12

Harnessing big data: A tsunami of transformation

Philip Evans

Let me start, if I may, with a story by Jorge Luis Borges (1954), the Argentinian poet and novelist. *On Exactitude in Science* is about an ancient lost kingdom, obsessed with cartography; the aristocrats cast one map after another of progressing levels of ambition until they launched the ultimate mapping project: to create the map of their kingdom on a scale of one-to-one. In Borges's story, the fragments of this failed effort can be found rotting in the corners of this lost empire.

And that is the end of the story! One paragraph long and classic Borges: full of metaphysics, and a meditation on the futility of human ambition. What I want to suggest to you in this chapter is that the image of Borges's map—a map on a scale of one-to-one, a map that is the same size as the reality that it represents—is the image we should have in our heads when we think about where technology is taking us.

Let me share with you some examples.

First, consider the Google self-driving car. It is aware of roads, traffic lanes, signals, and it is aware of other traffic. It is even aware of pedestrians and cyclists. In fact, in its first million miles the Google car only had two accidents, one of which was when somebody rammed it from the

rear while parked and the other when it was being driven manually by a Google engineer. This is not wildly futuristic technology: within just a few years, these cars will be on the market.

But that's not the only way that the world could become self-aware and self-mapping. Consider a car park that is aware of the vacancy of parking spaces. Each parking space is equipped with a small sensor, which costs about $25, and is powered by a battery that lasts about five years. Using low-powered radios, these sensors form a mesh network allowing the municipality to collect parking fines when somebody outstays their welcome.

The extensions of this technology are obvious. Using near-field communication technology, the car can communicate with the sensor in the parking space, so that drivers can pay their parking fee automatically. Moreover, since the owner of the car park or the municipality then knows the exact availability of spaces, they can broadcast to the world a universal map of parking. In some major cities, 40 per cent of the traffic is made up of people driving around in circles looking for parking spaces. When the location of empty parking spaces is universally visible, drivers can book them and owners can even auction them to the highest bidder. We could massively improve the efficiency with which these spaces are allocated and reduce traffic congestion.

And of course, obviously, when the Google car drives you to work, it can drop you off and go find a parking space for itself. The key point here is that in this future world, the physical distribution of cars in spaces on asphalt and the electronic rendering of what is going on, become coincident. Traffic and parking become their own map.

Consider a third example. Nature's map of humans is the chemical structure of DNA. It took something in the order of 10 years and $150 million to first map one human genome. In the intervening years, the cost of mapping the human genome has come down with extraordinary speed. Quite soon, we will be able to map the human genome for less than $100.

Back when mapping the human genome cost $150 million, it was a very expensive exercise in 'big science'. And by treating just one person's genome as representative of all humanity, they abstracted from the human variation that is the essence of medicine. But when genomic mapping takes 20 minutes, and costs $99 while you wait, it is no longer a matter of

abstract research—it is a matter of clinical medicine. Consequently, when you go to the doctor in the near future, the first thing they will do, if they haven't already, will be to map your genome.

The essence of medical practice in the future will become *statistical*, relating genomic data with other medical data: your medical record, symptoms, and ambient data from the environment. What used to be a process based solely on expertise, conducted within hierarchical organisations, where the final judgement was made by somebody with many years of expensive technical training, will become much more of a *statistical* exercise.

Already, Watson, the software technology originally developed by IBM to win the *Jeopardy* television contest in the United States is being applied to the task of medical diagnosis. Using statistical techniques, Watson is able to combine all of these data sets, including genomic data, and outperform 98 per cent of human practitioners in performing some radiological diagnoses. This doesn't render the doctor obsolete, but it fundamentally transforms how we need to think about medicine. In particular, our ability to aggregate, to standardise, to anonymise and to protect large data sets becomes crucial to our ability to use statistical methods in order to address these kinds of problems.

This is the challenge. It's a challenge in terms of the computer science and the mathematics, but it's also a challenge in terms of the institutions. In countries like the United States, where medicine is privatised, each hospital and each clinic thinks of medical data as proprietary. It's a 'switching cost'; it's a source of what they would call 'competitive advantage'. But preserving that data-based competitive advantage is of limited compatibility with large-scale data-mining of genomic and clinical information. As a result, we have a fundamental conflict emerging between the direction that the technology is taking us and many of the institutional arrangements that, in the public sector as well as the private, stand in the way. This is destined to become an enormous challenge over the next 10 years.

Consider my last example of how data can serve as infrastructure. Readers would be familiar with satellite maps of the United States at night, where you can see cities and roads lit up in the darkness. If you do the same kind of map for Africa, tragically it is the Dark Continent. The lack of infrastructure is one of the things holding back economic development in Africa. A few years ago in the Ivory Coast, Orange, the French telecommunications

company and monopoly provider of cell phone services in that country, launched a project where they collected metadata on cellular phone usage for a high fraction of the population over a nine-month period. What resulted is an immense data set: who talks to whom, how people move around. Orange then carefully anonymised the data and published it (Palchykov et al. 2014), encouraging researchers simply to see what they could find.

Over 80 research papers were written with the benefit of this extraordinary data set. Data on how people move around, for example, shed light on the spread of infectious diseases. Warning people to wash their hands or boil their drinking water is among the most important methods to combat the spread of infections. But it has long been known that word of mouth is the most effective way to spread such messages. So the Orange data sets revealed not only the network over which infection spreads (people's movements) but also the communication network through which countervailing propaganda can be disseminated.

But some researchers at IBM in Dublin realised that this same data set showed the commuter patterns in cities. They took the largest city in the Ivory Coast, Abidjan, and extracted the daily movements of people from their home to their workplace and back. They then asked themselves the question: what is the optimal design of a bus system, given daily commuting patterns?

Mathematically, it's a straightforward optimisation problem, but in computational terms, it's very difficult because of the size of the data sets. To manage this, the researchers set up a Hadoop cluster, linking a network's computers to work in parallel. After some days of computation, they arrived at a solution. It turned out it was possible to reduce by 10 per cent the average commuting times in Abidjan without adding a single bus. All because of the availability of heretofore invisible data.

In a country like the Ivory Coast, data is serving as infrastructure. This data, about how people use cell phones in a society where other kinds of infrastructure are largely absent, turns out to be a source on the basis of which all sorts of insights can be gleaned.

We have a new paradigm here. Traditionally, data has been the by-product of linear processes, used close to where it originates to make local improvements to the process. But now data can be aggregated over very large (possibly universal) scale, and we can optimise globally rather than locally.

Used this way, data becomes a universal enabler: general-purpose, large-scale, high fixed-cost, zero variable-cost. Like roads or telecommunications, it becomes *infrastructure*. Data as infrastructure is open-ended and its uses are unknowable before the fact. It makes possible experimentation, innovation and technical improvements in things like bus routing that could not have been anticipated.

What's the larger pattern? It is, I think, the interaction of four very large trends. The first is what people call the internet of things, the proliferation of sensors. For example, the aforementioned $25 devices implanted in the parking spaces or the sensors used in the Google car. Current estimates are that by the year 2030, there'll be something of the order of 100 trillion sensors in the world. There are already in the world today 140 sensors for every man, woman and child. As the cost of sensing falls and the volume proliferates, every device knows and reports on its own status.

Second, all that data accumulates. This results in an extraordinary growth in the world's stock of information, which is doubling every two years—the phenomenon of 'big data'.

But data is useless without insight. The third trend is breakthroughs in artificial intelligence or 'sense-making'. Machines learn not from explicit hand-crafted models, but by brute force from immense data sets. Correlation substitutes for causation.

The fourth big trend is mobility. The number of cell phones in the world is now roughly equal to the number of people. And because an increasing fraction of those are smart phones, they are themselves sensors feeding data into the network. They are thus a huge source of data. But they are a principal means by which insight can be consumed. You used to have to go somewhere to get insight—to a 'library', for example—now, you can do your search from your phone or even your watch. Insight is delivered at exactly the point where it is needed.

Together these four trends are what make the world self-aware and self-describing. They are really recent and they are mutually multiplying. They are driving a tsunami of transformation.

How do we organise to exploit these technologies, whether in the private sector or the public? We see hints by looking at a company that is native to this world: Google. Google's search system has a stacked and layered architecture. When you make a query, it is passed through 'layers'

of servers. It gets broken into its component pieces and finally draws answers from so-called index servers, each of which contains lists of every instance on the web of particular words (together with measures of the centrality of the source in the network of hyperlinks). These references are recombined, aggregated and returned through the server layers to the user. It takes a quarter of a second. An absolute miracle.

Now, notice a couple of things about this technical architecture. First of all, it is highly modular—divided into small and interoperable components. Modularisation is key to how we deal with complexity. Second, it is layered—each of these rows represents a different kind of functionality. At the bottom, infrastructure: 2.5 million servers holding lists of words; at the top, the customised, localised front-end facing the user. This layered architecture is fundamental, because it enables Google to implement what engineers call the 'end-to-end' principle: moving functions as far up the stack (towards the end-user) as is consistent with their efficient utilisation.

Therefore, if something needs to be customised, if it's experimental, or something where you're recombining resources, you move it as near to the top as possible. If, on the other hand, it's infrastructure—a list, a passive resource—then you move it as low in the infrastructure to achieve economies of scale and utilisation. The end-to-end principle enables Google (and, indeed, the entire internet itself) to finesse the fundamental trade-off between innovation and scale. You get scale, efficiency, utilisation of capital-intensive functions at the bottom of the stack, and you get experimentation and innovation at the top. And, by separating those two kinds of activities, a generative architecture is created that can scale massively and also accommodate experiments and customisation.

In practical terms, this means that if Google doubles in size, they can add another 2.5 million servers at the bottom of this architecture with little difficulty. Scalability at the bottom is essentially unbounded. And once Google has the architecture in place, they can produce new products, enjoying what economists would call economies of scope, by recombining the same resources into new products and services. And as they proliferate products and services at the top, they add scale to the bottom.

The top of the stack enables innovation. The innovation isn't necessarily done by Google itself or by any provider—it can be done by customers. One of the interesting stories of the last 10 years has been the way so much innovation has come from users themselves. For example, an engineer

called Paul Rademacher needed to move home and found himself going back and forth between Craigslist for the listings and Google Maps for the locations. It was all very cumbersome, but he had the idea of hacking into the JavaScript that these two internet sites used, and synthesising their results to answer real estate searches in an integrated fashion.

This became a business in its own right: housingmaps.com. And it was one of the first examples of what we now call a 'mashup': a web service that combines information from *other* web services in order to create new value. When Google got wind of Rademacher's business, perhaps their first instinct was to sue him. But in fact they hired him to create application programming interfaces (APIs), to enable people with rudimentary programming skills to build their own mashups on the Google platform. The first API was for Google Maps. Google has since published hundreds.

Worldwide, something like 10,000 APIs have been published, creating the possibility of $10{,}000^2/2$ pairwise mashups. Most of those are meaningless, but there are actually 8,000 mashup businesses. The interesting thing about them isn't that any one mashup is a tremendously radical thing, but that 10 years ago, you would have needed months of work and substantial programming skills. Therefore, you would have needed funding, a business plan and a venture capitalist to sponsor your work. The barriers to doing this were huge.

Thanks to APIs, exactly the same thing can be done in a few hours. The investment required for innovation decreases dramatically. A lot of these 8,000 mashups aren't businesses at all—they're things that people did for fun; because they wanted to show how smart they were, or for ideological or humanitarian reasons. When the cost of innovation is just one wet Sunday afternoon's worth of work then the models by which innovation happen fundamentally change. You don't need corporations. You don't need government departments. You don't even need venture capital.

That's just one example of how innovation happens at the top of the stack. There's many others. Take e-lancing, the practice of taking freelancing work through online networks. Think Uber and Airbnb—both enable people to buy and sell services. Think about a commons like Wikipedia. Think about developer communities, such as the iPhone and Android communities. Think about social networks. Think about peer-to-peer

networks such as BitTorrent or Bitcoin. All of these arrangements flourish at the top of the stack. Why? Because the scale is provided by platforms further down.

I will finish this chapter by summarising my key points in six headline propositions.

- Convergence of four big forces: sense-making (artificial intelligence), big data, the proliferation of sensors and mobility. Those four together create this self-describing world.
- Products become services, and services become systems.
- Data processing becomes infrastructure. This is the emergence of cloud computing.
- Less obviously, data itself becomes infrastructure: something that we build, release and allow the world to exploit.
- We see the emergence of new topologies of networked experimentation, innovation and customisation at small scale by very large numbers of people.
- And we see the emergence of horizontal architectures, stacks and platforms replacing traditional vertical architectures. Stacks replace value chains; maybe in the bureaucratic world, they replace departmental organisations. The world is repolarised from vertical to horizontal.

The managerial challenges that we all face—whether in the private sector or the public—are to grasp the scale of this change and exploit the opportunity to use information in fundamentally new ways.

References

Borges, J.L. 1954. *A Universal History of Infamy*. Buenos Aires: Emecé.

Palchykov, V., M. Mitrovic, H. Jo, J. Saramäki and R.K. Pan. 2014. 'Inferring Human Mobility Using Communication Patterns'. *Scientific Reports* 4(6174). doi.org.virtual.anu.edu.au/10.1038/srep06174

This text is taken from *Opening Government: Transparency and Engagement in the Information Age*, edited by John Wanna and Sam Vincent, published 2018 by ANU Press, The Australian National University, Canberra, Australia.

doi.org/10.22459/OG.04.2018.12