Customer data analytics: privacy settings for ‘Big Data’ business

Peter Leonard*

Why Big Data analytics is suddenly big (privacy) news

The ongoing US Federal Trade Commission inquiry into the business activities of so-called data brokers\(^1\) ignited a global privacy policy debate about the matching of disparate datasets to assemble profiles of individuals.

The internet privacy policy agenda in 2011 was dominated by global discussion about online behavioural advertising,\(^2\) a push in the United States for ‘do not track’ when online browsing, and a European debate about consumer consent requirements for internet cookies.\(^3\)

In 2012 the European Union again turned its focused inward and, in particular, to a debate about the European Commission’s proposal for a comprehensive rewrite of the privacy directive with a draft General Data Protection Regulation and draft General Data Protection Directive.\(^4\) In early 2012 there was a broader debate on consumer privacy issues that was first enlivened by the business activities of so-called data brokers,\(^1\) ignited a global privacy policy debate about the matching of disparate datasets to assemble profiles of individuals.

The internet privacy policy agenda in 2011 was dominated by global discussion about online behavioural advertising,\(^2\) a push in the United States for ‘do not track’ when online browsing, and a European debate about consumer consent requirements for internet cookies.\(^3\)

In 2012 the EU turned its focus inward and, in particular, to a debate about the European Commission’s proposal for a comprehensive rewrite of the privacy directive with a draft General Data Protection Regulation and draft General Data Protection Directive.\(^4\) In early 2012 there was a broader debate on consumer privacy issues that was first enlivened by the business activities of so-called data brokers,\(^1\) who ignited a global privacy policy debate about the matching of disparate datasets to assemble profiles of individuals.

1 Federal Trade Commission, FTC to Study Data Broker Industry’s Collection and Use of Consumer Data: Commission Issues Nine Orders for Information to Analyze Industry’s Privacy Practices, Media Release of 18 December 2012. The Media Release and orders are available at <http://www.ftc.gov/opa/2012/12/databrokers.shtm>. The nine ‘data brokers’ that received orders to provide responses from the FTC are Acxiom, Corelogic, Datalogix, eBureau, ID Analytics, Intellius, Peekyou, Rapleaf, and Recorded Future. The FTC said that it will use the responses it receives to prepare a study and to make recommendations on whether, and how, the data broker industry could improve its privacy practices.


Abstract

- At the heart of a current global debate as to privacy regulation and ‘Big Data’ debate lie four questions:
  - Can national privacy laws and regulation facilitate socially beneficial uses and applications of Big Data while precluding ‘Big Brother’ or unduly ‘creepy’, or otherwise socially or culturally unacceptable Big Data practices?
  - Can diverse national privacy laws and regulation be applied or adapted so as to accommodate socially beneficial uses and applications of Big Data, or is a more fundamental overhaul of privacy law required?
  - If fundamental design precepts of privacy regulation require adaptation or supplementation to address Big Data, can those changes be made without threatening broader consistency and integrity of privacy protections for individuals?

\(^{*}\) Peter G. Leonard B.Ec. (Hons) LLM (Syd.) is a TMT and Projects Partner of Gilbert + Tobin Lawyers, based in the firm’s Sydney, Australia office. He is a director of the International Association of Privacy Professionals Australia & New Zealand (appANZ) and National Chair of the Media and Communications Committee of the Business Law Section of the Law Council of Australia. He is a former Global Chair of the Technology Committee of the Business Law Section of the International Bar Association and a former director of the Internet Industry Association of Australia. The author acknowledges valuable assistance from comments on an earlier draft of this paper by many friends and privacy colleagues, including Malcolm Crompton, Jennifer Barrett-Glasgow, Carolyn Lidgerwood, and Teresa Troester-Falk, and the ever cheerful and questioning assistance of Natalie Lane. The author also acknowledges the influence of Greg Schneider and Mike Briers, who introduced the author to practicalities of business customer analytics. The paper is the better for their assistance, but may not reflect their views. The paper was last revised on 8 August 2013.

-- Can any adaptation or changes be made quickly enough to address growing citizen concerns about unacceptable or hidden Big Data practices?

- Responsible governance of data analytics affecting citizens, whether by businesses or government, requires a new dialogue and community understanding about appropriate transparency and ethical boundaries to uses of data analytics. This requires both businesses and government to acknowledge that for many citizens, privacy still matters and that many citizens have a deficit of trust as to uses of their personal information by government and at least some businesses.

- A nuanced debate about good Big Data and bad Big Data and transparency as to acceptable data analytics practices is necessary for a dialogue between citizens and data users. This paper examines how to build transparency and engender trust through good business practices in data analytics.

- The paper considers developing regulatory policy concerning the de-identification of personal information and how to embed privacy impact assessment practices and privacy by design and security by design principles as operational (administrative, security, and contractual) safeguards within data analytics service providers, governments, and businesses.

Obama Administration’s ‘Consumer Data Privacy’ paper (White House Privacy Paper)\(^5\) and the US Federal Trade Commission’s ‘Final Report on Protecting Consumer Privacy’ (FTC Consumer Privacy Paper).\(^6\) From mid-2012 media and policy attention in the United States moved to ‘Big Data’\(^7\) and to consumer data analytics as conducted by offline (‘bricks and mortar’) businesses of their customers, and the consumer analytics services offered by third-party providers, misleadingly grouped under the phrase ‘data brokers’.\(^8\) The debate was fuelled by US print media reports\(^9\) characterizing Big Data analytics as ‘big brother’, ‘creepy’, or ‘spooky’. Media reporting reinforced consumer unease and a perception that business data analytics principally involved hidden and deliberately secretive identification and targeting of individual consumers for ‘one to one’ marketing, including targeting by corporations with whom the individual has had no prior customer relationship. It is perhaps not surprising that this was a US led debate: the USA has not had economy wide collection and notification requirements in relation to personal information broadly defined or in relation to the notification of the collection and processing of personal information other than directly from the data subject.\(^10\) This probably led consumer advocates to focus upon the activities of the data intermediaries. The FTC Consumer Privacy Paper advocated the ‘transparency’ of the activities of ‘data brokers’, such as a central register point for finding out what information is held by whom, and possibly including a right of individual inquiry and correction.\(^11\)

The European Big Data debate commenced in early 2012 with a quite different focus, reflecting the long
standing, particularly European, concern about decisions made by automated data processing without significant human judgement—so called ‘automated individual decisions’ or ‘profiling’.12 The European profiling debate has a philosophical core: is the personal dignity and integrity of individuals compromised by decisions made by automated processes, when contrasted with decisions made by humans having regard to individual circumstances and constrained by human rights laws and also, perhaps, human empathy? The profiling debate in the United Kingdom also included a pragmatic, economic dimension. The Office of Fair Trading examined possibilities for geo-location based and ‘personalised pricing’: that is, ‘the possibility that businesses may use information that is observed, volunteered, inferred, or collected about individuals’ conduct or characteristics, such as information about a particular user’s browsing or purchasing history or the device the user uses, to set different prices to different consumers (whether on an individual or group basis) based on what the business thinks they are willing to pay.’13

Momentum for the ‘data broker’ debate in the USA was particularly fuelled by a New York Times newspaper report of June 2012 focused on the activities of the Acxiom Corporation.14 Journalist Natasha Singer asserted that ‘data broking companies’ are in possession of large amounts of data about consumers going far beyond age, race, and sex and relating to such matters as weight, height, marital status, education level, politics, buying habits and household health worries.15 The article went on to assert that many consumers would not understand the extent to which information is being compiled about them, as data broking companies ‘often were not required’ to notify a consumer that particular data were being collected or so used. The author noted that while some data brokers permit consumers to request details of information kept about them, unlike consumer reporting agencies that sell financial information about consumers for credit purposes, US data brokers are not required by law to show consumers their reports or allow them to amend any incorrect information. For a consumer to be able to make a request, the consumer would need to know which organizations hold data about them.

This New York Times article appears to have been the catalyst for a group of US legislators, led by Sen. Jay Rockefeller (D-WV) and Rep. Ed Markey (D-Mass), to commence their correspondence with nine third-party providers of consumer analytics services.16 All nine data analytics providers responded to requests from the legislators for details on their activities. For example, in addition to collecting data about consumers from sources such as telephone directories, mobile phones, government agencies, financial institutions, and directly from consumers themselves, several data analytics providers reported mining consumer information from social media sites such as Facebook and LinkedIn. In response, the legislators asserted that there was still a lot more which needed to be understood about how the ‘data broking industry’ operates, particularly in respect of how data are analysed. Their conclusion was that ‘many questions about how these data brokers operate have
been left unanswered, particularly how they analyze personal information to categorize and rate consumers.17 This in turn led to the US Federal Trade Commission initiating its own enquiries of data analytics service providers: as of August 2013, an investigation by the FTC remains underway.18

The debates outlined above occurred concurrently with four important international developments:

- Publication by the UK’s Information Commissioner’s Office (ICO) of a paper ‘Anonymisation: managing data protection risk, code of practice’ (hereafter, ICO Anonymisation Code) discussing practices for the de-identification of personal information that may be acceptable to the Information Commissioner.19 This paper was subsequently endorsed in discussion papers issued by other privacy regulators, including in Australia,20 Singapore,21 and Ontario.22

- Publication by the US Department of Health and Human Services of ‘Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act’ (HIPAA), which also included consideration of acceptable means and levels of assurance for the de-identification of sensitive health information about individuals. This guidance was intended to assist covered entities to understand the meaning of de-identification, the general process by which de-identified information is created, and the options available for performing de-identification in compliance with the de-identification standard in the HIPAA Privacy Rule. Under this standard, health information is not individually identifiable if it does not identify an individual and the covered entity has no reasonable basis to believe that the information can be used to identify the individual.23

- Discussion on commercial barriers to deployment of cloud computing across national borders derived from at best inconsistent, and at worst irreconcilably different, approaches to the definition of personally identifying information.24

- Government interest in data analytics for the activities of government and its agencies. This prompts particular sensitivities in many countries with echoes of George Orwell’s Big Brother, probably exacerbated by more recent revelations about the US PRISM surveillance activities by intelligence agencies.25 In any event, since the global financial downturn in 2007, this interest has been stimulated by budget pressure on national governments that has led to more intensive examination by governments of the possibilities of using data analytics across government for fraud prevention and to achieve cost efficiencies in the delivery

18 See Federal Trade Commission (n 1). The FTC is seeking details about the
nature and sources of the consumer information the data brokers collect, how they use, maintain, and disseminate the information; and the extent to
which the data brokers allow consumers to access and correct their information or to opt out of having their personal information sold.
of government services, as well as meeting consumer expectations on a single point of entry for the range of government services. Governments have also sought to implement more open government through the publication of open datasets for mash-ups and other third-party uses, in turn raising issues about how robustly effective technical de-identification has been in mitigating any risk of re-identification of individuals or other politically sensitive uses of open datasets.

Privacy regulation and the contextual nature of Big Data ‘creepiness’

At the heart of the Big Data privacy debate are four questions:

- Can national privacy laws and regulation facilitate socially beneficial uses and applications of Big Data while precluding ‘Big Brother’ or unduly ‘spooky’, ‘creepy’, or otherwise socially or culturally unacceptable Big Data practices?
- Can diverse national privacy laws and regulation be applied or adapted to accommodate socially beneficial uses and applications of Big Data or is a more fundamental overhaul of privacy law required?
- If fundamental design precepts of privacy regulation require adaptation or supplementation to address Big Data, can those changes be made without threatening the broader consistency and integrity of privacy protections for individuals?
- Can any adaptation or changes be made quickly enough to address growing citizen concerns about unacceptable or hidden Big Data practices?

Core privacy principles pose particular compliance challenges for Big Data analytics processes that use data about individuals. Because Big Data analytics often draws disparate datasets from multiple sources and looks experimentally for relationships or correlations that may generate useful insights, collecting of information (to sift for possible correlations) is the norm. Potential uses often cannot be identified and therefore the subject of meaningful notice to the individual until potentially useful insights are identified. These insights may in turn lead to new avenues of enquiry and analysis, and to further uses which uses had not originally been anticipated. Through the process of the correlation of disparate datasets that may have been reasonably considered effectively anonymized or de-identified, individuals may become identifiable. Any possible re-identification may only be effectively restrained by the business practices of the data analytics services provider and such restrictions (if any) as are placed upon those practices by contract with the data sources or by law. And as already been noted in relation to the ‘data broker’ debate, the entity conducting the analytics may be removed from a direct contractual or business relationship with the affected individual. This can lead to the result that accountability to the individual is attenuated or possibly non-existent. An adversely affected individual may be unable to impose the sanctions that are often available in a direct customer relationship, such as exercising their ability to take custom elsewhere, or to take unethical or unlawful business practices to the media or regulators. As the Centre for Information Policy Leadership notes, the compliance challenges for Big Data applied to individuals are broadly similar in regulated privacy jurisdictions:

- concepts of notice and consent are difficult to apply when uses are unknown or unpredictable;
- it will be difficult to meet the requirements of specification of purpose for collection and use articulated at the time of collection and before that use;
- there may not be a legitimate business purpose for the collection and proposed use of personal information for data analytics;

26 For examples of single point of entry, see Service Canada at <http://www.servicecanada.gc.ca/> and Singapore’s eCitizen portal at <http://www.ecitizen.gov.sg/Pages/default.aspx>.
• the data minimization principle (minimizing data collected to that reasonably necessary to fulfil the legitimate business purpose) is almost antithetical to Big Data analytics;
• the data retention principle (only retaining data for as long as is reasonably necessary) is antithetical to data accumulation and enrichment by the addition of further datasets for prospective and possibly unknown future uses; and
• in Europe, prohibitions against automated individual decisions, or ‘profiling’, may also attach to the use of insights derived through data analytics.

Other privacy analyses of privacy regulation and Big Data have explored these challenges and possible regulatory responses. The commonality of concerns around ‘bad’ Big Data has been partially obscured in discussion of regional differences and in the detail of technical legal analysis of the interpretation and application of particular national privacy laws to Big Data. There is a healthy and ongoing global debate as to how fundamental privacy concepts of notice and fully informed consent can be applied in a networked world.32 There is also useful discussion of continuing differences in national regulatory approaches to ‘personally identifying information’ and ‘personal information’ and how these differences may affect the implementation of transborder services such as global data centres and cloud computing. However, the debate around Big Data has generally not yet focused on how the business practices of Big Data analytics can be appropriately risk managed through the application of privacy by design principles. In the following sections of this paper we concentrate on this less analysed area. Inevitably this leads to discussion of the vociferous academic debate on the limits of technical de-identification and recent attempts by privacy regulators to engage with those limits while not foreclosing individual level data analytics.

Before descending into the detail of privacy by design in individual level data analytics, it is useful to remind ourselves of the limits of privacy law.33 The privacy interests of individuals are often given effect through privacy regulation and through legal sanctions and remedies—at least where these are available and affordable. However, citizen concerns are also given practical effect through the significant reputational damage, in particular adverse media coverage, suffered by governments and businesses that misjudge consumer sentiments and tolerance, regardless of the underlying law. Two recent debates illustrate how a lack of transparency over activities that appear to comply with present privacy laws can nonetheless create significant citizen concern and result in significant pressure in the form of calls for the extension of privacy regulation. The first is the ongoing debate about acceptable limits for surveillance activities by law enforcement agencies, most recently in relation to the online metadata mining conducted by US intelligence agencies in the PRISM programme. The second, more narrowly focused, was the backlash of users of Bloomberg Financier terminals to the use of information about their transaction activity by Bloomberg News financial journalists.34 Adverse sentiment is sometimes expressed as individuals not liking the ‘creepiness’ or ‘spookiness’ or ‘big brother’ aspects of such use. These adverse sentiments sometimes reside with unlikely candidates who otherwise consider that privacy was ‘last generation’ and who live their lives in the semi-public glare of Facebook pages and Instagram postings. Citizen concerns about lack of transparency, control, or accountability often cannot be readily mapped to these basic privacy principles or the implementation of these principles in national privacy laws. The application of privacy law to map a course through Big Data implementation may miss the mark of not identifying the highly contextual way in which ‘creepiness’ concerns arise and therefore may not sufficiently address those concerns. It is, therefore, not surprising that not infrequently Big Data customer analytics projects stall at a point where a chief marketing officer has successfully addressed the concerns of the chief information officer, the chief privacy officer, and the general counsel, but the chief executive or a consumer advocate within a corporation is then persuasive with her or his view that customers will not accept the ‘spookiness’ inherent in a proposed implementation.

Acxiom Corporation has expressed the ‘creepiness’ litmus test as ‘are your actions for the individual (not creepy) or to the individual (creepy). A creepy movie, story, or experience is usually about the unknown, the hidden motivation, the ulterior motive. Be as open as you possibly can about your interactions with individuals; use data responsibly to help the individual; provide descriptions of your processes; and describe how

you ensure personal data is kept safe.\textsuperscript{35} The creepiness factor can be highly contextual to a myriad of factors that include a particular transaction type, data collector, vendor–client relationship, geography, culture or cultural expectation, and so on. Many consumers understand that enabling geo-location on mobile devices for a particular app enables the provider of that app to target the content of offers to them based upon that location. Most consumers understand that they derive a benefit from a loyalty card in a value exchange with a vendor who will use that loyalty card data for customer analytics to target offers to that consumer. Such direct and proximate vendor–client relationships promote accountability: consumers may vote with their trade if the vendor betrays the customer’s expectations, whether those expectations are based on legal rights or not. A direct and proximate relationship also creates other accountability linkages: many consumers will draw no distinction between a vendor and the vendor’s sub-contractors—such as external data analytics providers—in relation to breaches of security or uses and abuses of personal information collected by the vendor, so woe betide the vendor that does not hold its sub-contractors to account. The term ‘data broker’ itself conjures a sense of lack of accountability and lack of transparency.\textsuperscript{36} Some vendors engender trust regardless of their level of legal compliance or transparency.\textsuperscript{37} Local banks may carry the burden of high expectations of customer confidentiality but may also derive the reward of their customers being more willing to disclose detailed personal information to them. Geographical proximity to the vendor can promote a sense of consumer empowerment: if consumers know that they can ‘march on the citadel’ to personally inflict reputational damage, their tolerance of disclosure and use of their personal information may be higher. Cultural expectations country by country, and within nations, as to what governments may do, will do, or should not do, continue to differ widely, notwithstanding the broad commonality of many privacy principles. The level of citizen trust in their body politic and organs of government is a function of diverse national cultural and socioeconomic histories.

In conclusion, it is difficult to generalize when ‘creepiness’ concerns will manifest themselves, but responsible governance for individual level data analytics should give due regard to such concerns, regardless of whether a proposed Big Data project complies with the letter of privacy regulation. With that caution, we now turn to consider privacy by design for Big Data.

Designing privacy into Big Data analytics

Good privacy compliance also requires planning operational processes to embed privacy in how an organization conducts its business, manages its contractors, offers products and services, and undertakes customer engagement. The now well-known seven ‘foundational principles of privacy by design’,\textsuperscript{38} developed by Ontario Information and Privacy Commissioner Dr Ann Cavoukian, provide the link between making responsible business governance decisions and embedding those decisions in good privacy compliance in technology and communication systems, accountable business practices and physical design. The principles are now commonly cited by privacy regulators when they state expectations of how organizations will evaluate the privacy impacts of any new acts, practices, or projects through the conduct of a privacy impact assessment, including an information security risk assessment.\textsuperscript{39} Given the extent of the reliance of privacy regulation upon consideration of context—what is reasonable in the circumstances—it is

\begin{itemize}
  \item \textsuperscript{35} Acxiom Corporation (n 14).
  \item \textsuperscript{36} The White House Paper suggested that consumers need to take some responsibility (to the extent they are able to) in respect of the initial sharing of data: for example, when considering the privacy settings they should have on social media accounts. However, once data have been shared, the White Paper identified that the consumer is left with little or no control over what their personal data is used for and pointed out that ‘as third parties become further removed from direct interactions with consumers, it may be more difficult for them to provide consumers with meaningful control over data collection. Data brokers, for example, aggregate personal data from multiple sources, often without interacting with consumers at all.’ Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy, The White House, February 2012, \texttt{<http://www.whitehouse.gov/sites/default/files/privacy-final.pdf>}, 13.
  \item \textsuperscript{37} A Ipsos MORI study of Public Understanding of Statistics examined how much trust the participants had in information provided by certain categories of people. Information provided by scientists was the most trusted (28 per cent trusted the information ‘a great deal’, 46 per cent ‘a fair amount’, compared to politicians: 1 per cent ‘a great deal’, 7 per cent ‘a fair amount’). As Marion Oswald notes, ‘emotional responses are very important in the assessment of risk. ‘If you do not trust the parties who manage the risk, you are not likely to have confidence that the risk is being safely managed’. Kieron O’Hara has said that ‘trust is an important risk and complexity management tool. . . . The stronger X’s trust, the higher the degree, and the greater the risk he is willing to take.’ See Marion Oswald ‘“Something bad might happen”: Lawyers, Anonymization and Risk’, UK Society for Computers and the Law blog posting of 19 June 2013, at \texttt{<http://www.scl.org/site.aspx?i=ed32520>}. \textsuperscript{38} Available at \texttt{<http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/>} together with extensive materials as to application of the privacy by design principles in various business contexts. See also Deidre Mulligan and Jennifer King, ‘Bridging the Gap between Privacy and Design’, (2012) 14/4 Journal of Constitutional Law.
  \item \textsuperscript{39} Many recent examples include the Office of the Australian Information Commissioner’s ‘Privacy Impact Assessment Guide’, revised April 2010 (<http://www.oaic.gov.au/publications/guidelines/Privacy_Impact_Assessment_Guide.html>), and the ‘Guide to information security: “Reasonable steps” to protect personal information’, April 2013, available at \texttt{<http://www.oaic.gov.au/publications/guidelines.html#privacy_guidelines>}. \end{itemize}
prudent to include the application of privacy by design principles, and a periodic review of practices against the principles, as good business practice.

The need for transparency creates particular challenges in the field of Big Data customer analytics. The business processes are themselves complex, unfamiliar to consumers, and difficult to explain. Uses of personal information may involve separate and limited profile-based marketing activities that use personal information and more extensive business process improvement using aggregated or otherwise anonymized information. Transparency may require the description of both the range of activities across a corporation and how and where safeguards are implemented to protect particular activities. And, as we have already noted, many Big Data projects require the accumulation of raw data from various sources and data streams for transformation and then the testing of various hypotheses during a ‘discovery phase’ before inferences are drawn and applied through an ‘applications phase’ with operational algorithms and production codes. This creates problems of how to intelligibly explain the differences between experimental and actual uses even before the legal requirements of notice and consent are addressed. Almost by definition, there is excessive data processing in the testing and rejection of hypotheses, and a desire to retain raw data for possible future testing rather than any currently known or anticipated future use. To state a possible future purpose as ‘customer analytics testing’ is vague and probably unintelligible to a data subject. Fully informed consent of the data subject is also often impractical where data are generated through routine activities and transactions and near-ubiquitous sensors (surveillance cameras, or geo-location by smart phones, computers embedded in cars, and other such devices).

Given the practical difficulties of compliance with core privacy principles as given effect through national privacy laws, it is not surprising that many Big Data applications are designed to use de-identified or anonymized information. Three questions then arise. First, is the data subject required to be informed that their personal information may be so used following de-identification—or to put the question another way, is de-identification to enable subsequent data analytics itself a use of personal information that properly should be the subject of a privacy notice? Second, should the manner of use of de-identified information itself be subject to privacy by design, or should privacy regulation end at the point at which a collecting party hands off transactional or other individual level data having taken all reasonable steps to de-identify that information? Third, given that the data analytics provider that is the recipient of the de-identified information is not in a direct relationship with the data subject, what steps if any should that recipient be required to take to inform an individual if, through the analytics processes utilized by that data analyst, that individual is re-identified by the analytics provider?

Underlying these questions is the practical problem of effecting enduring anonymization of atomized transactional information. Of course, as soon as individual level data are aggregated, the ability for subsequent re-identification is rapidly lost. Many data analytics applications can derive useful insights for data at various levels of aggregation that effectively preclude the subsequent re-identification of the individuals engaging in transactions so aggregated. But the derivation of useful insights often requires individual level data and these insights can be derived through the use of anonymized data without any intention or need of the data analyst to re-associate those insights with a particular identified or identifiable individual. For example, the development of hypotheses about the association of particular pathologies to particular genetics may require a comprehensive medical profile of a particular individual that could be used to re-identify a particular individual, but the re-identification of that individual is neither planned nor undertaken. If re-identification is not planned nor undertaken, but nonetheless possible, is the information still effectively anonymized? There is now a well accepted academic literature on the practical difficulties of ensuring enduring anonymization of many public datasets in the face of motivated re-identification efforts. This leads to advocacy for...
treat such de-identified but potentially re-identifiable datasets as personally identifying information (PII) and fully subject to privacy regulation in like manner to original personally identifying information. In many privacy jurisdictions this would lead to compliance requirements including collection notices and, at least in the case of health information, more rigorous restrictions on the collection and use of sensitive personal information including a requirement for express consent by the data subject and, in many privacy jurisdictions, limitations on collection other than from the data subject. Is the distinction between PII and non-PII too rigid or too confining for socially beneficial Big Data analytics?42

An alternative, more nuanced approach might be to develop a more clearly articulated, ethically based characterization of acceptable and unacceptable data analytics processes.43 For example, ‘safe harbours’ might be created to facilitate particular applications deemed of sufficient social benefit to warrant protection for potentially imperfect but good de-identification.44 However, ‘safe harbours’ require normative judgements to be made by regulators that can be politically or socially contentious and in any event will require frequent revisiting as new applications of analytics are developed and need to be considered as candidates for the safe harbour.

Another social policy approach has been to advocate new mechanisms to promote control by a data subject of uses of their personal information. For example, some academic commentators have moved beyond the European distrust of profiling to place a new emphasis upon ‘consumer equity’ or value exchange:

The exclusion of individuals from the benefits of the use of their data is manifest in two main ways.

First, online interactions are barter—like transactions


44 The HIPPA Guidance (n 23) is a good example. See also Khaled El Emam, Risky Business: Sharing Health Data while Protecting Privacy (Trafford Publishing, February 2013).


49 The policy balance required is well presented in Viktor Mayer-Schönberger and Kenneth Cukier, Big Data, a Revolution that will transform how we live, work and think (Houghton Mifflin Harcourt, 2013).
of ‘privacy by design’ requirements might be required to be applied to Big Data analytics in relation to ostensibly de-identified individual-level information where that individual-level information is quarantined form the public domain and therefore protected against concerted re-identification attack.

**Re-identification risk and data analytics**

Customer data analytics enables businesses to discover or infer previously unknown facts and patterns in a database based upon correlations that arise from the application of non-public algorithms to large collections of data, rather than statistical sampling or other cause and effect analysis. As we have already noted, sometimes the purpose of customer analytics is to identify particular customers for one-to-one marketing, but many customer analytics projects are not directed at targeted customer marketing and instead focus on the derivation of improvements in business processes or practices. This is frequently misunderstood in privacy analyses of customer data analytics, which tend to assume that at least the primary purpose of any customer analytics process is one-to-one marketing back to customers based upon their individual profile. However, other important applications of customer analytics include:

- **business monitoring and insights**: using business intelligence to monitor on-going business performance—for example, benchmarking against previous periods, against previous campaigns, against industry benchmarks and indices (such as brand development and customer satisfaction) to flag under- or over-performing areas of a business and providing dashboard style alerts. Insights might include identifying that particular in-flight campaigns or marketing activities are more effective than others and provide specific recommendations for how much marketing spend to shift to the more effective activities.

- **business optimization**: using analytics to automatically optimize parts of business operations, such as resource scheduling and distribution and inventory optimization, given current and predicted buying patterns coupled with local demographic, weather, and events data.

- **data monetization**: Packaging of analytic insights for sale to other organizations: for example, retailers may sell data insight packages to major suppliers of fast moving consumer goods to enable them to better plan their product range, positioning or presentation, or marketing in that retailers’ outlets.\(^{50}\)

An illustrative example is provided in the ICO Anonymisation Code:

A North American home-ware retailer (HW) was experiencing declining sales. In order to address this situation, HW needed to better understand its customers’ requirements so it could improve its sales.

HW identified that the analysis of historical point of sale transactional data (POS data) would enable it to better understand what customers were buying in HW stores. HW engaged a third party, Research Direct, to help undertake an analysis of its POS data. Due to payment card regulations, HW was prohibited from sending raw POS data (which included credit card payment details) to Research Direct. In order to comply with the payment card regulations, HW applied one way encryption to the credit card data (contained in the POS data) for the purposes of transferring data to Research Direct. Research Direct was then able to analyse, over time, purchases made using the same payment card (using the encrypted key) and therefore enriched HW’s understanding of its customers through analysis, like customer segmentations. Top-level findings were then shared with HW.

By using this method to anonymise its data, HW was able to analyse 82% of sales (the remaining 18% were cash purchases).

This enabled HW (a formerly struggling retailer) to accurately analyse its customers’ motivations when buying products and specifically what they bought over time.\(^{51}\)

Regardless of whether re-identification of individuals is an objective, the accretion and correlation of individual level data that initially are anonymized or de-identified may lead to the accreted information enabling identification of the individual concerned. This concern is well put by Ann Cavoukian and Jeff Jonas.\(^{52}\)

As more data, from more sources, assembles around a single individual—despite de-identification efforts—attempts to reliably protect identity [are] compromised. Imagine a folder that contains no references to the neighborhood you live in, the neighborhood where you work, your favorite coffee shop, and the make/model/year of your car. Without personal identifiers, could it be associated with you? As more and more individually benign facts are assembled, they collectively become strongly identifying; indeed, the

---


right set of such data can approach your driver’s license number in its ability to identify you.

Many of the analyses of the re-identification risks of Big Data have been of data ‘at large’ in the public domain through the release of purportedly anonymized datasets. In such circumstances motivated persons endeavouring to re-identify individuals are effectively unrestrained in their ability to match these datasets to other data or knowledge. This may be because some of the first well publicized uses of ‘Big Data’ were public health studies that combined private data with public datasets and thereby armed motivated analysts to re-identify some well-known individuals by correlation of public knowledge with purportedly anonymized individual level personal health information. The focus upon the possibilities of such re-identification has distracted attention from now more common business models of customer data analytics within controlled environments. Contrast the position in analysis of the same datasets in the course of secure, non-public, outsourced provision of data analytics services governed by contract between a data collector and the data analytics services provider and subject to requirements for security and other restrictions imposed by that contract and within the service provider’s business.

Re-identification risk therefore needs to be considered in the context of the particular analytics activity that is planned. Paul Schwartz and Dan Solove have developed what has been referred to as a graduated or ‘differential’ approach to PII and anonymization. Under this approach, PII would be identified on a risk matrix taking into account the risk, intent, and potential consequences of re-identification, rather than a bi-polar dichotomy between personally identifiable and un-identifiable data. As Omer Tene puts it, ‘a bi-polar approach based on labelling information either personally identifiable or not is unhelpful and inevitably leads to an inefficient arms race between de-identifiers and re-identifiers. In this process, the integrity, accuracy and value of the data may be degraded or lost, together with some of its potential societal benefits.’ A number of leading academics including Helen Nissenbaum and Paul Ohm suggest that privacy regulation should effectively require the data releaser (that is, the entity making available purportedly de-identified individual level data) to consider all the determinants of re-identification risk and to assess whether a threat to the data subject exists. Jane Yakowitz, has suggested a series of ‘simple, easy-to-apply rules’ designed to set clear boundaries for what is appropriate anonymization and by so doing reduce the transactional costs in data analytics while keeping good privacy. This approach entails clarifying what a data producer is expected to do in order to anonymize a dataset and avoid the dissemination of legally defined PII, creating a safe harbour for the data producer from privacy-related liability if the anonymization protocols are properly implemented, and punishing with substantial penalties any recipient of anonymized data who re-identifies a subject in the dataset for an improper purpose. Professor Yakowitz suggests, by way of example, that a data producer might be required to do two things in order to convert personally identifiable data into anonymized (non-PII) data: first, to strip all direct identifiers, and second, either to check for minimum subgroup sizes on a preset list of common indirect identifiers, such as race, sex, geographic indicators, and other indirect identifiers commonly found in public records, or to use an effective random sampling frame.

Omer Tene and Chris Wolf have further developed this view in a critique of the proposed European privacy framework:

The limitation of technology-based de-identification does not mean, however, that all data should henceforth come under the full remit of data protection laws, regardless of how remote the risk of re-identification. When using data internally or sharing them with third party service providers or business associates, organizations can put in place stringent administrative and legal safeguards in addition to technical de-identification, to greatly reduce the likelihood of re-identification. Moreover, organizations frequently have no interest in re-identifying specific individuals from de-identified datasets. Treating all such data as personal would create perverse incentives for organizations to forgo...
administered and legal safeguards and retain as much personal data as they can.\footnote{Omer Tene and Christopher Wolf, ‘The Definition of Personal Data: seeing the Complete Spectrum’ The Future of Privacy Forum While Paper, January 2013.}

A move away from the bi-polar world of labelling information personally identifiable or not risks creating unmanageable complexity in privacy regulation. However, considering re-identification risk in a graduated continuum between unbreakable de-identification and expressly identifying personal information provides incentives for organizations to map information flows and develop operational processes and safeguards that mitigate risk. The developing graduated or ‘differential’ approach to privacy risk management leads to four key questions for privacy regulators:

- Can a privacy risk management approach be applied across diverse current regulatory regimes and varying definitions of personal information and PII? This will be particularly important as the global management of information and use of cloud and other offshore data storage increases.
- How should a privacy impact assessor or a regulator assess the risk mitigation value of stringent administrative and legal safeguards and are these only relevant in addition to technical de-identification?
- Is there a subset of legal obligations that should apply to users of de-identified datasets about individuals to protect against re-identification risk?
- How can citizens be informed about customer data analytics in a way that is understandable and user-friendly and which accommodates the dynamic and unpredictable manner in which business insights may be discovered and then given operation in production data analytics?

Privacy regulation meets the reality of business Big Data analytics in the way that these four questions are required to be answered in business practices. The questions have not been definitively addressed by privacy regulators. The questions are being answered today through business practices that (if and when done well) embed privacy by design principles into operational (administrative, security, and contractual) safeguards within data analytics service providers and other corporations, as well as the use of technical de-identification. But because this area is new, there is no common industry practice for such safeguards and there is limited regulatory guid-

\footnote{The White House, 'Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy', 23 February 2012.}


\footnote{Omer Tene and Christopher Wolf, ‘The Definition of Personal Data: seeing the Complete Spectrum’ The Future of Privacy Forum While Paper, January 2013.}

\footnote{Ibid, at 18.}

\footnote{57 Omer Tene and Christopher Wolf, 'The Definition of Personal Data: seeing the Complete Spectrum' The Future of Privacy Forum While Paper, January 2013.}

\footnote{58 The White House, 'Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy', 23 February 2012.}


\footnote{60 Ibid, at 18.}

\footnote{57 Omer Tene and Christopher Wolf, ‘The Definition of Personal Data: seeing the Complete Spectrum’ The Future of Privacy Forum While Paper, January 2013.}

\footnote{58 The White House, 'Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy', 23 February 2012.}


\footnote{60 Ibid, at 18.}

\footnote{57 Omer Tene and Christopher Wolf, ‘The Definition of Personal Data: seeing the Complete Spectrum’ The Future of Privacy Forum While Paper, January 2013.}

\footnote{58 The White House, 'Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy', 23 February 2012.}


\footnote{60 Ibid, at 18.}

\footnote{57 Omer Tene and Christopher Wolf, ‘The Definition of Personal Data: seeing the Complete Spectrum’ The Future of Privacy Forum While Paper, January 2013.}

\footnote{58 The White House, 'Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy', 23 February 2012.}


\footnote{60 Ibid, at 18.}
’operational safeguards to prevent the data from being used to re-identify consumers.’ If it chose to share such data with third parties, Netflix stated that it would limit access ‘only to researchers who contractually agree to specific limitations on its use.’

Accordingly, as long as (1) a given dataset is not reasonably identifiable, (2) the company publicly commits not to re-identify it, and (3) the company requires any downstream users of the data to keep it in de-identified form, that data will fall outside the scope of the framework.

This clarification of the framework’s reasonable linkability standard is designed to help address the concern that the standard is overly broad. Further, the clarification gives companies an incentive to collect and use data in a form that makes it less likely the data will be linked to a particular consumer or device, thereby promoting privacy. Additionally, by calling for companies to publicly commit to the steps they take, the framework promotes accountability.61

In November 2012 the UK’s Information Commissioner’s Office issued the ‘Anonymisation Code of Practice’.62 The Code quickly became the benchmark regulator statement of the appropriate measures to manage re-identification risk, cited as useful guidance by privacy regulators in a number of jurisdictions including Australia, Singapore, and the European Union.63

The ICO first dealt with the threshold issue of whether de-identification of personal information was itself a ‘use’ of personal information that properly ought be the subject of a collection notice. The ICO stated that the United Kingdom Data Protection Act should not be interpreted as preventing the anonymization of personal data, given that anonymization safeguards individuals’ privacy and is a practical example of the ‘privacy by design’ principles that data protection law promotes.64 The ICO also stated that once information is appropriately anonymized, the information is no longer regulated information and disclosure of the information does not amount to a disclosure of personal data—even where the data are capable of being re-identified by the disclosing party.65

As noted by the ICO, the regulatory definition of ‘personal data’ can be difficult to apply in practice for two main reasons. First, the concept of ‘identify’—and therefore of ‘anonymize’—is not straightforward. Individuals can be identified by direct identification where someone is explicitly identifiable from a single data source, such as a list including full names, or indirect identification, where two or more data sources need to be combined for identification to take place. Second, the data releaser may be satisfied that its data intended for release does not, in itself, identify anyone. However, it is not uncommon that a data releaser cannot reliably predict whether other data are available that would allow re-identification by a third party to take place. ‘The DPA [Data Protection Act] does not require anonymisation to be completely risk free—you must be able to mitigate the risk of identification until it is remote.’66 The ICO Anonymisation Paper advocated re-identification testing be conducted as a type of ‘penetration’ (or ‘pen’) testing, attempting to re-identify individuals from an anonymized dataset or datasets to detect and deal with re-identification vulnerabilities. This testing would involve investigation into what other data, personal data or not, are publicly available or accessible by search and could be linked to the anonymized data and result in re-identification. The ICO suggested that a penetration test should meet three criteria. First, the test should attempt to identify particular individuals and one or more private attributes relating to those individuals. Second, the test should employ any method which is reasonably likely to be used by an intruder. Third, the test may use any lawfully obtainable data source which is reasonably likely to be used to identify particular individuals in the datasets.67 In determining whether re-identification is possible, the ICO proposed the use of the ‘motivated intruder test’ as used previously by the ICO and the administrative tribunal that hears English DPA and Freedom of Information Act appeals. The “motivated intruder” is taken to be a person who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been derived. This test is meant to assess whether the motivated intruder would be successful.68

61 Ibid, at 22.
62 ICO Anonymisation Paper (n 19).
63 See nn. 20, 21, and 22.
64 ICO Anonymisation Paper (n 19) at 7.
66 Ibid at 6.
67 Ibid at 42.
68 Ibid at 22. For a recent case, see The Information Commissioner (Appellant) v Magherafelt District Council 2012 UKUT 263 AAC, available at <http://www.oscc.gov.uk/Aspx/view.aspx?id=3536>. That case was a request under the Freedom of Information Act 2000 for details of disciplinary action taken against employees of Magherafelt District Council in Northern Ireland. The Council was a small authority with only 150 employees, all known to each other, in a district with a population of 39,500. The Council was likened to a family, with a high level of knowledge of each other’s affairs. The Tribunal used the motivated intruder test, saying that a motivated intruder is someone who has access to the internet and public documents and would use investigatory techniques such as making enquiries of people likely to have additional knowledge. The requestor was an investigative journalist, and so might have been highly motivated to identify individuals using other information available and common investigative steps. The Tribunal concluded that an investigative journalist ‘would have little difficulty in making the necessary enquiries which could lead to the identification of individuals subject to disciplinary proceedings’, particularly as the community was small and close-knit, and that identification would be all the more likely when the sanction was suspension or dismissal in that case. That Tribunal decision followed a
The ICO then turned to more difficult and novel issues. One issue was how to assess the risk of identification when anonymized information is disclosed to a closed group, as compared to public release. The second concerned the role of independent third-party analytics providers and when and how the use of a trusted third party was relevant to privacy risk management and mitigation.69

The ICO distinguished disclosure of anonymized information to a small group of parties, ‘such as researchers under contractual prohibitions limiting any further disclosure’, and publication to the world at large. The ICO considered the former to have significantly lower risks of disclosure and advocated a contextual approach, well expressed in the following:

It is good practice to try to look at identification ‘in the round’, i.e. all organisations disclosing anonymised data should assess whether any organisation or member of the public could identify any individual from the data being released—either in itself or in combination with other available information. The risk involved will vary according to the local data environment and particularly who has access to information. This means that anonymised data disclosed within a secure local environment, e.g. when disclosed to a particular research organisation, could remain anonymous even though if published, the likelihood of re-identification would mean that the anonymised data would become personal data.

There will clearly be borderline cases where, in reality, it will be difficult, or even impossible, to determine whether it is likely that re-identification will take place. The test in the DPA for determining whether information relating to a living individual is ‘personal data’ is based entirely on the identification or likely identification of the individual. The risk posed to individuals by disclosure, or the public benefit of this, are not factors that the DPA allows to be taken into account when determining whether or not information is personal data. In reality though, some types of data will be more attractive to a motivated intruder than others—and more consequential for individuals. In reality these factors should also inform an organisation’s approach to disclosure.

Clearly the identification of an individual can have a range of consequences depending on the nature of the data, the context in which it is disclosed and who it is about. The Information Commissioner would certainly be more concerned about a disclosure of personal data that is detrimental to an individual, than about an inconsequential one. The Information Commissioner will take the effect or potential effect into account should a case of re-identification or inappropriate data disclosure come to his attention.

In borderline cases where the consequences of re-identification could be significant e.g. because they would leave an individual open to damage, distress or financial loss, organisations should:

- seek data subject consent for the disclosure of the data, explaining its possible consequences;
- adopt a more rigorous form of risk analysis and anonymisation.

In some scenarios, data should only be disclosed within a properly constituted closed community and with specific safeguards in place.

In some particularly high-risk situations, it may not even be possible to share within a closed community.70

The Personal Data Protection Commission of Singapore proposed a similar practical or contextual approach:

Like many jurisdictions, the Commission will take a practical approach towards anonymisation and risks of identification. If the risk of re-identification is high, then the data will be considered personal data. If the possibility of re-identification is trivial, the Commission will consider the data anonymised.

In assessing the risks of re-identification when disclosing anonymised data, organisations should consider the risk of the receiving organisation being able to re-identify an individual. The likelihood that the receiving organisation would attempt to re-identify an individual from the anonymised data is also a consideration.71

The Australian Office of the Information Commissioner in April 2013 released Privacy Business Resource 3—De-identification of data and information—consultation draft April 2013 for public comment. Similar to the Singapore Advisory Guideline, the Australian draft acknowledges that the risk of re-identification need not be removed entirely, but stated:

De-identification can be a useful technique for protecting privacy and confidentiality when releasing information assets. Nevertheless, de-identification is not infallible; it may be possible in some circumstances to re-identify data or information by matching it with other datasets or information. However, where de-identification is administered to a high standard, together with appropriate risk management strategies, the risk of re-identification can be minimised.72

69 ICO Anonymisation Paper (n 19) at 16.

70 Ibid. at 19 and 20.


The Australian Commissioner discussed de-identification techniques and endorsed the UK ICO’s proposal of a ‘motivated intruder test’ and assessment of re-identification risk ‘in the round’. In assessing whether a risk of re-identification is ‘remote’ or ‘low’, the draft Australian De-identification Paper states that de-identification ‘administered to a high standard, together with appropriate risk management strategies’, may be taken into account in determining whether the risk of re-identification has been effectively minimized. The draft also suggested that as well as technical techniques of de-identification, it is appropriate to:

Consider other steps that can be taken to manage and minimise the risk of re-identification. This may include:

- requiring the data or information receiver to sign a contract limiting the use and distribution of the information or data, and enforcing the terms of that contract.
- limiting the access to information or data by, for example, allowing other agencies or organisations to view the data rather than providing a copy, or running an analysis of the data and providing the result rather than the raw data.

The European Commission’s Article 29 Working Party in its April 2013 Opinion 03/2013 on purpose limitation\(^73\) noted the absence of ‘comprehensive guidance on anonymisation at European level’ and stated that it was preparing a guidance document on open data which will address, among other things, some issues related to anonymization. The Opinion noted:

In practice, there is a very significant grey area, where a data controller may believe a dataset is anonymised, but a motivated third party will still be able to identify at least some of the individuals from the information released. Addressing and regularly revisiting the risk of re-identification, including identifying residual risks, therefore remains an important element of any solid approach in this area.

Partial anonymisation or partial de-identification may be the appropriate solution in some situations when complete anonymisation is not practically feasible. In these cases, various techniques (including pseudo-anonymisation, key-coding, keyed-hashing, using rotating salts, removal of direct identifiers and outliers, replacing unique IDs, introduction of ‘noise’, and others) should be used to reduce the risk that data subjects can be re-identified, and subsequently, that any measures or decisions can be taken in their regard. In addition, there will also often be a need to complement these techniques with other safeguards in order to adequately protect the data subjects. These include data minimisation, as well as appropriate organisational and technical measures, including effective ‘data silo’-ing, to ensure functional separation.

Directly identifiable personal data may be processed only if anonymisation or partial anonymisation is not possible without frustrating the purpose of the processing, and further provided that other appropriate and effective safeguards are in place.

The above analysis shows that anonymisation is a key tool in achieving functional separation, and although it is highly recommended, it does have its challenges and limits. The analysis also shows that once the first assessment has been completed in terms of the possibilities and limits of effective de-identification, the second step of applying additional safeguards will often need to follow.

As essential guidance, it should be kept in mind that the easier the data subject can be identified, the more additional safeguards will be needed. 

Among the appropriate safeguards which may bring additional protection to the data subjects, the following could be considered:

- taking specific additional security measures (such as encryption);
- in case of pseudonymisation, making sure that data enabling the linking of information to a data subject (the keys) are themselves also coded or encrypted and stored separately;
- entering into a trusted third party (TTP) arrangement in situations where a number of organisations each want to anonymise the personal data they hold for use in a collaborative project;
- restricting access to personal data only on a need-to-know basis, carefully balancing the benefits of wider dissemination against the risks of inadvertent disclosure of personal data to unauthorised persons. This may include, for example, allowing read only access on controlled premises. Alternatively, arrangements could be made for limited disclosure in a secure local environment to properly constituted closed communities. Legally enforceable confidentiality obligations placed on the recipients of the data, including prohibiting publication of identifiable information, are also important. It is important to note that in high-risk situations, where the inadvertent disclosure of personal data would have serious or harmful consequences for individuals, even this type of access or restriction may not be suitable.

**Conclusion**

What is apparent from comparing these regulatory approaches to anonymization is the conceptual similarity

---

between them. This of itself is unusual: privacy regulation in Europe, the USA, Singapore, and Australia rests on distinctly different theoretical foundations, yet the regulatory analysis as to re-identification risk, and the management and mitigation of that risk through implementation of appropriate safeguards, is remarkably convergent. Each regulator expects an assessment of the sensitivity of the data, the context and limits of its disclosure, and the implementation of appropriate risk mitigation measures. Although the standard for acceptable risk is variously stated, the statements are not dissimilar: ‘low’, ‘remote’, or ‘trivial’. The possibility of re-identification is contextually assessed—‘in the round’. Risk mitigation measures—being appropriately ‘robust’ safeguards—are to be implemented before purportedly anonymized data are made available to others. These robust safeguards may be a combination of technical, operational, and contractual safeguards.

The regulatory views also converge in not being prescriptive on particular safeguards, instead offering a menu board approach for consideration in a privacy and security impact assessment individual to that deployment as to the safeguards appropriate for a particular data analytics deployment.

The menu board of safeguards might include the use of trusted third-party arrangements; use of pseudonymization keys and arrangements for the separation and security of decryption keys; contractual limitation of the use of the data to a particular project or projects; contractual purpose limitations, for example, that the data can only be used by the recipient for an agreed purpose or set of purposes; contractual restriction on the disclosure of the data; limiting the copying of, or the number of copies of, the data; required training of staff with access to data, especially on security and data minimization principles; personnel background checks for those granted access to data; controls over the ability to bring other data into the environment (allowing the risk of re-identification by linkage or association to be managed); contractual prohibition on any attempt at re-identification and measures for the destruction of any accidentally re-identified personal data; arrangements for technical and organizational security, for example staff confidentiality agreements; and arrangements for the destruction or return of the data on completion of the project.

While these regulatory views are being developed and refined, the questions that the regulators are tentatively answering are already being addressed through business practices that, if and when done well, deploy technical de-identification and also embed privacy impact assessment, privacy by design and security by design principles into other operational (administrative, security, and contractual) safeguards within data analytics service providers, governments, and corporations. But because this area is new, there is no common industry practice for such safeguards, and sub-standard implementations continue and threaten to further erode citizen trust of Big Data. If bad practices and bad media further promote other businesses and government to be less transparent about their data analytics projects, the public perception of business and government colluding in secrecy will grow, prompting more prescriptive regulation. Big Data and the privacy regulatory and compliance response to it will be one of the most important areas for the development of operational privacy compliance for the next five years.

74 ICO Anonymisation Paper (n 19) at 37 and 38.