Using *Privacy by Design* to Achieve Big Data Innovation Without Compromising Privacy

www.privacybydesign.ca

Ann Cavoukian, Ph.D.
Information and Privacy Commissioner
Ontario, Canada

David Stewart
National Advanced Analytics Leader
Deloitte

Beth Dewitt
Manager and Privacy Specialist
Deloitte

June 10, 2014

ACKNOWLEDGEMENTS

The authors wish to acknowledge Megan Brister, National Privacy Leader, Deloitte and Michelle Chibba, Director, Policy & Special Projects, IPC. Our thanks also go to Catherine Thompson, Regulatory and Policy Advisor and David Weinkauf, Policy & Information Technology Officer at IPC and Michelle Gordon, Daniel Horovitz and Sylvia Kingsmill at Deloitte for their contributions to this paper.

Information and Privacy Commissioner
Ontario, Canada

2 Bloor Street East
Suite 1400
Toronto, Ontario
M4W 1A8
Canada

416-326-3333
1-800-387-0073
Fax: 416-325-9195
TTY (Teletypewriter): 416-325-7539
Website: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

Using *Privacy by Design* to Achieve Big Data Innovation Without Compromising Privacy

TABLE OF CONTENTS

Foreword

The argument that privacy stifles Big Data innovation reflects a dated, zero-sum mindset. It is a false dichotomy, consisting of unnecessary trade-offs between the benefits of Big Data and the protection of personal information within Big Data sets. In fact, the opposite is true—privacy drives innovation and it forces innovators to think creatively to find solutions that serve multiple functionalities. We need to abandon zero-sum thinking and adopt a positive-sum paradigm where both Big Data innovation and privacy may be achieved.

Did you know that you can't have quality Big Data results without privacy? That is because context is a key factor in Big Data. When Google Flu's ability to predict the spread of influenza was found to be overstated, the reason cited was missing information from the data subjects on why they were Googling flu-related search terms. Data collected directly from the individual with their knowledge and consent invariably increases the quality of the data under analysis.

The use of privacy tools within Big Data can allow for the protection of personal information while also allowing for analysis on that data. Some of these techniques are de-identification, data aggregation, and emerging technologies such as differential privacy and synthetic data which will be explained further in this paper. Even in Big Data scenarios where algorithms are tasked with finding connections within vast datasets, data minimization should also be considered as a tool for safeguarding personally identifiable information—it could help with finding the needle *without* the haystack.

Privacy is just as big as Big Data. The tools exist to systemically protect personal information and to bring about the benefits of Big Data. Together we can ensure that Big Data and 'Big Privacy' can both be accomplished in a win-win scenario.

Ann Cavoukian, Ph.D.

Information and Privacy Commissioner
Ontario, Canada

Big Data and Privacy are *not* mutually exclusive

Data is one of the most valuable assets of any organization. The intelligence that can be driven through the application of analytic techniques can provide essential insights that decision-makers will need to develop strategy, deliver growth and operational performance, and manage risk. Data is increasingly becoming the oxygen of modern business.

The amount of data generated by individuals, Internet-connected devices, and businesses is growing at an exponential rate. Financial services, retail, and healthcare organizations, for example, generate vast amounts of data during their interactions with vendors, patients, customers, and employees. Even more data is being created outside these organizations through Internet search queries, social media, mobile-device GPS location information, stock transactions, and more.

If we consider the traditional intelligence cycle, this is a continuous process. Raw data is collected, and this source data is then analyzed, transformed, and connected to other raw datasets. This processing enables existing knowledge to be applied to both analytic insights and new insights created by the data transformation process itself—and this, in turn, creates *information*. Subject matter experts process this new information, and their interpretation creates *intelligence*. This intelligence is then disseminated, and the next set of business priorities are determined.

There are currently 9.6 billion Internet-connected devices,[1] 1.3 billion mobile broadband connections,[2] and 1.2 zettabytes (10^{21}) of annual global IP traffic.[3] Every two days, our use of these devices creates roughly five exabytes (10^{18}) of data—as much as all the data created by humans from the dawn of civilization to 2003.[4] The result is what has now become known as the data revolution or the era of "Big Data."

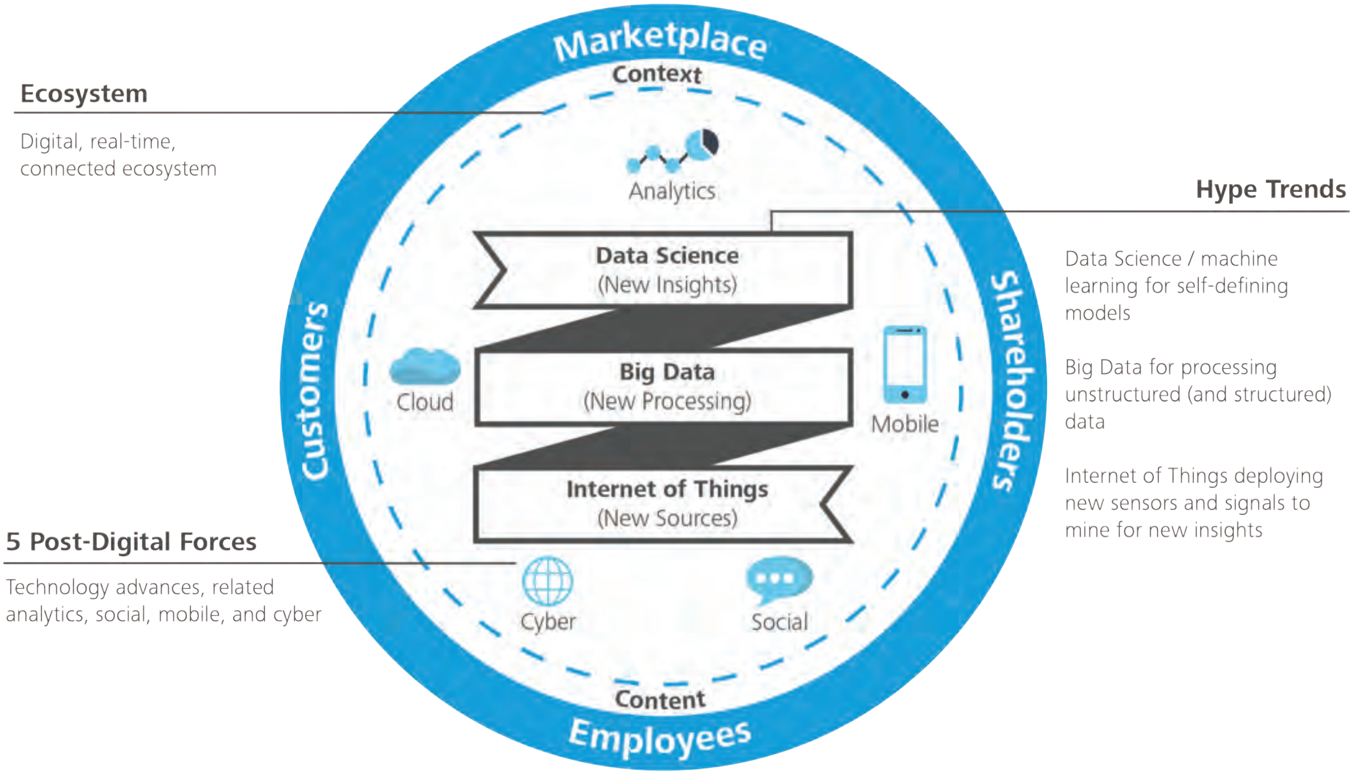
The term "Big Data" is used to describe a universe of very large datasets that hold a variety of data types. This has spawned a new generation of technology and information architecture to facilitate the fast processing speeds needed to analyze and extract value from these extremely large sets of data using distributed platforms. In common usage, "Big Data" is used to refer both to these vast datasets and also to the process of analyzing and extracting value from enormous amounts of data across multiple silos of information.

[1] IMS Research, "Internet connected devices approaching 10 billion, to exceed 28 billion by 2020," October 2012, http://imsresearch.com/press-release/Internet_Connected_Devices_Approaching_10_Billion_to_exceed_28_Billion_by_2020&cat_id=113&type=LatestResearch.

[2] GSMA, http://www.gsma.com/newsroom/gsma-research-demonstrates-that-mobile-industry-is-creating-a-connected-economy.

[3] Cisco, Global Cloud Index (2012–2017), http://www.cisco.com/c/en/us/solutions/service-provider/global-cloud-index-gci/index.html.

[4] M.G. Siegler, "Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up to 2003," TechCrunch, August 4, 2010, http://techcrunch.com/2010/08/04/schmidt-data.



Ecosystem

Digital, real-time, connected ecosystem

Hype Trends

Data Science / machine learning for self-defining models

Big Data for processing unstructured (and structured) data

Internet of Things deploying new sensors and signals to mine for new insights

5 Post-Digital Forces

Technology advances, related analytics, social, mobile, and cyber

Big Data plays an essential role in what Deloitte refers to as the Digital Enterprise—our evolving vision of where business is heading in the next few years. Already, mobile advances have put incredible technology in everyone’s hands. Social networks enable people to connect in ways never before possible. The cloud is drastically reducing the costs associated with hardware and data infrastructure, while data storage capabilities grow ever larger and ever cheaper. Data analytics can now make sense of vast amounts of data to provide valuable, actionable insights.

Data Analytics: Inspiring Disruption[5]

Understandably, organizations are keen to unlock data's potential for its business value. They are eager to find ways to use the data for making smarter decisions that will result in better service for their customers, improved efficiencies for their processes, and better outcomes against their strategies.

Recent and rapid advances in data processing speeds and analytical algorithms make it possible to process these large amounts of structured and unstructured data at very high speeds. Today's data analytics enables organizations to make connections, identify patterns, predict behaviour, and personalize interactions to an extent only dreamed of before.

As a result, data analytics is accelerating the pace of innovation and disrupting traditional business models. It allows retailers to deliver offers finely tailored to their customers' preferences and purchasing behaviour. It enables financial services firms to deliver proactive advice and product recommendations. It helps healthcare organizations improve diagnoses, treatments, and public health management. In some industries, competitors are sharing data to address common concerns, such as fraud, cyber security, and health and safety performance.

The public sector is also exploring the potential of data analytics through "open government" and "open data" initiatives. Such initiatives are making Big Data available to the public, often for the first time.[6] In part, these initiatives are designed to increase government transparency and encourage public engagement. Governments also hope that citizens and organizations will be able to use this data to develop new insights and innovations.[7] According to the Canadian federal government's Digital Canada 150 Strategy released in April 2014, "Canada will be one of the global leaders in applying 'big data' to change how we think about and carry out health care, research and development, as well as the myriad activities of business and government."[8]

As Deloitte sees it, the digital enterprise is about harnessing the art of the possible. It is about capitalizing on data stores and using the intelligence derived from them to be bold, to innovate, and to challenge our conventional ways of doing business—while protecting privacy at all times.

[5] https://www.deloitte.com/assets/Dcom-Luxembourg/Local%20Assets/Documents/Whitepapers/2014/dtt_en_wp_techtrends_10022014.pdf.

[6] The Ontario government, for example, provides access to forestry, hydrographic, transport, demographic and other data under its Open Data program (http://www.ontario.ca/government/government-ontario-open-data).

[7] See Cavoukian, A. Privacy and Government 2.0: The Implications of an Open World, May 2009. http://www.ipc.on.ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/?id=874.

[8] Digital Canada 150, http://www.ic.gc.ca/eic/site/028.nsf/eng/home.

Privacy is about *personal* information

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. Although there may be jurisdictional differences, personal information (also known as personally identifiable information or "PII") may be defined as any information, recorded or otherwise, relating to an identifiable individual. Almost any information, if linked to an identifiable individual, can become personal in nature, be it biographical, biological, genealogical, historical, transactional, locational, relational, computational, vocational, or reputational. Determining whether some information falls into the category of personal information requires consideration of context. If there is a reasonable possibility of identifying a specific individual—whether directly, indirectly, or through manipulation or data linkage—then privacy concerns arise.

However, not all data is personally identifiable, and thus not all data gives rise to privacy concerns. It is important to understand the distinctions between the different forms of non-personal data:

- *De-identified information* refers to 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.[1]
- *Aggregated information* refers to information elements whose values have been generated by performing a calculation across all individual units as a whole. While uncovering new treatment strategies, medical researchers might use aggregated patient data—e.g., a certain percentage of patients taking a particular combination of drugs who experienced adverse side effects—but have no way to connect this data to a specific individual.
- *Non-personal, confidential information* is information that often holds tremendous value and importance for organizations, such as business plans, revenue forecasts, proprietary research, or other intellectual property. The disclosure or loss of such confidential information can be of grave concern for organizations—and Deloitte often advises clients on how to prevent such losses—but it does not constitute a privacy breach because it does not involve the handling of *personal* information. Within Deloitte, all client data is safeguarded with the highest degree of protection, regardless of whether it constitutes personal or confidential information.

Some kinds of information are not so easily characterized as personal or non-personal information. One such example is metadata—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.[2] While context is key in making determinations about personal information, in the case of metadata it is especially important. The detailed pattern of associations revealed through metadata can be far more invasive of privacy than merely accessing the content of one's communications.

[1] See NIST, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, April 2010, p. E–1.

[2] See Ann Cavoukian, *A Primer on Metadata: Separating Fact from Fiction*, July 2013, http://www.ipc.on.ca/images/Resources/metadata.pdf.

Data Analytics, Innovation and Privacy: who wins?

With organizations increasingly undertaking data analytics activities to derive new insights, regulators, legislators, interest groups, and citizens have begun to voice concerns about the impact of this activity on privacy—from the misuse or unauthorized disclosure of personal information to data-based surveillance. Once taken for granted, fundamental protections afforded to individuals in the processing of their personal information—e.g., notice, consent, purpose specification, and limitation—are now increasingly being challenged by the nature of Big Data analytics. Some argue that our notion of privacy itself must change, and that the requirements of consent, purpose specification and use limitation act as a barrier to Big Data analytics.[9] These arguments represent dated, zero-sum thinking. A new solution is needed—one in which the interests and objectives of both sides can be met in a doubly enabling, "win-win" manner.

Shifting responsibility for personal information processing from individuals to organizations alone is not the answer. Doing so amounts to a form of "privacy paternalism,"[10] where organizations determine "what is best" for individuals, and those individuals are unable to contribute to any discussions involving the use or misuse of their personal information. If the history of privacy has taught us anything, it is that an individual's loss of control over their personal data leads to more privacy abuses, not less.

In fact, inadequate restraints and a paternalistic approach could lead to what privacy advocates fear most—ubiquitous mass surveillance, extensive and detailed profiling, sharpened information asymmetries, power imbalances, and, ultimately, various forms of discrimination. Thus, diluting notice and consent requirements weakens essential privacy protections, while diminishing limits on the specified purpose, collection, and use of personal data minimizes accountability instead of strengthening it.

Privacy requirements are not obstacles to innovation or to realizing societal benefits from Big Data analytics—in fact, they can actually foster innovation and win-win outcomes. By using privacy-enhancing technologies, such as strong de-identification techniques and tools, and applying appropriate re-identification risk measurement procedures, it is possible to provide a high degree of privacy protection, while ensuring a level of data quality that may be appropriate for secondary use in Big Data analytics.

In some cases, privacy principles can actually *improve* Big Data insights. In the Google Flu example, later studies showed that "Google's estimates of the spread of flu-like illnesses were overstated by almost a factor of two."[11] Why? The absence of context, a key element of privacy. Google's estimates were missing the reasons *why* were people searching for flu information: Did they have the flu? Did they know someone with the flu? Did they want to know how to avoid getting the flu? When the individual participant is directly involved in information collection, the accuracy of the information's context grows dramatically.

[9] See, e.g., Fred H. Cate, Peter Cullen, and Viktor Mayer-Schönberger, "Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines," December 2013.

[10] See Ann Cavoukian, Alexander Dix, Khaled El Emam, "The Unintended Consequences of Privacy Paternalism," March 2014, http://www.ipc.on.ca/images/Resources/pbd-privacy_paternalism.pdf.

[11] Tim Harford, "Big data: are we making a big mistake?" Financial Times Magazine, March 28, 2014, http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2yaNDIgbN.

Data analytics: inspiring disruption

Tech Trends 2014: Inspiring disruption[1], Deloitte's fifth annual report on the ever-evolving technology landscape, focuses on disruptive trends that are transforming business, government and society. Information technology continues to be dominated by five forces: analytics, mobile, social, cloud, and cyber. Disruptors are areas that can create sustainable, positive disruption in IT capabilities, business operations, and even business models.

As organizations find themselves challenged to improve their ability to sense and respond, cognitive analytics offers a powerful way to bridge the gap between the promise of Big Data and the reality of practical decision-making. Cognitive analytics will likely become more mainstream, where predefined rules and structured queries will be augmented with artificial intelligence, machine learning, and natural language processing to generate hypotheses drawn from Big Data.

Five disruptive forces

The 5 disruptive forces are embedded into mainstream processes and value chains:

Analytics
Mobility
Cloud
Social
Cyber

Enterprise adoption of the power of the crowd allows specialized skills to be dynamically sourced from anyone and anywhere—and as needed. Companies can use the collective knowledge of the masses to help with tasks from data entry and coding to advanced analytics and product development. The potential for disruptive impact on cost alone likely makes early experimentation worthwhile, but there are also broader implications for innovation in the enterprise.

Content and assets are increasingly digital, with audio, video, and interactive elements. They are also consumed across multiple channels—from mobile, social, and the web to in-store, on location, or in the field. Digital engagement is about creating a consistent, compelling, and contextual way of personalizing, delivering, and even monetizing the user's overall experience, especially as core products become augmented or replaced with digital intellectual property.

Wearable computing takes many forms, from glasses and watches to smart badges and bracelets. The potential is tremendous: hands-free, heads-up technology can help us reshape how we work, make decisions and engage with employees, customers and partners. "Wearables" introduce technology to situations where safety, logistics, or even etiquette constrained the use of laptops and smartphones. While consumer wearables are in the spotlight today, we expect business to drive acceptance and transformative use cases.

By 2020, there will be more than 40 zettabytes of data worldwide, and the vast majority of that data will be unstructured[2], coming from recent technological innovations such as the Internet, mobile connectivity, cloud computing, and social networking. In response, new technologies and processes have been developed to meaningfully analyze all of this data for business purposes. Mainstream business processes now collect and analyze data in ways not contemplated decades earlier. These forces are disruptors because they have changed, and continue to change, the way that business is conducted. However, while driving innovation, they also increase the potential for privacy risks.

[1] https://www.deloitte.com/assets/Dcom-Luxembourg/Local%20Assets/Documents/Whitepapers/2014/dtt_en_wp_techtrends_10022014.pdf.

[2] John Gants and David Reinsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," December 2012, IDC.

Privacy Risks associated with Big Data

The primary privacy challenge that Big Data poses is the risk of creating automatic data linkages between seemingly non-identifiable data. These linkages can result in a broad portrait of an individual—a portrait once inconceivable since the identifiers were separated in various databases.

Big Data can easily connect key pieces of data that connect people to things—rendering ordinary data into information about an identifiable individual and revealing 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.

Because the potential for Big Data to create data linkages is so powerful, some suggest that certain data be considered "super" data or "super" content.[12] This "super" data is a step above 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.

When a Big Data set is comprised of identifiable information, then a host of customary privacy risks apply.[13] For example, a large dataset of identifiable information could be subject to unauthorized disclosure, loss, or data theft; the larger the dataset, the more likely it will 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 can put public safety at risk. 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.[14]

Managing accountability with regards to outsourcing is another issue that arises 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 outsource this analysis and reporting.

Secondary use of data raises additional concerns. In general, organizations can only use individuals' personal information for the purposes identified at the time the information was collected ("primary purpose") 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.

[12] Kim Cameron, "Afterword," *Digital Enlightenment Yearbook 2013: The Value of Personal Data*, p. 293.

[13] To understand the full extent of privacy risks and mitigation strategies for any dataset involving personal information, it is recommended that organizations perform a Privacy Risk Assessment.

[14] Nudging is a technique which involves exploiting irrational human tendencies ("cognitive bias") in order to nudge people towards certain outcomes. For example, someone with a bias against scarcity will be automatically served an ad which states "while supplies last," while a person with a bias for following others will get an ad labelled "best selling." See Ryan Calo, "Digital Market Manipulation," University of Washington School of Law, Research Paper, no. 2013-27 (2013).

Consumer demands are creating privacy pressures, too

Deloitte's 2013 report, *Customer-focused growth: Rising expectations and emerging opportunities*, explored various trends transforming how businesses and consumers interact. In this report, we note that we will likely look back on Facebook's 2004 launch as a turning point in customer privacy. Suddenly users had a new, social incentive to voluntarily share all kinds of personal data online. Today, what people used to keep private—friends, vacation plans, favourite restaurants, and brands—has become very public. At the same time, Big Data's ability to track, store, and analyze this data and other web activity has grown exponentially, transforming media and marketing along with it.

Customers' growing demands for personalized offers and experiences are challenging companies' privacy policies and practices. Customers now want to receive targeted ads and other benefits based on the information they've disclosed (or think they've disclosed)—though privacy rules may not allow this.

Companies themselves are pushing the privacy envelope too—asking users for permission to access friends' status updates and photos, or offering additional benefits to those who provide additional personal information such as personal income. We expect to see companies revise and update their privacy policies and customer consents to reflect changes in customer behaviour and expectations and new practices in data collection, disclosure and use.

In our report, Deloitte advise that companies that want to harness Big Data will need to be transparent about their intentions and practices—and ensure that their value proposition for doing so encourages customers to give the required consents. We also caution organizations to use the additional customer consents they obtain with care: Customers must be aware they're being targeted and that their choice to opt in or opt out, if made, will be respected.

Don't let the risks keep you from innovating!

The risks of unauthorized access to data, should such access lead to actual disclosure or misuse of personal information, can result in severe consequences for any organization. These consequences can include reputational harm, legal action, damage to your brand or regulatory sanctions, disruption of internal operations—not to mention weakened customer loyalty that results in revenue and profit losses. According to TRUSTe's Consumer Privacy Confidence Index, 93 per cent of individuals worry about their privacy online, 45 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.

Yet just because these risks exist, organizations should not fear pursuing innovation through data analytics. Applying privacy controls and using privacy tools appropriately can dramatically reduce privacy risks and enable organizations to capitalize on the transformative potential of Big Data—while adequately safeguarding personal information.[15]

[15] TRUSTe, Consumer Privacy Confidence Index, 2014, http://www.truste.com/us-consumer-confidence-index-2014/.

Proactive Privacy/Protecting personal information **and** enabling innovation

We believe it is entirely possible to achieve privacy in the Big Data era, while also using data analytics to unlock new insights and innovations to move an organization forward.

In our view, compliance-based approaches to privacy protection tend to focus on addressing privacy breaches after-the-fact. As a result, they do not 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.

One of the most widely recognized approaches to proactive privacy is *Privacy by Design* (*PbD*), a framework developed in the late 1990s by co-author Dr. Ann Cavoukian, Information and Privacy Commissioner of Ontario. The *PbD* concept involves embedding privacy directly into the design specifications of technology, business practices, and networked infrastructure. Dr. Cavoukian developed *PbD* as a response to the ever-growing impact of information and communications technologies and large-scale networked data systems. It offers a useful framework for any organization trying to balance the desire to innovate with the need to preserve privacy by providing a "middle way" by which to achieve both.

PbD urges organizations to take a proactive approach to privacy. It makes privacy the default setting, incorporating privacy measures directly into IT systems, business practices, and networked infrastructure. In this way, *PbD* preserves privacy and personal control over one's information while providing organizations with a sustainable competitive advantage.

Implementing *PbD*—or proactive privacy thinking—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.[16] In 2010, *PbD* was unanimously passed as a framework for privacy protection by the International Assembly of Privacy Commissioners and Data Protection Authorities.[17] *PbD* has also been incorporated into suggestions for a consumer review board to review the ethical aspects of Big Data projects.[18]

[16] These include, *inter alia*, the U.S. White House, Federal Trade Commission, Department of Homeland Security, Government Accountability Office, European Commission, European Parliament and the Article 29 Working Party, among other public bodies around the world who have passed new privacy laws based upon the FIPPs. In addition, international privacy and data protection authorities unanimously endorsed *Privacy by Design* as an international standard for privacy.

[17] *Ibid*, *IPC/Ontario*, Resolution.

[18] Ryan Calo, "Consumer Subject Review Boards: A Thought Experiment," *Stanford Law Review Online 66* (2013): 97-102.

Privacy by Design Principles

The *PbD* concept is rooted in seven foundational principles designed to reconcile the need for robust data protection and an organization's desire to unlock the potential of data-driven innovation:

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

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***).

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).

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

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***).

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***).

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***).

Achieving Privacy and Enabling Innovation: Strategies to Deploy

The misleading view that privacy stifles innovation reflects a dated, zero-sum mindset. The notion that privacy must be sacrificed for innovation is a false dichotomy, consisting of unnecessary trade-offs. In fact, the opposite is true: privacy drives innovation. It forces innovators to think creatively to find solutions that will serve multiple functionalities.

A new "playbook" is needed. We need to abandon zero-sum thinking and adopt a positive-sum paradigm where both innovation *and* privacy may be achieved. Adopting *PbD* is a powerful and effective way to embed privacy into the "DNA" of an organization to establish a solid foundation for data analytics activities that support innovation without compromising personal information. We believe implementing *PbD* is a highly worthwhile goal for any organization.

There are numerous strategies organizations can use to advance privacy in data analytics. Data minimization, de-identification, and user access controls are three strategies that organizations can use today to develop valuable, data-driven business insights and innovations while safeguarding personal information.

1. Data minimization

Big Data analytics does not always involve the use of personally identifiable information. However, when it does, data minimization has the biggest impact on managing data privacy risks, by effectively eliminating risk at the earliest stage of the information life cycle.

According to this strategy, the starting point for designing Big Data analytical systems must be *no* collection of personally identifiable information—unless and until a specific and compelling purpose is defined. 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. In other cases, organizations may find that summary or aggregate data may be more than sufficient for their needs. Data minimization strategies also align with de-identification strategies (see below).

Data minimization tip: The first question you should ask about your data analytics process is whether personal information is required to be present in the data being analyzed. If the answer is no—in other words, no personal information is required—then privacy concerns do not arise. However, it is always recommended that organizations employ appropriate controls to ensure that *confidential* information is handled and stored appropriately. Leveraging key privacy principles can help organizations define what types of controls should be used.

Gaining real-time insight into airline performance management

A major airline engaged Deloitte to create a real-time performance management tool for its executives and decision-makers. The tool would integrate information from various sources—corporate information in particular, but also from selected external information—to provide a snapshot of overall performance across the airline's network.

To deliver this, we developed a tool that collected a wide range of non-personal information about each of the airline's locations: flight schedules and statuses, airline contact information, anonymized customer service ratings, employee statistics, and corporate financial data. We also collected anonymous external information, such as social media traffic, in order to provide the client with a means to monitor what its customers were saying in real time. The collected data was delivered in real-time using a highly visual "dashboard" application (accessible by touchscreen or tablet) that enabled client's executives to check on location performance at a glance, and drill down for additional information with a simple touch.

All of the data involved in this project was *non-personal* in nature. While it was important to ensure that confidential information was kept secure, there was no need to take additional steps to safeguard personal information as such information did not form part of the dataset.

2. De-identification

De-identification refers to a set of tools or techniques used to strip a dataset of all information that could be used to identify an individual, either directly or indirectly, through linkages to other datasets. These techniques include deleting or masking “direct identifiers,” such as names or social insurance numbers, and suppressing or generalizing indirect identifiers, such as postal codes or birthdates. While not personally identifying in and of themselves, indirect identifiers may be linked to other datasets that contain direct identifiers, and may thus be used to personally identify individuals. If done properly, de-identified data can be used for research purposes and data analysis—thus contributing new insights and innovations—while minimizing the risk of disclosure of the identities of the individuals behind the data.

Differential privacy and synthetic data

While de-identification tools and techniques have gained popularity over the years and have been developed into commercial products, there are some emerging research-level technologies that hold much promise for enabling privacy and utility to co-exist. Two of these technologies are differential privacy and synthetic data.

Differential privacy

Differential privacy[1] 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; based on that analysis, the software determines the level of “noise” to introduce into the result before releasing it. This distortion 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.

Synthetic data

Most differential privacy methods do not give researchers access to the dataset to analyze themselves. Not surprisingly, this limits the kinds of questions researchers can ask. To address this, some 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.[2] 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.[3]

[1] See Cynthia Dwork, “Differential Privacy,” *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)* 2, 2006, p. 1–12; Dwork, “A firm foundation for private data analysis,” *Communications of the ACM*, vol. 54, 2011.

[2] See Avrim Blum, Katrina Ligett, Aaron Roth, “A Learning Theory Approach to Non-Interactive Database Privacy,” *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing*, 2008, p. 609–618.

[3] See Justin Thaler, Jonathan Ullman, Salil Vadhan, “Faster Algorithms for Privately Releasing Marginals,” arXiv:1205.1758 [cs.DS]; see also Jonathan Ullman, Salil Vadhan, “PCPs and the Hardness of Generating Synthetic Data,” *Electronic Colloquium on Computational Complexity*, Technical Report TR10-07, February 2010, http://people.seas.harvard.edu/~salil/research/synthetic-Feb2010.pdf.

De-identification tip: De-identification can also help organizations use data without compromising secondary use requirements or restrictions. Removing personally identifiable information from datasets allows organizations to use their data stores while meeting commitments regarding the primary purpose for which the information was originally collected.

However, not all de-identification practices deliver the same level of de-identification rigour—and not all de-identification tools provide the same quality of outcomes to ensure a sufficiently low re-identification risk. The choice of which additional tools and techniques to use to de-identify a dataset will vary. Organizations using de-identification tools should guide their choice of tools through the use of strong de-identification frameworks. One excellent framework is Dr. Khaled El Emam's framework for de-identifying health data for secondary use.[19]

Enabling healthcare research through aggregated data

Medical professionals, researchers, and public health authorities must, by necessity, collect and handle an immense amount of personal information about their patients each day: patient ID numbers, health insurance numbers, medical records, and histories, and more. Protecting the privacy of this highly personal information is of the utmost importance; yet within that information could lie the secret to more effective diagnoses and treatments that may greatly benefit society at large.

For example, Deloitte was engaged by a government organization that had data from two distinct databases: Occupational Employment Health and Safety (EHS) and Employment Standards (ES). Combining the two databases would help their management team understand whether or not there was a correlation between employee's claims from a certain employer and how often that same employer was failing to uphold financial obligations to their employees (i.e., prompt payment of wages).

To assist the management team, Deloitte produced a dashboard and applied advanced analytics to showcase trends within each of the datasets over time, as well as a comparison of results from the two datasets.

We also developed a new de-identification algorithm that would mask protected employee information with proxy identifiers in order to permit data analysis on aggregated claims data. The algorithm masked individual alphanumeric identifiers (such as employee ID numbers, medical record numbers, salary earnings, and the dates of medically relevant events or claims), while suppressing all other protected employee information. The algorithm thus enabled Deloitte to show the management team an extensive analysis on these companies' employee data and provide them with the ability to focus on correlations within the data—without compromising any employee's personal information.[1]

[1] Deloitte Health Informatics LLC. *Deloitte De-Identification Algorithm*. 2012

[19] See Khaled El Emam, "De-identifying Health Data for Secondary Use: A Framework," October 2008, http://www.ehealthinformation.ca/documents/SecondaryUseFW.pdf.

3. User access controls

It is always important to safeguard personal information from unauthorized access. However, in the case of Big Data analytics, the size and variety of information being analyzed makes safeguarding of the data a vitally important concern. For networked computers, access control refers to the process of granting or denying specific requests to obtain and use information and related information processing services.[20] When combined with other "Security by Design" policies such as least privilege, need-to-know, least trust, and segregation of duties,[21] access control is an effective way to safeguard personal information.

A financial institution's customer database, for example, can contain a wealth of information about customers: their employer's name, their income, the identities of their spouse or children, their address, and more. However, very few people in the institutions require access to that information—or at least, all of it. It is important to develop levels of appropriate access to personal information on a need-to-know and least-privilege basis.

User access controls tip: Security does not equal privacy. While strong security is essential to privacy, the term privacy incorporates a much broader set of protections than security alone. 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. 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 a world where datasets are increasingly large. Organizations should regularly review and evaluate user access control protocols to ensure that controls are continually enhanced as systems evolve and analytics practices change.

[20] See NIST, *Glossary of Key Information Security Terms*, http://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7298r2.pdf, p. 2.

[21] See Ann Cavoukian, Mark Dixon, "Privacy and Security by Design: An Enterprise Architecture Approach," September 2013, http://www.ipc.on.ca/site_documents/pbd-privacy-and-security-by-design-oracle.pdf.

Improving workplace safety in the mining industry

Our client, a large international miner who had invested heavily in safety processes, structures, controls, and culture support, was still experiencing an unacceptable level of severe incidents, injuries, and fatalities. The company approached Deloitte and engaged our Safety Analytics services to provide an objective, fact-based assessment of current performance—and to identify key relationships and root causes that might be invisible to safety managers.

We analyzed a large, complex array of related and unrelated safety data, including employee records, training data, production data, and asset performance data. Because the data included personal information, it was vital that we took steps to ensure that individuals' privacy was protected right from the outset. Certain information, such as treatment protocols and health records, was simply not collected. Employee IDs were fully masked, to eliminate the risk of re-identification. Heeding our own advice, the analytics team itself was subject to rigorously applied security clearance and access privileges to ensure that no one had more access to the data than was required. Therefore, we removed any unnecessary access privileges to further protect privacy.

The outcome of the project was a number of insights that highlighted a series of relationships and potential root causes that were driving the client's current safety performance. Indeed, from the insights gained, the mining company was able to improve its safety measures and over time show a reduction in the number of safety incidents.

Deploying these strategies—data minimization, data de-identification, and user access controls—can have an immediate, positive impact on an organization's ability to protect the privacy of the personal information it holds. By reducing the overall amount of data collected, rigorously de-identifying the data, and restricting user access to it, organizations can provide privacy assurance proactively, while preserving their ability to use Big Data to gain new insights into their business.

Innovation and Privacy:
You *can* have it all!

Organizations will continue to apply data analytics to Big Data in order to advance their strategic goals and better serve their customers. However, that doesn't mean that privacy must be abandoned—far from it. Through careful planning and application of privacy techniques and principles, such as those embodied in *Privacy by Design*, organizations can use data for its desired business effect, while at the same time safeguarding personal information.

Strong leadership is required to make privacy a clear priority. Smart design and implementation decisions are needed to embed privacy into an organization's DNA. Careful monitoring and evaluation will ensure that measures put in place today meet the needs of tomorrow's data challenges.

At Deloitte, we've made this commitment in the form of our National Discovery and Analytics Lab, which is supported by our team of privacy practitioners. The Lab is a world-class facility driven to deliver best practice data analytics, data privacy and data security.

We are also committed to helping our clients think about innovative ways to undertake analytics, while implementing sound privacy practices to protect personal information and confidential business data.

The Big Data era is here for good. However, this does not mean we must sacrifice privacy or shackle innovation. Through careful planning and application of privacy techniques and principles, such as those embodied in *Privacy by Design*, organizations can use data for its desired business effect while at the same time protecting the personal information contained in the data. It is possible to have it all.

Office of the Information and Privacy Commissioner,
Ontario, Canada
2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8
Telephone: 416-326-3333
Fax: 416-325-9195
E-mail: info@ipc.on.ca
Website: www.ipc.on.ca

Deloitte
2 Queen Street East, Suite 1200
P.O. Box 8
Toronto, Ontario
M5C 3G7 Canada
Telephone: 416-601-6150
Fax: 416-601-6151
Email: deloitteanalytics@deloitte.ca
Website: www.Deloitte.ca/analytics

The information contained herein is subject to change without notice. Deloitte and the IPC shall not be liable for technical or editorial errors or omissions contained herein.

Privacy by Design: www.privacybydesign.ca

June 2014