The collision between Big Data and privacy law

Stephen Wilson
Constellation Research Inc.

Summary: We live in an age where billionaires are self-made on the back of the most intangible of assets – the information they have about us. The digital economy is awash with data. It's a new and endlessly re-useable raw material, increasingly left behind by ordinary people going about their lives online. Many information businesses proceed on the basis that raw data is up for grabs; if an entrepreneur is clever enough to find a new vein of it, they can feel entitled to tap it in any way they like. However, some tacit assumptions underpinning today's digital business models are naive. Conventional data protection laws, older than the Internet, limit how Personal Information is allowed to flow. These laws turn out to be surprisingly powerful in the face of 'Big Data' and the 'Internet of Things'. On the other hand, orthodox privacy management was not framed for new Personal Information being synthesised tomorrow from raw data collected today. This paper seeks to bridge a conceptual gap between data analytics and privacy, and sets out extended Privacy Principles to better deal with Big Data.

Introduction

'Big Data' is a broad term capturing the extraction of knowledge and insights from unstructured data. While data processing and analysis is as old as computing, the term 'Big Data' has recently attained special meaning, thanks to the vast rivers of raw data that course unseen through the digital economy, and the propensity for entrepreneurs to tap that resource for their own profit, or to build new analytic tools for enterprises. Big Data represents one of the biggest challenges to privacy and data protection society has seen. Never before has so much Personal Information been available so freely to so many.

Big Data promises vast benefits for a great many stakeholders (Michael & Miller 2013: 22-24) but the benefits may be jeopardized by the excesses of a few overly zealous businesses. Some online business models are propelled by a naive assumption that data in the 'public domain' is up for grabs. Many think the law has not kept pace with technology, but technologists often underestimate the strength of conventional data protection laws and regulations. In particular, technology neutral privacy principles are largely blind to the methods of collection, and barely distinguish between directly and indirectly collected data. As a consequence, the extraction of Personal Information from raw data constitutes an act of *collection* and as such is subject to longstanding privacy statutes. Privacy laws such as that of

ISSN 2203-1693. Copyright © 2014 The Author(s). http://doi.org/10.7790/ajtde.v2n3.54

Australia don't even use the words 'public' and 'private' to qualify the data flows concerned (*Privacy Act* 1988).

On the other hand, orthodox privacy policies and static data usage agreements do not cater for the way Personal Information can be synthesised tomorrow from raw data collected today. Privacy management must evolve to become more dynamic, instead of being preoccupied with unwieldy policy documents and simplistic technical notices about cookies.

Thus the fit between Big Data and data privacy standards is complex and sometimes surprising. While existing laws are not to be underestimated, there is a need for data privacy principles to be extended, to help individuals remain abreast of what's being done with information about them, and to foster transparency regarding the new ways for personal information to be generated.

When online innovations cross the line

Personal Information is the lifeblood of most digital businesses today and yet consumers commonly find digital practices to be 'creepy' (Tene & Polonetsky 2013: 61-68). Individual reactions depend on experience and computer literacy but different proportions of people are disturbed by phenomena such as tailored advertising appearing in their browser soon after they've sent an email about the very subject of the ads, or by an online social network automatically naming people in photos, or an Internet map service displaying reservation details alongside a hotel when a user simply searches for that property. Subjective reactions are useful pointers to privacy transgressions, and yet 'intuitions and perceptions of "creepiness" are highly subjective and difficult to generalize as social norms are being strained by new technologies and capabilities' (Tene & Polonetsky 2013: 59). And so we need objective tests of privacy breaches in order to codify and enforce reasonable rights.

Consider 'Pay-as-You-Drive' car insurance, a new type of cover, with premiums scaled according to how you drive. By analysing data from automobile 'black boxes', an insurance company can tell not only how far the car has gone (so that infrequent drivers can enjoy discounted premiums) but can also detect how fast it has been going and in what neighbourhoods it has been parked. Higher risk driving behaviours can attract extra levies or other forms of disincentive. Across the whole community, these types of innovative products can bring shared benefits, with more precise risk management, fine-grained risk figures as compared with traditionally coarse actuarial data, lower prices for lower risk customers, and eventually, lower average costs for everyone.

GPS signals could be used to inform 'smart' car insurance offerings, yet explicit vehicle tracking arouses privacy fears. Therefore, some Pay-as-You-Drive systems promise not to use GPS, and instead draw only on seemingly more innocuous speed and time measurements. And yet, the privacy picture is not so simple, for there remain alternative and less obvious ways to work out where a driver is.

Recent research in the United States (Dewri et al 2013) has shown that vehicle speed and time measurements, when combined with map data, can be used to infer the location of a car at any time with just the same precision that GPS coordinates provide. Therefore 'customer privacy expectations in non-tracking telematics applications need to be reset and new policies need to be implemented to inform customers of possible risks' (Dewri et al 2013: 267).

It is not known if insurance companies are in fact exploiting automobile black box data in this way, but the temptations of Big Data prove time and time again to be irresistible. If time and speed data can be accessed by third parties and linked to maps or other data sets to extract insights about drivers, it may only be a matter of time before this routinely happens.

When businesses go too far with advanced data analytics and leave users feeling violated or betrayed, then everyone suffers. Disillusioned customers won't just abandon the firms that have squandered their trust; they will also lose confidence in cyberspace more broadly and withdraw from other new and worthwhile services. The society-wide promises in e-government and e-health programs will be compromised if the proportion of citizens participating in them is reduced for want of privacy (Wilson et al 2005: 11-16).

The Big Business of Big Data

It's not for nothing that people use the term 'data mining'. Raw data is often likened to crude oil (Singh 2013), meaning the riches to be extracted from an undifferentiated ore-like matrix spread across cyberspace, comparable to the ground beneath our feet explored by traditional prospectors.

Consider photo data, for instance, and the rapid evolution of tools for monetising it. Techniques for extracting value from images range from the simple metadata embedded in digital photos which record when, where and with what sort of camera they were taken, through to increasingly sophisticated pattern recognition and facial recognition algorithms (Sawant et al 2011). Image processing and image analysis can extract places and product names from photos and automatically pick out objects. Biometric facial recognition can identity faces by re-purposing biometric templates that originate from social network users tagging their friends for fun in entirely unrelated images. Therefore it has become entirely feasible for social media companies to work out what people are doing, when and where, and who they're doing it with, thus revealing personal preferences and relationships, without anyone explicitly 'liking' anything or 'friending' anyone.

The ability to mine photo data defines a new digital gold rush. Like petroleum engineering, image analysis is very high tech. There is extraordinary R&D going on in face and object recognition. Information companies like Facebook and Google (whose fortunes are made on nothing other than information) and digital media companies like Apple have invested enormously in their own R&D, and also in acquiring start-ups in this space. For example Google acquired the cloud photo storage service Picasa in 2004 (for an unknown price); Facebook bought photo sharing network Instagram in 2012 for approximately US$1 billion; and in late 2013, instant photo messaging service Snapchat turned down an offer from Facebook for approximately US$3 billion. These extraordinary investment decisions are not explained merely by users having fun taking photos and tagging them, but rather by the potential for extracting monetisable intelligence from images.

So, more than data mining, Big Data is really about data *refining*: the transformation of unstructured facts and figures into fresh insights, decisions and value.

Business models for monetising photo data are still embryonic. Some entrepreneurs are beginning to access photo data from online social networks. For example 'Facedeals', a proof of concept from advertising invention lab Redpepper, provides automated check-in to retail stores by way of face recognition; the initial registration process draws on images and other profile information made available by Facebook (with the member's consent) over a public Application Programming Interface (API see http://redpepperland.com/lab/details/check-in-with-your-face). It is not clear if Facedeals accesses the online social network's actual biometric templates, but nothing in Facebook's privacy and data use policies restrains the company from providing or selling the templates (Facebook: 2014). But as we shall see, international privacy regulations do in fact restrict the uses that can be made of the by-products of Big Data, should they be personally identifiable. Facebook has been taken to task by regulators for stretching social data analytics beyond what members reasonably expected to occur (Johnston & Wilson 2012:59-64). I believe more adverse surprises like this await digital businesses in retail, healthcare and other industries.

Big Data and the Law

It's often said that technology has outpaced the law, yet by and large that's just not the case when it comes to international privacy law. Technology has certainly outpaced the intuitions of consumers, who are increasingly alarmed at what Big Data can reveal about them behind their backs. However, data privacy principles set down by the OECD in 1980 (OECD 1980) still work well, despite predating the World Wide Web by decades. Privacy laws are strengthening everywhere (Greenleaf 2011). Outside the U.S., rights-based privacy law has proven effective against many of today's more worrying business practices (Johnston &

Wilson 2012, OAIC 2012). Digital entrepreneurs might feel entitled to make any use they like of data that comes their way, but in truth, 30-year-old privacy law says otherwise.

Information innovators ignore international privacy law at their peril. In this article, I will show why, by reviewing the possibly surprising definitions of Personal Information, and that technology-neutral privacy principles are as relevant as ever.

Personal Information technicalities

Technologists in general know that the devil is in the details, and that technicalities matter. They need to know the definition of Personal Information, because the technicalities have the power to surprise.

'Privacy' can be a difficult topic. Indeed, a leading privacy scholar has said 'Privacy is a concept in disarray. Nobody can articulate what it means' (Solove 2006: 477). On the other hand, the smaller field of *data* privacy (also referred to as data protection) is tightly defined. Throughout this paper, privacy means data privacy. Different privacy regimes worldwide variously use the terms 'Personal Data', 'Personal Information' and 'Personally Identifiable Information'. Loosely speaking and for our purposes here, the terms are interchangeable. The important thing about all these terms is they do not require the data in question to be personally *identifying*; there is nothing in the respective privacy regimes that requires 'Personal Information' to point uniquely to any individual. Data privacy then focuses simply on regulating the handling and processing of a special class of information.

The U.S. General Services Administration (GSA) defines Personally Identifiable Information (PII) as 'information that can be used to distinguish or trace an individual's identity, either alone or *when combined with other personal or identifying information* that is linked or linkable to a specific individual' (*emphasis added*)(GSA 2014).

Recently updated Australian privacy legislation defines Personal Information as 'information or an opinion about an identified individual, or an *individual who is reasonably identifiable*:
(a) whether the information or opinion is true or not; and

(b) whether the information or opinion is recorded in a material form or not' (emphasis added)(*Privacy Amendment Act* 2012).

It is notable that the definition of Personal Information in Australian law and that of Personally Identifiable Information used in the American government are so close (for the rest of this paper, I will treat the terms interchangeably and use the acronym PII to refer to both). The Australian and American definitions both embody a precautionary approach. Items of data can constitute PII if other data can be combined to identify the person concerned. Note carefully that the separate *fragments* of data are each regarded as PII rather than the whole data that eventually might identify someone precisely. Laypeople may presume that PII stands for Personally Identifying (rather than Identifiable) Information. The difference is subtle but very important. The definition means that some data items can – and should – be classified as PII before they are ever actually identified rather than after, with due consideration to the context of the data flows and the potential for identification. This is only prudent; if personal data needs to be safeguarded, then it is best this be done before the data is identified and it becomes too late.

Classical data privacy controls

To set it apart from security, it is often said that privacy is more about 'control' than confidentiality or secrecy. So what does that mean?

The OECD Privacy Principles (OECD 1980) were developed in the 1970s to deal with the emerging threats of computerisation. Even in the decade before that, the burgeoning databases of governments, police forces and insurance companies were seen as a danger to civil liberties (Boeth 1970); the cover of Newsweek magazine on July 27, 1970 screamed 'IS PRIVACY DEAD?' Since that time, over 100 countries have legislated data privacy protections based on the OECD principles or the more advanced European Union Privacy Directive (Greenleaf 2011).

Despite this strong global trend, American legislators have declined to enact broad-based privacy law, although particular sectors, such as U.S. healthcare, feature some of the strictest data protection rules anywhere in the world. The Fair Information Practice Principles (FIPPs) of the United States have been adopted here and there, and reflect most of the OECD principles, though crucially not all.

Conventional rights-based privacy principles are relatively straightforward. They neatly sidestep philosophical complexities like the 'self', data ownership and the increasingly nuanced difference between public and private domains. Instead data privacy principles essentially require data custodians to be careful, restrained and transparent in how all PII is handled. As one US law professor put it recently, '"privacy" is shorthand for the regulation of information flows' (Richards 2014).

In the context of Big Data, two of the standard international privacy principles stand out:

Firstly, the Collection Limitation Principle requires that organisations collect only the Personal Information they need for legitimate and transparent purposes. Collection Limitation is about discipline rather than prohibition. Privacy regulations do not stop businesses collecting the information they truly need; rather, it requires that businesses justify what they collect. The starting point for the much talked about practice of 'Privacy by

Design' should be conscious Collection Limitation. IT designers should habitually analyse their PII requirements and from there carefully specify systems that collect the PII which is needed, and nothing more.

Secondly, the Openness Principle requires that data custodians set out for all to see what PII they collect, why they need to collect it, how, when and where they collect it, and who else the PII may be shared with.

These privacy principles apply to all Personally Identifiable Information, whether it is collected directly by form or questionnaire, or indirectly through data analytics. And so we come to a central challenge in Big Data privacy: open-ended data analytics can lead to brand new discoveries which cannot be fully envisaged and outlined at the time raw data was collected. With the best will in the world, a Big Data company might not be able to spell out in its Privacy Policy what PII it expects to collect (via data mining) at all points in the future. Do we want the potential benefits of Big Data to be inhibited by the finality of the Collection Limitation principle? If there are shared benefits in the possibility of new PII being uncovered in raw data, how can the privacy promises of Collection Limitation and Openness be honestly kept? Answers may in part lie in organisations keeping up an open dialogue with their users and customers instead of trying to freeze a privacy understanding at the time raw data is gathered.

Big Data 'spills'

Here's another reason raw data is like crude oil: when it 'spills', it's expensive to clean up. On several occasions, data analytics and other innovative information business practices have led to major privacy breaches and costly non-compliance actions, in ways that have surprised the practitioners. And further surprises are in store for companies that do not grasp the meaning of PII and international privacy law.

Google finds that 'public' can still be private

While they drive around photographing towns and cities, Google's StreetView cars listen for Wi-Fi hubs and collect the geographical coordinates of any transmitters they find. Google collects Wi-Fi landmarks for its geo-location database, which underpins its maps and other important services.

On their own, the names and locations of Wi-Fi hubs (technically known as SSIDs, or 'Service Set Identifiers') have never raised significant privacy concerns, but in 2010 it became apparent that StreetView cars were also inadvertently collecting unencrypted Wi-Fi network traffic, which inevitably contained PII including user names, banking details and even passwords. Privacy commissioners in Australia, Japan, Korea, the Netherlands and elsewhere found Google was in breach of respective data protection laws. The company investigated, and responded that one developer writing StreetView Wi-Fi mapping software was responsible for conducting something of an isolated experiment. The collection of network traffic was said by Google to be inadvertent. The company apologized and destroyed all the content that had been gathered (OAIC 2012).

The nature of this particular privacy offense confused some commentators and technologists. Some have argued against me that Wi-Fi data in the public domain is not private, and categorically therefore could not be private. Such a line of thinking holds that Google is within its rights to do whatever it likes with such data, but the reasoning fails to grasp the technicality that data protection laws in Europe, Australia and elsewhere do not essentially distinguish 'public' from 'private'. If data is identifiable, then various privacy rights attach to it, irrespective of how it is collected.

Facebook over-automated photo tagging

Photo tagging is a popular feature of photo sharing services that helps users to better organise their albums. Tagging as offered by Facebook creates biometric templates that mathematically represent an individual's facial features, allowing other photos to be identified as being of the same person. When Facebook makes automatic 'tag suggestions', its facial recognition algorithms have been running in the background over all photos in the network's enormous database, making putative matches and flagging deduced user names against the analysed photos. When a photo containing an identified individual is next displayed to a member, the tag suggestion is displayed and the viewer is invited to confirm it. According to the definition of Personally Identifiable Information, when Facebook's software adds a name to a hitherto anonymous photo record, Facebook turns that record into PII; the tagging process therefore *collects* PII, albeit indirectly.

European privacy regulators in mid-2012 found that collecting biometric data in this way without consent was a breach of privacy laws (HmbBfDI 2012). By late 2012, German and then Irish authorities forced Facebook to shut down facial recognition and tag suggestions for all its European operations, and to delete all biometric data collected to that time. This was quite a show of force over one of the most powerful companies of the digital age.

So it doesn't much matter if data miners generate PII almost out of thin air, using sophisticated data processing algorithms; they are still subject to privacy principles, such as Openness and Collection Limitation. Crucially, until 2012, Facebook's privacy policy and data usage policy had not even mentioned that biometric facial recognition templates were created by tagging, let alone that they were subsequently used to automatically identify people in other photos.

Target gets intimate with its female customers

In 2012, the New York Times revealed a carefully designed customer relations program at the department store Target that set out to identify customers who were likely to be pregnant, in order to subsequently direct-market lucrative early childhood products (Duhigg 2012). Because the U.S. has no data privacy laws governing the private sector generally (Greenleaf 2011) the rights and wrongs of retailers divining such intimate insights about their customers are difficult to arbitrate in that country. But in Australia, it would likely to be unlawful. Health information here is included in legislation amongst 'Sensitive' Personal Information. This category includes information (or opinion) about an individual's health or genetics, as well as their racial or ethnic origin, political opinions, membership of a political association, membership of a professional association, trade association or trade union, religious beliefs or affiliations, philosophical beliefs, sexual preferences or practices or criminal record (Privacy Act 1988:17). Special conditions apply to Sensitive PII. Most relevant for Big Data is that Sensitive PII in Australia can only be collected with the prior informed consent of the individual concerned. Should Australian stores wish to use Big Data techniques as Target did in the U.S., they may need to disclose up front the possibility of health information being extracted from shopping data and obtain customers' express consent for that to occur. Data miners need to be aware that rights-based privacy laws set a low bar for privacy breaches: simply collecting Sensitive PII may constitute a technical breach of the law even before that PII is used for anything or disclosed. For example, Google's breach of Australian law in the case of the StreetView Wi-Fi incident was confined to the Collection Principle; no use or further disclosure of the Wi-Fi data was found to have occurred (OAIC 2012).

More privacy shocks are likely to come

Digital businesses are availing themselves of an ever richer array of signals created automatically as we go about our lives online. Wearable technologies in particular, such as augmented reality eyewear and personal fitness monitors, measure a number of parameters more or less continuously, and typically transmit the data back to their manufacturers for processing and recording. Ostensibly the users of these devices benefit by way of automated reports, but it is worrying that the privacy policies of many technology businesses tend to be silent on what they plan to do with the by-products of the processing. Natural language processing of spoken commands, face recognition, object recognition and 'quantified self' monitoring equipment all generate rich traces and metadata about the devices' users, with enormous commercial value.

At the same time, informaticians are discovering ever more clever ways to de-anonymise us in cyberspace (that is, to undo the state of anonymity that users expect or have been led to believe applies). In one of the more spectacular recent examples, self-described 'DNA hackers' at MIT's Whitehead Institute for Biomedical Research in 2012 worked out how to combine publicly available genealogical data with anonymously donated DNA samples in the 'Thousand Genomes' research program, to identify a number of those donors (Bohannon 2013). This was despite reassurances given to donors in the informed consent form that 'it will be very hard for anyone who looks at any of the scientific databases to know which information came from you' (Thousand Genomes 2008).

Do these developments mean 'privacy is dead' after all? No. The fact is that *anonymity* is threatened by information technologies, but anonymity or secrecy is not the same thing as privacy. The act of undoing anonymity creates new named data records and thus represents an act of PII collection subject to privacy regulations in a great many jurisdictions. The de-anonymising of Thousand Genomes donors for instance can be seen to be a breach of certain privacy regulations, independent of the conditions of the original raw data collection (Wilson 2013). If law-abiding Big Data businesses are alert to the broad definition of PII and to the technology neutrality of data privacy regulations, then they can reduce the chances of more shocks to come.

Extended privacy principles to deal with Big Data

Big Data represents a qualitative shift in how business is done in the digital economy rather than just a quantitative change. The term 'Big Privacy' is used by some to describe an organised response to Big Data. Former Information and Privacy Commissioner for Ontario, Ann Cavoukian for example has written 'Big Privacy is Privacy by Design writ large' (Cavoukian & Reed 2013:6) implying that dealing with Big Data essentially requires just a redoubling of existing privacy measures. To the contrary, I contend Big Data demands some new ways of safeguarding PII flows, with an update to traditional privacy principles.

As we have seen, to many technologists' evident surprise, principles-based privacy laws have proven powerful in constraining Big Data processes, even though scenarios like facial recognition in social networks could not have been envisaged 30 years ago when the OECD first formulated its privacy principles. When we appreciate that generating PII out of raw data is a form of indirect collection of PII, orthodox privacy principles apply and can restrain what may be done with Big Data's outputs. The Collection Limitation Principle is perhaps the most fundamental privacy control; it is after all, the first of the OECD Privacy Principles (OECD 1980). And yet transparency is crucial too. Traditional privacy management entails telling individuals what information is collected about them, when it is collected and why. With Big Data, even if an information company wants to be completely transparent, it may not be able to say today what PII it's going to synthesise tomorrow. Any promise of openness with Big Data cannot be made once and forgotten; it needs to be continually revisited.

There is a bargain at the heart of most social media businesses today in which PII is traded for a rich array of free services. Sophisticated Internet users may know 'there is no such thing as a free lunch' and that the things they take for granted online – like search, maps, cloud email and blogging – are funded through the monetisation of PII. And there is nothing intrinsically wrong with business models that extract valuable PII from anonymous raw data; however privacy requires transparency. Today's service-for-PII bargain is mostly opaque, with personal data harvested seamlessly and covertly, with nary a mention in privacy policies.

The fact that there is a real price to pay, one way or another, for things like online social networking has led some privacy advocates to call for overt user-pays models of digital service delivery. The new 'Respect Network' for example, founded by a team of personal cloud and sharing economy advocates, aims to provide a social logon button which is 'not based on advertising or surveillance' but rather which is underwritten sustainably by crowd funding and subscriptions (Blum 2014). It remains to be seen of course just how many users are sufficiently worried about privacy and, moreover, aware of how their PII is exploited to make the switch to user-pays digital services. For others, ignorance is not bliss. Social network members in general deserve to be told about the PII exchange (including details of what information-based businesses do with all this PII) so they can make up their own minds about it.

Going beyond classic data privacy principles

While existing privacy principles are surprisingly powerful, they are limited by virtue of being framed for the static data collection and processing capabilities that characterised information and communication technology until recently. A traditional privacy policy properly sets out what PII is collected, why it is collected, when, where and how it is collected, and to whom it may be disclosed. A privacy policy reflects a business model in which PII flows into and about a business and tangible benefits result. A fair privacy policy accurately reflects an exchange of PII for value. But with Big Data, the PII-value exchange is shifting all the time, both quantitatively and qualitatively. For data protection to remain a good fit for evolving Big Data practices, certain privacy principles could be updated as follows.

Super transparency

If basic data privacy means being open about what PII is collected and why, then privacy into the future, where business models and data mining techniques are evolving rapidly, should take transparency further. As well as telling people what information is collected and why, businesses should be candid about the business models, the emerging Big Data tools, and what sort of results data mining is expected to return. In keeping with better visibility, users should be offered ways to opt in and out and in again, depending on how they gauge the returns on offer from Big Data participation.

Engage customers in a fair deal for PII

The nascent digital economy is distorted to some extent by savvy digital citizens modifying their behaviours to protect themselves in ad hoc ways against online exploitation. Most pointedly, 30% of Australians have been found to have given a false name and 32% have given false personal details in an effort to protect their personal information (OAIC 2013:30). Many resort to covering their tracks with encrypting browsers like Tor or by maintaining multiple email addresses so when they register for disparate services, it's harder to join up their activities. There's nothing wrong with having multiple digital personae, but being forced to concoct them in order to hide from prying eyes is an unfair burden and ultimately counter-productive. Consumers and companies alike will do much better by engaging in a more overt deal which sets out what PII is really worth and offers a fair price for it, whether that is by way of tangible products, services, discounts or explicit payment.

Dynamic consent

As we've seen, when Sensitive PII like health information is collected – whether directly or indirectly – the prior consent of the individual is required. If such collection is going to be through the mining of relatively innocuous data like shopping habits, then we have a dilemma. Around the time of raw data collection, businesses could try to disclose all conceivable future outcomes of data mining, yet they may understandably be reluctant to confront customers with dense data usage agreements full of possibly hypothetical scenarios. Alternatively, as and when data custodians develop new data mining techniques and find they are able to extract fresh heath information, they could seek permission at that time.

It will take artful user interface design to present individuals with the types of PII that might be extracted about them, in a meaningful way, such that they can make sound decisions about whether or not to grant permission. There is a logical contradiction in the letter of the Collection Limitation Principle. Consider a hypothetical scenario where a Big Data company develops a way to predict from your travel patterns that you are at risk of a contagious disease, and they would like to bring that possibility to your attention. Strictly speaking, any determination about an aspect of someone's health (even if wrong) is a piece of Sensitive Personal Information, and working it out (i.e. collecting it) without consent is not permitted by Australia's Privacy Act. So by the time the company brings the risk of contagion to your attention, the company has already breached the law.

This Catch-22 could be resolved by giving users a purposefully vague indication of 'what we might know about you' in a sort of graphical dashboard. Such a user interface could serve to remind the user of the raw data that is available to the company, and illustrate how particular processing can extract more detailed information about them, such as diseases, without yet being specific. If the user perceives benefits from such processing, they could indicate their consent to proceed. In some cases, the quality of 'what might be known' about a user may vary in proportion to the amount of raw data the user is willing to give permission for. In the hypothetical case of predicting disease from travel data, the confidence limits on the prediction might be improved if the user agreed to provide more history or to link other data into the calculation; conceivably the user could be presented with graduated interactive controls over the amount of data they agree may be factored into the Big Data process and indications of the differential benefits to be expected.

Conclusion: Making Big Data privacy real

A Big Data dashboard like the one described could serve several parallel purposes in aid of progressive privacy principles. It could reveal dynamically to users what PII can be collected about them through Big Data; it could engage users in a fair and transparent exchange of value-for-PII transaction; and it could enable dynamic consent where users are able to opt in to Big Data processes, and opt out and in again, over time, as their understanding of the PII bargain evolves.

Big Data holds big promises, for the benefit of many. There are grand plans for population-wide electronic health records, new personalised financial services that leverage massive retail databases, and electricity grid management systems that draw on real-time consumption data from smart meters in homes, to extend the life of aging 'poles and wires' while reducing greenhouse gas emissions. The value to individuals and operators alike of these programs is amplified as computing power grows, new algorithms are researched, and more and more data sets are joined together. Likewise, the privacy risks are compounded. The potential value of Personal Information in the modern Big Data landscape cannot be represented in a static business model, and neither can the privacy pros and cons be captured in a fixed policy document. New user interfaces and visualisations like a 'Big Data dashboard' are needed to bring dynamic extensions to traditional privacy principles, and help people appreciate and intelligently negotiate the insights that can be extracted about them from the raw material that is data.

References

Blum, D. 2014. 'Update on Personal Clouds' at *https://info.respectnetwork.com/update-on-personal-clouds/* retrieved August 19, 2014.

Boeth, R. 1970. 'The assault on privacy; Snoops, Bugs, Wiretaps, Dossiers, Data Banks – and Specters of 1984'. *Newsweek,* July 27, 15-20.

Bohannon, J. 2013. 'Genealogy databases enable naming of anonymous DNA donors'. *Science* 339(6117): 262.

Cavoukian, A; Reed, D. 2013. 'Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design' at *http://www.privacybydesign.ca/content/uploads/2013/12/pbd-big_privacy.pdf* retrieved August 22, 2014.

Dewri, R; Annadata, P; Eltarjaman, W; Thurimella, R. 2013. 'Inferring Trip Destinations from Driving Habits Data'. *12th ACM Workshop on Privacy in the Electronic Society WPES*, Berlin, 267–272.

Duhigg, C. 2012. 'How Companies Learn Your Secrets'. *New York Times* at *www.nytimes.com/2012/02/19/magazine/shopping-habits.html* retrieved August 22, 2014.

Facebook. 2014. 'Data Use Policy'. *https://www.facebook.com/full_data_use_policy* retrieved August 18, 2014.

Greenleaf, G. 2011. 'The influence of European data privacy standards outside Europe: implications for globalisation of Convention 108'. *International Data Privacy Law*. ips006.

GSA General Services Administration. 2014. 'Rules and Policies - Protecting PII - Privacy Act', *at* *http://www.gsa.gov/portal/content/104256* retrieved August 18, 2014.

HmbBfDI Hamburg Commissioner for Data Protection and Freedom of Information. 2012. 'Proceedings against Facebook resumed' at *https://www.datenschutz-hamburg.de/fileadmin/user_upload/documents/PressRelease-2012-08-15-Facebook_Proceedings.pdf* retrieved August 22, 2014.

Johnston, A; Wilson, S. 2012 'Privacy Compliance Risks for Facebook'. *IEEE Technology and Society Magazine*. 31(2), 59-64.

Michael, K; Miller, K. 2013. 'Big data: new opportunities and new challenges'. *IEEE Computer*, 46(6), 22–24.

OAIC Office of the Australian Information Commissioner. 2012. 'Google Street View Wi-Fi Collection: Statement from Australian Privacy Commissioner, Timothy Pilgrim' at *http://www.oaic.gov.au/news-and-events/statements/privacy-statements/google-street-view-wi-fi-collection/google-street-view-wi-fi-collection-statement-from-australian-privacy-commissioner-timothy-pilgrim* retrieved August 23, 2014.

ISSN 2203-1693. Copyright © 2014 The Author(s). http://doi.org/10.7790/ajtde.v2n3.54

OAIC Office of the Australian Information Commissioner. 2013. 'Community Attitudes to Privacy survey' at *http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-reports/Final_report_for_WEB.pdf*.

OECD Organisation of Economic Cooperation and Development. 1980. 'Guidelines on the Protection of Privacy and Transborder Flows of Personal Data' at *http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm#guidelines* retrieved August 29, 2014.

Privacy Act 1988, at *http://www.austlii.edu.au/au/legis/cth/consol_act/pa1988108/*.

Privacy Amendment (Enhancing Privacy Protection) Act 2012, at *http://www.comlaw.gov.au/Details/C2012A00197*.

Richards, N. 2014. 'Privacy is Not Dead—It's Inevitable' *Boston Review*. *www.bostonreview.net/blog/neil-m-richards-privacy-not-dead* retrieved August 23, 2014.

Sawant, N; Li, J; Wang, J. 2011. 'Automatic image semantic interpretation using social action and tagging data'. *Multimedia Tools and Applications* 51(1), 213-246.

Singh, A. 2013. 'Is Big Data the New Black Gold?'. 2013. *Wired*. *http://www.wired.com/2013/02/is-big-data-the-new-black-gold/* retrieved August 18, 2014.

Solove, D. 2006. 'A taxonomy of privacy'. *University of Pennsylvania Law Review*, 477-564.

Tene, O; Polonetsky. J. 2013. 'A Theory of Creepy: Technology, Privacy and Shifting Social Norms'. *Yale JL & Tech*. 16: 59-134.

1000 Genomes Project. 2008. 'Consent to Participate' at *http://www.1000genomes.org/sites/1000genomes.org/files/docs/Informed%20Consent%20Form%20Template.pdf* retrieved August 23, 2014.

Wilson, S; Connolly, C; Denney-Wilson, E. 2005. 'Patient Privacy and Security - Not a Zero Sum Game! *Australasian Epidemiologist*. 12(1), 11-16.

Wilson, S. 2013. 'Legal Limits to Data Re-Identification'. *Science* 339(6120), 647.

ISSN 2203-1693. Copyright © 2014 The Author(s). http://doi.org/10.7790/ajtde.v2n3.54