When the Big Data Surge Crashes Against the Big Privacy Rocks: The Conflict Between the Commercial Value of Information and Increasing Privacy Regulation

The term “Big Data” is relatively new but has already been credited with many accomplishments. A search engine can identify the hot spots for a flu outbreak faster than the Centers for Disease Control by isolating 40+ key search terms that correlate to the sickness. A major retailer can perform detailed analysis of its sales history to identify the consumer products that it needs to bolster in its inventory three weeks before a hurricane hits. A credit analytics firm can predict which patients may need reminders to remember to take their medicine based on seemingly irrelevant factors such as how long they have lived at the same address, whether they are married, whether they own a car, and how long they have been in the same job. An employer can develop sophisticated log-in activity profiles for each of its employees in order to detect potential data security breaches through aberrant location, time of day, frequency, or other log-in specifics. These are just a few of the examples of Big Data uses. But, by the way, what exactly is Big Data?

I. What is "Big Data"?

In their recent book, “Big Data: A Revolution That Will Transform How We Live, Work, and Think,” Viktor Mayer-Schönberger and Kenneth Cukier explain that Big Data "refers to things one can do [with data] at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more.” The book further explains that the Big Data surge is fueled by the extraordinary growth in digital information across mobile applications, social networks, web tracking, search data, smart grids, and a myriad of other applications. In one study, Martin Hilbert of the University of California determined that the amount of data in the world is doubling roughly every three years. So, in 2007, the world had more than 300 billion gigabytes of stored data, and by the end of 2013, the amount of stored data in the world should be around 1,200 billion gigabytes (or 1,200 exabytes). Approximately 98% of this data is in digital form. It is difficult to grasp how much data this really is, but by one illustration, if 1,200 billion gigabytes of data were placed on CD-ROMs and stacked together, the CD-ROMs would stretch to the moon in five separate piles.

Big Data offers big possibilities across healthcare, financial services, consumer products and services, information technology, and other industry sectors. It may seem that the only limits on Big Data are our own

---

1 Advanced copy of publication from the June 2013 edition of the International Association of Privacy Professionals (IAPP) Privacy Advisor.
imaginations. But that is not the case. Data privacy regulation is a key barrier to Big Data analysis. Data privacy laws specifically aim to regulate the collection, use, disclosure, and processing of data about an identified or identifiable individual ("personal data"), and therefore are designed to interfere with the type of information flows associated with Big Data. This article provides an introduction to the commercial privacy implications of Big Data and identifies issues that companies should consider when crafting their privacy solutions related to Big Data. Note that there are other critical dimensions that will not be discussed here, such as government's relationship with its citizens.

II. How Do Privacy Regulations Impact Big Data?

There is no one data privacy law or uniform approach to privacy regulation in the world today. For example, in the United States alone, we have adopted at the state and federal levels a number of different privacy laws and regulations, including: (i) industry-specific privacy laws in healthcare, financial services, telecommunications, and other areas; (ii) activity-specific privacy laws such as those related to credit reporting and online data collections; and (iii) data-specific privacy laws such as those related to data security and breach notification. We have also conferred broad privacy enforcement authority on the Federal Trade Commission ("FTC"), the Federal financial regulatory authorities, the Department of Health and Human Services, the Federal Communications Commission, and other federal authorities, as well as State Attorneys-General and other state authorities. Plaintiffs' firms in the U.S bring an increasingly aggressive new dimension to privacy enforcement through individual privacy claims and class actions. By contrast, outside the US, many jurisdictions have comprehensive data privacy laws that apply to all data about an identified or identifiable individual, regardless of industry, activity, or the specifics of the personal data, and also have put in place government agencies, often called "data protection authorities," that are specifically charged to enforce data privacy laws. Other related laws can also apply, such as electronic communications and telecommunications requirements, professional confidentiality and secrecy obligations, and more.

Analysis of how privacy laws impact Big Data in any particular case requires specific analysis of industry, geography, data types, and other factors. However, there are several relatively common aspects of data privacy regulations that may restrict Big Data analysis, including the following.

a) Expanding scope of regulated personal data. Privacy regulators are increasingly establishing broader definitions of what constitutes regulated personal data. For example, the FTC issued the report "Protecting Consumer Privacy in an Era of Rapid Change" in March 2012 ("FTC Report"). This report identifies a set of privacy principles that should apply to any data that is "reasonably linkable to a consumer, a computer, or a device." The rationale for the broad application of privacy principles to data about a computer or a device is, for example, that a business does not need a consumer's name to track that individual. Instead, such tracking can be done using a cookie, precise geolocation, or a mobile device's unique device identifier. The FTC has correspondingly expanded the scope of regulated personal data under the Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501 et seq and its implementing regulations ("COPPA"), to regulate unique device identifiers, geolocation information, IP addresses, and other data about a mobile phone or other device.

Outside of the United States, data protection authorities are also taking a broader view of the scope of regulated personal data. For

The expanding scope of regulated personal data impacts Big Data in several key ways. First, Big Data analysis that relies on unique device identifiers or other computer-linked data to help correlate different information sets may be directly subject to privacy regulation (including the full suite of privacy notice, consent, and other requirements as discussed below). Second, to the extent that Big Data analytics reach beyond the "controlled" information in databases and applications into so-called "unstructured data," which may account for 80% of all business information, such unstructured data itself may be privacy regulated. Most organizations find it challenging to achieve privacy compliance for core databases and other structured data, so the prospect of privacy compliance with regard to unstructured data can be daunting. Third, to the extent that a business seeks to avoid privacy concerns by de-identifying data, the broader definitions of personal data may make such a de-identification exercise more time consuming and challenging.

b) Data subject notice and choice. Virtually all privacy laws establish requirements for the owner or controller of the information ("data controller") to provide a privacy notice to the individual data subject concerned ("data subject"), and in some cases, to provide the data subject with a privacy choice. In terms of timing, provision of the notice is generally required at the time of collection, or sooner, and in a context that the data subject can understand. For example, the FTC Report articulates a comprehensive privacy framework for consumer personal data that includes three top-line requirements: (i) privacy-by-design; (ii) simplified choice; and (iii) greater transparency. Transparency and simplified choice must be offered at the time of data collection, except that choice is not required where the purposes of use and disclosure qualify as "commonly accepted" practices, such as product fulfillment, legal compliance, and first-party marketing. In addition, among other provisions, affirmative (opt-in) consent should be obtained before collecting sensitive personal data (such as financial or health, children's or geolocation data) and before any personal data is used for a purpose different than specified at the time of collection.

Similarly, the privacy provisions in the Gramm-Leach-Bliley Act, 15 U.S.C. §§ 6801 et seq., and its implementing regulations ("GLBA") require financial institutions to provide a privacy notice to consumers of financial products and services at the time the consumer relationship is established and on an annual basis thereafter. The privacy notice must explain how the financial institution collects, uses, discloses, and protects that information, and must provide consumers with an opportunity to opt-out of certain disclosures. Likewise, the Health Insurance Portability and Accountability Act, 42 U.S.C. § 1320d et seq., 45 C.F.R. Parts 160, 162, and 164 ("HIPAA") generally
establishes data subject consent or authorization requirements for marketing and certain other purposes beyond treatment, payment, and healthcare operations.

The EC Directives are founded on the principle that privacy is a fundamental human right, and accordingly establish a comprehensive set of privacy requirements for all personal data collection and processing, including among others, robust notice and various choice and consent obligations. In April 2013, the EU Working Party opined that personal data must be collected for “specified, explicit and legitimate purposes only” (i.e., the privacy notice provided at time of collection must be clear on the purpose of use and processing) and that personal data may not generally be used or “processed in a way that is incompatible with the purposes specified at collection.” In other words, the data controller cannot use, disclose, or otherwise process personal data for a new purpose that is incompatible with the original purposes of collection, where the application of the standard of incompatibility changes depending on the sensitivity of the personal data and other factors.

In many instances, notice and consent may create significant obstacles for Big Data analysis. Big Data focuses on taking data captured in one context for one purpose, and re-using it in another context for a secondary purpose. Often, the secondary purpose may not be clearly understood by the business (or at least the privacy compliance team) at the time of collection. Use or disclosure of the data for the secondary purpose may therefore be inconsistent with applicable privacy rules, unless the business undertakes an effort to re-notify and/or re-consent the data subjects. Depending on the context, such attempts at privacy compliance activities may not be feasible after initial data collection, or at a minimum, may generate a “drop out” rate where a portion of data subjects refuse to provide consent or otherwise object. Given that Big Data typically leverages modern analytics and computational power to analyze complete data sets rather than sampling a smaller set of data, privacy requirements of notice and choice may directly impact the effectiveness of such analysis.

c) Data subject rights (access, correction, matching procedures, and automated decisions). Privacy laws confer various rights on data subjects to obtain access to personal data that a business maintains about them and to require correction or other actions regarding such data. For example, the Fair Credit Reporting Act, 15 U.S.C. § 1681 et seq. ("FCRA"), imposes various obligations on "consumer reporting agencies" to provide individuals with access to their own consumer reports, an accounting of disclosures, and opportunities to correct or otherwise challenge the accuracy of that information. Businesses that furnish information to consumer reporting agencies, as well as users of consumer reports, are also subject to various obligations to validate or correct information when inaccurate, to provide adverse action notices that inform data subjects about their rights, and other obligations. The FCRA’s broad definition of "consumer reporting agency" has long been viewed as a trap for the unwary. Generally stated, a "consumer reporting agency" is any business or person that (i) for fees or other compensation, (ii) regularly engages in the practice of assembling or evaluating “non-experience” information about consumers, (iii) for the purpose of disseminating such information to third parties for use in connection with the evaluation of the consumer for credit, employment,
insurance, or other "permissible purposes." Data brokers, financial institutions, and other data controllers often must carefully structure their operations to avoid becoming a consumer reporting agency, including in some contexts obtaining representations and warranties from recipients to whom data will be disclosed that the recipients will **not** use the data for credit evaluation or other FCRA-covered purposes.

The Hong Kong Personal Data (Privacy) Ordinance (Cap. 486) ("Hong Kong Ordinance") sets forth a restriction on the use of a "matching procedure" without the data subject's express consent. A "matching procedure" is generally defined as: (i) a comparison of two sets of personal data, each of which is collected for different purposes; (ii) where each comparison involves the personal data of 10 or more data subjects; (iii) the comparison is not carried out by manual means (e.g., it is carried out by an automated process designed and applied for performing the comparison process); and (iv) the end result of the comparison may be used, whether immediately or at any subsequent time, for the purpose of taking adverse action against any of the data subjects concerned. See Section 30 of the Hong Kong Ordinance. The Hong Kong Ordinance also contains various other obligations regarding the provision of access to data subject, and honoring rights of correction, erasure and other rights. See Sections 18 to 28 of the Hong Kong Ordinance.

The EC Data Protection Directive, among other relevant provisions, establishes rights for data subjects to not be subject to legally significant decisions based solely on automated processing intended to evaluate performance at work, creditworthiness, reliability, or other personal aspects. The EC Data Protection Directive also establishes, as defined in greater detail in national laws, rights for data subjects to obtain access to personal data, including in some instances an accounting of disclosures and uses, as well as an opportunity to correct and object to the processing of personal data. See Articles 12, 14, and 15 of the EC Data Protection Directive.

Data subject rights such as those described above raise hurdles for Big Data analysis. First, Big Data has the potential to produce uncanny predictions about behavior, but often this is accomplished through complex calculations of complete data sets to isolate correlations between otherwise apparently irrelevant data, rather than identifying causal links between prior behavior and the predicted behavior. In practice, in a small data world under FCRA, a consumer exercising the right of access could understand how unpaid credit card bills may affect credit rating, but would the consumer (or a regulatory authority) understand how a complex analysis of the data subject's social media network of "friends" provides an accurate correlation to creditworthiness? Second, given that Big Data utilizes information from various sources to produce predictions and correlations, the data controllers that store or use that consolidated information may have corresponding duties to produce substantially more personal data in response to data subject access requests, and likewise may need to address more complex correction, objection, and erasure requests. Third, provisions such as the "matching procedure" restriction under the Hong Kong Ordinance and the "automated decisions" provision under the EC Data Protection Directive may make it difficult to use Big Data in decisions that have a substantial legal effect on affected data subjects.
d) Disclosures to third parties. Privacy laws impose various obligations on data controllers before they disclose personal data to: (i) sourcing providers, business associates, and other third parties that access or process the information on behalf of the data controllers ("data processors"); and/or (ii) third parties that receive and use the data for their own purposes ("third party data controllers"). Generally speaking, data controllers are obligated to impose requirements of non-disclosure and non-reuse on their data processors, as well as obligations to provide appropriate data security and other controls. Data controllers are generally obligated to meet certain thresholds before disclosing personal data to third party data controllers, including in some cases obtaining data subject consents.

A key threshold determination for a data controller is whether its third party recipient, under applicable privacy laws, is acting as its data processor, or is acting as a third party data controller. The distinction is often not an easy-to-apply, bright-line rule. For example, under the EC Data Protection Directive, the recipient is closer to a data processor if it acts only under the instructions of the data controller with respect to handling personal data, whereas the recipient is closer to a third party data controller if it determines the "purposes or means" of processing the personal data. Historically, service providers have been treated as data processors, but as clarified by the EU Working Party, a service provider that processes personal data for its own purposes can functionally demonstrate actual control over the data such that it can inadvertently become a third party data controller, even if a master services agreement or companion agreement identifies the service provider as a "processor." See Working Party Opinion on the Concepts of "Controller" and "Processor" (Opinion 1/2010).

In a Big Data world, service providers can quickly become inadvertent third party data controllers. For example, a medical device provider may consider itself a data processor acting on behalf of its customer, a healthcare provider, with regard to the collection and handling of personal data about the patient's use of the device (the primary purpose). However, the medical device provider may become an inadvertent third party data controller if it utilizes the personal data internally for its own Big Data analysis, or provides personal data about usage of its device to other third parties for their use in Big Data analysis (i.e., use or disclosure for secondary purposes). The privacy challenges can be significant if a service provider inadvertently becomes a third party data controller. Among other concerns, the service provider itself becomes directly subject to substantially more privacy requirements than it likely could have originally anticipated, and the service provider's customer is unlikely to have properly addressed privacy requirements with respect to disclosures to the service provider as a data controller. See, e.g., Working Party Opinion on the Processing of Personal Data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT) (Opinion 10/2006) (whereby a financial payment intermediary became an inadvertent third party controller when it decided to disclose data to the US government without specific authorization from its financial institution customers).

e) Data de-identification. By definition, privacy laws only apply to information that falls within the scope of personal data. As such, a key issue is whether any personal data at issue can be suitably de-identified such that it falls outside the scope of applicable privacy
laws. For example, HIPAA establishes a standard that regulated
personal data is de-identified when "there is no reasonable basis to
believe that the information can be used to identify an individual." 45
C.F.R. § 164.514(a). There are two methods to comply with this
standard. The first requires a formal determination by a qualified
statistician who, applying statistical and scientific principles and
methods for rendering information not individually identifiable,
determines that the risk is very small that the information could be
used, alone or in combination with other reasonably available
information, by an anticipated recipient to identify an individual who is
a subject of the information. The statistician must document the
methods and results of the analysis that justify such determination
(”HIPAA Statistical Method”). In performing this analysis, the
statistician must take into account available technology and the
overall information environment. 45 C.F.R. § 164.514(b)(1). Thus,
even as technology and the overall information environment change
over time, the statistician must account for such changes in the
analysis, and reach a conclusion that the risks of re-identification are
"very small" in order for the personal data to be properly "de-
identified."

The second method involves the removal of eighteen specified patient
identifiers, including but not limited to, patient name, location (other
than state or 3-digit ZIP codes with populations greater than 20,000),
email address, telephone number, Social Security Number, and the
Significantly, the eighteenth identifier that must be removed is “any
other unique identifying number, characteristic, or code.” This is a
broad definition that will change over time as technology and the
overall information environment change. Moreover, even after
removing all of the designated identifiers, the data controller must not
have actual knowledge that the remaining data could be used alone
or in combination with other information to re-identify the data subject.
45 C.F.R. §164.514(b)(2)(ii).

Outside the United States, privacy laws can also establish strict
standards for anonymization and de-identification. For example, the
EU Working Party has articulated a standard that whether data has
been anonymized requires an analysis that takes into account "all the
means likely reasonably to be used by the controller or any other
person" to link the data to an individual. Thus, in a situation where
the data controller deletes names and other identifiers, but appends
the data with key coding so that data profiles can be compared for Big
Data analytics purposes, there still would be an open question of
whether a recipient third party data controller would hold regulated
personal data. The analysis would depend on various factors,
including risks of an external hack of the original data controller, the
likelihood that someone within the data controller’s organization would
provide the key, and the feasibility of re-identification through indirect
means. See EU Working Party Opinion on the Concept of Personal

In a Big Data world, given the commingling and consolidation of
different data sources, the potential for re-identification by third parties
is significant. Such re-identification could have cascading
consequences regarding privacy compliance across all of the issues
cited in this discussion, including notice/choice, data subject rights,
data security and breach notification, and cross-border data transfers.
f) Cross-border data transfer restrictions. Certain privacy laws include specific restrictions on the cross-border transfer of personal data outside the local jurisdiction to countries that do not maintain adequate or equivalent data protection laws. For example, the EC Data Protection Directive provides that personal data may not be transferred to third countries unless the recipients are located in a jurisdiction that provides adequate protection or certain exceptions are satisfied. See Articles 25 and 26 of the EC Data Protection Directive. In a Big Data world, a data controller may share information with data processors and/or third party data controllers in non-local jurisdictions, which may in turn flow data downstream to other recipients with whom the data controller has no pre-existing privity of contract or other controls. (or even knowledge of such disclosures). Such scenarios pose significant challenges for a data controller seeking to address cross-border data transfer restrictions.

III. What Are the Privacy Solutions for Big Data Initiatives?

As is evident from the discussion above, there is no one-size-fits-all privacy solution for all Big Data initiatives. Each initiative needs to be examined in light of the industry, data types, geography, and other activities at issue. Having said that, privacy professionals for data controllers and other Big Data participants may wish to bear in mind the following overall recommendations:

a) Stay close to the business teams. Early notification of Big Data plans can help mitigate the chances of direct conflicts between "Big Data" and "Big Data Privacy" by providing privacy professionals with a better opportunity earlier in the process to address applicable privacy laws. This can be critical given how Big Data uses, if not properly anticipated, can give rise to challenging privacy requirements such as obligations to re-notify and re-consent data subjects for prior-collected personal data.

b) Enhance privacy notices/consents. Within the constraints of applicable privacy laws, consider enhancing or expanding the language in existing privacy notices and privacy choices to reflect secondary purposes for analytics and related Big Data activities. This is preferably addressed once the business has formulated its Big Data plans, but there may be some privacy benefit even before that stage depending on the terms, if any, in existing privacy statements related to secondary uses of data and analytics.

c) Improve controls on data usage and information management. Privacy compliance for Big Data, including proper actions to address data subject access, accounting of disclosures, correction, erasure, and other rights, requires data controllers to maintain rigorous controls over data usage and information management across the full life cycle of information, including: (i) data intake (including direct collections from data subjects and indirect collections from other third parties); (ii) data usage (e.g., credit or similar decisions, matching procedures/automated decisions, and usage of health, financial or other sensitive personal data usage); (iii) data disclosure to data processors and third party data controllers; (iv) information security and data breach notification, and (v) data retention and secure deletion. Among other dimensions, data privacy compliance will place increasing demands on information technology and security systems, and should be a factor that supports investment decisions related to
any information technology and security resources that may be used for 2 to 5 years.

d) **Enhance policies and procedures on third party disclosures.** Consider adopting or enhancing internal policies to require advice from privacy compliance before disclosure of personal data to data processors as well as third party controllers. Key issues will include usage of data for secondary purposes, information security and incident response, and if de-identification is a strategy, obtaining assurances about non re-identification.

e) **Consider de-identification as a strategy.** Strategies to manage privacy risks can include de-identification techniques to the extent possible. Key concerns include the specific standards for de-identification under applicable privacy rules, as well as the risks of re-identification given the expanding array of data sources, analytics, and downstream recipients in the Big Data world.

f) **Address cross-border data transfer restrictions.** Special care should be taken to address data movement within the group of companies and external disclosures that involve cross-border data transfers. Consider leveraging the company's existing data cross-border data transfer solutions (e.g., the US-EU Safe Harbor, the European Commission Standard Contractual Clauses, Binding Corporate Rules, and other approaches), as well as the solutions adopted by relevant data processors and third party data controllers.

Beyond responding to existing privacy laws, data controllers should actively participate in the development of new privacy laws regulations. There are literally hundreds of privacy bills currently contemplated at the state and federal levels in the United States, and significant legislative and regulatory developments are underway worldwide. For example, the European Commission is contemplating an update to the EC Data Protection Directive that would result in the implementation of an EC Data Protection Regulation. The current draft EC Data Protection Regulation contains a series of greatly expanded privacy rights, including a "right to be forgotten," that directly implicate Big Data. Legislatures and privacy regulators must receive thoughtful input from Big Data participants in order to avoid adopting privacy laws that unnecessarily restrict the healthcare, financial services, consumer choice, and other benefits that Big Data offers to all of society.