HAS THE INTERNET CHANGED SCIENCE?

The ocean of information available on the web is challenging the standard model of hypothesis-driven science. Yet that model has always borne little relation to the mucky reality of scientific research

ELIZABETH PISANI

It's a glorious Saturday morning, a day for kayaking or sitting in a sunny courtyard with a coffee and the silly bits of the *FT*. But here I am, stuck in a bunker at the British Library with a bunch of Generation Y geeks telling me that my nice, tidy world of hypothesis, experiment and knowledge generation is about to end.

I am attending the Science Online conference, in which the usual scientific lexicon of sample size calculations, placebo-controlled trials and statistical significance tests is nowhere to be seen. The talk is of scraping and mining, terabytes and petabytes, of algorithms. It's the language of Big Data—the ocean of information being generated by ever-larger telescopes, ever-cheaper genetic sequencing techniques and ever more Facebook users. As Royal Society president Martin Rees has written (*Prospect*, November 2010), Big Data will allow us to mine and mash our way to unexpected discoveries and insights. It allows us to ask new questions, ones that we couldn't have asked when science depended on the work of a few people in a single lab working in a limited area of knowledge with just a few gigabytes of processing power. Some people say that Big Data also changes the *way* that we ask questions. Gone are the days of hypothesis-driven science as we know it. Nowadays, it's all about pattern recognition.

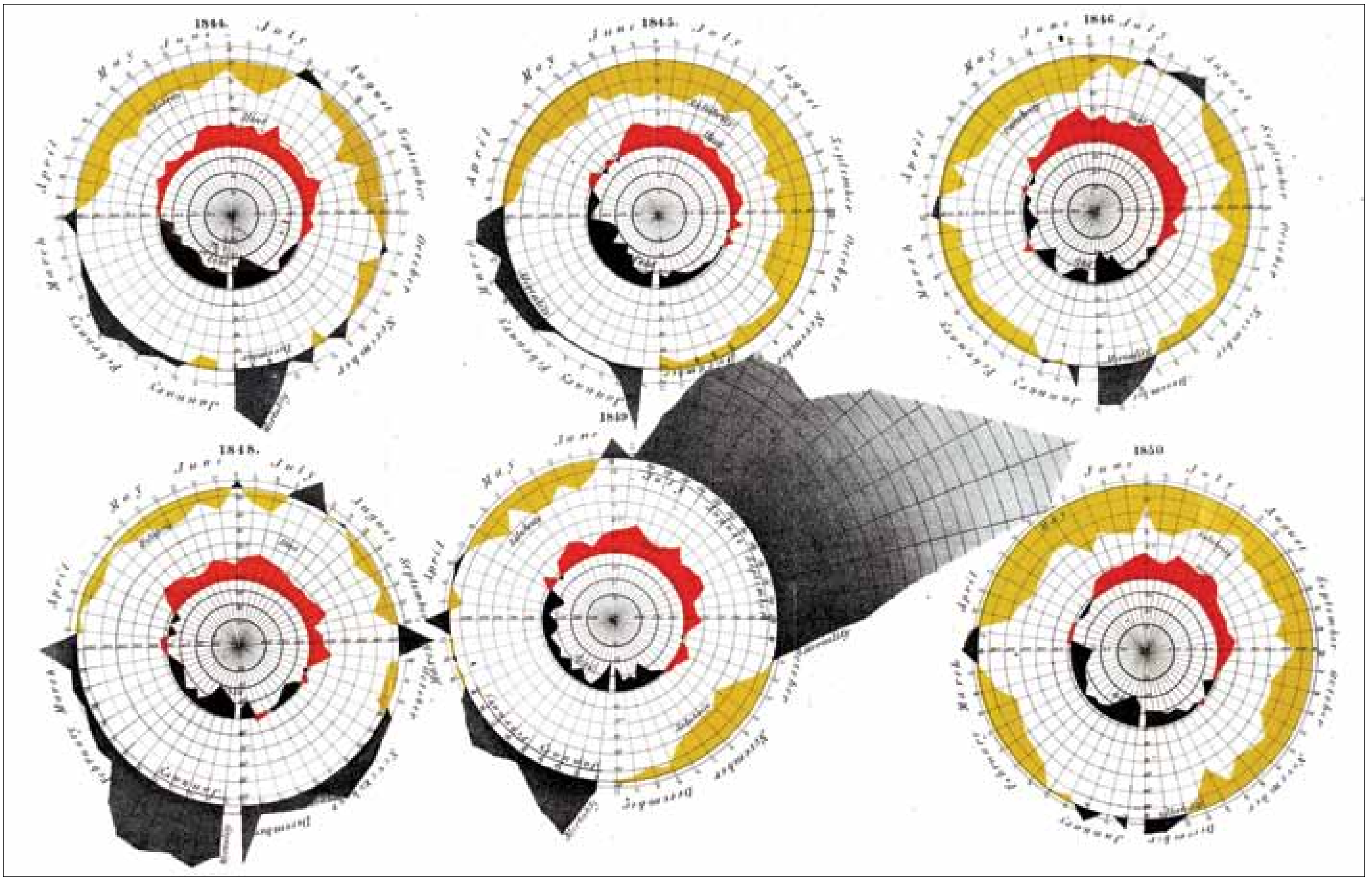
David McCandless, a mildly geeky writer and designer who runs the blog Information is Beautiful, is making a presentation to the Science Online attendees. He displays a graph that runs from January to December. The line bumps along for the first few months until the largish, double-humped peak in the late spring and early summer. It drops off in the autumn and hits another sharp peak just before Christmas. He challenges the audience to guess what the graph shows. Chocolate sales, perhaps? Greeting cards? With a flourish, he adds the headline to the slide: "Peak Break-Up Times."

Relationships melt down because of the stress of spending time together over the holidays, McCandless theorises, and the tension of meeting families. The data is gleaned from "scraping" over 10,000 random Facebook status updates for the phrases "break up" or "broken up." His obvious excitement at this result is shared by an appreciative audience.

Then a woman behind me sticks up her hand. She's of an age to be a slave to the school run. "Couldn't it be that the phrases are not to do with relationships, but with the end of terms?" she asks him. "Breaking up for the Easter hols next week. Broke up for Christmas last Tuesday." There's a brief pause; McCandless deflates. Some of his attendees are probably thinking: that's what you get for doing hypothesis-free science. I'm thinking: if we "scraped" just the changes in the relationship status box, we could remove that possibility.

When *Wired* magazine declared in 2008 that the petabyte age would sweep away scientific method, the comments board flamed with indignant rebuttals. The topic has been smouldering ever since. Why are the old guard so threatened by the

Elizabeth Pisani is a writer, journalist and epidemiologist



© WELLCOME LIBRARY, LONDON

Detail from a data map showing deaths from cholera in London in the 1840s—an early example of "big data" collection

idea of science-by-algorithm? Epidemiologists like me collect information about disease outbreaks, risk behaviours and the environment and use it to find threats to public health. It is legitimate to worry that computers will throw up spurious associations and send us down the wrong track. But I suspect our real fear is that Google might do our jobs better than us.

Take flu. The US Centres for Disease Control and Prevention (CDC) tracks flu trends through reports from health service providers. The sooner it knows there's an outbreak, the more quickly it can move to contain it with vaccination and prevention campaigns. CDC epidemiologists start analysing the data as soon as it comes in, but it takes two or three weeks to churn out a meaningful analysis. A couple of years ago, dataheads at Google decided to see if they could do better. They filtered 50m searches to determine which ones correlated best with flu outbreaks of the previous five years. They found 45 common search queries relating to flu, its symptoms and treatment, which taken together correlated closely with outbreaks. And since Google monitors trends in real time, its tracking beats the CDC by at least a fortnight.

The epidemiological world was at first sceptical, making much of the fact that searches about Oscar winners and high school basketball also correlated well with outbreaks, simply because the Oscars and basketball coincide with the flu season. Without a clear hypothesis based on a well-defined theoretical model, some argue, all this pattern recognition amounts to throwing data at the wall and seeing what sticks.

Roni Zeiger, a medic and chief health strategist at Google, disagrees. "With Flu Trends, we didn't throw data at the wall. What happened was that two thoughtful engineers saw how CDC did things and wondered whether we could get the same results more efficiently," he said. "We had a very specific hypothesis." He's right, of course. Except that most health researchers live in a world dominated by the fascism of the randomised controlled trial. In this rarefied world, the idea that we can use the Google searches of people worried about having flu to track the virus doesn't count as a hypothesis.

Before we book the musicians to wail at the wake of hypothesis-driven science, we might peer a little more closely at it. A cursory glance at the history of quantitative science (and the huge body of qualitative research) suggests that many great discoveries had different origins. I talk to Simon Schaffer, a professor of the history of science at Cambridge and he senses my existential angst. "Your questions betray a certain nostalgia for the experimental ideal: very small groups of very clever men dream up clever predictions. Then, because they've read Karl Popper, they send out younger, poorer men to collect data to try and disprove their predictions. But that never ever happened."

I don't admit that *Objective Knowledge* by Popper, the Austro-British philosopher of science, graces my bedside table. It has been there for a while; the high priest of hypothesis-driven science wrote a dull book. I've found it hard to see ▶

why his Idea-Experiment-Analyse-Refute/Confirm-New Idea model became the dominant narrative of science.

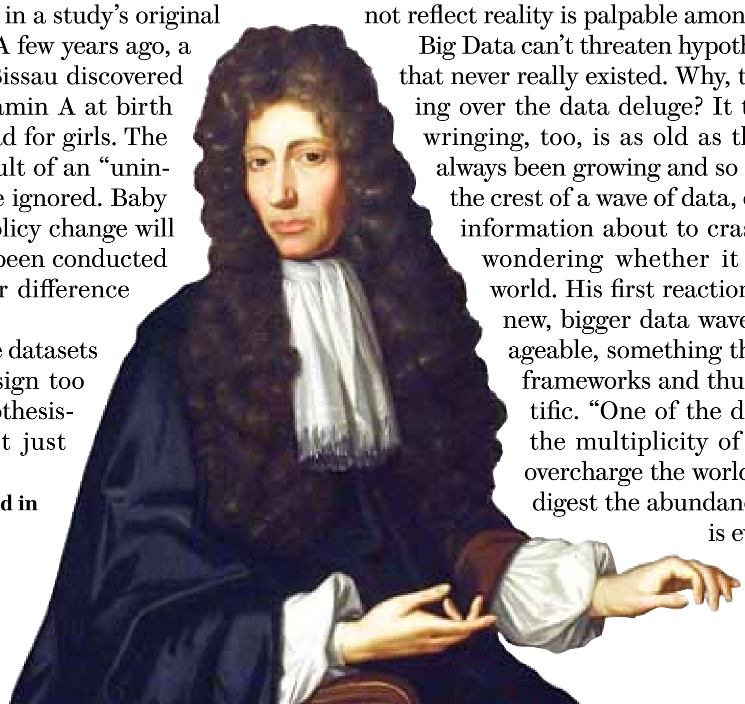
"If a proposition is scientific, it ought to be verifiable. That was Karl Popper's normative account and it's very convincing, but it completely fails to describe what people like Darwin, Pasteur, Newton or Boyle actually did. They did much more ducking and weaving, manipulating and selecting," said Schaffer. Dominic Kwiatkowski, an Oxford-based geneticist who studies the correlations between genes and disease, agrees that Popper's model doesn't reflect reality. "A hypothesis doesn't come out of the head of some chap sitting in a leather armchair with a whisky. It comes out of existing data." So what appears to be hypothesis-free data in an initial, exploratory trial can be constructed into a hypothesis for another, narrower trial, he explains. Chickens and eggs.

Scientists themselves have done little to disabuse the public of the view that they have thought-bubble moments of brilliance which they then toil to confirm. That's in part because the myth is tidier than the truth. "We retro-fit that idea of hypothesis-driven science in part because scientists are too embarrassed to admit that they were stumbling around in the data and stubbed their toe on a finding," said Chris Hilton, senior archivist at the Wellcome Library, which specialises in the history of science. In the biomedical sciences, where we worship at the altar of the randomised controlled trial, the supremacy of the hypothesis is written into our codes of conduct; you are forbidden not to have one. When bright-eyed epidemiology students ask me about "fishing" (our more organic term for data mining), I have to tell them it is *streng verboten* to trawl through their data until they net some association that will be statistically significant and thus give them a "result." We protect against this wickedness by requiring researchers to tell us what questions they will be answering before they have enrolled a single person in a clinical trial.

There was a good reason for this requirement: it was supposed to prevent Big Pharma making a success story out of every study by restricting the analysis to whichever subset gave them the result they wanted: those patients who had a high white blood-cell count before the trial, or those recruited on the second Tuesday after a full moon. But the result is that important findings that aren't in a study's original hypothesis are easily disregarded. A few years ago, a Danish team working in Guinea Bissau discovered that a new fashion for giving Vitamin A at birth appears to be good for boys and bad for girls. The findings were dismissed as the result of an "unintended experiment" and thus to be ignored. Baby girls may die as a result, but no policy change will be recommended until a trial has been conducted on the specific question of gender difference and Vitamin A supplements.

"Fishing" is only a problem if the datasets are too small or the sampling design too weak to support the results. "Hypothesis-free doesn't mean rigour-free. It just means that you don't have to state your bets beforehand," says Kwiatkowski.

If there is one thing that has contributed more than anything else to entrenching the myth of hypothesis-driven science, it is cash



Scientific pioneer Robert Boyle believed in experimentation for its own sake

The supremacy of the peer-reviewed paper as the currency of science further entrenches the standard approach. Laboratory life is messy—experiments in human populations even more so. What goes on record, though, are not the lab books and study diaries that tell of the spilled samples, minor explosions and police raids that we've all worked around. What goes on record is the peer-reviewed paper. Introduction, Hypothesis, Methods, Findings, Discussion; the well-tailored garment that covers the unwashed reality of science.

But if there is one thing that has contributed more than anything else to entrenching the myth of hypothesis-driven science, it is cash. And, perhaps, the second world war. It was then that we first saw teams of scientists get together in taxpayer-funded labs, collaborating on giant research projects that had a clear, predetermined goal—most famously the Manhattan Project at the Los Alamos laboratory in New Mexico, which gave us the atomic bomb. The war created an infrastructure for Big Science as well as an ideological commitment to fund it; it also entrenched the idea that even large teams could work towards answering a well-defined research question. "The Popperian model is basically a science funding model, and it blossomed at the same time as public science," said Joe Cain, a historian and philosopher of biology at University College London. "Funding became available, and what's easy to fund is a nice, tidy five-page testable hypothesis."

Kwiatkowski suggests that this pushes people into unnecessarily narrow spaces. "Developing a hypothesis is a way of fabricating a justification for an experiment, because if you say at the start that you don't know what you'll find, you get marked down [by funding committees]." The frustration of brilliant people trying to fit their work into a model that does not reflect reality is palpable among many scientists.

Big Data can't threaten hypothesis-driven science if that never really existed. Why, then, the hand-wringing over the data deluge? It turns out that hand-wringing, too, is as old as the hills. Science has always been growing and so every scientist sits on the crest of a wave of data, eyeing the tsunami of information about to crash down on him and wondering whether it will wash away his world. His first reaction is to pooh-pooh the new, bigger data wave as chaotic, unmanageable, something that we cannot fit into frameworks and thus inherently unscientific. "One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world," thundered Barnaby Rich in 1613. He himself contributed 26 books to the multiplicity and eventually gave his name to the Barnaby Rich effect: "a high output of scientific writings accompanied by complaints on the excessive productivity of other authors."

Scientists have not been good at turning their instruments on themselves. It was not until the early 1960s that Derek de Solla Price measured scientific output, showing it had been increasing exponentially for over 300 years. As long as that continued, he wrote in 1962 in *Little Science: Big Science*, it will always be true that most of what is known has been determined in living memory. Science is "always exploding, always on the brink of its expansive revolution. Scientists have always felt themselves to be awash in a sea of scientific literature that augments in each decade as much as in all times before."

Simon Schaffer agrees. "Historians say it was ever thus," he said. "but really, it *has* always been like this in meteorology, in epidemiology, in many of the sciences." He points to the work of John Graunt in the 17th century as the first example of Big Data. Graunt collected mortality rolls and other parish records and, in effect, threw them at the wall, looking for patterns in births, deaths, weather and commerce. "Having observed that most [parishes], tho constantly took in the weekly bills of mortality, made little other use of them... I thought, that some other uses might be made of them," he wrote in 1676. He scraped parish rolls for insights in the same way as today's data miners transmute the dross of our Twitter feeds into gold for marketing departments. Graunt made observations on everything from polygamy to traffic congestion in London, concluding: "That the old Streets are unfit for the present frequency of Coaches... That the opinions of Plagues accompanying the Entrance of Kings, is false and seditious; That London, the Metropolis of England, is perhaps a Head too big for the Body, and possibly too strong."

Graunt and his colleague William Petty used their results in the service of the state. In one of his "Essays on Political Arithmetick," Petty took death rates collected for another purpose, stirred them with a couple of wild assumptions on population, and seasoned them with a dash of prejudice to conclude that British hospitals were much less likely to kill their patients than French ones, where "Half the said numbers did not die by natural necessity but by the evil administration of the hospital." In a precursor to the World Bank's habit of pricing productivity lost by ill-health, Petty goes on to calculate the cost of the unnecessary deaths, valuing the French at £60 each, "being about the value of Ariger Slaves (which is less than the intrinsik value of People at Paris)."

At around the same time as Graunt, Isaac Newton winkled information on tides out of records kept by ships engaged in the slave trade. More than 300 years ahead of today's data miners, he mashed the shipping data together with information from astronomical and meteorological data sets to give us the *Philosophiæ Naturalis Principia Mathematica*. "Absent the slave trade, you don't have the *Principia*," notes Schaffer. No scrapable data sets, no foundation for modern physics.

The tradition of trawling for data continued through the Victorian age, most literally in the Challenger expedition of the 1870s. *HMS Challenger* sailed the world's oceans dredging its way to the discovery of some 4,700 species of marine life. Though the expedition's aim was to collect new information, there were no preconceptions about what that information might be. "It was a matter of: keep dropping the nets and let's see what we come up with. Then take it home and organise it," said Joe Cain. The expedition took three and a half years, the analysis a further 19. Cain draws parallels with today's Big Science. "You collect petabytes of data in an instant, and then it takes ten years to work out what it all means." The bigger the data sets, the more we rely on algorithms (a series of mathematical steps which generate a solution to a problem) to do the grunt work of the analysis. But algorithms can only propose; it is still up to the scientist to interpret.

SEE SCIENCE & TECHNOLOGY ON PAGE 64 »
Sonia Livingstone on how digital media is affecting family life

Four decades ago, Derek De Solla Price predicted an upheaval as the tectonic plates of Big Science crashed against the Popperian model of scientific life. To be a brilliant scientist, you have to be a bit of a maverick. "The scientist tends to be the man who, in doing the word-association test, responds to "black" not with "white" but with "caviar," he wrote. A misty-eyed view, perhaps. But he was probably right that this is not the sort of person who easily fits into a Big Science project that runs on tramlines towards a predetermined goal.

His words remind me once more of the 17th century, and a dispute between Thomas Hobbes and Robert Boyle about what constituted acceptable knowledge. Boyle was a practical ▶

man, part of a clique (later formalised as the Royal Society) who favoured experiment for its own sake, just to see what might happen. Hobbes believed that experimentation was at best entertaining and at worst grubby; knowledge arrived at by philosophising was inherently better than that arrived at by experimenting. Both men embody traits that are rather British. On the one hand, the tendency to favour intellectual work over manual. On the other, the desire, among those who could afford it, to become curious generalists.

Many would say that in the Popper years, at least, Hobbes has been in the ascendancy. We reward the thought bubble at the expense of the grunt on the lab bench. Cain points to the example of Rosalind Franklin, who had an important but still disputed role in the discovery of the double helix structure of DNA. “She was a numbers nerd,” says Cain. “Without her you don’t have a dataset to work on, to speculate and build theories about. But it’s the speculators and theorists who win Nobel prizes.” (Franklin’s untimely death prevented her being considered for a Nobel in any case.) And in this example, as in so many others, James Watson and Francis Crick’s prizewinning thought bubble came after the data, not before it as Popper’s model suggests, just as the Google flu model is built on five years of sneezing and subsequent mouse-clicks.

The petabyte age has forced funders to rethink how they invest in science. “Big Data isn’t new. But the current interest in it might help us to reorganise the image of what scientists do. It makes trouble for the idea that science begins north of the eyebrows,” says Schaffer. The Wellcome Trust, one of the world’s largest charitable funders of health research, recently announced a new funding model which does not require recipients to design a study around a specific hypothesis. And speculative research just produced a Nobel prize in physics for two scientists at the Manchester University, who used sticky tape to prise a layer of the super-conductor graphene out of the graphite found in pencils. That, in turn, led Royal Society president Martin Rees, whose five-year term in the job ends in December, to lobby for more funding of “‘open-ended’ research projects.

A big advantage of Big Data research is that algorithms, scraping, mining and mashing are usually low cost, once you’ve paid the nerds’ salaries. And the data itself is often droppings produced by an existing activity. “You may as well just let the boffins go at it. They’re not going to hurt anyone, and they may just come up with something useful,” said Cain.

What’s more, people have to go through the apprenticeship of writing the scrape-and-crunch algorithms if they are to develop the skills we really need; the ability to look critically at other people’s models and come up with alternative explanations. There’s no danger in throwing up the “relationship meltdown” graphs, but it’s worth having someone posit the “school holiday” alternative if you want to avoid spending every Christmas alone. “What’s wrong with Big Data is what’s always wrong with induction. You can’t know anything about the evidence you’ve got. You need judgement,” said Schaffer.

One of the useful things the data miners may come up with are hypotheses for others to test in a more deductive way. And that has implications for how we reward scientists.

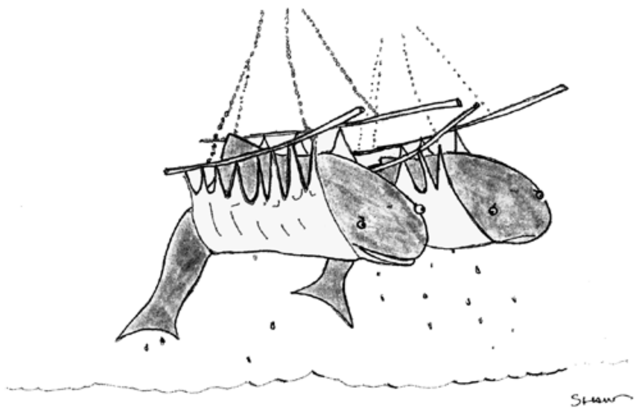
> **It was not until the early 1960s that scientific output was measured, and found to have been increasing exponentially for over 300 years**

“If an individual, through novel data analysis, generates 100 hypotheses and ten turn out to be confirmed by other scientists, then certainly they should be rewarded,” said Google’s Roni Zeiger. “I think there ought to be an evolution to rewarding people for impact.”

It’s a good idea, and one that opens the cupboard door on one of the scientific world’s older skeletons. We still measure impact and dole out funding on the basis of papers published in peer-reviewed journals. It’s a system which works well for thought-bubble experiments but is ill-suited to the Big Data world. We need new ways of sorting the wheat from the chaff, and of rewarding collaborative, speculative science. Until things settle down, we will see a lot of chaff. “It’s common sense that there are discoveries to be made as huge data sets come together. But those discoveries will be the product of a highly rigorous process,” says Kwiatkowski. He notes that such rigour is not yet common: “The beginning of any revolution generates a bunch of crap. At the end of the 1990s, people said the internet would never be good for anything except selling cosmetics. Now we’ve got Google. Hypothesis-free doesn’t mean model-free. You need a model, you need interpretation, even where you don’t have a hypothesis.”

In his appeal for rigour, Kwiatkowski sounds not unlike the father of Big Data. Writing in 1665, John Graunt reminded himself not to throw data at the wall without having some kind of model which would lead to useful interpretation. “Finding some truths and not commonly believed opinions to arise from my meditations on these neglected papers, I proceeded further, to consider what benefit the knowledge of the same would bring the world; that I might not engage myself in idle and useless speculations.”

It was, indeed, ever thus. P



“Now will you ask for directions?”