1. Introduction

‘Big Data’ is a misleading term that has arisen to describe a new phenomenon of data. Big Data encompasses “data mining” and “predictive analytics,” but it is much more than this. Big Data is in many ways a data revolution. As one expert has put it, “the real revolution is not in the machines that calculate data but in data itself and how we use it.”

While there is no rigorous and widely accepted definition of Big Data, the term came to existence due to the sheer volume of information capable of collection and storage by computers. There simply isn’t a term that adequately describes this volume. Big data is often used alongside cloud computing, and refers to both the collection of giant amounts of data as well as the software developed to extract information and exploit big data for a range of uses. We will refer to big data software as ‘BD Tools’ for clarification purposes to distinguish it from, collection of information.

The legal, ethical and technical issues that surround big data will make it as challenging to govern and regulate (if indeed this path is chosen) as the Internet itself. In the early days there was great excitement about the freedom the Internet offered, this included freedom to choose a new identity, or to disguise your identity. Your gender, race, sexuality, and age could be re-invented. One popular early internet image was that of a dog behind the computer, revelling in not being identified as a dog.

---

It was thought that the Internet offered the ability for any user to assume a different identity, perhaps even to be anonymous, to not be identified at all. The ability to re-invent one’s identity and to access vast amounts of information online became a concern to governments and organisations. Initially, and perhaps unsurprisingly, non-democratic regimes led the way in regulating what citizens could view. Former U.S. President Bill Clinton said China’s desire to regulate the Internet was “like trying to nail jello to a wall.” However, as the Internet evolved it has become perhaps the most regulated technology of all. No longer is the regulation, mass capture and retention of online data limited to non-democratic regimes, these tools have also become irresistible to Western democratic governments as well.

New cartoons based on the famous dog image now depict a new reality – where corporations and governments are able to learn significantly more about you precisely because you use the Internet.
The scenario in this cartoon above would require less than a percent of the data available about the computer and the dog.

Big Data will bring a new era of change to how data is collected and how it is used. It is not simply the case that more will be available about the dog’s identity (though clearly there will be significantly more information available). The famous Internet dog cartoon might look something more like this in the era of Big Data:
Big Data doesn’t necessarily care about acquiring more information about one dog’s identity, or even identifying the dog itself. Big Data often isn’t collected for a particular use or with a specific purpose in mind, and it may be more useful for this reason (although use for a purpose other than the one it was collected for often creates methodology and validity issues, hence the traditional prohibition on doing this in Privacy Principles such as those in the Australian Privacy Act).

Big Data collects vast amounts of often unconnected and disparate data, and applies Machine Learning pattern matching and other techniques to discern patterns in it, and propose associations. Attempting to predict patterns of behaviour however, is both the most extreme promise of big data (in the form of “predictive analytics”), and also its most potentially controversial over-reach (since association does not necessarily mean causation, and causation is a much better basis for prediction).\(^3\)

Data does not need to be fully categorised or identified ahead of time for it to be taken up and harnessed by Big Data tools; equally, it can selectively strip or discard certain attributes during uptake.

---

For this reason, Big Data is touted as having the useful feature of enabling “de-identification” and “re-identification”, and also of enabling the extraction of valuable and useful insights from de-identified data sets.

De-identification is the removal, stripping or obfuscation of directly identifying elements from a data record or set such that the result is not immediately identifiable as associated or linked with a particular individual.\(^4\)

Re-identification is the ascertainment of a link between an individual and a set of data which had either deliberately (after de-identification) or by circumstance, initially appeared not to enable making any such link with a particular person. Big data methods offer the means for automating associative rule generation using Machine Learning algorithms, which derive rules from very large data sets. Among the many functions offered by these techniques are improved tools for identification or re-identification on a probabilistic basis, and others for pattern recognition, voice recognition, face recognition and similar content analysis tools that can be potentially tied back to metadata collections. For instance, if you combine face recognition at a shop cash register with face recognition at the entry to a shopping centre sited next to a passive mobile phone ping generator, you may be able to re-identify a person’s path travelling through the centre using only the (intended to be non-identified) TMSI temporary session identifier.\(^5\) The result is that re-identification or de-anonymisation is now tractable to cheap and effective means, where once it may have been prohibitive.\(^6\)

It is often believed that, because the primary objective of certain initial data collection processes was not identification, or dealing in identifiable records, and since big data technologies can work with de-identified data, that big data practices fall outside of the privacy regime.

The problem with this analysis is that it fails to highlight a glaring problem with de-identification, or the use of data not initially tied to an identifiable individual. While de-


\(^5\) Personal correspondence, author Vaile with investigative journalist in Brisbane, 13 March 2014.

identification in the collection of data is possible (identifying elements can be stripped out or otherwise rendered less useful for identification before uptake), and some data records are in the first place not easily identified with an individual, the more data is collected (i.e. the more Big Data tools come into play), the more likely re-identification becomes.

In other words, tools and cross-correlated data sets can be applied to the collected data in the above figure, and with the tetrillion points of data, re-identifying the dog is not only possible, it is almost certain. Not only does re-identification of a de-identified record become possible with sufficiently advanced Big Data tools and data sets, but the identification could take place at the DNA level of the dog.

Why then, is de-identification and re-identification a problem for law? At their core, many privacy laws only come into play when data or information is linked to a person, or is capable of the identification of a person. If big data sets or records don’t fall within the relevant legal definition of “personal information”, “personal data” (or its equivalent) then they may not be subject to privacy law in a given jurisdiction, so a wide latitude can be given for their use and transfer. On the other hand, if big data tools can in effect re-identify previously unidentified or de-identified data, this may bring data previously outside of the privacy regime, (or related legal regimes) back into coverage.

Although many important privacy issues warrant close analysis and rigorous debate in the context of big data, this article has a modest role. We seek to look at the definitions of personal data or information (or their equivalent) in Australia, the United States and the European Union, in the context of identification, identifiability and Big Data.

The basic question becomes: to what degree does the ability to de-identify or re-identify information bring Big Data practices within the scope of privacy legislation?

The answer to this question is of interest to an increasing range of industries and activities, from online behavioural advertising and medical research through to communications ‘metadata’ surveillance and beyond.\(^7\)

---

We will initially focus on metadata in the telecommunications sector as a key example, before examining other non-metadata material, such as photographs exposed to automated face recognition based on pattern matching and Big Data tools. From these examples, it is worth noting that not all Big Data is metadata, and not all metadata is of interest to us in this exercise. This article initially focuses on those items of telecommunications metadata that are of interest when manipulated by big data tools because they can shift the underlying activity in or out of regulatory scope, in this case, categorisation as, identifiable (regulated), non-identifiable or de-identified (not regulated).

Metadata of any sort typically has semantic connections and meaning embedded, like partly digested and ready to process nutrients, or like a data set that suggests its own database structure. Communications metadata is fast, cheap, and easy to use with big data tools, highly semantically meaningful in its raw state, and open immediately to significant uses, experimentation or rule extractions. Far from being a sort of second-rate, minimalist, marginally-nutritious “undigested roughage” of the communications diet, metadata is sometimes a natural and preferred feed stock for Big Data tools, such as online communications mass surveillance.\(^8\)

As noted above, Big Data tools are valuable as a method to re-identify previously unidentified or de-identified communications and traffic metadata. While underlying infrastructure has been steadily adding individual data items, which are often not considered sufficient to identify anyone on their own, big data tools (BD TOOLS?) and data models take these new items and apply the functionality which enables identification, and can thus render the formerly unidentified as identifiable.

This paper compares the various definitions of “personal data” (or its equivalent) used in Australia, the European Union and the United States, and considers the significance of the differences in definition and scope, and how the definitions are changing or resolving in how they characterise various communications metadata. We explore how certain items of metadata associated with online or mobile communications do or do not fit within the definitions in key privacy instruments.

We start by discussing Big Data and BD tools in the context of various forms of data. Next, we examine various definitions of “personal data” and its equivalent, and make comparisons between the different definitions. Then we examine different factors leading to “identifiability”. We move on to communications metadata, its basis for identification, and

---

the contexts in which it might satisfy the definitions of “personal data” or equivalents. Lastly, we look at implications for categorisation, and for privacy in general.

2. Metadata and Big Data

Big Data can use both structured and unstructured data, and of structured data, both the core message or document content (like the audio of a phone call, the pixels of an image, or the text in an SMS) and the surrounding traffic data or ‘metadata’ which is associated with the message or document.

Metadata is “data about data”. The term can apply to both ‘structural’ metadata, that documents the design specifications of data structures and the containers of data (this is the ISO 11179 "traditional" usage⁹), and also to other forms such as ‘descriptive’ metadata which refers a well-defined scheme applying to aspects of individual data content. In this paper we use metadata in the latter sense, since this can be embedded in, readily associated with, or wrapped around the data content, rather than being held externally as a form of system documentation. Examples of this central form of metadata include the “meta tags” associated with a web page, photographic metadata standards from particular industry sources, the metadata field tags in data structures like XML, and others discussed below.

2.1 Metadata from mobile or fixed telephones

The classic telephone metadata includes date and time of a call’s start and end, and number dialled from and to. It was the mass collection of this metadata that triggered renewed interest in the topic in 2013.¹⁰

There are a large number of potential identifying metadata items generated by modern devices like smartphones when interacting with the telephone system. These include IMEI, IMSI, TMSI, SIM card and telephone number, and are discussed further below.

Some of these are retained or logged by the local tower access point, and some are propagated further up the channel, and be ‘becoming more easily accessible remotely’.

2.2 Metadata from Internet connectivity

When it comes to Internet connections, whether using a mobile phone platform or a traditional wired broadband network, there are a range of other potential metadata identifiers, including IP address (assigned by ISP and perhaps re-used), MAC address...

---


(assigned by network device manufacturer to the network interface), and higher level items like browser history, browser fingerprint, and perhaps unique device fingerprint.

In addition, each connection and request to an online service such as a web page generates session metadata, including URL or DNS addresses. With modern Web based Cloud Software as a Service, there are a large number of clicks on say an email editing page, and each of these may contain information of significance to either identity of parties involved or the content (though much of this is obfuscated).

2.3 Metadata as Big Data

As noted above, metadata can be the ideal input into Big Data systems, because it is embedded with semantic meaning. For instance, the analysis of a telephone call audio content, as well as requiring a warrant in some jurisdictions, also requires a great deal of human or computer attention and pattern matching and meaning extraction.

By contrast, the calling and receiving number metadata, and the time of call start and end (the traditional telephone “traffic data”) can instantly be slotted into an analysis system which recognises both their data type and their significance. When analysed against a large set of other similar data, plus the association available from industry and other sources, and previous insights, a quite substantial array of derivative new insights can be extracted with minimal effort integrating and understanding what the data means.

In addition, communications metadata is also tractable to a wide variety of re-identification strategies. See below.

2.4 Identifiability and other forms of data

As noted above, non-metadata information is also useful to support identifiability, depending on what tools, and what training or comparison data sets, are available.

Some Australian government phone systems are already using voice recognition to authenticate identity. The identification or re-identification comes from servers at the call centre; if the data were to be made available within government it may render raw voice data as an identifier, after processing and comparing against massive templates and data sets. [CITE]

As mobile computing takes off, the built in cameras on smartphones will come to be very widely used, and there is a prospect that routine communications of photographs may become tractable to re-identification tools.

2.5 Big Data

Data is the lifeblood of most digital enterprises today, and the preference is for personal data. Many social media business models are fuelled by a generally one-sided bargain for
personal data, where the customer (or in many cases, like say, Facebook, the user – the customer of this advertising business being the advertiser, not the user) accepts access to an online service on the basis of un-informed consent to unspecified collection, dealing and distribution of their personal information. Data analytics and data mining can pull personal data almost out of thin air. And so department stores now have the power to predict when female customers have become pregnant, and social network operators have the power to link candid photos to profile photos via facial recognition and thus infer identities and relationships. Although enthusiasm for predictive analytic tools is reaching fever pitch, some business commentators are starting to fear that digital government and business may have already gone too far in their enthusiasm for Big Data, sacrificing the trust of their users for short term commercial or political gain. Others are raising doubts about many of the claims of methodological efficacy made or implied by proponents.

Big Data promises great benefits for many stakeholders. For example, town planners can extract detailed patterns in the way people and goods move about our urban environments; maintenance engineers can predict wear and tear from how infrastructure is being used; retailers can detect how people actually behave in stores without relying on self-reported ‘likes’ and complaints; law enforcement officials can pick out suspicious communications suggestive of criminal preparations; medical professionals and epidemiologists can discover deep patterns in health and lifestyle data to help better manage the entire population’s future well-being.

Data analytics isn’t nicknamed “data mining” for nothing. The raw material of Big Data – namely all the ones and zeroes coursing beneath us in the digital environment – is like crude oil, insofar as it contains enormous potential riches to be extracted from an undifferentiated matrix. Consider photo data for instance, and the rapid evolution of tools


15 Mayer-Schönberger and Cukier, op. cit.

16 For a sceptical view of the “data is the new oil” metaphor, see John Naughton, “The web giants pumping us for data,” The Observer, 1 September 2013. At: http://www.theguardian.com/technology/2013/sep/01/big-data-corporations-information
for monetizing it. These tools range from simple metadata embedded in digital photos which record when, where and with what sort of device they were taken, through to increasingly sophisticated pattern recognition and facial recognition algorithms applied to the content. Image analysis can extract places and product names from photos, and automatically pick out objects. It can identity faces by re-purposing biometric templates that originate from social network users tagging their friends for fun in entirely unrelated images. Image analysis lets social media companies 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. These conclusions or associations are potentially embedded in the datasets as new content or often as new forms of metadata around particular documents.

The ability to mine photographic data (often undermining traditional copyright norms) promises a new digital gold rush. Like petroleum engineering, image analysis is very high tech. There is extraordinary research and development going on in face and object recognition. The advertising “infomopolies” 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 Research and Development and in acquiring start-ups. They pay remarkable sums for photo companies like Picasa,\(^\text{17}\) Instagram\(^\text{18}\) and Snapchat\(^\text{19}\): not merely because photos and tagging photos are fun, but because the potential for extracting intelligence from images and their context of use and transmission, appears limitless.

Big Data goes beyond data mining and is really about data refining, as suggested by Figure 1, transforming unstructured information into fresh insights, decisions and value.


Business models for “monetizing” 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 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 API. It is not clear if Facedeals accesses the biometric templates, but nothing in Facebook’s privacy and data use policies restrains them from providing or selling the templates. Facebook has been taken to task for stretching big data beyond expectations of those posting material to their service over years past; more legal surprises await for digital businesses in retail, healthcare and other industries.

3. Definitions of information capable of identifying a person

Different jurisdictions use different terminology to describe the type of information about people which falls within the coverage of their privacy and data protection rules.

These terms are “personal information” (PI), “personal identifying information” (PII), and “personal data” (PD). The definitions of the terms are important as they reveal clues as to whether Big Data is capable of being caught under various privacy and information data legislative instruments. Of particular interest is metadata such as online IP address, but non-metadata is of course also sometimes capable of identifying a person.

---

3.1 Australia: "Personal information" (PI) in the Privacy Act 1988 (Cth)

The Australian Law Reform Commission’s Report on Privacy recommended that the Privacy Act that the definition of “personal information” be changed. Recommendation 6–1 proposed that “The Privacy Act should define ‘personal information’ as ‘information or an opinion, whether true or not, and whether recorded in a material form or not, about an identified or reasonably identifiable individual.”[21]

Prior to changes made in 2012 (in force in March 2014), the definition of personal information in s6 was as follows:

> personal information" means information or an opinion (including information or an opinion forming part of a database), whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion [emphasis added]

After the amendments in Privacy Amendment (Enhancing Privacy Protection) Act 2012 (Cth), the definition of “personal information was subtly changed by the omission of the words qualifying identity be ascertainable “from the information or opinion”. The definition inserted now reads:

> personal information means 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.

The new definition was stated to be adopted to be more consistent with APEC Privacy Framework[22] and other international instruments (i.e. OECD Guidelines, Council of Europe Convention and EU Directive) such that international jurisprudence will be more relevant to the Privacy Act 1998 (Cth).[23]

The effect of this short omission is to make it clearer that the question of identifiability can take into account information other than the data item itself. Various forms of log and related data would appear to be within contemplation, so long as doing so bears the risk of

---

reasonably identifying an individual. If there were ever any doubt that the steps one could take to identify a person starting from a piece of information could extend to referring to other information, this has probably been removed.

This rather Spartan definition is expanded upon by Guidelines from the regulator.

**OAIC Guidelines for the Australian Privacy Principles**

The *APP Guidelines* from the OAIC offer guidance at B.71: Whether an individual is ‘reasonably identifiable’ from particular information about that individual will depend on:

1. the nature and extent of the information
2. the circumstances of its receipt
3. whether it’s possible to identify the individual using resources available to the person or entity that holds or has access to the information

B.71-72 address the question “in the eyes of who should the information be reasonably identifiable?” The answer is the person or entity that holds or has access to the information.

The guidelines use the example of a licence plate number. A lay person is unlikely to have the resources to identify the owner of the license plate and thus would not be holding personal information. However, a car registration agency would most likely be able to use their database to identify the owner, which would make the licence plate personal information.

At B.76, the OAIC Guidelines note that individuals are not reasonably identifiable where steps required to identify the individual are overly expensive or resource intensive, and would seem to strongly suggest ‘reasonably identifiable’ depends on the capability of the person or entity holding the information.

In a related set of Guidelines on mobile privacy, OAIC notes that what constitutes personal information will vary “depending on what can reasonably be ascertained in a particular circumstance, but may include:

- photographs
- Internet Protocol (IP) addresses, Unique Device Identifiers (UDIDs) and other unique identifiers in specific circumstances

---

• contact lists, which reveal details about the contacts themselves and also a user’s social connections
• voice print and facial recognition biometrics, because they collect characteristics that make an individual’s voice or face unique location information, because it can reveal user activity patterns and habits”. 25

It seems clear from this that a range of metadata items like IP address may fit within the new definition with no fundamental difficulty. One commentary suggests examples when IP Addresses may constitute personal information:

1) where a person has a permanent IP address
2) when a record previously made by the website operator exists and is accessible, or
3) when (dynamic) IP address generated by dial-up ISPs can be linked to the log books of the ISP (only permissible in limited circumstances, and a warrant may be necessary)

Characteristics of the website must also be taken into account in determining whether or not an IP address can be deemed personal information (e.g. when access to the site is via a proxy server or anonymising software, it may be impossible to identify an individual by any means, thus the IP address would not be classified as personal information). 26

A recent decision by the regulator confirms that the new definition is broad enough to catch information which needs to be processed alongside external information sources (though it hints that this may have already been the case), in paragraph 44:

The main textual difference between the current and the new definitions is that the new definition will not explicitly require that an individual be identifiable ‘from the information or opinion’. Information in a government document can qualify as ‘personal information’ even though some additional action – such as an internet search or data-matching – must be undertaken to identify an individual. However, it is

---


probable that the current definition applies in those circumstances, and that the change will merely resolve any doubt.27

3.2 US: “Personally Identifying Information” (PII)

While the Privacy Act of 1974 regulates the collection of personal information by federal government agencies, there is no overarching federal law regulating private entities; some states have their own laws, such as California’s Online Privacy Protection Act of 2003. There thus appears to be no single term in the US, though in general, as will be seen, the definition is narrower than in Australia or the EU.

The most often used term is “personally identifiable information” (PII), but there are variations on terminology and meaning. So for instance the U.S. Health Insurance Portability and Accountability Act (HIPAA) refers to “individually identifiable” information, and includes information “1) that identifies the individual; or 2) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.”

The latter phrasing of this medical records law has some parallels to PI in the Australian context, but this is not widely used outside of health insurance, and it appears in practice that even HIPAA works on a fixed set of 18 data items which are considered identifiable elements, and everything else falls outside of this category.28 (Most of these would fall outside the metadata category.)

The most widely used and mainstream definition for PII appears to be the Government Accountability Office expression of an amalgam of the definitions of PII from the Office of Management and Budget:

“PII is — any information about an individual maintained by an agency, including

(1) any information that can be used to distinguish or trace an individual’s identity, such as name, social security number, date and place of birth, mother’s maiden name, or biometric records; and

(2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.”

This makes no mention of communications or location metadata, or similar data types.

*California Senate Bill 1386*, a data breach law, is another example: its definition of personal information includes Social Security numbers, driver’s license numbers, financial accounts, but not, for example, email addresses or telephone numbers (perhaps in part because of its special focus).

The *Children’s Online Privacy Protection Act of 1998* (COPPA), by contrast, states that Personal Information is:

Individually identifiable information about an individual collected online, including a first and last name; a home or other physical address including street name and name of a city or town; an e-mail address; a telephone number; a Social Security number; any other identifier that the Commission determines permits the physical or online contacting of a specific individual; or information concerning the child or the parents of that child that the Web site collects online from the child and combines with an identifier described in this paragraph.

This would appear to exclude IP addresses or other metadata, unless directly linked by the collecting site to one of the named identifiers. As discussed above, “de-identification” has been a widely practiced method of moving information out of the PII category in the US, using various data obfuscation methodologies. This is apparently on the assumption that there are only a fixed small set of data items which can be used to identify a person, and if you deal with them, what you have left is no longer identifiable. This assumption is coming under increasing scrutiny, in part because the de-identification methods relied on simplistic fixed data sets prone to breaking when faced with the more sophisticated, open-ended tools of big data. As stated by experts Narayanan and Shemitikov, “The versatility and power of re-identification algorithms imply that terms such as “personally identifiable”

---


and “quasi-identifier” simply have no technical meaning. While some attributes may be uniquely identifying on their own, any attribute can be identifying in combination with others.”

However, this caution from the technical community may not have weakened the attachment in business and some US government circles to the helpful notion that there can be sets of metadata, or indeed other data which are somehow not capable of supporting identifiability, despite their origin with a person.

3.3 Europe: “Personal Data” (PD) definitions

Across the Atlantic, the 1981 Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, also known as Convention 108, offers in Article 2(a) an important definition of “personal data”:

“personal data” means any information relating to an identified or identifiable individual (‘data subject’) 

The OECD Privacy Framework, in its July 2013 first revision of the original 1980 Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data (Privacy Guidelines), contains a definition of “personal data” on page 13, paragraph 1(b) in identical terms to Convention 108 above. (Australia is a member of the OECD, and its privacy law derives from this source, as does much of Europe’s.)


Personal data shall mean any information relating to an identified or identifiable natural person (“data subject”); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

---

33 Narayanan and Shmetikov, op cit.
The broad definition in the Directive is refined in its recital 26: "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person".

The term was considered in the case of Bodil Lindqvist (Case C-101/01). The European Court of Justice decided that "referring, on an internet page, to various persons and identifying them by name or by other means, for instance by giving their telephone number or information regarding their working conditions and hobbies, constitutes the processing of personal data". Above provides an analysis of the definition of “personal data”, part 3. The analysis breaks down the definition into four building blocks. These are:

1. “any information”
2. “relating to”
3. “an identified or identifiable”
4. “natural person”

Example No. 15 on pages 16-17 outlines the Working Party’s consideration of whether IP addresses constitute “data relating to an identifiable person.” The Working Party concludes that “unless the Internet Service Provider is in a position to distinguish with absolute certainty that the data correspond to users that cannot be identified, it will have to treat all IP information as personal data, to be on the safe side.” The meaning of “directly or indirectly identifiable” is covered on page 12-13.

The 2007 Article 29 opinion also considers when an individual is identifiable in light of "all the means likely reasonably to be used either by the controller or by any other person to identify the said person". This requires a range of factors to be considered, including the purpose of the information, the structure of the processing, the advantage to the data controller, the interests of the individual and the risk of technical or organisational failures. The extent of reasonable means are set broadly, so a third party in possession of, for instance, an IP address might apply through the courts to an ISP in order to identify the name and address of the subscriber attached to an IP address. An IP address in the hands of any party must be regarded as potentially personal data. On its face this broad interpretation


39 This draws on material from Linklaters, Technology, Media & Telecommunication News, “EU – What is Personal Data?” 1 October 2008, at: <http://www.linklaters.com/Publications/Publication1403Newsletter/PublicationIssue20081001/Pages/PublicationIssueItem3513.aspx> (accessed 20/3/14)

may seem like a controversial position. However, when one considers how easy it is in the era of Big Data to identify, or re-identify an individual, this no longer seems like a broad interpretation, but a critical one. 41

Pseudonymised, coded and ‘anonymous’ data is also considered. A pseudonym is an identifier that uniquely applies to one entity, but is not linked to other identifiers such that the person can be easily linked to the record. Coded data is similar, and widely used in medical research: the traditional identifying items like name and address are removed and stored in a secure lookup table, and a code with random or non-significant content is added in their place, with access to the lookup table being the only way to re-associate the data with a known individual. 42 ‘Anonymous’ data is assumed to not be reasonably able to be linked to an identified individual; there are generally no lookup tables, although there may be random record identifiers.

If this pseudonymisation or coding can be reversed to identify the individual, by the data controller or anyone else, it is likely to be “personal” data. Attempts to anonymise and de-identify data are increasingly looked on with caution, in part because of demonstrated failures, and in part because of the advent of more sophisticated big data methods of re-identification. 43

UK law is generally consistent with EU principles in this area, although there is a UK Court of Appeal case, Durant, 44 which adds two extra requirements to block 2 above, “relates to the individual”, including required connection with “biographical significance”. The Article 29 Working Party apparently regards Durant as compromising the protection afforded to individuals by the Data Protection Directive without these two extra tests for PD, and as a

---


44 Durant v Financial Services Authority [2003] EWCA Civ 1746.
result the European Commission considers the UK to have failed to implement the Directive adequately.\textsuperscript{45}

Moving up to the current time, recent moves to reform existing EU law on data protection now appear confirmed.\textsuperscript{46} In Article 4 (2) of the new regulation,\textsuperscript{47} “Personal data” means any information relating to a data subject. “Data subject” means:

an identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.

By explicitly including both location data and online identifiers, it seems clear that a wide range of communications metadata may fall within the latest EU definition.

### 3.4 Comparison of the different definitions

As noted above, the extended new Australian PI definition, information about “an individual who is reasonably identifiable”, not actually identified, and not necessarily identifiable from the information itself, appears in text and practice to be broader than most versions of PII, because of the more explicit and implicit understanding that information other than the data in question can play a role in determining identifiability for Australian purposes.

For instance, it is clearer that Internet device IP addresses could typically be PI, and other metadata identifiers could also fit in some circumstances. The characterisation in the US of say an IP address as PII is less certain and often contested.

Given the various European definitions of PD noted above, it is difficult to compare the more stable Australian PI definition, but it generally appears to be very similar in concept, in particular by accepting that recourse to external data sources can be had in the course of reasonable efforts to identify. This may include traffic and metadata like IP address; both presumably contemplate access to other stores of data, perhaps even including face recognition based on large data sets.

\textsuperscript{45} For more on this topic, see generally Lee A. Bygrave, \textit{Data Privacy Law: An International Perspective} (Oxford University Press, 2014)


There appear to be, as with PI, significant differences in scope between PD (as used in Europe) and PII (as mostly used in the US).

As to the question of whether this can cover communications metadata like IP address or any of the other temporary or permanent identifiers, the European usage which accepts that IP addresses can be PD “contrasts sharply to the approach taken in the United States under laws such as COPPA where, a decade ago, the FTC considered whether to classify even static IP addresses as personal information, but ultimately rejected the idea out of concern that it would unnecessarily increase the scope of the law. In the past few years, however, the FTC has begun to suggest that IP addresses should be considered PII for much the same reasons as their European counterparts. Indeed, in a recent consent decree, the FTC included within the definition of “non-public, individually-identifiable information” an “IP address (or other “persistent identifier”).” And the HIPAA Privacy Rule treats IP addresses as a form of “protected health information” by listing them as a type of data that must be removed from PHI for de-identification purposes. 48

However, courts are more reluctant to do so. For example, the Irish High Court held in April 2010 that an IP address does not constitute “personal data” when being collected by record companies for the purpose of detecting copyright infringement. And a U.S. federal district court in Washington State also held that an IP address is not PII because it identifies a computer rather than a person.”49

4. Significance of various factors to "identifiability"

There are a number of other factors which need to be considered when addressing “identifiability”. The OAIC Guidelines (noted in section 3) stated that other attributes could contribute to identifiability. It is also clear that from a technical perspective, it is no longer possible to say that certain metadata items, like IP addresses, can be assured to be not identifiable. Whether they are “reasonably identifiable” in practice will depend on the context. The main factors are listed below.

4.1 Cost

The OAIC APP Guidelines at B.76 note that individuals are not “reasonably identifiable” where steps required to identify the individual are overly expensive or resource intensive,


and would seem to strongly suggest ‘reasonably identifiable’ depends on the capability of the person or entity holding the information.

One factor to take into account here is the phenomenon in IT folklore known as Moore’s Law, which estimates the time it takes for the cost of a given IT product to fall in half (typically 18 months to two years). While there are signs that certain IT areas are reaching a physical limit, the history of the past 40 years has been that features do indeed regularly, typically in a two year period, tend to drop roughly half in price.

To the extent that this continues (and given the advances in software behind big data appear to be unhindered by the physical limits at the atomic level now constraining some hardware refinement, we can expect some continuation), it can be assumed that the relative expense of re-identification techniques using data will also continually fall. Something which may be not identifiable because of cost this year may become identifiable at reasonable costs within a few years. It is likely that this may keep on happening, and the threshold for expense for identifiability will fall over time, bringing more and more of the previously impracticable identifiability problems within easy reach of a solution.

4.2 Technical feasibility

As above, the technical feasibility of re-identification will typically also improve, especially at present when so much innovation is occurring under the rubric of Big Data.

This is not just the incremental evolutionary improvement within the same technology often expected of Moore’s law for hardware, it may also be related to relatively radical new processes, algorithms, data sets, and techniques for pattern matching and re-identification.

In addition, as more sensors are built in to devices, such as accelerometers, the range of parameters which can be recorded and potentially analysed increases.

4.3 Current practice by various entities: who can access the means of identification?

Some communications and online metadata useful for re-identification, is not easily accessed unless by legal compulsion. However, there appear to be a wide range of such means and in any case it appears that some carriers and ISPs are prepared to be persuaded by less than legal compulsion: see for instance the unenforceable operation of s313(1) of the

---


Telecommunications Act (Cth) in Australia as a basis for a de facto Internet censorship filter based on crime prevention, certain litigation against carriers and ISPs by copyright holders seeking mass traffic information, and the widely reported cooperation with NSA spying activities by major US carriers.

This means that there are potentially quite a substantial number of entities who might demand otherwise secret metadata from the source for the purposes of re-identification, and thus the raw metadata may be identifiable in their hands.

4.2 Reliance on information other than the item itself

As noted above, the new Australian definition of PI drops the requirement that the identity be ascertainable from the information itself, although the OIAC Guidelines seem to suggest that this was arguably also the intent of the definition before 2014. In either case it is clearly permissible to rely on external information for the identification step.

A potentially wide range of external tools and data sets may be viable for use to support identifiability by advanced means such as big data, including of course logs and traffic data retained in various places around the Internet. At the most basic level, to massive samples of for instance more or less unique “device fingerprints” based on the detail of every customised or personalised combination of software, hardware and firmware which can be disgorged from the typical web browser.

It is likely that in the face of growing regulatory moves to enable “Do Not Track” rules in the US, the online behavioural advertising industry there (and its global offshoots here) may be moving away from more obvious, blockable discrete tracking tools like Cookies and Bugs to the passive collection of such nevertheless quite functional browser or device fingerprint data, but this in turn may continue to become easier to successfully use to support identifiability. [CITE] This is also likely to be the case in the EU. The position in the US is less clear, and more likely to require use of a specified set of data items, held by the data controller, to do the identification. So while the use of tracking data goes more underground becomes ever more capable, it may be that under US law PII remains unable to address this new technical capacity, often based on big data.

5. Implications of the differing definitions

5.1 Communications metadata which is ‘Personal’ in Australia or Europe but not ‘personal’ in the US?

There are scenarios where information arising from use of a phone or internet connected device would be considered PI in Australia may not be considered PII (and thus not be treated as ‘personal’).
The sort of digital device, online traffic or metadata item to which this may apply include: browser ‘cookies’ and other online behavioural advertising tracking elements or ‘bugs’, a UDID (universal device identifier), ‘geo-location’ information (such as GPS coordinates), the telephone identifiers IMSI (International mobile subscriber identity) and TMSI (Temporary mobile subscriber identity - interesting in that it is supposed to be temporary and hence more secure), an email addresses not tied to your name, medical research file code names (where the lookup table is accessible), a wireless or wired network card’s MAC (Media Access Control) address, and the IMEI (International Mobile Station Equipment Identity) of a mobile phone.

If ‘reasonable’ means exist to re-identify such data items, even if the means are not necessarily available to everyone, then they may fit as PI or PD. They are less likely to be considered PII.52

For example, in the US-based browser cookie and bug detector plugin Ghostery, a comprehensive list of detectable items and the companies that insert them is maintained as part of its customisation options. They recognise this by way of their category names for Data Collected under Privacy Information. In the case of Google Analytics, this is what the entry shows:

**Anonymous** (Ad Views, Analytics, Browser Information, Cookie Data, Date/Time, Demographic Data, Hardware/Software Type, Interaction Data, Page Views, Serving Domains)

**Pseudonymous** (IP Address (EU PII), Search History, Location Based Data, Device ID (EU PII))

**PII** (Phone Number)53

Note the recognition that certain items claimed to be “Pseudonymous” would be considered PII (more correctly, PD) in Europe. Note also the use of the term ‘Anonymous’ for a range of data items which could, with the right forensic tools or the right Big Data resources at the server end, arguably be linked back to a person or a device (or could form part of a package of items used to create a more or less unique “device fingerprint” or “mobility trace”54 that

---

52 On the technical ambiguity of the US definition, see Narayanan and Shmatikov op. cit.
would ultimately identify the device and thus potentially the user). Finally note the one item of PII recognised, phone number.

5.2 Assumptions deriving from US law and practice may not travel well

The consequence of the divergent definitions between the US on one hand and Australia and the EU on the other, is that some of the language usage, industry assumptions and practice in the US may not travel well to the latter jurisdictions.

This is potentially a source of significant confusion, because the US is home to many of the technology companies developing big data tools, and key sectors using such tools, like online behavioural advertising (Facebook, Google and the other 2,000 or so participants identified by monitors like Ghostery), or government metadata surveillance (as practiced by the NSA and telecommunications and Internet firms operating under their direction).

Consequently, some of the earlier and most intensive discussions of the issues arising in these technologies and uses start and develop in the US, typically based on assumptions of US law and the key term PII.

While being US based does not mean that the companies are not subject to regulations outside of the US, it may lead to US centric notions of privacy, and the use of PII as the foundation for discussion, rather than PI or PD. This will lead to different conclusions compared to discussions founded on a more receptive model for identifiability, as found in PI and PD definitions: it would be obvious that the scope of what was ‘personal’ would be broader.

6. Big Data Incidents

This section describes three incidents involving data mining and Big Data. While these examples highlight existing privacy problems with Big Data, and its ability to cross the line between non-identifiable and identifiable data sets, they only represent the tip of the iceberg of potential examples likely to emerge as the tools become more widely used and entrenched.

These examples mainly involve ‘content’ data: communications data streams, image content, transaction details, with some communications metadata also involved, such as SSID, internal record identifiers, or transaction logging information. (Examples involving metadata alone could be found in other corporate settings, and in online surveillance efforts.)

6.1 Google Finds that Public can still be Private

While they drive around photographing towns and cities, Google’s StreetView cars listen for Wi-Fi access points, and collect the geographical coordinates of any they find. Google
collects such Wi-Fi for its geo-location databases, though some reports suggest this was for location based services other than Google Maps.55

On its own, the broadcast identifier of a Wi-Fi access point (known as the SSID, or Service Set Identifier) and its location is not necessarily identified or linked to a person, and as such may not initially count as personal information. (It is unclear whether the capacity of Google to apply other elements of its Big Data archives to re-identification of SSIDs, which later changes to its terms of use in 2012 permitted to be used in this way, was pursued.) In any case it turned out that at least some of the StreetView software was (allegedly inadvertently) also collecting Wi-Fi network traffic, some of which contained potentially identifiable or personal elements like user names, banking details and even passwords. Privacy regulators in Australia56, Japan, France, Germany, Korea, the Netherlands and elsewhere57 found Google was in breach of respective data protection law.

The company asserted that the collection of this wireless communications traffic was inadvertent and mistaken, and that the relevant software code was written by a developer as an experiment and should not have been included in the production StreetView software (though given that the “sniffing” program was apparently in operation globally for up to five years before discovery, and was feeding data into the core geo-location data sets, this explanation raised further questions). Google eventually apologized and destroyed the Wi-Fi traffic data that had been gathered.

Google’s obstruction of the investigation, which led to a US$25,000 fine by the US Federal Communications Commission and complaints elsewhere, made it difficult to ascertain the underlying facts. While this was less than the EU100,00 fine from the French data protection authority levied for the interception of private WiFi data,58 or the million Euro fines from France and Spain for its contemporaneous March 2012 Big-Data-friendly policy

change which saw users agree to aggregation of their data from its over 60 services,\(^59\) it suggests a lack of a reciprocal commitment to transparency of its own actions (perhaps noteworthy for a Big Data pioneer expecting others to accept being the subject of potentially intrusive information gathering).

It could be argued on behalf of Google that the Wi-Fi access point data was broadcast and present in the open, not in some private zone, and was collected by vehicles travelling the public streets; but Data Protection laws in Europe, Australia and elsewhere do not necessarily distinguish “public” from "private" spaces. In fact the word “private” doesn’t even appear in Australia’s *Privacy Act 1988*.\(^60\) If data is identifiable for the purposes of the Act, then privacy rights generally attach to it irrespective of how it is collected (subject to the exceptions in the fine print of this increasingly complex Act).

### 6.2 Facebook was too Clever with Photo Tagging

Photo tagging is a feature of photo sharing services that helps users organise their albums. Biometric face tagging as offered by Facebook creates biometric templates that mathematically extract and represent facial features, allowing other similar photos to be identified. When Facebook makes automatic “Tag Suggestions”, its facial recognition algorithms run in the background over all photo albums making putative matches; when a photo thus identified is next displayed to a member, the tag suggestion is displayed and the member invited to confirm it. By the definition of Personal Information (PI) when Facebook’s software adds a name to an hitherto unidentified photo record, it turns that record into PI; the tagging process therefore *collects* PI (the confirmation step by another user decreasing uncertainty, but not making the initial identification).

European privacy regulators in mid-2012 found that collecting biometric data in this way without consent was a serious privacy breach. By late 2012 the authorities had forced Facebook to shut down facial recognition and tag suggestions for all its EU operations, and to delete all biometric data collected to that time.\(^61\) (Facebook was by late 2013 in New York showing renewed interest in automated recognition and tagging of video\(^62\) posted by users.) This was a confrontation between European data protection regulators and one of the

---

\(^{59}\) Leo Mirani “Google’s sneaky new privacy change affects 85% of iPhone users—but most of them won’t have noticed”, *Quartz*, April 3, 2014. At: http://qz.com/194032/ (accessed 6/4/14)


most powerful companies of the digital age, a pioneer in the development of Big Data methods and business models, over the implementation of an iconic example of Big Data identification capability.

Until 2012, Facebook’s Privacy Policy and Data Usage Policy had not even mentioned that biometric facial recognition templates were created during tagging, let alone that they were subsequently used to automatically identify people in other photos. Even if data miners generate personal information almost out of thin air, using sophisticated data processing algorithms, they are still subject, at least in Australia and Europe, to Privacy Principles such as Openness, and Limitation of Collection and Use.

6.3 Target Gets Too Close to its Female Customers

In 2012, the New York Times published an investigation by reporter Charles Duhigg into data mining practices at Target. He exposed a carefully designed customer relations program that set out to identify customers who were likely to be pregnant, in order to lock them into a lucrative period of early childhood product sales. Target had hired statisticians to research and prove Big Data techniques that mined younger females’ buying habits.

Internationally, we are not aware that such bold data mining has been attempted, possibly because it would likely be unlawful. In Australia amongst other places, health information is specially categorized in legislation as “Sensitive” PII, and special conditions apply under Australian Privacy Principle 3. Most important for Big Data is that Sensitive PII can only be collected with the informed consent of the individual concerned. Should Australian stores want to use Big Data techniques, they may need to disclose up front the possibility of health information being mined from their shopping data, and obtain customers’ express consent for the algorithms to be run. We note that rights-based privacy laws set a low bar for privacy breaches: simply collecting Sensitive PII may be a serious breach, even before it is used for anything or disclosed.

---


64 See https://www.facebook.com/full_data_use_policy. (accessed 4/4/14)


6.4 DNA Hackers

Digital businesses are availing themselves of an ever richer array of signals created automatically as we go about our lives online. Worryingly, the privacy policies of many technology businesses tend to be silent on what they plan to do with the by-products of Natural Language Processing, face recognition, object recognition and similar “digital bread crumbs”.

At the same time, informaticians are discovering ever more clever ways to de-anonymise us in cyberspace. 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 leverage publicly available genealogical data to decipher the names of unidentified DNA donors in the Thousand Genomes research program.  

7 Is Privacy Dead – or Do we Need “Big Privacy”?  

Do these developments mean “privacy is dead” after all? No. The fact is that anonymity is threatened by information technologies, but anonymity or secrecy are not the same thing as privacy. As we have shown above, when erstwhile non-identified data becomes identifiable by data analytics processes such as biometric facial recognition or DNA “hacking”, the outcome is personal information, and may fall under the purview of data protection laws.

Many cavalier online businesses are propelled by a naive assumption that all data in the “public domain” is up for grabs; they err on the side of abandon. Technocrats may think the law has not kept pace with developments, but they are often caught out by conventional data protection regulations. The extraction of PI or PII from raw data may be interpreted as a collection and as such is subject to long standing data protection statutes. On the other hand, orthodox privacy policies and freeze-frame consent mechanisms do not cater for the way PII can be conjured in future from raw data collected today using BD tools, the frequent reluctance to specify uses when initially collected, and other threats to ‘informed consent’ as the basis for a data subject’s capacity to exercise some control over collection, use and disclosure of what turns into their PI. And so big data does demand a particular approach to privacy. The question remains, which approach.  

---


Ontario Information and Privacy Commissioner Dr Ann Cavoukian and Drummond Reed coined the phrase “Big Privacy” as a response to the challenge of Big Data. They describe “Big Privacy” as “Privacy by Design writ large”. What exactly “privacy by design” might entail in practice remains largely unexplored. Another way of approaching this might be to ask if we have reached the point where we will need to figure out how to regulate algorithms.

There may be some hope. In 2013 large Silicon Valley firms (Google, Amazon, Microsoft) pioneering in the field met to discuss issues around Big Data. There was a consensus that Big Data and its tools had moved into what they could only describe as “a creepy zone”. There is an indication that industry recognises that there are major privacy concerns, and some prospect that they wish to respond in a responsible and respectful manner (though as marketers, there is always the concern that they may alternatively be interested more with creating the perception of responsibility than with the reality, especially in the light of attempts to distance themselves from a public backlash against the NSA mass surveillance operations while often in effect relying on continuing mass surveillance operations themselves).

Constellation Research Group has proposed a privacy compact which they are encouraging Big Data implementers to adopt. This involves the following principles (using the US PII term), which offer one approach to resolving the potential conflict between the new potentials and the new risks from Big Data:

---


“Respect and Restraint: More than ever, in the face of Big Data’s temptations, remember that privacy is essentially about restraint. Privacy is not only about what we do with PII; just as important is what we choose not to do.”

“Super transparency: If basic data privacy means being open about what PII is collected and why, then advanced privacy means going further, telling customers more about the business models and 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 (even to the extent of lying) to protect themselves in ad hoc ways against online exploitation. Many resort to covering their tracks with encrypted browsers like Tor or maintaining multiple email addresses so when they register at 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 businesses is an unfair burden, and ultimately counter-productive. Consumers and businesses alike will do much better by engaging in a real deal that sets out in good faith how PII is truly valued and leveraged.

“Innovate in Privacy in relation to business models as well as Privacy Enhancing Technologies: There’s a common refrain that ‘privacy hampers innovation’. Too often this suggestion is a cover for a more cynical philosophy where some digital businesses reserve the right to do as they like with any personal data that comes their way. We believe that true innovation lies in business practices which leverage PII while honouring the OECD Privacy Principles. If a business can justify data mining in terms of beneficial by-products, then it should be open about its model, and bring customers into its confidence.”

This preference for transparency and more effective choices for data subjects, rather than secrecy and lack of control, may offer a good starting point, one from which industry participants can embrace the sentiment behind these elements (perhaps with some prodding from civil society, if not from government). If they do, the current privacy framework may be workable with refinements to existing regulation and enforcement. If industry doesn’t

---

73 This appears to be a reaction to impressions that many Big Data and mass data surveillance proponents assert without hesitation that “if it can be done, it must be done”: a dangerous approach for an open-ended, self-evolving technology like the Machine Learning at the heart of Big Data tools.

74 S. Wilson, “‘Big Privacy’ Rises to the Challenges of Big Data”, Constellation Research Inc. 2014.
embrace this somewhat more respectful model, or one engage with similar issues from the emerging risks for individual data subjects arising out of Big Data’s capacity for ever-greater identifiability, then growing pressure for further regulation, or perhaps even a new framework, will follow as the basis for trust comes under increasing threat.

Nobody wants Big Data to become Big Brother, but given the ease with which data previously outside the realm of the ‘personal’ can now become identifiable, and given the scope for conflict between traditional models of privacy and personal information security protection and Big Data’s potentially insatiable and often indiscriminate appetite for data, big efforts may be needed to avert such a dystopian merger.

---