Chapter 2
Start with Privacy by Design in All Big Data Applications

Ann Cavoukian and Michelle Chibba

2.1 Introduction

The evolution of networked information and communication technologies has, in one generation, radically changed the value of and ways to manage data. These trends carry profound implications for privacy. The creation and dissemination of data has accelerated around the world, and is being copied and stored indefinitely, resulting in the emergence of Big Data. The old information destruction paradigm created in an era of paper records is no longer relevant, because digital bits and bytes have now attained near immortality in cyberspace, thwarting efforts to successfully remove them from "public" domains. The practical obscurity of personal information—the data protection of yesteryear—is disappearing as data becomes digitized, connected to the grid, and exploited in countless new ways. We've all but given up trying to inventory and classify information, and now rely more on advanced search techniques and automated tools to manage and "mine" data. The combined effect is that while information has become cheap to distribute, copy, and recombine; personal information has also become far more available and consequential. The challenges to control and protect personal information are significant. Implementing and following good privacy practices should not be a hindrance to innovation, to reaping societal benefits or to finding the means to reinforce the public good from Big Data analytics—in fact, by doing so, innovation is fostered with doubly-enabling, win–win outcomes. The privacy solution requires a combination of data minimization techniques, credible safeguards, meaningful individual participation in data processing life cycles, and robust accountability measures in place by organizations informed by an enhanced and enforceable set of

A. Cavoukian (✉) • M. Chibba
Faculty of Science, Privacy and Big Data Institute, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada
e-mail: ann.cavoukian@ryerson.ca; michelle.chibba@ryerson.ca

S. Srinivasan (ed.), *Guide to Big Data Applications*, Studies in Big Data 26,
DOI 10.1007/978-3-319-53817-4_2

universal privacy principles better suited to modern realities. This is where Privacy by Design becomes an essential approach for Big Data applications. This chapter begins by defining information privacy, then it will provide an overview of the privacy risks associated with Big Data applications. Finally, the authors will discuss Privacy by Design as an international framework for privacy, then provide guidance on using the Privacy by Design Framework and the 7 Foundational Principles, to achieve both innovation and privacy—not one at the expense of the other.

2.2 Information Privacy Defined

Information privacy refers to the right or ability of individuals to exercise control over the collection, use and disclosure by others of their personal information (Clarke 2000). The ability to determine the fate of one's personal information is so important that the authors wish to bring to the attention of the readers, the term "informational self-determination" which underpins the approach taken to privacy in this chapter. This term was established in 1983 in Germany when the Constitutional Court ruled that individuals, not governments, determine the fate of their personal information. Since this time, in December 2013, the United Nations General Assembly adopted resolution 68/167 (UN 2016), which expressed deep concern at the negative impact that surveillance and interception of communications may have on human rights. The General Assembly affirmed that the rights held by people offline must also be protected online, and it called upon all States to respect and protect the right to privacy in digital communication.

Information privacy makes each of us 'masters' of the data that identifies each of us – as individual, citizen, worker, consumer, patient, student, tourist, investor, parent, son, or daughter. For this, the notions of empowerment, control, choice and self-determination are the very essence of what we refer to as information privacy. As 'custodians' of our information, we expect governments and business can be trusted with its safekeeping and proper use.

There have also been references to statements such as "If you have nothing to hide, you have nothing to fear." (Solove 2007) Privacy is not about secrecy. It is about the freedom to exercise one's right to decide who to choose to share the personal details of one's life with. Democracy does not begin with intrusions into one's personal sphere—it begins with human rights, civil liberties and privacy—all fundamental to individual freedom.

Sometimes, safekeeping or information security is taken to mean that privacy has been addressed. To be clear, information security does not equal privacy. While data security certainly plays a vital role in enhancing privacy, there is an important distinction to be made—security is about protecting data assets. It is about achieving the goals of confidentiality, integrity and availability. Privacy related goals developed in Europe that complement this security triad are: unlinkability, transparency and intervenability. In other words, information privacy incorporates a much broader set of protections than security alone. We look to the work on

'contextual integrity' (Dwork 2014) that extends the meaning of privacy to a much broader class of transmission principles that cannot be presumed unless warranted by other context-specific parameters influenced by other actors and information types. Privacy relates not only to the way that information is protected and accessed, but also to the way in which it is collected and used. For example, user access controls protect personal information from internal threats by preventing even the possibility of accidental or intentional disclosure or misuse. This protection is especially needed in the world of Big Data.

2.2.1 Is It Personally Identifiable Information?

Not all data gives rise to privacy concerns. An important first step for any Big Data application is to determine whether the information involved falls under the definition of personally identifiable information (PII). Privacy laws around the world include a definition of personal information and it is this definition which is integral to whether or not the rules apply. Although there are privacy laws around the world, each with a definition of personal information, we will use the NIST definition, where personal information (also known as personally identifiable information) may be defined as any information, recorded or otherwise, relating to an identifiable individual (NIST 2010). It is important to note that almost any information (e.g. biographical, biological, genealogical, historical, transactional, locational, relational, computational, vocational, or reputational), may become personal in nature. Privacy laws and associated rules will apply to information if there is a reasonable possibility of identifying a specific individual—whether directly, indirectly, or through manipulation or data linkage.

Understanding the different forms of non-personal data helps to better understand what constitutes personal information. One example is de-identified or anonymous information, which will be dealt with in more detail later in this chapter. NIST defines de-identified information as records that have had enough personal information removed or obscured in some manner such that the remaining information does not identify an individual, and there is no reasonable basis to believe that the information can be used to identify an individual (NIST 2015). As an illustration, under a U.S. law known as the Health Insurance Portability and Accountability Act (HIPAA), a set of standards exist to determine when health-care information is no longer 'individually identifiable' or de-identified (HHS 2012). If this standard is achieved, then the health-care information would not be subject to this law governing the privacy of health care information. Another example is the EU General Data Protection Regulation (GDPR) that similarly, excludes anonymous information (EU Commission 2015). Of interest, however, is that this European law introduces the concept of "pseudonymization" defined as the processing of personal data in such a way as to prevent attribution to an identified or identifiable

person without additional information that may be held separately.[1] For research and statistical purposes, certain requirements under the GDPR are relaxed if the personal data is pseudonymized, which is considered an appropriate safeguard alongside encryption (Official Journal of the European Union 2016).

Another form is when personal information is aggregated. Aggregation refers to summary data that have been generated by performing a calculation across all individual units as a whole. For example, medical researchers may use aggregated patient data to assess new treatment strategies; governments may use aggregated population data for statistical analysis on certain publicly funded programs for reporting purposes; companies may use aggregated sales data to assist in determining future product lines. Work has also been done on privacy-preserving data aggregation in wireless sensor networks, especially relevant in the context of the Internet of Things (Zhang et al. 2016). By using aggregated data, there is a reduced risk of connecting this information to a specific person or identify an individual.

Lastly, while personal information may be classified as confidential, not all confidential information should be governed under privacy rules. Confidential information includes information that should not be publicly available and often holds tremendous value and importance for organizations, such as strategic business plans, interim revenue forecasts, proprietary research, or other intellectual property. The distinction is that while the theft or loss of such confidential information is of grave concern for an organization it would not constitute a privacy breach because it does not involve personal information—rather, it is business information.

The growth in Big Data applications and other information communication technologies have added to the challenges of definition of personal information. There are times when information architectures, developed by engineers to ensure the smooth functioning of computer networks and connectivity, lead to unforeseen uses that have an impact on identity and privacy. These changes present challenges to what constitutes personal information, extending it from obvious tombstone data (name, address, telephone number, date of birth, gender) to the innocuous computational or metadata once the purview of engineering requirements for communicating between devices (Cameron 2013; Mayer et al. 2016).

Metadata, for example, is information generated by our communications devices and our communications service providers as we use landline or mobile phones, computers, tablets, or other computing devices. Metadata is essentially information about other information—in this case, relating to our communications (Mayer et al. 2016). Using metadata in Big Data analysis requires understanding of context.

[1]NIST (2015) defines 'pseudonymization' as a specific kind of transformation in which the names and other information that directly identifies an individual are replaced with pseudonyms. Pseudonymization allows linking information belonging to an individual across multiple data records or information systems, provided that all direct identifiers are systematically pseudonymized. Pseudonymization can be readily reversed if the entity that performed the pseudonymization retains a table linking the original identities to the pseudonyms, or if the substitution is performed using an algorithm for which the parameters are known or can be discovered.

Metadata reveals detailed pattern of associations that can be far more invasive of privacy than merely accessing the content of one's communications (Cavoukian 2013a, b). Addresses, such as the Media Access Control (MAC) number that are designed to be persistent and unique for the purpose of running software applications and utilizing Wi-Fi positioning systems to communicate to a local area network can now reveal much more about an individual through advances in geo-location services and uses of smart mobile devices (Cavoukian and Cameron 2011). Another good example in the mobile environment would be a unique device identifier such as an International Mobile Equipment Identity (IMEI) number: even though this does not name the individual, if it is used to treat individuals differently it will fit the definition of personal data (Information Commissioner's Office ICO 2013).

No doubt, the mobile ecosystem is extremely complex and architectures that were first developed to ensure the functioning of wireless network components now act as geo-location points, thereby transforming the original intent or what might be an unintended consequence for privacy. As noted by the International Working Group on Data Protection in Telecommunications (IWGDPT 2004) "The enhanced precision of location information and its availability to parties other than the operators of mobile telecommunications networks create unprecedented threats to the privacy of the users of mobile devices linked to telecommunications networks." When a unique identifier may be linked to an individual, it often falls under the definition of "personal information" and carries with it a set of regulatory responsibilities.

2.3 Big Data: Understanding the Challenges to Privacy

Before moving into understanding the challenges and risks to privacy that arise from Big Data applications and the associated data ecosystem, it is important to emphasize that these should not be deterrents to extracting value from Big Data. The authors believe that by understanding these privacy risks early on, Big Data application developers, researchers, policymakers, and other stakeholders will be sensitized to the privacy issues and therefore, be able to raise early flags on potential unintended consequences as part of a privacy/security threat risk analysis.

We know that with advances in Big Data applications, organizations are developing a more complete understanding of the individuals with whom they interact because of the growth and development of data analytical tools, and systems available to them. Public health authorities, for example, have a need for more detailed information in order to better inform policy decisions related to managing their increasingly limited resources. Local governments are able to gain insights never before available into traffic patterns that lead to greater road and pedestrian safety. These examples and many more demonstrate the ability to extract insights from Big Data that will, without a doubt, be of enormous socio-economic significance. These challenges and insights are further examined in the narrative on the impact of Big Data on privacy (Lane et al. 2014).

With this shift to knowledge creation and service delivery, the value of information and the need to manage it responsibly have grown dramatically. At the same time, rapid innovation, global competition and increasing system complexity present profound challenges for informational privacy. The notion of informational self-determination seems to be collapsing under the weight, diversity, speed and volume of Big Data processing in the modern digital era. When a Big Data set is comprised of identifiable information, then a host of customary privacy risks apply. As technological advances improve our ability to exploit Big Data, potential privacy concerns could stir a regulatory backlash that would dampen the data economy and stifle innovation (Tene and Polonetsky 2013). These concerns are reflected in, for example, the debate around the new European legislation that includes a 'right to be forgotten' that is aimed at helping individuals better manage data protection risks online by requiring organizations to delete their data if there are no legitimate grounds for retaining it (EU Commission 2012). The genesis of the incorporation of this right comes from a citizen complaint to a data protection regulator against a newspaper and a major search engine concerning outdated information about the citizen that continued to appear in online search results of the citizen's name. Under certain conditions now, individuals have the right to ask search engines to remove links with personal information about them that is "inaccurate, inadequate, irrelevant or excessive." (EU Commission 2012)

Big Data challenges the tenets of information security, which may also be of consequence for the protection of privacy. Security challenges arise because Big Data involves several infrastructure layers for data processing, new types of infrastructure to handle the enormous flow of data, as well as requiring non-scalable encryption of large data sets. Further, a data breach may have more severe consequences when enormous datasets are stored. Consider, for example, the value of a large dataset of identifiable information or confidential information for that matter, that could make it a target of theft or for ransom—the larger the dataset, the more likely it may be targeted for misuse. Once unauthorized disclosure takes place, the impact on privacy will be far greater, because the information is centralized and contains more data elements. In extreme cases, unauthorized disclosure of personal information could put public safety at risk.

Outsourcing Big Data analytics and managing data accountability are other issues that arise when handling identifiable datasets. This is especially true in a Big Data context, since organizations with large amounts of data may lack the ability to perform analytics themselves and will outsource this analysis and reporting (Fogarty and Bell 2014). There is also a growing presence of data brokers involved in collecting information, including personal information, from a wide variety of sources other than the individual, for the purpose of reselling such information to their customers for various purposes, including verifying an individual's identity, differentiating records, marketing products, and preventing financial fraud (FTC 2012). Data governance becomes a sine qua non for the enterprise and the stakeholders within the Big Data ecosystem.

2.3.1 Big Data: The Antithesis of Data Minimization

To begin, the basis of Big Data is the antithesis of a fundamental privacy principle which is data minimization. The principle of data minimization or the limitation principle (Gürses et al. 2011) is intended to ensure that no more personal information is collected and stored than what is necessary to fulfil clearly defined purposes. This approach follows through the fully data lifecycle where personal data must be deleted when it is no longer necessary for the original purpose. The challenge to this is that Big Data entails a new way of looking at data, where data is assigned value in itself. In other words, the value of the data is linked to its *future and potential* uses.

In moving from data minimization to what may be termed data maximization or Big Data, the challenge to privacy is the risk of creating automatic data linkages between seemingly non-identifiable data which, on its own, may not be sensitive, but when compiled, may generate a sensitive result. These linkages can result in a broad portrait of an individual including revelations of a sensitive nature—a portrait once inconceivable since the identifiers were separated in various databases. Through the use of Big Data tools, we also know that it is possible to identify patterns which may predict people's dispositions, for example related to health, political viewpoints or sexual orientation (Cavoukian and Jonas 2012).

By connecting key pieces of data that link people to things, the capability of data analytics can render ordinary data into information about an identifiable individual and reveal details about a person's lifestyle and habits. A telephone number or postal code, for example, can be combined with other data to identify the location of a person's home and work; an IP or email address can be used to identify consumer habits and social networks.

An important trend and contribution to Big Data is the movement by government institutions to open up their data holdings in an effort to enhance citizen participation in government and at the same time spark innovation and new insights through access to invaluable government data (Cavoukian 2009).[2]

With this potential for Big Data to create data linkages being so powerful, the term "super" data or "super" content has been introduced (Cameron 2013). "Super" data is more powerful than other data in a Big Data context, because the use of one piece of "super" data, which on its own would not normally reveal much, can spark new data linkages that grow exponentially until the individual is identified. Each new transaction in a Big Data system would compound this effect and spread identifiability like a contagion.

Indeed, to illustrate the significant implications of data maximization on privacy we need only look at the shock of the Snowden revelations and the eventual repercussions. A top EU court decision in 2015 declared the longstanding Safe Harbor

[2]There are many government Open Data initiatives such as U.S. Government's Open Data at www.data.gov; Canadian Government's Open Data at http://open.canada.ca/en/open-data; UN Data at http://data.un.org/; EU Open Data Portal at https://data.europa.eu/euodp/en/data/. This is just a sample of the many Open Data sources around the world.

data transfer agreement between Europe and the U.S. invalid (Lomas 2015). The issues had everything to do with concerns about not just government surveillance but the relationship with U.S. business and their privacy practices. Eventually, a new agreement was introduced known as the EU-U.S. Privacy Shield (US DOC 2016) (EU Commission 2016). This new mechanism introduces greater transparency requirements for the commercial sector on their privacy practices among a number of other elements including U.S. authorities affirming that collection of information for intelligence is focussed and targeted.

The authors strongly believe that an important lesson learned for Big Data success is that when the individual participant is more directly involved in information collection, the accuracy of the information's context grows and invariably increases the quality of the data under analysis. Another observation, that may seem to be contradictory, is that even in Big Data scenarios where algorithms are tasked with finding connections within vast datasets, data minimization is not only essential for safeguarding personally identifiable information—it could help with finding the needle without the haystack by reducing extraneous irrelevant data.

2.3.2 Predictive Analysis: Correlation Versus Causation

Use of correlation analysis may yield completely incorrect results for individuals. Correlation is often mistaken for causality (Ritter 2014). If the analyses show that individuals who like X have an eighty per cent probability rating of being exposed to Y, it is impossible to conclude that this will occur in 100 per cent of the cases. Thus, discrimination on the basis of statistical analysis may become a privacy issue (Sweeney 2013). A development where more and more decisions in society are based on use of algorithms may result in a "Dictatorship of Data", (Cukier and Mayer-Schonberger 2013) where we are no longer judged on the basis of our actual actions, but on the basis of what the data indicate will be our probable actions. In a survey undertaken by the Annenberg Public Policy Center, the researchers found that most Americans overwhelmingly consider forms of price discrimination and behavioral targeting ethically wrong (Turow et al. 2015). Not only are these approaches based on profiling individuals but using personal information about an individual for purposes the individual is unaware of. The openness of data sources and the power of not just data mining but now predictive analysis and other complex algorithms also present a challenge to the process of de-identification. The risks of re-identification are more apparent, requiring more sophisticated de-identification techniques (El Emam et al. 2011). In addition, while the concept of "nudging" is gaining popularity, using identifiable data for profiling individuals to analyse, predict, and influence human behaviour may be perceived as invasive and unjustified surveillance.

Data determinism and discrimination are also concerns that arise from a Dictatorship of Data. Extensive use of automated decisions and prediction analyses may actually result in adverse consequences for individuals. Algorithms are not neutral, but reflect choices, among others, about data, connections, inferences,

interpretations, and thresholds for inclusion that advances a specific purpose. The concern is that Big Data may consolidate existing prejudices and stereotyping, as well as reinforce social exclusion and stratification (Tene and Polonetsky 2013; IWGDPT 2014; FTC 2016). This is said to have implications for the quality of Big Data analysis because of "echo chambers"[3] in the collection phase (Singer 2011; Quattrociocchi et al. 2016).

2.3.3 Lack of Transparency/Accountability

As an individual's personal information spreads throughout the Big Data ecosystem amongst numerous players, it is easy to see that the individual will have less control over what may be happening to the data. This secondary use of data raises privacy concerns. A primary purpose is identified at the time of collection of personal information. Secondary uses are generally permitted with that person's consent, unless otherwise permitted by law. Using personal information in Big Data analytics may not be permitted under the terms of the original consent as it may constitute a secondary use—unless consent to the secondary use is obtained from the individual. This characteristic is often linked with a lack of transparency. Whether deliberate or inadvertent, lack of openness and transparency on how data is compiled and used, is contrary to a fundamental privacy principle.

It is clear that organizations participating in the Big Data ecosystem need to have a strong privacy program in place (responsible information management). If individuals don't have confidence that their personal information is being managed properly in Big Data applications, then their trust will be eroded and they may withdraw or find alternative mechanisms to protect their identity and privacy. The consequences of a privacy breach can include reputational harm, legal action, damage to a company's brand or regulatory sanctions and disruption to internal operations. In more severe cases, it could cause the demise of an organization (Solove 2014). According to TRUSTe's Consumer Privacy Confidence Index 2016, 92 per cent of individuals worry about their privacy online, 44 per cent do not trust companies with their personal information, and 89 per cent avoid doing business with companies that they believe do not protect their privacy (TRUSTe/NCSA 2016).

Despite the fact that privacy and security risks may exist, organizations should not fear pursuing innovation through data analytics. Through the application of privacy controls and use of appropriate privacy tools privacy risks may be mitigated, thereby enabling organizations to capitalize on the transformative potential of Big Data—while adequately safeguarding personal information. This is the central

[3] In news media an echo chamber is a metaphorical description of a situation in which information, ideas, or beliefs are amplified or reinforced by transmission and repetition inside an "enclosed" system, where different or competing views are censored, disallowed, or otherwise underrepresented. The term is by analogy with an acoustic echo chamber, where sounds reverberate.

motivation for Privacy by Design, which is aimed at preventing privacy violations from arising in the first place. Given the necessity of establishing user trust in order to gain public acceptance of its technologies, any organization seeking to take advantage of Big Data must apply the Privacy by Design framework as new products and applications are developed, marketed, and deployed.

2.4 Privacy by Design and the 7 Foundational Principles

The premise of Privacy by Design has at its roots, the Fair Information Practices or FIPs. Indeed, most privacy laws around the world are based on these practices. By way of history, the Code of Fair Information Practices (FIPs) was developed in the 1970s and based on essentially five principles (EPIC n.d.):

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.
3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

FIPs represented an important development in the evolution of data privacy since they provided an essential starting point for responsible information management practices. However, many organizations began to view enabling privacy via FIPs and associated laws as regulatory burdens that inhibited innovation. This zero-sum mindset viewed the task of protecting personal information as a "balancing act" of competing business and privacy requirements. This balancing approach tended to overemphasize the significance of notice and choice as the primary method for addressing personal information data management. As technologies developed, the possibility for individuals to meaningfully exert control over their personal information became more and more difficult. It became increasingly clear that FIPs were a necessary but not a sufficient condition for protecting privacy. Accordingly, the attention of privacy protection had begun to shift from reactive compliance with FIPs to proactive system design.

With advances in technologies, it became increasingly apparent that systems needed to be complemented by a set of norms that reflect broader privacy dimensions (Damiani 2013). The current challenges to privacy related to the dynamic relationship associated with the forces of innovation, competition and the global adoption of information communications technologies. These challenges have been

mirrored in security by design. Just as users rely on security engineers to ensure the adequacy of encryption key lengths, for example, data subjects will rely on privacy engineers to appropriately embed risk-based controls within systems and processes. Given the complex and rapid nature of these developments, it becomes apparent that privacy has to become the default mode of design and operation.

Privacy by Design (PbD), is a globally recognized proactive approach to privacy. It is a framework developed in the late 1990s by co-author Dr. Ann Cavoukian (Cavoukian 2011). Privacy by Design is a response to compliance-based approaches to privacy protection that tend to focus on addressing privacy breaches after-the-fact. Our view is that this reactive approach does not adequately meet the demands of the Big Data era. Instead, we recommend that organizations consciously and proactively incorporate privacy strategies into their operations, by building privacy protections into their technology, business strategies, and operational processes.

By taking a proactive approach to privacy and making privacy the default setting, PbD can have a wide-ranging impact across an organization. The approach can result in changes to governance structures, operational and strategic objectives, roles and accountabilities, policies, information systems and data flows, decision-making processes, relationships with stakeholders, and even the organization's culture.

PbD has been endorsed by many public- and private-sector authorities in the United States, the European Union, and elsewhere (Harris 2015). In 2010, PbD was unanimously passed as a framework for privacy protection by the International Assembly of Privacy Commissioners and Data Protection Authorities (CNW 2010). This approach transforms consumer privacy issues from a pure policy or compliance issue into a business imperative. Since getting privacy right has become a critical success factor to any organization that deals with personal information, taking an approach that is principled and technology-neutral is now more relevant than ever. Privacy is best interwoven proactively and to achieve this, privacy principles should be introduced early on—during architecture planning, system design, and the development of operational procedures. Privacy by Design, where possible, should be rooted into actual code, with defaults aligning both privacy and business imperatives.

The business case for privacy focuses on gaining and maintaining customer trust, breeding loyalty, and generating repeat business. The value proposition typically reflects the following:

1. Consumer trust drives successful customer relationship management (CRM) and lifetime value—in other words, business revenues;
2. Broken trust will result in a loss of market share and revenue, translating into less return business and lower stock value; and
3. Consumer trust hinges critically on the strength and credibility of an organization's data privacy policies and practices.

In a marketplace where organizations are banding together to offer suites of goods and services, trust is clearly essential. Of course, trust is not simply an end-user issue. Companies that have done the work to gain the trust of their customers cannot risk losing it as a result of another organization's poor business practices.

2.4.1 The 7 Foundational Principles

Privacy by Design Foundational Principles build upon universal FIPPs in a way that updates and adapts them to modern information management needs and requirements. By emphasizing proactive leadership and goal-setting, systematic and verifiable implementation methods, and demonstrable positive-sum results, the principles are designed to reconcile the need for robust data protection and an organization's desire to unlock the potential of data-driven innovation. Implementing PbD means focusing on, and living up to, the following 7 Foundational Principles, which form the essence of PbD (Cavoukian 2011).

Principle 1: Use proactive rather than reactive measures, anticipate and prevent privacy invasive events *before* they happen (*Proactive* not Reactive; *Preventative* not Remedial).

Principle 2: Personal data must be automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact (Privacy as the *Default*). Data minimization is also a default position for privacy, i.e. the concept of always starting with the minimum personal data possible and then justifying additional collection, disclosure, retention, and use on an exceptional and specific data-by-data basis.

Principle 3: Privacy must be embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. Privacy is integral to the system, without diminishing functionality (Privacy *Embedded* into Design).

Principle 4: All legitimate interests and objectives are accommodated in a positive-sum manner (Full Functionality—*Positive-Sum* [win/win], not Zero-Sum [win/lose]).

Principle 5: Security is applied throughout the entire lifecycle of the data involved—data is securely retained, and then securely destroyed at the end of the process, in a timely fashion (End-to-End Security—*Full Lifecycle Protection*).

Principle 6: All stakeholders are assured that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification; transparency is key (*Visibility* and *Transparency*—Keep it *Open*).

Principle 7: Architects and operators must keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options (*Respect* for User Privacy—Keep it *User-Centric*).

2.5 Big Data Applications: Guidance on Applying the PbD Framework and Principles

While the 7 Foundational Principles of PbD should be applied in a holistic manner as a broad framework, there are specific principles worthy of pointing out because they are what defines and distinguishes this approach to privacy. These are principles 1 (Proactive and Preventative), 2 (By Default/Data Minimization), 3 (Embedded in Design) and 4 (Positive-sum). Although the two examples provided below are specific to mobile apps, they are illustrative of the Privacy by Design approach to being proactive, focussing on data minimization and embedding privacy by default.

2.5.1 Being Proactive About Privacy Through Prevention

Privacy by Design aspires to the highest global standards of practical privacy and data protection possible and to go beyond compliance and achieve visible evidence and recognition of leadership, regardless of jurisdiction. Good privacy doesn't just happen by itself—it requires proactive and continuous goal-setting at the earliest stages. Global leadership in data protection begins with explicit recognition of the benefits and value of adopting strong privacy practices, early and consistently (e.g., preventing data breaches or harms to individuals from occurring in the first place).

> Your app's main purpose is to display maps. These maps are downloaded by a mobile device from your central server. They are then later used on the device, when there may be no network connection available. You realise that analytics would be useful to see which maps are being downloaded by which users. This in turn would allow you to make targeted suggestions to individual users about which other maps they might want to download. You consider using the following to identify individuals who download the maps: i) the device's IMEI number; ii) the MAC address of the device's wireless network interface; and iii) the mobile phone number used by the device. You realise that any of those identifiers may constitute personal data, so for simplicity you decide not to take on the responsibility of dealing with them yourself. Instead, you decide to gain users' consent for the map suggestions feature. When a user consents, they are assigned a randomly generated unique identifier, solely for use by your app. (Excerpted from Information Commissioner's Office ICO 2013)

2.5.2 Data Minimization as the Default Through De-identification

Personal information that is not collected, retained, or disclosed is data that does not need to be protected, managed, or accounted for. If the personal information does not exist, then it cannot be accessed, altered, copied, enriched, shared, lost, hacked, or otherwise used for secondary and unauthorized purposes. Privacy by Design is premised on the idea that the starting point for designing information technologies and systems should always be maximally privacy-enhancing. The default configuration or settings of technologies, tools, platforms, or services offered to individuals should be as restrictive as possible regarding use of personally identifiable data.

When Big Data analytics involves the use of personally identifiable information, data minimization has the biggest impact on managing data privacy risks, by effectively eliminating risk at the earliest stage of the information life cycle. Designing Big Data analytical systems at the front end with *no* collection of personally identifiable information—unless and until a specific and compelling purpose is defined, is the ideal. For example, use(s) of personal information should be limited to the intended, primary purpose(s) of collection and only extended to other, non-consistent uses with the explicit consent of the individual (Article 29 Data Protection Working Party 2013). In other cases, organizations may find that summary or aggregate data may be more than sufficient for their needs.

> Your app uses GPS location services to recommend interesting activities near to where the user is. The database of suggested activities is kept on a central server under your control. One of your design goals is to keep the amount of data your app downloads from the central server to a minimum. You therefore design your app so that each time you use it, it sends location data to the central server so that only the nearest activities are downloaded. However, you are also keen to use less privacy-intrusive data where possible. You design your app so that, by default, the device itself works out where the nearest town is and uses this location instead, avoiding the need to send exact GPS co-ordinates of the user"s location back to the central server. Users who want results based on their accurate location can change the default behaviour. (Excerpted from Information Commissioner's Office ICO 2013)

De-identification strategies are considered data minimization. De-identification provides for a set of tools or techniques to strip a dataset of all information that could be used to identify an individual, either directly or indirectly, through linkages to other datasets. The techniques involve deleting or masking "direct identifiers," such as names or social insurance numbers, and suppressing or generalizing indirect identifiers, such as postal codes or birthdates. Indirect identifiers may not be

personally identifying in and of themselves, but when linked to other datasets that contain direct identifiers, may personally identify individuals. If done properly, de-identified data can be used for research purposes and data analysis—thus contributing new insights and achieving innovative goals—while minimizing the risk of disclosure of the identities of the individuals behind the data (Cavoukian and El Emam 2014).

This is not to suggest, of course, that data should be collected exclusively in instances where it may become useful or that data collected for one purpose may be repurposed at will. Rather, in a big data world, the principle of data minimization should be interpreted differently, requiring organizations to de-identify data when possible, implement reasonable security measures, and limit uses of data to those that are acceptable from not only an individual but also a societal perspective (Tene and Polonetsky 2013).

2.5.3 Embedding Privacy at the Design Stage

When privacy commitments and data protection controls are embedded into technologies, operations, and information architectures in a holistic, integrative manner, innovation and creativity are often by-products (Cavoukian et al. 2014a, b). By holistic, we mean that broader contexts should always be considered for a proper assessment of privacy risks and remedies. An integrative approach takes into consideration all stakeholder interests as part of the development dialogue. Sometimes, having to re-look at alternatives because existing solutions are unacceptable from a privacy perspective spurs innovative and creative thinking. Embedding privacy and data protection requires taking a systematic, principled approach—one that not only relies on accepted standards and process frameworks, but that can stand up to external reviews and audits. All of the 7 Foundational Principles should be applied with equal rigour, at every step in design and operation. By doing so, the privacy impacts of the resulting technology, process, or information architecture, and their uses, should be demonstrably minimized, and not easily degraded through use, misconfiguration, or error. To minimize concerns of untoward data usage, organizations should disclose the logic underlying their decision-making processes to the extent possible without compromising their trade secrets or intellectual property rights.

The concept of "user-centricity" may evoke contradictory meanings in networked or online environments. Through a privacy lens, it contemplates a right of control by an individual over his or her personal information when online, usually with the help of technology. For most system designers, it describes a system built with individual users in mind that may perhaps incorporate users' privacy interests, risks and needs. The first may be considered libertarian (informational self-determination), the other, paternalistic. Privacy by Design embraces both. It acknowledges that technologies, processes and infrastructures must be designed not just for individual users, but also structured by them. Users are rarely, if ever, involved in every design decision

or transaction involving their personal information, but they are nonetheless in an unprecedented position today to exercise a measure of meaningful control over those designs and transactions, as well as the disposition and use of their personal information by others.

User interface designers know that human-computer interface can often make or break an application. Function (substance) is important, but the way in which that function is delivered is equally as important. This type of design embeds an effective user privacy experience. As a quid pro quo for looser data collection and minimization restrictions, organizations should be prepared to share the wealth created by individuals' data with those individuals. This means providing individuals with access to their data in a "usable" format and allowing them to take advantage of third party applications to analyze their own data and draw useful conclusions (e.g., consume less protein, go on a skiing vacation, invest in bonds) (Tene and Polonetsky 2013).

2.5.4 Aspire for Positive-Sum Without Diminishing Functionality

In Big Data scenarios, networks are more complex and sophisticated thereby undermining the dominant "client-server" transaction model because individuals are often far removed from the client side of the data processing equation. How could privacy be assured when the collection, disclosure, and use of personal information might not even involve the individual at all? Inevitably, a zero-sum paradigm prevails where more of one good (e.g., public security, fraud detection, operational control) cancels out another good (individual privacy, freedom). The authors challenge the premise that privacy and data protection necessarily have to be ceded in order to gain public, personal, or information security benefits from Big Data. The opposite of zero-sum is positive-sum, where multiple goals may be achieved concurrently.

Many security technologies and information systems could be designed (or redesigned) to be effective while minimizing or even eliminating their privacy-invasive features. This is the positive-sum paradigm. We need only look to the work of researchers in the area of privacy preserving data mining (Lindell and Pinkas 2002). In some cases, however, this requires broadening the scope of application from only information communication technologies (ICTs) to include the "soft" legal, policy, procedural, and other organizational controls and operating contexts in which privacy might be embedded.

De-identification tools and techniques are gaining popularity and there are several commercially available products. Nonetheless, furthering research into de-identification continues (El Emam 2013a, b). Some emerging research-level technologies hold much promise for enabling privacy and utility of Big Data analysis to co-exist. Two of these technologies are differential privacy and synthetic data.

Differential privacy is an approach that injects random noise into the results of dataset queries to provide a mathematical guarantee that the presence of any one individual in the dataset will be masked—thus protecting the privacy of each individual in the dataset. Typical implementations of differential privacy work by creating a query interface or "curator" that stands between the dataset's personal information and those wanting access to it. An algorithm evaluates the privacy risks of the queries. The software determines the level of "noise" to introduce into the analysis results before releasing it. The distortion that is introduced is usually small enough that it does not affect the quality of the answers in any meaningful way—yet it is sufficient to protect the identities of the individuals in the dataset (Dwork 2014).

At an administrative level, researchers are not given access to the dataset to analyze themselves when applying differential privacy. Not surprisingly, this limits the kinds of questions researchers can ask. Given this limitation, some researchers are exploring the potential of creating "synthetic" datasets for researchers' use. As long as the number of individuals in the dataset is sufficiently large in comparison to the number of fields or dimensions, it is possible to generate a synthetic dataset comprised entirely of "fictional" individuals or altered identities that retain the statistical properties of the original dataset—while delivering differential privacy's mathematical "noise" guarantee (Blum et al. 2008). While it is possible to generate such synthetic datasets, the computational effort required to do so is usually extremely high. However, there have been important developments into making the generation of differentially private synthetic datasets more efficient and research continues to show progress (Thaler et al. 2010).

2.6 Conclusion

There are privacy and security risks and challenges that organizations will face in the pursuit of Big Data nirvana. While a significant portion of this vast digital universe is not of a personal nature, there are inherent privacy and security risks that cannot be overlooked. Make no mistake, organizations must seriously consider not just the use of Big Data but also the implications of a failure to fully realize the potential of Big Data. Big data and big data analysis, promise new insights and benefits such as medical/scientific discoveries, new and innovative economic drivers, predictive solutions to otherwise unknown, complex societal problems. Misuses and abuses of personal data diminish informational self-determination, cause harms, and erode the confidence and trust needed for innovative economic growth and prosperity. By examining success stories and approaches such as Privacy by Design, the takeaway should be practical strategies to address the question of 'How do we achieve the value of Big Data and still respect consumer privacy?' Above all, Privacy by Design requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric!

References

Article 29 Data protection working party (2013). *Opinion 03/2013 on purpose limitation*. http://ec.europa.eu/justice/data-protection/index_en.htm. Accessed 2 August 2016.

Blum, A., Ligett, K., Roth, A. (2008). A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing* (pp. 609–618).

Cameron, K. (2013). Afterword. In M. Hildebrandt et al. (Eds.), *Digital Enlightenment Yearbook 2013*. Amsterdam: IOS Press.

Cavoukian, A. (2009). *Privacy and government 2.0: the implications of an open world*. http://www.ontla.on.ca/library/repository/mon/23006/293152.pdf. Accessed 22 November 2016.

Cavoukian, A. (2011). *Privacy by Design: The 7 Foundational Principles*. Ontario: IPC.

Cavoukian, A. (2013a). *A Primer on Metadata: Separating Fact from Fiction*. Ontario: IPC. http://www.ipc.on.ca/images/Resources/metadata.pdf.

Cavoukian, A. (2013b). Privacy by design: leadership, methods, and results. In S. Gutwirth, R. Leenes, P. de Hert, & Y. Poullet (Eds.), *Chapter in European Data Protection: Coming of Age* (pp. 175–202). Dordrecht: Springer Science & Business Media Dordrecht.

Cavoukian, A., & Cameron, K. (2011). *Wi-Fi Positioning Systems: Beware of Unintended Cosnequences: Issues Involving Unforeseen Uses of Pre-Existing Architecture*. Ontario: IPC.

Cavoukian, A., & El Emam. (2014). *De-identification Protocols: Essential for Protecting Privacy*, Ontario: IPC.

Cavoukian, A., & Jonas, J. (2012). *Privacy by Design in the Age of Big Data*. Ontario: IPC.

Cavoukian, A., Bansal, N., & Koudas, N. (2014a). *Building Privacy into Mobile Location Analytics (MLA) through Privacy by Design*. Ontario: IPC.

Cavoukian, A., Dix, A., & El Emam, K. (2014b). *The Unintended Consequences of Privacy Paternalism*. Ontario: IPC.

Clarke, R. (2000). *Beyond OECD guidelines; privacy protection for the 21st century*. Xamax Consultancy Pty Ltd. http://www.rogerclarke.com/DV/PP21C.html. Accessed 22 November 2016.

CNW (2010). Landmark resolution passed to preserve the future of privacy. Press Release. Toronto, ON, Canada. http://www.newswire.ca/news-releases/landmark-resolution-passed-to-preserve-the-future-of-privacy-546018632.html. Accessed 22 November 2016.

Cukier, K., & Mayer-Schonberger, V. (2013). The dictatorship of data. *MIT Technology Review*. https://www.technologyreview.com/s/514591/the-dictatorship-of-data/. Accessed 22 November 2016.

Damiani, M. L. (2013). Privacy enhancing techniques for the protection of mobility patterns in LBS: research issues and trends. In S. Gutwirth, R. Leenes, P. de Hert, & Y. Poullet (Eds.), *Chapter in european data protection: coming of age* (pp. 223–238). Dordrecht: Springer Science & Business Media Dordrecht.

Department of Commerce (US DOC) (2016). *EU-U.S. privacy shield fact sheet. Office of public affairs, US department of commerce*. https://www.commerce.gov/news/fact-sheets/2016/02/eu-us-privacy-shield. Accessed 22 November 2016.

Dwork, C. (2014). Differential privacy: a cryptographic approach to private data analysis. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, big data, and the public good: Frameworks for engagement*. New York: Cambridge University Press.

El Emam, K. (2013a). Benefiting from big data while protecting privacy. In K. El Emam (Ed.), *Chapter in risky business: sharing health data while protecting privacy*. Bloomington, IN: Trafford Publishing.

El Emam, K. (2013b). In K. El Emam (Ed.), *Who's afraid of big data? chapter in risky business: Sharing health data while protecting privacy*. Bloomington, IN, USA: Trafford Publishing.

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making*, *11*:46. http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-11-46. Accessed 22 November 2016.

EPIC (n.d.). Website: https://epic.org/privacy/consumer/code_fair_info.html. Accessed 22 November 2016.

EU Commission (2012). *Fact sheet on the right to be forgotten*. http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf. Accessed 22 November 2016.

EU Commission (2015). *Fact sheet—questions and answers—data protection reform*. Brussels. http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm. Accessed 4 November 2016.

EU Commission (2016). *The EU data protection reform and big data factsheet*. http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf. Accessed 22 November 2016.

Fogarty, D., & Bell, P. C. (2014). Should you outsource analytics? *MIT Sloan Management Review*, *55*(2), Winter.

FTC (2012). *Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers*. https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf Accessed August 2016.

FTC (2016). *Big data: A tool for inclusion or exclusion? Understanding the Issues*. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf. Accessed 23 November 2016.

Gürses, S.F. Troncoso, C., & Diaz, C. (2011). *Engineering privacy by design, Computers, Privacy & Data Protection*. http://www.cosic.esat.kuleuven.be/publications/article-1542.pdf. Accessed 19 November 2016.

Harris, M. (2015). *Recap of covington's privacy by design workshop. inside privacy: updates on developments in data privacy and cybsersecurity*. Covington & Burlington LLP, U.S. https://www.insideprivacy.com/united-states/recap-of-covingtons-privacy-by-design-workshop/. Accessed 19 November 2016.

HHS (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPPA) privacy rule*. http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed 2 August 2016.

Information Commissioner's Office (ICO) (2013). *Privacy in Mobile Apps: Guide for app developers*. https://ico.org.uk/media/for-organisations/documents/1596/privacy-in-mobile-apps-dp-guidance.pdf Accessed 22 November 2016.

International Working Group on Data Protection in Telecommunications (IWGDPT) (2004) *Common position on privacy and location information in mobile communications services*. https://datenschutz-berlin.de/content/europa-international/international-working-group-on-data-protection-in-telecommunications-iwgdpt/working-papers-and-common-positions-adopted-by-the-working-group. Accessed 22 November 2016.

International Working Group on Data Protection in Telecommunications (IWGDPT) (2014). Working Paper on Big Data and Privacy: Privacy principles under pressure in the age of Big Data analytics. *55th Meeting*. https://datenschutz-berlin.de/content/europa-international/international-working-group-on-data-protection-in-telecommunications-iwgdpt/working-papers-and-common-positions-adopted-by-the-working-group. Accessed 22 November 2016.

Lane, J., et al. (2014). *Privacy, big data and the public good: frameworks for engagement*. Cambridge: Cambridge University Press.

Lindell, Y., & Pinkas, B. (2002). Privacy preserving data mining. *Journal of Cryptology, 15*, 177–206. International Association for Cryptologic Research.

Lomas, N. (2015). *Europe's top court strikes down safe Harbor data-transfer agreement with U.S. Techcrunch*. https://techcrunch.com/2015/10/06/europes-top-court-strikes-down-safe-harbor-data-transfer-agreement-with-u-s/. Accessed 22 November 2016.

Mayer, J., Mutchler, P., & Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academies of Science, U S A, 113*(20), 5536–5541.

NIST. (2010). *Guide to protecting the confidentiality of personally identifiable information (PII). NIST special publication 800–122*. Gaithersburg, MD: Computer Science Division.

NIST (2015). *De-identification of Personal Information. NISTR 8053*. This publication is available free of charge from: http://dx.doi.org/10.6028/NIST.IR.8053. Accessed 19 November 2016.

Official Journal of the European Union (2016). *Regulation (EU) 2016/679 Of The European Parliament and of the Council*. http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf. Accessed 19 November 2016.

Quattrociocchi, W. Scala, A., & Sunstein, C.R. (2016) *Echo Chambers on Facebook. Preliminary draft, not yet published*. Available at: http://ssrn.com/abstract=2795110. Accessed 19 November 2016.

Ritter, D. (2014). *When to Act on a correlation, and when Not To. Harvard Business Review*. https://hbr.org/2014/03/when-to-act-on-a-correlation-and-when-not-to. Accessed 19 November 2016.

Singer, N. (2011). The trouble with the echo chamber online. *New York Times online*. http://www.nytimes.com/2011/05/29/technology/29stream.html?_r=0. Accessed 19 November 2016.

Solove, D. J. (2007). I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review, 44*, 745.

Solove, D. (2014). Why did in bloom die? A hard lesson about education privacy. Privacy + Security Blog. TeachPrivacy. Accessed 4 Aug 2016. https://www.teachprivacy.com/inbloom-die-hard-lesson-education-privacy/

Sweeney, L. (2013) *Discrimination in online ad delivery*. http://dataprivacylab.org/projects/onlineads/1071-1.pdf. Accessed 22 November 2016.

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *New Journal of Technology and Intellectual Property, 11*(5), 239–272.

Thaler, J., Ullman, J., & Vadhan, S. (2010). PCPs and the hardness of generating synthetic data. *Electronic Colloquium on Computational Complexity, Technical Report*, TR10–TR07.

TRUSTe/NCSA (2016). *Consumer privacy infographic—US Edition*. https://www.truste.com/resources/privacy-research/ncsa-consumer-privacy-index-us/. Accessed 4 November 2016.

Turow, J., Feldman, L, & Meltzer, K. (2015). *Open to exploitation: american shoppers online and offline. A report from the Annenberg Public Policy Center of the University of Pennsylvania*. http://www.annenbergpublicpolicycenter.org/open-to-exploitation-american-shoppers-online-and-offline/. Accessed 22 November 2016.

United Nations General Assembly (2016). *Resolution adopted by the General Assembly. The right to privacy in the digital age (68/167)*. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/68/167. Accessed 4 November 2016.

Zhang, Y., Chen, Q., & Zhong, S. (2016). Privacy-preserving data aggregation in mobile phone sensing. *Information Forensics and Security IEEE Transactions on, 11*, 980–992.

http://www.springer.com/978-3-319-53816-7

Guide to Big Data Applications
Srinivasan, S. (Ed.)
2018, XVII, 565 p. 205 illus., 155 illus. in color.,
Hardcover
ISBN: 978-3-319-53816-7