Centre For Information Policy Leadership
Hunton & Williams LLP

Big Data and Analytics:
Seeking Foundations for Effective Privacy Guidance

A Discussion Document
February 2013

Introduction

Analytics[1] promises to revolutionize business, science, research and education. Powerful algorithms help identify individuals in need of social services, detect fraud, predict the effects of natural disasters, recognize patterns in scientific research and discover trends in consumer demand. Analytics play a role in addressing concerns across all aspects of society – from understanding biology at the molecular level to managing energy resources and improving education.

While use of analytics can be traced to the late 1800s,[2] today its use is fueled by "big data" — vast stores of information gathered from traditional sources (e.g., public record data, health data, financial and transactional data) and from new collection points (e.g., web data, sensor data, text data, time and location data and data gleaned from social networks). Big data is characterized by the variety of its sources, the speed at which it is collected and stored, and its sheer volume.[3] While traditionally analytics has been used to find answers to predetermined questions, its application to big data enables exploration of information to see what knowledge may be derived from it, and to identify connections and relationships that are unexpected or were previously unknowable. When organisations employ analytics to explore data's potential for one use, other possible uses that may not have been previously considered often are revealed. Big data's potential to yield unanticipated insights, the dramatically low cost of information storage and the rapidly advancing power of algorithms have shifted organisations' priorities to collecting and harnessing as much data as possible and then attempting to make sense of it.

---

[1] The term "analytics" refers to the use of information technology to harness statistics, algorithms and other tools of mathematics to improve decision-making. Paul Schwartz, "Data Protection Law and the Ethical Use of Analytics," available at http://www.huntonfiles.com/files/webupload/CIPL_Ethical_Underpinnings_of_Analytics_Paper.pdf. (Last visited 7 January 2013).

[2] In the late 1800s, the science of data analysis was used mainly in industrial machinery development. Frederick Winslow Taylor initiated time management exercises, supported by analytics, and Henry Ford applied analytics to determine the pacing of the assembly line.

[3] "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population." Merv Adrian, "Big Data," *Teradata Magazine* 1:11, www.teradatamagazine.com/v11n01/Features/BigData/. "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition, and Productivity,* May 2011.

Although many applications do not involve personal information, in some cases the power of analytics, rich data stores and the insights they can yield raise risks to privacy. In some instances, data used in analytics is personally identifiable. In others, application of analytics to anonymous or non-personally identifiable data can reveal the identity of an individual or insights about him or her. Analytic models and algorithms and the data to which they are applied may vary in quality and integrity. While the outcomes of analytic processes can raise privacy concerns even when algorithms and data are appropriate for their intended use, algorithms and data whose quality is suspect can yield faulty results that may seriously compromise privacy or individual rights. Analytics may be applied to data that has been precluded by law from processing for certain purposes or to arrive at certain prohibited decisions.[4] Big data and analytics support automated processes that may arrive at decisions about an individual, raising important questions about self-determination, personal autonomy and fairness. They may also yield predictions about individuals that may be perceived as invasive or as precluding his or her choices.

While long-established principles of fair information practices provide guidance, analytics, processing technology and big data challenge the way we apply them. Policymakers, users of data and data protection authorities must, therefore, consider carefully how the principles are honestly and effectively applied to analytics. Moreover, it is important that laws and regulations take into account analytics as a distinct data-processing method. Prohibitions in law against automated decision-making can functionally preclude the use of modern analytics entirely. Legal requirements that require explicit consent for any processing of data — even data that has been de-identified – can also impede the use of analytics.

This paper provides context for "Guidance for Big Data and Analytics: Protecting Privacy and Fostering Innovation" — an industry-sponsored initiative led by the Centre for Information Policy Leadership.[5] The goal of this effort is to develop a governance framework for the use of analytics that protects privacy and promotes innovative uses of big data. The paper offers three real-life examples of uses of analytics, and for each describes the data that powers it, the process by which the data is analyzed and the algorithm is applied, its benefits, and how the risks it raises are mitigated. It sets out the challenges to analytics raised in current law, regulation and traditional notions of fair information practices. Finally, it articulates goals for effective guidance that would address the realities of how analytics and big data work, how they may challenge

[4] Use of data about race, for example, is prohibited in credit scoring. The U.S. Equal Credit Opportunity Act provides that, "It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction—
(1) on the basis of *race*, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract);
(2) because all or part of the applicant's income derives from any public assistance program; or
(3) because the applicant has in good faith exercised any right under this chapter." (Emphasis added.) U.S Code Title 15 › Chapter 41 › Subchapter IV › § 1691.

[5] Paul Schwartz, "Data Protection Law and the Ethical Use of Analytics," available at http://www.huntonfiles.com/files/webupload/CIPL_Ethical_Undperinnings_of_Analytics_Paper.pdf. (Last visited 7 January 2013).

privacy, and their importance for innovation, economic growth and the advancement of societal goals.

While analytics may pose risks, the failure to use it to solve longstanding problems in healthcare, research, education and development deprives individuals and society of the potential benefits of big data. Ideally, thoughtful guidance that takes into account the realities of big data and the nature of analytic processing will empower organisations to use analytics in a robust and responsible manner to arrive at long-sought solutions. Developing an approach to governance that balances individual interests in privacy and the need to protect fundamental rights and freedoms with the potential of this processing power will make it possible to realize the important, and in some cases still unanticipated, benefits of big data and analytics.

The Power and Promise of Analytics

The following are examples of analytics applied in network security, healthcare and education. Each will discuss the data that is used, the process by which the data is reviewed and an algorithm is derived, the application of the algorithm, the risks raised and how they are mitigated, and how the organisation determines whether and to what extent the results of analytics will be used.

Case 1: Intel – Big Data Analytics to Improve Network Security

Security professionals manage enterprise system risks by controlling access to systems, services and applications; defending against external threats; protecting valuable data and assets from theft and loss; and monitoring the network to quickly detect and recover from an attack. Big data analytics is particularly important to network monitoring, auditing and recovery. Intel's Security Business Intelligence uses big data and analytics for these purposes.

Most of the data analyzed for network security comes from log files of every event that occurs on a network system. Log files may include records of attempts to access a website or to download a file, system logins, email transmissions and authentication attempts. The vast amount of data generated by log files enables researchers to identify malfunctions, attacks or suspicious activity on the system. Intel Security gathers log file data from servers, clients, network devices, specific applications and specialised sensors. It also collects contextual information that helps security experts to interpret the events captured in log files. Because an enterprise system can generate five billion events per day, big data analytics is instrumental in making sense of network activity.

Data compiled in log files and contextual information is maintained in a variety of formats and must be put into a consistent format and entered into a system for analysis. For each network event, data is extracted, put in standard formats and loaded into a data warehouse. Formatting of data is an automated function that can process 11 billion new network events and more than one million events per second during periods of peak activity.

Because this volume of data is too large to be processed effectively, security experts distill samples of data that represents normal network behaviour to make anomalies and threats more easily detectable. By condensing data in this way, one can model anticipated threats based on identified network activity trends, geographic regions with disproportionate threat activity, and other network characteristics that signal an attack. Based on this analysis, one can create predictive models to identify new potential threats. Analytics models are continually refined based on feedback data, making possible faster responses to actual threats, more accurate predictions and real-time detection of potential attacks.

Prior to the use of big data analytics in network security, scheduled network analysis would be performed to assess the health of a network. Today, systems like Intel's Security Business Intelligence enable real-time processing and analysis of data to identify (1) safe traffic — network activity that is known not to be dangerous or associated with threats to the system and related trends; (2) high-risk traffic — activity that is dangerous or associated with threats to the system and related trends; and (3) predictive trends. All network activity is compared against these models; each individual network event can be flagged as safe, threatening or suspicious as compared to the trends identified. The activity may be blocked, noted or allowed. Understanding of the accuracy of those decisions supports further refinement of the model.

Intel Security maintains a privacy plan to address the collection and use of data. To mitigate the risk that individuals may be identified through such accumulated data, personnel access is restricted to appropriate areas of the system through a formal process that establishes whether an individual is authorized to see certain data. In addition, data may be de-identified — in some cases data sources are inherently de-identified because credentials are not associated with the access request, while in other cases logs deal primarily with identity.

Case 2: Merck – Reducing Patient Readmission Rates

Vree™ Health, a subsidiary of Merck & Co. Inc., applies analytics to big data to address patient care issues and to reduce hospital readmission rates.[6] The focus of Vree Health's TransitionAdvantage™ service is patients hospitalized for heart attack, heart failure or pneumonia. The leading causes of hospital readmission have been identified as failure to provide patients with necessary information upon discharge, lack of follow-up with patients, medication management issues and insufficient coordination of care. By helping hospitals identify issues that may arise after patients have left the hospital and promoting patient compliance with post-discharge care plans, the project aims to reduce admissions that might occur within 30 days of patient discharge.

[6] One in five Medicare patients is readmitted to the hospital within 30 days of discharge. Stephen F. Jencks, M.D.; Mark V. Williams, M.D.; Eric A. Coleman, M.D., "Rehospitalizations among Patients in the Medicare Fee-for-Service Program," *The New England Journal of Medicine* 360 (April 2, 2009): 1418–1428. Beginning in 2012, the Centers for Medicare & Medicaid Services has imposed significant penalties on hospitals reporting the highest rates of readmission for patients diagnosed with heart attacks, heart failure and pneumonia.

Vree Health uses data collected throughout the course of patient care — when the patient completes hospital admission forms, during the hospital stay and at the time the patient leaves the hospital.[7] Vree Health also uses data collected by its representatives during follow-up calls over the 30 days after discharge and data generated when patients interact with the resources available through the cloud-based web platform[8] or mobile application, over the phone with an operator or via its interactive voice response system. This information is combined with data from more traditional sources, including weight, changes in diet, medications, other clinical data, demographic details and data gathered from third-party data sources such as Centers for Medicare & Medicaid Services for comparison with historical controls of patient populations.

Data is preprocessed to correct any errors and to format it for analysis. Once cleaned, it is incorporated into the data warehouse where it may be organised into subcategories so that it can be more readily accessed for use in specific areas of research. For example, a data scientist and clinician could partner to investigate a data set that focuses on congestive heart failure and might exclude patients with heart attack or pneumonia. However, even though data may have been sub-categorized in this way, it is still possible to leverage the entire data set to discover broader population trends (i.e., within regions, within/between hospitals). By analyzing relationships among the data, researchers identify factors that are likely to lead to readmission. A variety of data sets may be reviewed: researchers may analyze data, for example, across all participating hospitals, across all the hospitals within a single region or within patient cohorts defined by demographic or disease category.

Partnering medical facilities may apply identified trends to information about patients to address conditions that could lead to hospital readmission. Patients who exhibit certain characteristics or behaviours that would indicate that readmission is likely can be provided with specified services or assistance. For instance, a patient who does not check in with her primary care physician within seven days of discharge — a factor indicating increased likelihood of readmission — could be contacted by email or phone.

Use of sensitive health information raises risks that Merck takes measures to address. Vree Health pays particular attention to issues related to obtaining consent to the use of the data, to the use of de-identified rather than identified information (and when the use of each is appropriate), and to the risks raised by the application of algorithms derived from this data.

Consent forms state that the health information it collects is de-identified and used for analysis to support research to enhance care for the individual patient and others. Merck periodically reviews their disclosures to assess their readability and effectiveness. It gives patients the ability to "deactivate" their consent, but even where consent is withdrawn, the de-identified information is still used for public health research, for meta-studies and to maintain the integrity of the knowledge discovery research. Patients admitted to a participating hospital other than the facility

[7] Approximately 40 to 50 pieces of data are generated by each of 30 to 40 patient discharges each day at 5,000 to 6,000 hospitals — as many as 12 million pieces of data per day.

[8] http://www.vreehealth.com/vreehealth/home. (Last visited 9 February 2013).

of their initial hospital stay are asked again to provide consent. The terms of consent provided by the patient are attached to the data; if the terms change, a new consent is requested.

Because identified data is not necessary to discover trends and build analytic models, Merck uses de-identified data for knowledge discovery. However, de-identification is not irreversible. Data may be re-identified for use by clinicians who may need to know the identity and health issues of those under their care. Data incorporated into Vree Health's patient profiles is only that germane to their episodic hospital discharge. If consent is granted, this information is shared with the patient's family caregiver (i.e., a parent or adult son/daughter), the primary care physician or specialist, the transition liaisons and the nurse call centre

Merck recognizes the potential for harm that may result from inaccurate or untrustworthy predictive models. Models and algorithms are scrutinized and validated before the interventions they suggest are applied to individual patients. Merck refines models and algorithms as more data becomes available and researchers arrive at new insights. By incorporating data about interventions and their effect, for example, researchers update and improve prediction models.

Case 3: IBM – Analytics to Reduce the Student Dropout Rate

Analytics applied to education data can help schools and school systems better understand how students learn and succeed. Based on these insights, schools and school systems can take steps to enhance education environments and improve outcomes. Assisted by analytics, educators can use data to assess and when necessary re-organise classes, identify students who need additional feedback or attention, and direct resources to students who can benefit most from them.[9]

Alabama's Mobile County public school system is the largest in the state, comprising 63,000 students and 95 schools. Forty-eight percent of students left school prior to graduation — a rate significantly higher than the national average. With the goal of reducing dropout rates, IBM worked with Mobile County Public Schools to apply analytics to education data to help the school system identify which students were at risk of dropping out, and which interventions would help at-risk students. Based on these insights, educators developed an individualised response to each student's problems.

Working closely with Mobile County Schools to develop the analytic models, IBM used information that had been collected about students over the course of their schooling. It included administrative and academic data that had been gathered from each of the system's schools, including data about attendance and test scores, and demographic information including neighbourhood, race, gender and socio-economic status. Some data was collected specifically for analytic research; other data were available from legacy systems and existing databases. This information was combined for analysis with aggregated, de-identified data related to population

[9] Toon Calders and Mykola Pechenizkiy, "Introduction to the Special Section on Educational Data Mining," *SIGKDD Explorations*, 2. http://www.kdd.org/sites/default/files/issues/13-2-2011-12/V13-02-02-Calders%28introduction%29.pdf. (Last visited 9 February 2013).

and ethnicity from external sources, including the U.S. Department of Education[10] (on general trends nationwide) and various state-level government agencies, including the Alabama State Department of Education. Data from these outside sources was also used to allow researchers to test findings and to assess progress as compared to similar school districts.

Data was cleaned and formatted for analysis. Redundancies were removed and relevant data was retrieved from the database or warehouse. Data were examined to determine the relationship between a dependent variable (e.g., whether a child will drop out in the future) based on independent variables (e.g., parents' education and income; child's neighbourhood, test scores, absences).[11] These relationships were used to create a model for predicting which students were at risk of dropping out. Each student's data was entered into an algorithm to yield a score representing the student's risk of leaving school.

Mobile County addressed a variety of concerns raised by their use of student data for analytics. While consent to school officials' use of personally identifiable information contained in student education records for legitimate education purposes is not required,[12] the Mobile County Public Schools obtained parental consent to use student information and maintained conservative disclosure and access policies.[13] The system designed by Mobile County Public Schools and IBM also provided parents with access to their children's information through personal computers or hand-held devices.

Project administrators also recognized that sharing data beyond the school system could create vulnerabilities and lead to unintended consequences.[14] Identifiable information was available to only parents, guardians and those within the school system. Unidentified information was made available to other institutions under the guidelines outlined on the school system's website.[15]

Project administrators also recognized that the results of analytics could be misunderstood or misused. False positives (that misidentify students as being at risk who are not) may seem

---

[10] E.g., http://nces.ed.gov/programs/coe/indicator_sde.asp. (Last visited 9 February, 2013).

[11] This is referred to as *regression analysis*. The regression model is developed based on observed data for many children over a period of time. The regression model is used for trend analysis and to flag at-risk students who fall within the danger-zone of the dropout model.

[12] Family Education Privacy Act, 20 U.S.C.A. §1232g.

[13] Details of student record policies are outlined on the school system's "Student Records" section of the website, http://www.mcpss.com/?DivisionID=2150&DepartmentID=2021&ToggleSideNav=ShowAll. (Last visited 9 February 2013).

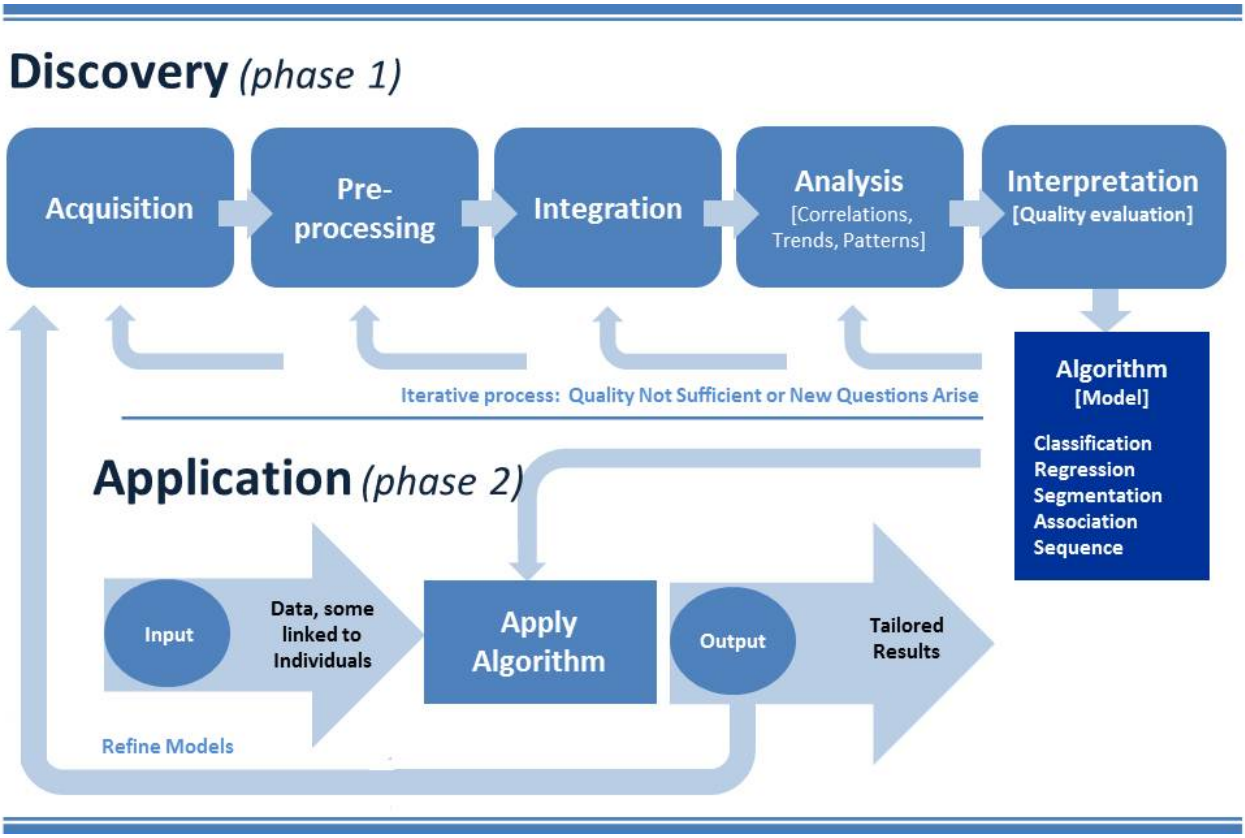
[14] For example, requests for the data outside the education system are made by interested parties like law enforcement trying to determine which students were present at school on specific dates.

[15] http://www.mcpss.com/?DivisionID=2150&DepartmentID=2021&ToggleSideNav=ShowAll. (Last visited 9 February 2013).

harmless because teachers or administrators will intervene to offer support. However, placing students in different categories (e.g., such as remedial course tracks) can impede their mobility within a school system (e.g., they will remain in remedial courses when they are ready for more advanced work). While the predictions may help educators more accurately place students in programs that will lead to their academic success, it is important that teachers and administrators understand the limitations of the predictions. Users were, therefore, instructed on how to understand and make appropriate decisions based on the findings of the analysis.

The Process of Analytics

While these examples highlight the range of applications for analytics and big data, they also reveal the steps involved in analytic processes. This section articulates those steps and suggests that analytics comprises a two-phase process that includes *knowledge discovery* and *application.*



Knowledge Discovery

Analytics may be applied to large data sets to determine what insights they may yield. This *knowledge discovery*[16] phase involves gathering data to be analyzed, pre-processing it into a format that can be used, consolidating it for analysis, analyzing it to discover what it may reveal and interpreting it to understand the processes by which the data was analyzed and how conclusions were reached. These steps are explained in more detail below.

1. Acquisition – Data acquisition involves collecting or otherwise acquiring data for analysis. Acquisition requires access to information and a mechanism for gathering it, such as website tracking, sensors, application logs, search inquiries and responses to intake forms. Data may be gathered to answer specific questions to undergo analysis, but existing data from inside and outside an organisation also may be incorporated. In Case 3, IBM obtained data directly from student records and legacy databases maintained by the Mobile County school system, as well as from sources compiled by other public institutions.

2. Pre-processing – Data is structured and entered into a consistent format that can be analyzed. Pre-processing is necessary if analytics is to yield trustworthy, useful results. In Case 1, Intel compiles a vast amount of data gathered at a variety of collection points and places it in a standard format for analysis.

3. Integration – Integration involves consolidating data for analysis. It entails 1) retrieving relevant data from various sources for analysis; and 2) eliminating redundant data or clustering data to obtain a smaller representative sample. In Case 2, Merck incorporates clean data into its data warehouse and further organises it to make it readily useful for research. The enormous quantity of data collected by Intel in Case 1 requires its distillation into manageable samples.

4. Analysis – Knowledge discovery involves searching for relationships between data items in a database, or exploring data in search of classifications or associations. Analysis can yield descriptions (where data is mined to characterize properties) or predictions (where a model or set of models is identified that would yield predictions). Merck researchers analyze relationships among data to identify broad trends within populations that reveal factors likely to lead to patient readmission.

5. Interpretation – Analytic processes are reviewed by data scientists to understand results and how they were determined. Interpretation involves retracing methods, understanding choices made throughout the process and critically examining the quality of the analysis. It provides the foundation for decisions about whether analytic outcomes are trustworthy, and supports the risk analysis necessary for accountability and responsible use of

[16] Knowledge discovery differs from more traditional applications of analytics in which data is explored in search of an answer to a particular question or to make some determination.

analytics. Based on this interpretation, organisations can determine whether and how to act on them.

For the sake of simplicity these steps are set out here as occurring linearly. However, they are often implemented in a more cyclical way. For example, when a data scientist interprets the results of knowledge discovery, he or she may notice that the quality of the data used was not sufficient, or that errors occurred in the data integration phase that could affect the reliability of the findings. At that point he may make adjustments and conduct the analysis again, with the goal of attaining better or more reliable results. The data scientist could also decide that certain insights or trends revealed by the data suggest other questions. In such a case the scientist could decide to revisit the data set to pursue other avenues of inquiry.

The product of the knowledge discovery phase is an algorithm. Algorithms can perform a variety of tasks:

- *Classification algorithms* categorize discrete variables (such as classifying an incoming email as spam);
- *Regression algorithms* calculate continuous variables (such as the value of a home based on its attributes and location);
- *Segmentation algorithms* divide data into groups or clusters of items that have similar properties (such as tumours found in medical images);
- *Association algorithms* find correlations between different attributes in a data set (such as the automatically suggested search terms in response to a query); and
- *Sequence analysis algorithms* summarize frequent sequences or episodes in data (such as understanding a DNA sequence to assign function to genes and proteins by comparing it to other sequences).

Application

Associations discovered amongst data in the knowledge phase of the analytic process are incorporated into an algorithm and applied. While in some cases algorithms are applied to non-personally identifiable data, they may also be applied to information that pertains to individuals to derive insights about them. An organisation may, for example, classify individuals according to certain criteria, and in doing so determine their suitability to engage in a particular activity. It may predict what individuals may buy or where they may travel, and on that basis decide what to market to them. In the application phase organisations reap the benefits of knowledge discovery. Through application of derived algorithms, organisations make determinations upon which they can act.

Challenges in Current Law, Regulation and Fair Information Practices

The data environment anticipated by current legal regimes and guidance differs markedly from the one we know today. Long-recognised instruments of data protection and guidance assume a world in which individuals provide consent to the use and processing of their data based on disclosures made by organisations; data controllers reasonably specify what data they need to collect and how they will use it; and organisations are motivated to minimize the data they collect and hold.

Today, the capturing of information that fuels big data is ubiquitous. Data is gathered, generated, stored and processed in a highly interconnected environment. It is collected through sensors, surveillance video and online social network interactions. Web surfing, tweets, social media posting, email, text messages, instant messages, real-time chat messages and audio messages generate data. Individuals freely make information about themselves public through social networks. In just about everyone's pocket or purse is a mobile device that tracks and captures time and location data. The deluge that results from the collection and generation of information yields the big data that, coupled with analytics, holds such potential for innovation, economic growth and societal good.

Application of analytics to big data does not conform well to traditional legal approaches because big data does not result from one-on-one interaction between the data controller and the individual. Big data instead pulls in information from disparate sources. Its value derives not only from its volume, but also from its varied and expansive scope — big data brings together an enormous pool of information that initially may seem unrelated. The more varied and representative the data is, however, the more likely the possibility of identifying unexpected relationships and the truer and more accurate the results of analysis. Moreover, analytics often explores data in nontraditional ways. While it may be applied to find answers to specific, predetermined questions, it may also explore data to see what insights it may hold. Analytics may be applied to data over and over again, identifying relationships and trends, raising and seeking answers to new questions.

As a result, some longstanding notions of fair information practices and provisions of existing law and guidance raise significant challenges for organisations that want to apply analytics to big data.

Consent: Nearly all data protection law and guidance is based on concepts of consent. In the United States, notice and consent are fundamental to fair information practices as articulated by the Federal Trade Commission,[17] and the U.S. Department of Commerce has recently reaffirmed consent as central to information privacy protection.[18] The European Directive requires that data

[17] "Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers," http://www.ftc.gov/os/2012/03/120326privacyreport.pdf. (Last visited January 7, 2013).

[18] "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy," http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. (Last visited 7 January 2013).

controllers establish a legal basis for processing data and provide for consent as a mechanism by which organisations meet that requirement.[19]

Principles of fair information practices, and the guidance in law and best practices derived from them, require that organisations inform individuals, among other things, about the collection of information and how it will be used. Based on that disclosure, individuals may consent to the use of the data. In Case 2 and Case 3, Merck and IBM are able to obtain from individuals explicit consent to the use of data that pertains to them. In a growing number of instances, however, such consent may not — and will not — be feasible. Users of data for analytics may not be able to locate individuals to obtain consent, particularly when carrying out longitudinal studies that may span a significant period of time. Analytic processing can involve the use of increasingly large data sets obtained from such diverse sources that obtaining consent may not be practicable. Moreover, given that analytics entails a knowledge discovery phase that allows for exploration of data to determine what insights it may yield — and thus how and for what purposes it may ultimately be used — providing the disclosure necessary for fully informed consent may not be feasible. And because analytics often involves iterative processes, data controllers may not be reasonably expected to return to data subjects to obtain consent for each application of analytics.

Consent also may not be appropriate in cases where the analytics process supports activities that are recognized to provide broadly accepted public benefits (e.g., scientific or healthcare research). It is often the case in such instances that to derive the optimal and most accurate results from analytics, the research data must be as complete and representative as possible. Allowing individuals to opt out of the use of data for such purposes could compromise the ability of analytic research to arrive at the most accurate and broadly applicable results.[20]

Legitimate business purpose: Another way in which controllers may establish a legal basis to process data is by establishing what is referred to as "a legitimate business purpose." Data may be used for such a purpose "except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject." How organisations practically establish

---

[19] EU Data Protection Directive 95/46/EC. "Member States shall provide that personal data may be processed only if:
(a) the data subject has unambiguously given his consent; or
(b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or
(c) processing is necessary for compliance with a legal obligation to which the controller is subject; or
(d) processing is necessary in order to protect the vital interests of the data subject; or
(e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data is disclosed; or
(f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject, which require protection under Article 1.

[20] The limitations of consent in instances where individuals are not in a position to freely grant it are reflected in Article 29 WP, Opinion 8/2001 on the processing of personal data in the employment context, adopted September 13, 2001, WP 48 p. 3, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2001wp48en_en.pdf.

that data is to be used for a legitimate business purpose for the application of analytics to big data requires further clarification.

Purpose specification: Fair information practices, and many of the laws that are derived from them, require that data controllers specify for what purpose data will be used. Because big data may serve purposes that can be revealed only through the knowledge discovery phase of analytics, organisations either will not be able to describe in their notices to what purpose data will be put, or will be forced to articulate that purpose so broadly as to lack meaning.

Data minimisation: The fair information practice principle of data minimisation requires that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose and retain the data only for as long as is necessary to fulfill that purpose. In light of the potential to reveal knowledge that large, varied and extensive data sets may hold, organisations, equipped with the ability to store data cheaply, will be motivated to maximize their data holdings rather than minimize them. The failure to do so may limit or preclude their ability to harness the power of big data and analytics to address important economic, scientific and social issues.

Organisations may also need to retain data for follow-up purposes — to track where data used in analytics originated and how it was used, to test the validity of findings derived from data and to identify where any errors may have been introduced.

Prohibitions against automated individual decisions and profiling: The European Directive precludes the use of data to arrive at what it refers to as "automated individual decisions" about individuals.[21] The proposed EU data protection regulation similarly provides that individuals have the right not to be subject to decisions about him that have legal effects and are based on profiling that results from automated processing.[22] These broad prohibitions would prevent organisations from using analytics for a wide range of beneficial purposes. While these provisions are clearly designed to protect the individual, a more nuanced approach could allow for use of analytics for valuable purposes while still providing appropriate safeguards.

[21] The Directive limits the use of automated processing to arrive at decisions about the individual in Article 15, which grants "the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him." The Article provides for limited exceptions in cases of certain contracts and in certain instances when authorized by law. EU Data Directive 95/46/EC.

[22] Article 20 of the proposed regulation states that "[e]very natural person shall have the right not to be subject to a measure. . . which is based solely on automated processing intended to evaluate certain personal aspects relating to this natural person or to analyse or predict in particular the natural person's performance at work, economic situation, location, health, personal preferences, reliability or behaviour." Proposal for a Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Com (2012) 11 final.

Goals for Analytics Guidance

1. Recognize and reflect the two-phased nature of analytic processes.

Traditional methods of data analysis usually involve identification of a question and analysis of data in search of answers to that question. Use of advanced analytics with big data upends that approach by making it possible to find patterns in data through knowledge discovery. Rather than approach data with a predetermined question, researchers may analyze data to determine what it can tell them. The results of this analysis may be unexpected. Moreover, this research may suggest further questions for analysis or prompt exploration of data to identify additional insights, through iterative analytic processing.

Knowledge discovery represents the first phase of analytic processing. The second involves applying to data the algorithms derived through knowledge discovery. Application of the algorithm enables insights about individuals: based on results of this application organisations may make decisions or predictions about the individual that they may act upon.

Guidance must recognize the differences in these two phases and the diverse challenges each presents. The knowledge discovery phase may require a different approach to guidance than the second, reflecting 1) the need to understand what data reveals; 2) how data is explored and algorithms are derived; and 3) the ways in which risks can be mitigated in this stage through de-identification and enhanced security. For the most part, the knowledge discovery phase does not involve analysis of a particular individual's data (which may be de-identified or pseudonomised), and does not result in decisions about him or her.

In contrast, the algorithms derived in the first phase may be applied to information about the individual to better understand him, predict his behaviour or make a decision about his qualification to participate in a certain activity or enjoy certain benefits. This phase of the analytic process can involve application of an algorithm to personally identifiable information, or may yield insights about an individual. Because the individual is implicated in this phase of analytics, different protections are warranted.

2. Provide guidance for companies about how to establish that their use of data for knowledge discovery is a legitimate business purpose.

Some jurisdictions allow for processing of data for a legitimate business purpose, but provide little guidance about how organisations establish legitimacy and demonstrate it to the appropriate oversight body. Guidance for analytics would articulate the criteria against which legitimacy is evaluated and describe how organisations demonstrate to regulators or other appropriate authorities the steps they have taken to support it.

3. Rely upon fair information practices, but apply them in a manner appropriate to the processing of big data for analytics.

While use of analytics and big data may challenge the way we apply them, principles of fair information practices should remain a cornerstone of guidance. In the context of analytics, fair information practices may be weighted — an organisation may consider the goals of one principle; if that principle cannot be feasibly applied, the organisation may consider how those goals can be achieved through robust application of another principle.

Fair information practices may also be applied based on a model in which the intended use of data triggers an organisation's requirements under fair information practices. Because the application of analytics to big data practically challenges traditional notions of consent, the manner in which data is used could provide a more workable way to establish an organisation's obligations. Under such a governance model, organisations consider potential uses for data (e.g., product fulfillment, internal administration, marketing) and meet certain requirements based on that use.[23]

4. Emphasize the need to establish accountability through an internal privacy programme that relies upon the identification and mitigation of the risks the use of data for analytics may raise for individuals.

Because analytic processing challenges how fair information practices are applied, it is important that organisations implement an internal privacy program that involves credible assessment of the risks data processing may raise, and practical, effective steps to mitigate those risks. Risk mitigation may involve de-identification and pseudo-nominisation of data, as well as other controls to prevent re-identification of the original data subject. It could entail evaluating data sources and data quality, and matching the level of quality to the nature and sensitivity of the data and the analytic inquiry. Risk assessment and mitigation could also include examination of the analytic process itself to identify where errors, flawed assumptions or faulty data might have been introduced, or to retrace methods and understand choices made throughout the process.

5. Take into account that analytics may be an iterative process using data from a variety of sources.

While this paper describes the steps involved in analytics, analytics is not necessarily a linear process. Insights yielded by analytics may be identified as flawed or lacking, and data scientists may in response re-develop an algorithm or re-examine the appropriateness of the data for its intended purpose and prepare it for further analysis. Knowledge discovery may reveal that data could provide additional insights, and researchers may choose to explore them further. Data used for analytics may come from an organisation's own stores, but may also be derived from public records. Data entered into the analytic process may also be the result of earlier processing.

---

[23] "A Use and Obligations Approach to Protecting Privacy: A Discussion Document," The Business Forum for Consumer Privacy, December 2009, found at http://www.huntonfiles.com/files/webupload/CIPL_Use_and_Obligations_White_Paper.pdf. (Last visited 7 February 2013).

Guidance for analytics must recognize that processing of data may not be linear, and may involve the use of data from a wide array of sources. Principles of fair information practices may be applicable at different points in analytic processing. Guidance must be sufficiently flexible to serve the dynamic nature of analytics and the richness of the data to which it is applied.

6. Reinforce the importance of appropriate data security measures.

While the need to implement security for data is well established, the large volume of data used for analytics heightens the importance of security measures. Moreover, analytic processing of big data can often yield sensitive insights about individuals. Security requirements appropriate to the risks raised by the volume and sensitivity of the data and the results of processing will be essential to effective guidance.

7. Foster interoperability across diverse jurisdictions.

Privacy values, concepts of data protection and the legal regimes instituted to further them differ across jurisdictions. Notions of privacy and the appropriate use of information in Japan may differ markedly from those in Canada. While the privacy and data protection regimes of most jurisdictions are based in a civil law tradition, those of the United States and the United Kingdom, for example, are founded in common law. While some countries and regions rely on law, others depend on industry compliance with guidance provided in international agreements and industry best practices, coupled with government oversight.

While privacy and the rules that protect it remain culturally based, the need for interoperability between diverse regimes has been emphasized in recent public policy documents.[24] To be useful, guidance for analytics must acknowledge local differences, but also facilitate the smooth, global movement of data. It must serve as a bridge between various regimes and not create impediments that would raise challenges, either to organisations implementing analytics for big data or to governments and regulators seeking to encourage its responsible use.

Conclusion

Analytics and big data hold growing potential to address longstanding issues in critical areas of business, science, social services, education and development. If this power is to be tapped responsibly, organisations need workable guidance that reflects the realities of how analytics and the big data environment work. Such guidance must be grounded on the consensus of international stakeholders — data protection authorities and regulators, business leaders, academics and experts, and civil society. Thoughtful, practical guidance can release and enhance the power of data to address societal questions in urgent need of answers. A trusted dialogue to arrive at that guidance will be challenging, but cannot wait.

[24] *See*, for example, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy," The White House, February 2012, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. (Last visited 7 February 2013).