Privacy Commissioner
Te Mana Matapono Matatapu

Privacy and Big Data

PRESENTATION BY PRIVACY COMMISSIONER JOHN EDWARDS

Monday 2 September 2014, Ministry of Social Development

The seductive power of big data

"Torture the data, and it will confess to anything."

You may be familiar with this statement by the late distinguished economist and Nobel Prize Laureate, Ronald Coase.

It is especially pertinent against a universal backdrop of the inevitable wider use and re-use of data in business and in government.

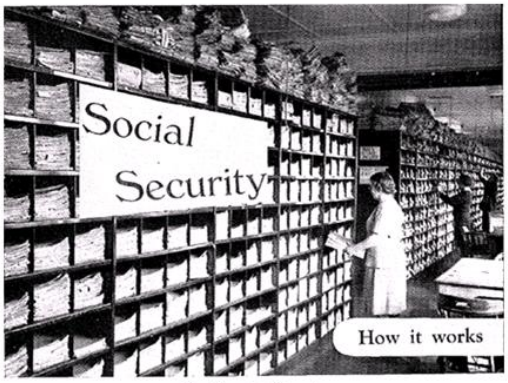
For the public sector, the potential to design better policies and target resources more efficiently means that the re-use of data is as much a responsibility as an opportunity.

After all, taxpayers have paid for the collection of their own information, and the government should be seeking to maximise the return of value to New Zealanders from it – at the very least by evaluating the effects of different interventions. It is a compelling argument for government to maximise the value of data it collects as a by-product of the services it delivers. This is a 2014 question.

Data past

But let's contrast that with what the collection, storage and use of 'data' looked like in the 1950s.

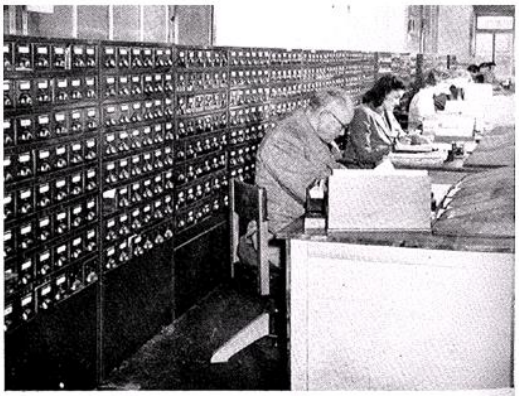
Take these images from the Department of Social Security. It was a time of large filing cabinets and also of the advent of using 'punch cards' – a quite revolutionary innovation in terms of using individual level data for operational purposes.



Keeping the Records

In the 1950s, having personal data on individual files and cards meant personal information was quite difficult to use,

access and accidentally release. Research that used the data was very difficult and it seldom used individual level data, apart from sampling the odd case note.


Part of the National Index

In the 60 years since, the costs of collecting, storing and using individual data have reduced immensely. It is now very cheap to use and access digital data on individuals. Digital datasets of unique individuals with records of their interactions with government now make research so much easier.

While increasing processing power, data storage capacity and use of rapidly evolving analytical tools are boons for maximising the value of data, they also have big implications for privacy.

As our lives become ever more enumerated or dissected, privacy and privacy law must ensure we have control over personal information about ourselves and maintain our belief in our ability to make autonomous decisions.

Paying the People. The Old Method—Writing Payment Warrants by Hand

New Zealand's data future: strategic direction

We are fortunate in New Zealand that we already have a robust legal framework for managing data with the Statistics Act and the Privacy Act as the pillars of that framework.

The aim is for value, inclusion, trust and control to be maintained in that framework for the opportunities and challenges ahead in a world where large scale data collection and analysis are ubiquitous.

That is the vision of the New Zealand Data Futures Forum as set out in its report *Harnessing the economic and social power of data*.

The New Mechanized Method—Punching Powers-Samas Cards

It is encouraging to see that the Data Futures Forum recognises that privacy is an integral part of the national conversation about our use of data.

The Privacy Act deals particularly well with the desire to realise value from data from recombination and reuse. But there are areas where the Act can be strengthened.

Strengthening Privacy Act controls

For example, under the Act, there is currently no explicit prohibition on the re-identification of data from which identifying information has been removed. It's food for thought that a prohibition of this nature could potentially increase public confidence in the safe use of "de-identified" or "anonymised" data.

Similarly, further work could be undertaken on strengthening individual rights to have information about them *deleted*, again increasing their confidence that information provided is not necessarily available forever and able to be combined with yet to be created data sets.

The right to be forgotten

Many of you will be aware of the debate over 'the right to be forgotten' ruling by the European Court of Justice in May this year. That decision was based on a Spanish legal requirement on the *relevance* of personal data.

I am not a big fan of the term 'right to be forgotten' because it means different things to different people. It can mean data portability, so that if you decide to leave (say) Facebook, you should be able to take your links and content with you, and have that deleted from the site and not retained by the proprietors of that site.

It could mean a right to anonymity, or to its cousin, obscurity. And it can mean the removal of content from public search.

On the back of the European Court of Justice ruling, Google in Europe has been receiving more than 10 thousand requests a day that it forgets about some kind of personal information, that it breaks links to information that is no longer relevant.

Since then, the European Union Justice Commissioner Martine Reicherts has accused Google and other internet search engines of abusing this ruling by blocking search results as a way of undermining planned stronger EU data protection reforms.

She warned:

> *This ruling does not give the all-clear for people or organisations to have content removed from the web simply because they find it inconvenient.*

Ms Reicherts also said:

> *"Search engines such as Google and other affected companies complain loudly. But they should remember this: Handling citizens' personal data brings huge economic benefits to them. It also brings responsibility. These are two sides of the same coin, you cannot have one without the other."*

This conversation which has begun in Europe, and which is gathering momentum around the world, is a response to this increasingly ubiquitous hoarding and storing of personal information.

If every aspect of your life is being tracked and collected in databases that never forget, then more and more of your information is heading out of your control at every moment. And privacy is all about helping people keep control over their own information in the face of technological innovations that lessen that control.

The security expert Bruce Schneier said, with reference to the internet-enabled world, "we are embarking on a great experiment of never forgetting".

But, at this stage, there is still debate as to what a right to be forgotten actually requires, let alone how it might be implemented.

Granting individuals stronger rights to have information deleted has the potential to improve our ability to obtain value from data, because it could give individuals greater confidence that when they provide information, they have not necessarily given it away forever.

The relationship between open data and personal information needs to be explored more fully, with the aim of developing a concept of "degrees of openness".

By definition, 'open data' is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike. At the moment, the open data model consists of only one of two option – published data that is available to everyone. But it is theoretically and practically possible to make data available on limited terms which meet analysts needs while minimising the intrusion on individual privacy. A good model for this is the government's Integrated Data Infrastructure which applies conditions and safeguards to the data.

I'll talk about the Integrated Data Infrastructure later.

Regulating open data

Government efforts to provide open data have sought to exclude personal information, although what constitutes personal information is not necessarily well understood. This is not a sustainable approach – not least because agencies are already publishing information that has not been appropriately anonymised.

There may now be a case for an independent 'data council' to promote the ethical and safe use of data as recommended by the Data Futures Forum.

While I play a role in setting the limits on what can be done with personal information, my role is limited to personal information, and does not include a mandate to promote the wider use of data. That role is not, as yet, part of any organisation's mandate.

There is also an important distinction between privacy and ethics in data use and there is no body charged with providing advice on what constitutes ethical practice outside of National Ethics Advisory Council's role in the health sector.

A case which you may have heard about recently involved Facebook admitting to manipulating nearly 700,000 of its users' news feeds to see whether it would affect their emotions. The social network was roundly condemned by many social scientists for breaching ethical guidelines for informed consent.

Two key features

There are two key features of our privacy law that make it a potential model for managing privacy in a data driven future:

1. The Privacy Act's definition of personal information is broad enough to encompass de-identified and pseudonymous information.

2. It also provides broad exceptions to principles on collection, use and disclosure where information will be used in a form in which individuals will not be identified.

What is the effect of that?

The upshot is that if agencies have a lawful purpose for collecting personal information and do not intend to use it in a form in which individuals will be identified, they can use and re-use it without having to obtain detailed consents that apply to all those future uses.

But agencies still have responsibilities to collect only data that they have a lawful use for, to store it securely, and to delete it when they no longer have a use for it.

An individual's confidence in the use of de-identified data is in part determined by their belief that they will not be able to be re-identified. Protection against re-identification is important because it can be surprisingly easy to identify individuals in supposedly de-identified data.

De-identified data and the Massachusetts case study: risks to individuals

While the rewards of re-using data are many and obvious, there are past lessons that serve as a warning to hubris about the lure of big data as a policy panacea.

Here's a well known American example.

In the mid 1990s, the Massachusetts Group Insurance Commission decided to release anonymised health data on state employees. Its aim was to help health researchers to improve healthcare. Obvious identifiers such as name, address and social security number were removed from the data.

The Massachusetts governor of the time, William Weld, assured the public that the Group Insurance Commission had protected patient privacy by deleting identifiers.

A graduate student in computer science at MIT, Latanya Sweeney, requested a copy of the data, and got to work. She knew that Governor Weld lived in Cambridge, Massachusetts, a city of 54 thousand residents and seven postal codes.

For 20 dollars, she purchased the complete electoral rolls for Cambridge. These included the name, address, postal code, birth date and sex of every voter. Only six people in Cambridge shared the governor's birthdate. Only three of those six were men, and of them, only he lived in his postal code. Dr Sweeney then had the governor's detailed health records, including diagnoses, prescriptions and details of hospital visits, delivered to his office.

Latanya Sweeney has continued to research in this area, and has revealed that our intuitive beliefs about how difficult it is to identify an individual from a supposedly anonymous set of data are often misplaced.

Among her findings, she has demonstrated that 87 percent of the American population can be identified by birth date, sex and postal code. This is particularly startling when you keep in mind that the average postal code in the United States has a population of around 7,500.

To put that in the New Zealand context, the average population of a Statistics New Zealand mesh block is about 90. For the next largest statistical unit, the "area unit", the average population is 2,100. On the upside, we have quite strict rules about the use of electoral rolls and other official population registers.

As more data sets are linked together there are an increased number of vectors for identifying a target. So while you might not have information about a target's birthday, you might know what they studied at university, or how many children they have, or whether they have been convicted of an offence.

Britain's Care Data fiasco: jeopardising objectives

As well as the risk to individuals, the ability to identify individuals within a large data set can jeopardise a project's objectives. This is a feature of current debates in Britain over the government's decision to make detailed National Health Service data available to research through its Care Data initiative.

Care Data is a National Health Service initiative to take patient data from GP records and upload them to the country's Health and Social Care Information Centre databases. The aim is to combine this with existing hospital records in that national database to provide a picture of healthcare being delivered and to identify areas where more work or investment might be needed.

Privacy campaigners have pointed out that where the details of treatments were in the public domain, it would be possible to identify the individual's and 'read across' their broader NHS record. The public outcry forced the government to delay the launch of Care Data. An online tool developed by a private company has already been shut down. There's been a Parliamentary select committee hearing. There's even a YouTube Downfall parody.

Now some experts are saying that Care Data could be on hold for as long as two to three years because of the nervousness over issues of privacy, consent, anonymity and commercial exploitation of people's health data.

The controversy surrounding Care Data is also likely to contribute to a further weakening of patient trust in the British health sector more broadly, and make it more difficult for the launch of similar initiatives in the future.

A notable feature of the Care Data controversy is that it is not possible to reduce the debate to a simple choice between scientific progress on the one hand, and the protection of privacy on the other.

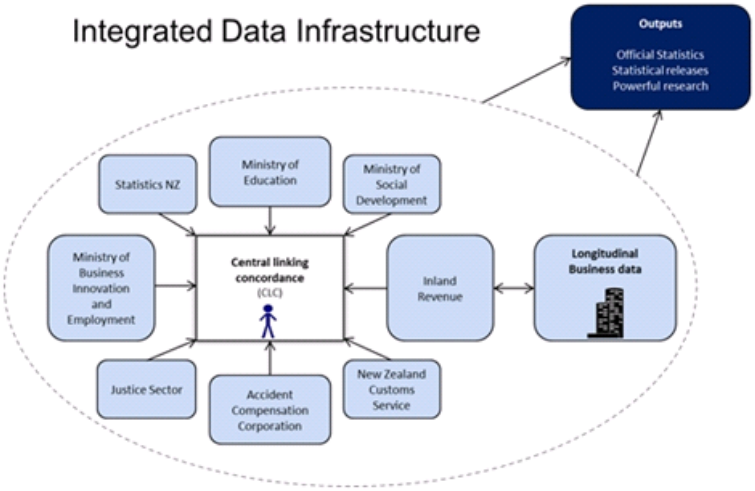
While the usability of data and privacy are in tension, the project cannot succeed without privacy. Individual patients must support the project politically as voters and taxpayers, and by not opting out. In order to achieve this, patients must trust that the data is well-governed, and access is appropriately controlled.

New Zealand's Integrated Data Infrastructure

In New Zealand, we've seen the government move to aggregate data sets across departments, with the government's decision to significantly expand Statistics New Zealand's Integrated Data Infrastructure (IDI).

This hasn't been accompanied by a significant controversy to date, despite the scale and breadth of the data set once the currently planned integrations have been completed.

My office has cautiously supported the expansion of the IDI, although not in an unqualified way.



The IDI expansion represents something of a grand bargain – consolidate the collection of integrated data sets in one place under a strong governance regime, including a very clear statutory mandate and controls – or the alternative is the piecemeal integration of data by various government agencies under potentially inconsistent controls.

Our confidence is based on our experience and knowledge of Statistics New Zealand's existing management of the IDI. This includes its long established culture of confidentiality and respect for data subjects which has led to its acceptance in the community as a trusted custodian of statistical data.

This happens within a robust framework in the Statistics Act which requires Statistics New Zealand to control the products of research using micro data while ensuring that individuals will not be identified.

Predictive risk analysis

The idea that we can prevent or reduce harm and improve health outcomes by early intervention is a compelling one.

Predictive risk analysis has the potential to have a positive impact on people's lives – and even save lives – as well as being a more cost effective approach than the proverbial 'ambulance to the bottom of the cliff'.

The science - predictive risk modelling - is defined as an automated algorithm which harvests data from a variety of sources in order to generate a probability that something is likely or not likely to happen.

Auckland University research

You will probably be aware of recent Auckland University research into how predictive modelling could be used to target early intervention for vulnerable children. The research was commissioned by MSD and the university researchers used linked MSD datasets, de-identifying all the data so they could not identify individuals.

The Auckland University researchers had developed a predictive risk model for children in a cohort who had contact with the benefit system before the age of two. These children accounted for 83 percent of all children who were recorded as suffering maltreatment by the age of five.

The research showed that predictive risk modelling had a fair-to-good power of predicting which of these young children would be abused. The research team described this as on a par with the predictive strength of mammograms for detecting breast cancer in women who show no symptoms of the disease.

It is hard to argue against this kind of information sharing and the use of predictive risk analysis techniques when trying to protect children. In this case, the research looked retrospectively at children who were already in the system. It, according to the researchers, obtained a fair-to-good power of prediction. But how accurate would such an algorithm be in identifying vulnerable children not already captured by the system? And how would it compare to potentially less costly but similar tools such as actuarial risk assessments that are widely used internationally?

The challenge of advocates of predictive risk modelling is for it to prove itself as being better than alternative ways of identifying the kids at risk. And remember that if we get it wrong, there is the potential to cause a great deal of harm to families who are bound to be affected by the error.

Building a predictive model

Building a predictive model requires very large amounts of data; with more data, there is greater accuracy. But the precision of predictive models is inevitably constrained by the range and accuracy of information contained in data sets.

Take for example, the number of predictive risk models and applications in healthcare. These have increased dramatically where once there were a handful of risk models. And one of the challenges facing health providers is choosing a model that is the right fit. There is continuing debate about the cost effectiveness of a number of screening programmes, and the potential for them to channel patients down potentially costly and invasive care pathways on the basis of relatively low probabilities.

A range of factors need to be considered carefully when choosing whether to ‘make or buy’ a predictive model, including the outcome to be predicted, the accuracy of the predictions made, the cost of the model and its software, and the availability of the data on which the model is run.

Getting the right model is necessary for getting the right outcome - otherwise skewed results can lead to inaccurate results, misapplication of resources, dangerous outcomes and the loss of public confidence.

Data errors or data weighting mistakes where wrong data elements are inappropriately deemed important can also lead to failures. You cannot arrive at a correct result if the data set is flawed.

In our submission to the Social Services Committee on the Vulnerable Children’s Bill last year, we said the proposal for Child Harm Prevention Orders in the Bill had the capacity to fail the accuracy of information test. The proposal attempted to predict the likelihood of harm to a child through an actuarial assessment tool.

That tool had not yet been developed but the office’s concern was that using information to predict future behaviour would be speculative. The Bill’s regulatory impact statement recognised that the uncertainty of risk prediction meant that Child Harm Prevention Orders would inevitably be imposed on individuals who would not have gone on to commit offences.

Remember that Child Harm Prevention Orders would not just have applied to individuals convicted of an offence. It would also have been applied to individuals who were considered more likely to commit a particular offence, and where evidence of past behaviour was less certain, and in places conflicted. Evidence from well established actuarial tools in use for managing convicted sex offenders suggested that a significant number of individuals subject to orders would not have gone on to offend.

There’s danger in becoming over confident in or over reliant upon what is essentially a mathematical system that is trying to predict future behaviour. A wrongful targeted Child Harm Prevention Order would have affected the individual and his or her family, including potentially the children the CHPO was intended to protect.

The government appeared to have taken note of our submission and others, and the Child Harm Prevention Order proposal was dropped from what is now the Vulnerable Children’s Act.

I’d now like to end with another quote from the economist, Ronald Coase. When asked about his politics, Coase said:

> *"I really don't know. I don't reject any policy without considering what its results are. If someone says there's going to be regulation, I don't say that regulation will be bad. Let's see. What we discover is that most regulation does produce, or has produced in recent times, a worse result. But I wouldn't like to say that all regulation would have this effect because one can think of circumstances in which it doesn't."*

As a regulator, I thought this was fitting. I’d like to think that privacy regulation and big data is one of those circumstances in which the outcome shouldn’t be worse - because the dangers of being wrong are self evident.

ENDS