of people who died, unknown victims of the spreadsheet error. They think the death toll is at least 1,500 people.

So the next time there's a pandemic, let's make sure we have our spreadsheets in order. After all, as Luca Pacioli, the father of accounting, warned us more than five hundred years ago, without a good spreadsheet, you will grope your way forward, "and you may meet great losses".

One thousand five hundred deaths. Relative to the scale of the whole pandemic, this is just a sliver of the total tragedy. But as the needless price of bad data management, they are great losses indeed.

This essay is adapted from an episode of my podcast, "Cautionary Tales". It was published in the Financial Times on 24 June 2021.

The paperback of "How To Make The World Add Up" is now out. US title: "The Data Detective".

"One of the most wonderful collections of stories that I have read in a long time… fascinating."- Steve Levitt (Freakonomics)

“If you aren’t in love with stats before reading this book, you will be by the time you’re done.”- Caroline Criado Perez (Invisible Women)

I’ve set up a storefront on Bookshop in the United States and the United Kingdom – have a look and see all my recommendations; Bookshop is set up to support local independent retailers.