Title: Big Data Analytics: Risks and Responsibilities

Short Title: Big Data Risks

Submission type: Original Article

Authors:
K. Krasnow Waterman
CEO, LawTechIntersect, LLC, New York, NY 10022, USA
Visiting Fellow, MIT, Computer Science & AI Lab, Cambridge, MA, 02139, USA
Paula J. Bruening
Senior Counsel, Global Privacy Policy, Intel, Washington, DC 20004

Abstract: Big data analytics sifts through mountains of data to identify or predict facts about individuals and to use those facts in decisions ranging from which products to sell them to whether to provide them medical treatment.

- Given the present state of technology, there are risks associated with big data analytics: source data may be misunderstood or contain errors and analytics processes may introduce new error or be less exact than intended.
- Data protection and the application of principles of fair information practices promote the responsible management and use of information about individuals and guard against risks that data and its use may raise.
- Given the potential harm to individuals - from the denial of credit or care to the elimination of educational opportunities - we addresses the corporate, social, and ethical responsibility to engage the appropriate professionals in making decisions to use the resulting predictions as well as the responsibility to implement a robust risk management process.

Keywords: analytics, big data, corporate accountability, data protection, fair information practices, risk management

I. Introduction

Big data analytic processing holds tremendous potential. The ability to successfully utilize analytics with big data promises solutions for health care, education, scientific research, economic growth and delivery of social services. While the possibilities are well recognized, observers also point out the risks inherent in big data, analytic processing, the models they yield, and the application of their predictions. These are especially significant when data pertains to individuals or when analytic processing yields faulty or incorrectly interpreted results that may have negative consequences.

Data protection and the application of principles of fair information practices promote the responsible management and use of information about individuals and guard against risks that data and its use may raise. They encourage data practices that protect information against misappropriation, loss or misuse and that ultimately protect individuals from the harm that may result.[1]

The principle of Data Integrity states that personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date. This principle promotes the use of data of a quality commensurate to the purpose for which it being used. The Data Integrity Principle is reflected in the EU data protection directive, which requires Member States to provide that personal data is, *inter alia* "accurate and, when necessary, kept up to date." It further requires that "every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified." Similarly, the US Privacy Act requires that government agencies "maintain all records which are used by the agency in making any determination about any individual with such accuracy, relevance, timeliness, and completeness as is reasonably necessary to assure fairness to the individual in the determination[.]" The principle is also found in international guidance on privacy and data protection and industry best practices. The quality, currency and suitability of data that the Data Integrity principle fosters is particularly important in big data analytics.

Some of the greatest potential of big data analytics lies in its ability to yield predictions and deep insights about individuals. Even when the processing and the information on which it relies are trustworthy, the results can have profound consequences, affecting one's ability to exercise certain life choices or to take advantage of certain opportunities. In doing so, analytics may yield predictions or arrive at decisions about individuals that raise important questions about self-determination, personal autonomy and fairness. It is particularly important, then, that organizations understand the sources, formatting and limitations of the data they use if they are to mitigate risks to individuals and act within the boundaries of applicable law.

Use of big data and analytics raises risks not only to individuals, but also to the brand and reputation of the organization. Assessing and mitigating the risk posed by such complex processing requires a commitment of resources at a level similar to that dedicated to managing

[1] When fair information practices are articulated in law, they are usually enforced by regulatory agencies. In Europe, independent data protection authorities supervise compliance, hear complaints, and enforce law. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Article 28 provides that Member States should establish such authorities and articulates their roles and responsibilities.

In the United States, the Federal Trade Commission may bring an enforcement action under Section 5 of the FTC Act, which prohibits "unfair or deceptive acts or practices in affecting commerce." 46 USC 5. The FTC may also bring an action under laws and regulations including the Fair Credit Reporting Act ("FCRA") 15 USC 1681 et seq., and the privacy provisions of the Graham Leach Bliley Act ("GLBA") 15 USC Sections 6801-6809. Under its FCRA authority the FTC has been studying and causing the correction of inaccurate personal data underlying credit ratings. http://www.ftc.gov/sites/default/files/documents/reports/section-319-fair-and-accurate-credit-transactions-act-2003-fifth-interim-federal-trade-commission/130211factareport.pdf other significant risks within the organization. It requires diverse skill sets that in combination empower decision makers to understand the structure and challenges of big data, the mathematics and science of analytics and modeling, and the economic and socio-political implications of analytics for individuals, markets and society at large.

II. The Phases of the Big Data Process

In its recent paper, "Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance,"[2] the Centre for Information Policy Leadership describes analytic processing of big data as involving two phases*: knowledge discovery* and *application*.

In *knowledge discovery,* data scientists acquire and analyze data to determine what insights it may yield. Knowledge discovery involves gathering data, pre-processing it into a useable format, consolidating it, and analyzing it to discover what it may reveal. A final "interpretation" step involves reviewing how the model was determined, the choices that were made about data throughout each of the previous steps, the processes by which the data was analyzed, and how conclusions were reached. This review enables an organization's decision-makers to evaluate how trustworthy and reliable a model is and whether it should be used. The Centre paper notes that understanding the knowledge discovery process is particularly necessary for organizations that adhere to an accountability approach to data governance,[3] whereby the organization holds itself out as responsible and answerable for understanding and mitigating the risks their use of data raises.

In the *application* phase, correlations discovered amongst data in the knowledge discovery phase are incorporated into an algorithm and applied to make predictions and/or business decisions. Thus, in the application phase organizations reap the benefits of knowledge discovery.

III. Risks in the Knowledge Discovery Phase

The knowledge discovery phase of data analytics raises risks of producing inaccurate results. These risks are introduced through flaws in the data itself as well as limitations inherent in analytic processing. The risks are compounded by the challenges that define "big" data, known as the "5V's" – volume, variability, velocity, veracity, and value. Many of the following examples reflect problems where even known remediation techniques cannot be applied to so much (volume) diverse (variability) data given the speed with which it is accumulated and the

[2] "Big Data and Analytics: Seeking Foundations for Effective Guidance," (2013) Centre for Information Policy Leadership white paper 2/2013 < > accessed 24 November 2013

[3] An accountability approach to data governance requires that an organization demonstrate commitment to accountability, implement data privacy policies linked to recognized external criteria such as law, regulation or accepted industry best practices, and implement mechanisms to foster adherence to those policies and responsible decision-making about the management and protection of data. For a comprehensive review of accountability see "Data Protection Accountability: The Essential Elements," (2009) Centre for Information Policy Leadership white paper 9/2009 <http://www.huntonfiles.com/files/webupload/CIPL_Galway_Accountability_Paper.pdf> accessed 24 November 2013Windows User Page 3 4/3/14

time available before a response is needed (velocity). Inaccurate results (low veracity) in knowledge discovery will yield predictions or classifications that are incorrect or misleading (low value).

A. Risks that Arise in the Data Environment: Collection and Aggregation

Big data proponents argue now that analytic tools are able to work on entire, massively large datasets, flaws in the underlying data do not significantly affect outcomes. They argue that technology could only handle smaller data sizes, margin of error was an issue because samples rather than entire data sets were analyzed and results were extrapolated to describe the whole.[4] But, practical experience shows that significant swaths of these faults can exist in the data and programmed tests for data quality can miss them. This results in matches not being identified, most often resulting in underrepresentation of one characteristic or group and overrepresentation of another.

1. *Corruption of Collected Data*

Some flaws in data result from the initial collection of the data. The same problems that corrupt an email and make it unreadable or cause a dropped phone call can occur when large quantities of data are collected or transmitted. Such flaws can result in data not being recorded at all, not being readable, or being modified or changed in an unexpected way.

2. *Flaws in Data Entry*

Data may be entered inaccurately, so that data elements are placed in the wrong fields. This may occur when effective controls have not been implemented. Over time, this has an impact similar to missing one response bubble on a standardized test answer page and putting all the remaining responses in the wrong rows. These sorts of flaws result in the wrong data or no data in data fields.

3. *Flaws Resulting from Merging of Legacy System Data*

Data to be analyzed may be derived from multiple legacy systems that store and format data differently. The vast majority of data used in business, health, and government has been collected over a long period of time and through a variety of systems. This “legacy” data is not all stored in the same software product and does not necessarily include standard data elements (for example, same categories of data, sizes of field). Sometimes the authors of different systems have used the same name for a data field but have different standards for what data the field includes. Problems can occur when such mismatched data is aggregated for use in analytics. Similar problems may arise when using data collected across different countries or regions of the world.

When date of birth, for example, is collected in the United States and Europe, it will be formatted according to different conventions. A March 6, 2003 birthday will be designated 3/6/2003 in the

[4] Victor Mayer-Schoenberger, and Ken Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think.* (Houghton-Mifflin 2013) pages 19-31

US, and 6/3/2003 in Europe. These would be machine read as different dates, leading to errors in analysis if not recognized and properly adjusted.

The problems arising from legacy data systems are exacerbated because traditionally no universal standard has defined data fields. In one project managed by our author, birthdays, presumptive death dates, and the dates on which businesses were started were inappropriately aggregated into a field designated "Date of Birth." When data is collected from sources around the globe, this problem can be aggravated by nuances in language translation. As a result, data may be wrongly included because it "looks right" and then wrongly interpreted because it has unanticipated meaning.

4. *Compound information*

Another common problem results when data is combined from systems that handled a piece of compound information, such as an address or name, differently. In some systems an entire address is stored in one field and in others the address is broken into multiple fields (for example, street, city, state/province). Most typically, when data sets in which the entire address in one field are aggregated into a bigger database, the entire address is pulled into the street field and the city and state fields end up empty.

This same problem is equally true and very common with people names, where there is wide cultural divergence in the number and order of names people use. Consider, for example, the difference between the convention of First, Middle, Last versus the conventions which use matronym and patronym versus those which also add honorifics and origin. Different systems authors have chosen to capture any of these options. For example, the name of the famous artist "Frida Kahlo" was actually her second middle name and patronymic; her full name was "Magdalena Carmen Frida Kahlo Calderon." If not identified and fixed, these flaws preclude the analytic tool from finding matches it should have or cause it to undercount the number of records about a particular subject.

B. Risks Arising in Analytic Processing

Analysis of the data can introduce other risks. These may arise from an incomplete understanding of the data, misunderstandings about what the data being analyzed represents, or the analytic processes themselves.

1. Understanding What Data Represents

A review of the results of an analysis must place the outcome in the context of the input. For example, depending upon the browser people used and the date on which a search was conducted, a sample of 6 million Internet searches may differ significantly in gender, geography, or religious distribution. A 2011 sample of Google searches would have included 45% women, while a sample of Bing searches would have included 58% women; thus the choice of search engine would have resulted in a 13% gender difference in the data.[5] Also in 2011, internet usage

[5] "A Tale Of Two Studies: Google vs. Bing Click-Through Rate," (*The Moz Blog*, 6 December 2011) < http://www.seomoz.org/blog/a-tale-of-two-studies-google-vs-bing-clickthrough-rate > accessed 5 December 2013

peaked in the Middle East during Ramadan,[6] but was the lowest in Britain on Christmas.[7] Before stating conclusions, therefore, one would need to know the date(s) of the sample searches and determine whether any holiday (or other event) occurred that would have skewed the demographic distribution. In May 2012, Internet Explorer was still the most commonly used browser in North America, China, and Australia, while Chrome dominated Eastern Europe and most of Latin America. Thus one would need to know which browser was the source of the sample searches to be able to reflect the likely geographic distribution of the users.[8] Failure to take into account the ongoing change in search engine and browser market may result in incorrect assertions of what is learned about whom.

Similarly, "Pre-processing" activities involve decisions that change the data being analyzed and therefore the nature of the result. Consider, for example, the collection of Enron emails that was released by the Federal Energy Regulatory Commission. The number of emails released exceeded 1 million, however, when three distinguished research entities pre-processed the data to remove duplicates, blanks, etc., their databases ranged in size from ~250,000 to ~600,000 emails and from 149 to 161 users represented in the data.[9] Discrepancies like these will flow through to affect different statistics about what the data contains, possibly significantly undercounting items in the smallest dataset and over-counting them in the larger one.

2. *Selecting the Appropriate Analytic Tool*

Analysis of data itself is not a settled science; opinions differ about which tool is best and what data to use. For example, one very successful practitioner asserts that results are improved by adding more independent data rather than continuing to improve the analytic algorithm.[10] This was in the context of describing work his Stanford University students were doing on the

[6] "High Internet Activity over Ramadan period in Saudia Arabia says Effective Measure," (*AMEinfo.com,* 25 August 2011) < http://www.ameinfo.com/273720.html > accessed 30 May 2013

[7] "Christmas is quietest online day of the year," *The Guardian* (*Technology News,* 26 December 2011) < http://www.telegraph.co.uk/technology/news/8972916/Christmas-is-quietest-online-day-of-the-year.html > accessed 5 December 2013

[8] Charles Arthur, "No, Google's Chrome isn't the world's leading browser - yet: see our map," *The Guardian* (*Technology Blog*, 20 May 2012) <http://www.guardian.co.uk/technology/blog/2012/may/22/google-chrome-isnt-world-leading-browser> accessed 5 December 2013

[9] K Krasnow Waterman, "Knowledge Discovery in Corporate Email: The Compliance Bot Meets Enron," pp. 47-48 (SM Thesis, Massachusetts Institute of Technology 2006) <http://dspace.mit.edu/bitstream/handle/1721.1/37574/85813548.pdf?sequence=1> accessed 5 December 2013

[10] Anand Rajaraman, "More data usually beats better algorithms," (*Datawocky*, 8 March 2008) <http://anand.typepad.com/datawocky/2008/03/more-data-usual.html> accessed 5 December 2013 (Note: Dr. Rajaram is known both for his implementation successes and as the co-author of *Mining of Massive Datasets* (Cambridge University Press, 2013.)

competition to produce a recommendation algorithm for NetFlix.[11] In that context, it's easy to understand that one could improve predictions if it were possible to add in the movies the same persons borrowed from the library or rented from a local video store (the primary alternatives at the time) to the ones they acquired through NetFlix; it is equally easy to understand that the predictions made without that additional data will be less accurate.

To provide a sense of how many tools might be used to solve a problem, at the end of the first year of the NetFlix competition, the leading team was using a combination of 107 algorithms. And, to understand their value, NetFlix found that combining only two of them produced most of the error reduction they were seeking. NetFlix, expressly noted that a more accurate result was possible, and then made a cost-benefit decision to implement only the two.

Even then, the algorithms had limitations – they could "only" handle 100 million records, rather than the needed 5 billion – and had to be adjusted. It is important to note the additional risk that, while some analytic processes will produce a clear error message when reaching the scale they can manage, others have been known to only process what they can and return a conclusion without indication of the volume not processed.

3. *Understanding Output: Estimates and Errors*

The global growth of data has been described in size as multiples of the Library of Congress and in speed by reference to companies that collect 1 million transactions in an hour. The volume and velocity of big data drive the need to increase the scale and the speed at which analysis occurs. Today, the analytics community addresses these challenges through extending such capabilities as parallel processing – breaking a large problem into smaller ones that are computed at the same time by different processors; machine learning – programs that allow the computer to optimize or improve something in future computing based upon experience from prior computation; and natural language tools – methods to recognize and use pieces of text based on their semantics or context. As with anything one seeks to do newly or faster, the potential for error is present, and some of these technologies have built-in features to try to identify, correct, and limit the magnitude of errors. Also, in order to meet speed requirements, some of these technologies focus on estimating an answer rather than trying to produce a wholly accurate one.

C. Quantifying the Risk in Prediction

The quality of the prediction depends on all of the quality of all of the knowledge discovery steps that precede it. As described above, each of these steps involves risks of error and misunderstanding. Any prediction must be provided in the context of the potential risks and a quantified margin of error.

For example, the results of analytics, such as described in the fourth step of knowledge discovery, are usually presented accompanied by a value that represents statistical confidence - the % range of believed accuracy (for example, ±9%). This value represents the likelihood that

[11] Xavier Amatriain & Justin Basilico, "Netflix Recommendations: Beyond the Five Stars (Part 1)," (*Netflix TechBlog*, 6 April 2012) <http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html> accessed 5 December 2013

the same result will be achieved when the data is analyzed using different tools or methods. So, if a prediction states that 78% of the population is/will do something, say, visit a national park within next 12 months with a 9% confidence, then it is possible that as many as 87% (78% *plus* 9%) will make such a visit, but also possible that as few as 69% (78% *minus* 9%) of the population will. If there were 5% errors in each of the three preceding steps of the knowledge discovery process the likelihood that the prediction is accurate could be reduced by 15% (minus 5% each for the collection, aggregation, and pre-processing steps) to make the prediction only 54% likely (the previous low estimate of 69% minus 15%) – barely better than the flip of a coin. This, of course, assumes that the appropriate math for compounding these errors in knowledge discovery is addition. If the errors have some sort of overlapping quality and the proper math involves fractions, the likely correctness of the prediction is not quite so low. But, if the compounding of these flaws has some sort of geometric impact, then the correct math would involve some multiplication of the error values, driving the likely correctness of the result well below 50%.

For these reasons, it is critical that the prediction be accompanied by a clear explanation of the risks at each stage, the mathematical method of compounding them, and the believed aggregate risk.

IV. Risk in the Application

Risks arise in the application phase of the analytic process primarily in two ways. First, application of predictions may lead to incorrect or misleading results. For example, if an analytic model yields a prediction that 78% of the population will take some action or behave in a certain way, it is still the case that 22% of the population falls outside the boundaries of that prediction. For 22% of the population, that prediction is wrong and its application to them could have negative consequences.

However, even when the quality of the data science is high and the application yields accurate, reliable results, the predictions and inferences yielded by analytics may be deemed too invasive. Moreover, results may raise concerns that analytic models may be applied in ways that will compromise the individual's autonomy or inappropriately limit his or her choices.

What is considered "too invasive" is often a subjective assessment that varies from individual to individual and from culture to culture. However, it is useful to consider different data uses to see where lines may be drawn. Individuals may have little concern about use of analytics to anticipate what clothing or furniture they may be interested in purchasing, but they may object to applying analytics to anticipate sexual orientation, medical status, or propensity to develop a disease or medical condition.

Other cases may be less clear. For example, analytics may enable companies to identify different pricing for customers depending upon their means (perhaps eliminating access to discounts for the more affluent) or context (charging higher prices for services provided to businesses than to individuals). Analytics could also identify an employee's susceptibility to an infectious disease prevalent in a foreign location that might factor against his eligibility for a work assignment.

The risk is compounded when big data and analytics are used in a way that could potentially limit an individual's access to information, resources, or choices. Big data may yield insights, for example, about suitability for certain kinds of education or predictions about an individual's success in a particular course of study. While such insights might be deemed useful, their use to track students or preclude them from pursuing a particular career could be deemed unacceptable. Similarly, analytics used with big data related to health care may be welcomed when it enhances the understanding of a person's propensity for disease and how it might most effectively be addressed. However, it may raise serious concerns for individuals when that same analysis is used to assess his or her eligibility for health insurance coverage or for certain medical treatments.

Finally, organizations using big data and analytics must guard against overreach and consider the cumulative effect of analytics and big data. Their use for certain discrete purposes may raise few concerns. When taken together, however, they may raise questions of personal autonomy and begin to have broad consequences that may be deemed contrary to societal values. Anticipation of such effects could prompt public and government scrutiny leading to regulation that could constrain the use of big data for positive purposes.

V. Big Data, Analytics and Corporate Governance

Questions about big data and analytics raise risks that can have three components – risk of error; legal impact; and ethical breach.

1. Quality of the Analysis

As discussed above, the gathering and processing of data for analytics can introduce errors and distortions that compromise analytic models. Close monitoring of data sources, collection practices and integration, and the resulting models will be necessary to ensure that analytic algorithms are of appropriate quality for their intended uses. Auditing will be necessary across the knowledge discovery phase and the application phase, as both activities involve processes that can raise risks.

In reviewing the knowledge discovery aspect of analytics, companies should ask questions fundamental to management oversight in any field – who, what, when, where, how. From whom was the data acquired? What is in the data? How was it cleaned, formatted and integrated? What geographic area(s) or population(s) is represented in the data? Is there anything about the period over which it was gathered from individuals that should be accounted for? Was the algorithm tested to confirm that the inference or prediction is actually yielded and can be trusted?

The answers to these questions should be articulated in such a way that knowledgeable, responsible personnel are able to intelligently review and understand them. In turn, this analysis and review form the basis for decisions about whether or not the outcome of analytic processes should be applied to individuals and relied upon.

2. Legal Impact

A concept of privacy protection is to ensure that sensitive personal information is not used in ways that cause socially unacceptable harm. This may be regulated directly as privacy protection or as prohibitions of the harms caused. Organizations will need to monitor carefully their use of big data and inferences derived from data in light of both sorts of privacy-protecting requirements. For example, in the United States, a bank or mortgage company would be subject to the requirements and proscriptions of the Equal Credit Opportunity Act,[12] the law that prohibits the use of data on race and gender for purposes of determining an individual's creditworthiness. Such organizations are precluded from considering these characteristics in evaluating whether or not an individual qualifies for a mortgage loan. As data has been used traditionally, to ensure compliance appropriate personnel will review practices and procedures to verify that such information is not considered.

Such straightforward compliance is complicated by big data analytics. While a responsible organization may take steps to remove gender and race fields from data sets, the analytics applied to predict credit-worthiness may inadvertently infer gender and race from other data (e.g., zip code, retailers frequented, product preferences) not proscribed by law. In other words, even though *de jure* gender and race discrimination does not occur because such factors are not directly included in the analysis, the risk of equally impermissible *de* facto discrimination remains because these characteristics may be inferred indirectly and influence decision-making, contrary to legal requirements. Organizations will need to take care to foster compliance with existing law, which in most cases does not contemplate the nuanced way in which big data can reveal inferences, patterns and predictions about individuals. They will need to carefully consider their decisions about using certain kinds of data and algorithms that technically may fall within the bounds of law but violate its intent.

3. *Ethical Questions*

While issues about analytics and big data are most often framed in the context of privacy, this process in fact raises many larger questions related to issues of personal autonomy and individual choice. Application of analytics to big data can upset traditional notions of a "fair playing field" by so empowering organizations with insights and predictions as to leave individuals with little or no power to negotiate.

Analysis of data for insurance underwriting provides an important example. If each individual's policy is based on such comprehensive knowledge and accurate predictions about a currently healthy individual that he or she is simply denied coverage or charged premiums that would cover any and all predicted illnesses and provide for the company's profits, what would our commonly held ideas about "insurance" mean in light of such a potential practice? Analytics and big data also raise questions about the consequences to individuals when faulty data, often provided by a third party, are used to make predictions or arrive at decisions. What happens when inaccurate information – often generated in jest by others on social networks – about an individual's alcohol or drug use is used to establish correlations that affect their education, employment or financial opportunities? Even when inferences or correlations about people are accurate, for how long should such an assessment about a person be considered applicable or relevant?

Organizations and governments will need to be mindful of the consequences – sometimes unintended – of their big data and analytic processing decisions and carefully take into account not only the integrity of their processes but also their responsibility to consumers, their brand and society at large.

4. Who Decides?

Some literature suggests that the decision-making functions associated with big data analytics and application will be provided by an individual within the company well versed in the technical aspects of analytics, referred to as an "algorithmist."[13] However, the, the diversity, complexity and far-reaching nature of the questions – technical, ethical, and legal – that must be answered to arrive at appropriate decisions, and the high stakes for subject individuals and the company suggest that a broader set of skills and personnel be brought to bear to resolve these issues.

More appropriately, a team of personnel should be tasked with reviewing, interpreting and determining how the outcomes of analytic processes were attained, are understood, and are used. Such a team could be modelled on the internal review boards found in hospitals and research institutions and on risk management committees in corporations. It would bring together personnel with knowledge of statistics, computer science and mathematics; chief privacy officers and others familiar with law and public policy; and staff that deals with questions of ethics and corporate and organizational responsibility. As in any other risk management activity, senior management and the Board of Directors hold the fiduciary responsibility for insisting upon, reviewing, questioning, and understanding regularly scheduled, intelligible, summary reporting from this team.

VI. Conclusion

Big data and analytics represent a new era in the ability of organizations to tap the potential of the information economy. But this new capability is not without hazards. The ability to ingest massive volume, variety, and velocity of data for analytics does not eliminate risk – if not properly addressed such ability can often compound it, creating the risk of data use outside the bounds of law, regulation and ethical practice. To derive the greatest benefit from big data and analytics, institutions will need to understand and address the implications of choices about data and analytic tools. They will need to carefully assess the integrity of their analytic processes and the accuracy of their findings, and to consider the legal consequences for privacy and data protection of applying the outcomes of analytic models to information about individuals. Organizations concerned about brand, reputation and the privacy interests of their customers must bring to bear broad expertise to evaluate their processes and decisions and to foster trusted, ethical outcomes.

[13] Mayer-Schonberger and Cukier, (note 4) pages 180-182. Mayer-Schonberger and Cukier argue that algorithmists would be experts in the areas of computer science, mathematics and statistics and would be responsible for reviewing big-data analyses and predictions. They compare the role of the algorithmist to that of an accountant in providing impartial opinions about big data analytics and its applications. Further, algorithmists working within companies would be charged with protecting the interests of the company and the individuals affected by big-data analysis.