CHATHAM HOUSE

Chatham House, 10 St James's Square, London SW1Y 4LE
T: +44 (0)20 7957 5700 E: contact@chathamhouse.org
F: +44 (0)20 7957 5710 www.chathamhouse.org

Charity Registration Number: 208223

Transcript

Tackling Global Problems with Big Data

Viktor Mayer-Schönberger

Professor of Internet Governance and Regulation, Oxford Internet Institute, University of Oxford

Kenneth Cukier

Data Editor, *The Economist*

Chair: Professor Angela Sasse

Head of Information Security Research, University College London

25 March 2013

The views expressed in this document are the sole responsibility of the author(s) and do not necessarily reflect the view of Chatham House, its staff, associates or Council. Chatham House is independent and owes no allegiance to any government or to any political body. It does not take institutional positions on policy issues. This document is issued on the understanding that if any extract is used, the author(s)/ speaker(s) and Chatham House should be credited, preferably with the date of the publication or details of the event. Where this document refers to or reports statements made by speakers at an event every effort has been made to provide a fair representation of their views and opinions, but the ultimate responsibility for accuracy lies with this document's author(s). The published text of speeches and presentations may differ from delivery.

Angela Sasse:

Good afternoon. My name is Angela Sasse. I'm professor of human-centred technology and head of information security at University College London. I'm here today to chair a presentation and the subsequent questions-and-answers by the two authors of the book which is here displayed on the table.

Ken Cukier is with *The Economist*, he's the data editor there – and he told me he's actually also a member of Chatham House. Viktor Mayer-Schönberger is professor at the Oxford Internet Institute. I was a big fan of his previous book, *Delete*, on privacy. Anybody who goes: what's this privacy stuff all about, why do they make such a fuss about it? I always recommend people read it.

But this book has been an extraordinary success already. You made the [*New York*] *Times* bestseller list last week and it's also, I've been told, the number-one bestselling business book in China, which is not something I would have necessarily expected. So without further ado, would you please come up and make your presentation.

Kenneth Cukier:

Good afternoon. I'm Kenneth Cukier and this is Viktor Mayer-Schönberger. We are both going to give a presentation – we're going to alternate – at the same time. We're going to talk about 'big data': how it's going to transform the way we work, the way that we live and the way that we think.

To explain what is happening and where things are headed, let me start with a story. The story goes like this. It starts with the flu.

Every year, the winter flu kills tens of thousands of people, but in 2009 a new strain of flu was discovered and there was a fear that it was going to kill tens of millions of people. There was no vaccine available and the best health authorities could do was to slow its spread, but to do that they needed to know where it was. In the US, the Centers for Disease Control (CDC) tracks new flu cases based on what doctors report, but collecting the data and analysing it takes time. So the CDC's picture is always a few weeks out of date, which is an eternity when a pandemic is underway.

At around the same time, engineers at Google developed an alternative way to predict the spread of the flu – not just nationally but down to different regions in the United States. They used Google searches. Google handles more than 3 billion searches a day and it saves all of them. So what Google did is they took 50 million of the most common search terms that Americans use and compared where and when these terms were searched for with 'flu', going back five years. The idea was to predict the spread of the flu through searches alone – and they struck gold. After crunching through almost half a billion mathematical models, Google identified 45 search terms that predicted the spread of the flu with a high degree of accuracy.

What you're looking at here is the official data of the CDC and alongside it Google's predicted data from its search queries. But where the CDC had a two-week reporting time lag, Google could spot the spread of the flu almost in real time. Importantly, Google's method does not involve distributing mouth swabs or contacting physicians' offices. Instead, it is built on big data: the ability to harness data to produce novel insights and new forms of economic value, new forms of goods and services.

Viktor Mayer-Schönberger:

Now, it is tempting to think of 'big' data in terms of its size. Our world is awash with data – the amount of digital data is growing fast, doubling a little over two years. We see the sea of data everywhere – including, for example, in the sciences. When the Sloan Digital Sky Survey telescope began in 2000, the telescope gathered more data in its first few weeks than had been amassed in the entire history of astronomy. Over 10 years, the collected astronomy data exceeded 140 terabytes of information. But a successor telescope due to come on stream in 2016 will acquire that quantity of information every five days.

Internet companies too are drowning in data. Twitter messages exceed 400 million a day; YouTube has more than 800 million monthly users, with an upload rate of an hour of video every second. On Facebook over 10 million photos are uploaded every hour, and Google processes a petabyte of data per day – that's around 100 times the quantity of all printed material in the US Library of Congress.

The quantity of data in the world altogether is estimated to be right now around 1.2 zettabytes, of which only two per cent are non-digital. If you look at it, that's just… basically a large number.

But just looking at big data and saying this is all about 'big', this is just about the absolute amount of data, I think would mischaracterize what is at stake and what big data is all about. In contrast, we suggest that big data is more than just about the volume. We believe there are three defining and reinforcing qualities that characterize what big data is really about: more, messy and correlations.

First: more. That means we can collect and analyse far more data relative to a particular problem or phenomenon than ever before, when we were limited to working with just a small sample. It's not the absolute size necessarily but the relative size that counts. That gives us a remarkably clear view of the granular, the details, that conventional sampling, for example, of data can't assess. We can let the data speak, and that often reveals insights that we never would have thought of. So that's first: more.

The second quality of big data is its embrace of messiness. Looking at vastly more data permits us to loosen up our desire for exactitude. Remember, when we only quantified a little, we had to treat what we did bother to measure as precisely as possible. In contrast, big data is often messy and of varying quality. But rather than going after exactitude in measuring and collecting just small quantities of data, with big data we'll accept some messiness. We'll often be satisfied with a sense of general direction rather than striving to know a phenomenon down to the last inch, the penny, the atom. We don't give up on exactitude entirely; we just give up our singular devotion to it. What we lose in accuracy at the micro-level, we gain in insight at the macro-level.

These two shifts lead, we suggest, to a third and a more important change: a move away from the age-old search for causality. Instead of asking 'why', of looking for often-elusive causal relationships, in many instances we can simply ask today: 'what?' And often, that is good enough. That is very hard for us human beings to stomach, because as humans we are conditioned to understand the world as a series of causes and effects. It makes the world comprehensible. It's comforting, it's reassuring – and oftentimes, it's just plain wrong. If we fall sick after having eaten at a new restaurant, our hunch – our causal hunch – will tell us it was the food, even though it's far more likely that we got the stomach bug from shaking hands with a colleague. It's these quick, quasi-causal hunches that often lead us down the wrong path.

With big data, we have a powerful alternative available. Instead of looking for the causes that are often elusive, we can go for correlations – that is, for uncovering connections and associations between variables that might not otherwise be known. Correlations help Amazon and Netflix to make recommendations. Correlations are at the heart of Google's translation services and its spellchecker. They do not tell us why – Google has no idea why people have been searching for the search terms that predicted the flu. They don't tell us why, but they tell us what – at a crucial moment and in time for us to act.

Take another example, a case in point. Premature babies are particularly prone to infections. The problem is that if you take their vital signs four times, six times a day, then by the time you find out that a baby has an infection and symptoms manifest themselves, it's often too late. So what you want to do is to find out that a baby has an infection very early on. So far we couldn't do that, but researchers in Canada, at the University of Toronto, around Dr Carolyn McGregor, have been able to make a breakthrough. What did they do? They took very modern sensors to measure 16 real-time data streams – like blood oxygenation level, heart rate, the heart electrocardiogram and so forth – amassing a thousand data points a second, and then capturing all of that over days and days, over quite a number of premature babies. With this data they did big data correlational analysis and they were able to find a pattern that often would predict the onset of an infection, 24 hours before the first symptoms would manifest themselves. That 24 additional hours saves lives of these very young, prematurely born babies.

Interestingly, the pattern that indicates that an infection is coming a day later is not that the vital signs go haywire – doctors might have caught that earlier – but that the vital signs stabilize. Who would have thought that, in the medical profession? If you look at vital signs that stabilize you usually, as a doctor, go home for the night because you think that the baby is doing well, rather than expecting an emergency and an onset of an infection. It was with big data that this could be revealed.

This is the quintessential big data analysis, featuring more, messy and correlation. The data was much more than we typically processed before. The data was so vast that it wasn't necessarily all in clean form – sometimes the sensor would come off for a couple of minutes, and that was okay. Analysis embraced messiness. And most importantly, the findings were correlations. They didn't answer why things were happening. They didn't answer or give the biological mechanisms that were at work. But they answered what was happening, and that was good enough.

Kenneth Cukier:

Often, big data is portrayed as a consequence of the digital age, but that misses a point. What really matters is that we're taking things that we never really thought of as informational before and rendering it into a data format. We're 'datafying' it, to coin a term. Once it's in the form of data, we can use it, process it, store it and analyse it to extract new value from it.

Think of location. We have longitude, we have latitude. We then developed the GPS system. Now we've got smartphones in our pocket. In a way, if you will, we have datafied our location – and at all times.

Think of it in terms of books. We had the temple of Delphi, with mottos etched in the top. Then we had books, and we scanned them – but by scanning books, we digitized books. We didn't datafy it – it wasn't searchable, there was no way we could run an analysis on it. Only when we datafied the text could we actually extract some value from it.

An example is: right now we have a technique called text mining. What that means is we could take every single medical journal in English, every single article that's been published for the last century, and run correlations on it and look for trends that we might not have spotted. For example, people are looking to see if there's adverse side effects that we didn't know about from different drugs in the medical literature – or promising drug targets, candidates that we want to research, that were thrown out for the study at hand but would be useful now. If any of you were to go into a room and try to read 100,000 medical journal articles, you'd never get these sort of answers. But buried in here, you can imagine, once we've datafied text, we can extract this value from it – latent value, there all the time, and now we have the tools to find it.

What else can we do? Researchers in the UK have done a big data analysis to visualize what topics were discussed and how frequently at the British House of Commons over the past 75 years. That's what you're looking at here. You can see how concerns on the environment are fairly recent as a development and how crime and national security, rather than defence and foreign affairs, have seen a long-term and sustained increase in interest.

Or you can think about it as posture. We typically don't think of the way that we sit as informational, but it is. The way that you're sitting right now, and the way that you're sitting, and you, and you – it's all different. It's a function of your weight, the distribution of your weight, your leg length, your posture, lots of different things. So if I were to put 100 sensors into each chair right now, I could create an index based on who you are that's relatively unique to you – sort of like a fingerprint. If you will, I'll have datafied your posture. What's the application? Researchers in Tokyo are actually doing this and putting it into a car seat. Now if an unauthorized driver is behind the wheel of a car, the car would know and maybe cut the engine unless a password was entered to start it again. So you could prevent car thefts.

What else could you do with it? What if we had 100 million vehicles on the road, all of which had the sensors inside the car seat? We might be able to find out the telltale shifts in posture 30 seconds prior to an accident that predict an accident. If we did that, what we might have datafied, if you will, is driver fatigue. The service would be to have an internal alarm to basically alert the driver that they have shifted, they have slumped, and that they should wake up. Maybe the steering wheel would vibrate or there would be a chime inside. That's the sort of thing that we can do once we datafy aspects of the world that are, if you will, informational but never rendered into a data format before.

Datafication is also a core by-product of social media platforms. Facebook has datafied our friendships and the things that we like. Twitter datafies our stray thoughts, our whispers. LinkedIn datafies our professional contacts. Once things are in data form, they can be transformed into something else.

Traditionally, data was processed for its primary purpose with little thought given about novel reuses – but this is changing. The core economic point of big data is that a myriad of reuses of information is possible that can unleash new services and improve existing ones. So the value of data shifts from the reason it was collected and the immediate use on the surface to the subsequent uses that may not have been apparent initially but are worth a lot.

Just as we are datafying aspects of the body for preemies, so too delivery vans. When a car breaks down, parts just don't fail all at once. Typically what happens is it's getting worn and it's announcing itself. We know this as drivers because the car drives a little bit funny, or we hear something that's just not quite right. So if we add sensors to the engine, we might be able to find the telltale shifts in temperature or in vibration that would predict that a part is going to fail prior to it failing, because we know what a normal part should be like and we can see what an out-of-range data signature looks like, and therefore we can see that something is going to happen. We would know in advance to take it into a service station and fix the part prior to the breakdown on the road. This is not science fiction – it's being done today by UPS.

There is a company in Seattle called INRIX that uses data from 100 million cars to predict the traffic flow in many cities around the world. By reusing its old data it found a strong correlation between road traffic and the health of local economies. Think about this: what it does as a business is to identify how easy it is to get from one place to another – it's a satellite navigation system – turn by turn. So it's predicting how long it will take you to go from point A to point B. But by collecting all this data and reusing it, what it's able to do is find the correlation with the health of local economies. It's a reuse. On top of that, one hedge fund licenses INRIX's data and it looks for the traffic in front of a large, national retailer over the weekends, because that correlates strongly with its sales. By looking at this data it can then trade the company's shares prior to its earnings announcements, because it has a lens in on its economic activity.

Britain too has been a pioneer in open data. What they have done in the NHS is they have revealed the data on their prescriptions. By matching the prescriptions in one area called statins, which is for lowering blood pressure, relative to where the prescription was made, they were able to identify some interesting findings. Specifically, what you're looking at is, of course, a map of England, and you can see that in some areas – in ways that we can't explain – the percentage of statins that are being prescribed is 35 per cent and over for branded statins that are very expensive. Other places, sometimes just a nearby county, it's actually less than 20 per cent.

We don't know why this is the case. Is it because there's a drug rep somewhere knocking on the door and giving freebies and kickbacks? We don't know. In fact, probably not. Maybe, maybe not – it's just as likely that maybe elderly doctors are doing what they're familiar with and so prescribe what they're comfortable with in the world prior to generics, what they're used to working with and their patients. It may be simply as innocent as that. But what's important is by mapping this information and being able to visualize and see it, we can take steps. The NHS was very interested in these findings because now they know there are some savings to be had. In a world of austerity for all governments, these sorts of techniques are very useful for public services.

Viktor Mayer-Schönberger:

So if data has all this hidden value, it will bring extraordinary benefits – but it unfortunately also has dark sides. As we have just heard, so much of data's value remains hidden, ready to be unearthed by secondary uses. This puts big data unfortunately on a direct collision course with how we currently protect informational privacy – that is, through telling individuals at the point of collection why we are gathering their data and asking for their consent – the so-called system of notice and consent. But in a big data age, we simply do not know when we collect the data for what purposes we'll be using it in the future. So as we try to reap the benefits of big data, our core mechanism of privacy protection is rendered ineffective.

But that is not big data's only dark side. A new problem emerges: algorithms will be predicting human behaviour that we are likely to do – how we will behave rather than how we have behaved – and penalizing us for it before we even have committed the infraction. If you think of *Minority Report*, that's exactly what we are thinking as well. Isn't it wonderful to be able to predict a behaviour in the future and then to stop it by punishing the person before he or she has committed the crime? Isn't prevention always better than punishment after the fact? At least you don't have a victim.

Yet we believe that such use of big data would be terribly misguided. For starters, predictions are never perfect – they only reflect the statistical probability. So we would punish people without certainty, negating a fundamental tenet of justice. Worse, by intervening before an illicit act has taken place and punishing the individuals involved, we essentially deny them human volition – our ability to live our lives freely and to decide whether and when to act. In a world of predictive punishment, we never know whether or not somebody would have actually committed the predicted action. We would not let fate play out, but hold people responsible on the basis of big data analysis that can never be disproven.

Let's be careful here. The culprit here is not big data itself but how we use it. The crux is that holding people responsible for actions they have yet to commit is using big data correlations to make causal decisions about individual responsibility.

So what can we do to control big data's dark sides? To begin with, we must not deny that big data has dark sides. We can only safely reap the benefits if we also expose its evils and discuss them openly.

In the book we suggest, first, that information privacy in the big data age needs a new, modified foundation. In the big data age, privacy controlled by the individual will likely have to be augmented by direct accountability of the data users.

Second and perhaps more importantly, on the dangers of punishing people based on predictions rather than actual behaviour, we suggest that we have to expand our understanding of justice. The big data age will require us to enact safeguards for human free will, for human volition, as much as we currently protect procedural fairness. In short, government must never hold an individual responsible for what they're only predicted to do.

Third, most big data analysis is far too complex for the individuals affected to comprehend. If we want to protect privacy and to protect individuality in the big data age, we may need help – professional help. Much like privacy officers aid in ensuring privacy measures are in place, we envision a new caste of experts – call them algorithmists – who understand the complexity of big data. We envision that they are experts in big data analysis, act as reviewers of big data applications and predictions. We see them take the vow of impartiality, confidentiality and professionalism like civil engineers do, or doctors.

Of course, big data requires more than these individual rights safeguards to fulfil its amazing potential. For instance, we may need to ensure that data isn't held by an ever-smaller group of big data holders, much like previous generations rose to the challenge posed by the robber barons that dominated the US railways and steel manufacturing in the 19th century. We may need to constrain the reach of nascent data barons and to ensure big data markets stay competitive.

Kenneth Cukier:

So we have seen the risks of big data and how to control them, but there is yet another challenge. It is one that is not unique to big data but that in the big data age society needs to be extra-vigilant to protect against, and it's what we call 'the dictatorship of data'. It's the idea that we may fetishize the data and endow it with more meaning than the data itself merits. As big data starts to play a part in all areas of life, this tendency to place trust in data and cut off our common sense may only grow.

The war in Vietnam was a war based on a data point. That is a simplification, but it's one that might rivet the mind. That data point was the body count. It was used to measure progress when the situation was far, far more complex. So in the big data age it will be critical that we do not follow blindly the path that big data seems to set.

Big data will help us. It is going to help us understand the world better. It will improve how we make decisions – from what medical treatments work to how we educate our children, to how a car can drive itself. But it also brings new challenges and new dangers. Just as there is a vital need to learn from the data, we also need to carve out a space for the human, for our reason, for our imagination, for acting in defiance of what the data says – because the data is always just a shadow of reality and therefore always imperfect, always incomplete. As we walk into the big data age, we need to do so with humility and humanity. Thank you.