Issues in Information Systems
Volume 17, Issue IV, pp. 1-7, 2016

BIG DATA AND ETHICS: EXAMINING THE GREY AREAS OF BIG DATA ANALYTICS

Gwendolyn White, Xavier University, whiteg@xavier.edu
Thilini Ariyachandra, Xavier University, ariyachandrat@xavier.edu

ABSTRACT

With the rise in technologies that support the collection of big data, organizations are scrambling to analyze data to enhance decision making. As the richness of data gathered increases, organizations are provided eyes into greater amounts of details on components of transactions than ever before. This has left many wondering how ethical concerns of data have continued to diminish especially when considering privacy and security. This research identifies and discusses the major ethical themes applicable to big data analytics from past literature. It described ethical questions to consider when dealing with the grey areas of big data analytics.

Keywords: Information Technology (IT), Ethics, IT and Ethics, Analytics and Privacy, Security and Big Data Analytics

INTRODUCTION

Big data usage has significantly increased in the past ten years and ethical guidelines are still in development. Ethics has been studied in information systems and information technology but ethics in big data are still in its infancy. This paper reviews existing literature in the rapidly growing field of big data and makes recommendations on developing ethics guidelines for big data.

The worldwide cache of data is expected to grow from 32 zettabytes[1] to 40 zettabytes in six years (Davis, 2014). New sources like mobile, social media, web and machine generated data will make up 85% of new data (Bearden, 2014). With these increases, high-value data will double and 60% of the information delivered to decision makers can be used to enhance business (IDC, 2016). Due to the increase of computing power and storage capacity, big data in its raw form is more than a structured database – it has created unanticipated uses for analysis. In addition, the quality of big data will be improved (Innovation Enterprise). Trends in 2016 are pointing towards Big Data becoming more main stream and not just for big business. In the next few years organizations will have big data applications integrated into their programs to retrieve data regarding their organizations. Over the next few years the amount of big data will increase because video audio and images will become part of existing data.

Currently 70% of organizations purchase data but that is expected to increase to 100% by 2019 (Information Management, 2015). In addition, data-as-a-service as a business model will enter the marketplace, similar to other existing services for organizations. (Davis, 2014). Big data has the potential to transform education, business, healthcare, transportation and other industries. New positions in analytics have emerged including data scientist, data analyst, and data explorer. However, this transformation comes at a cost of staffing problems. There are many analytics positions that go unfilled. 43% of organizations surveyed by A.T. Kearns consulting stated that 10% of digital/analytics positions go unfilled. In the next five years there will be at least a 33% increase for data analytics talent (AT Kearns Consulting). It is expected that over 181,000 positions in data analytics will be available by 2018 (Information Management, 2015). New major and degrees are being developed to provide an employment force for data analytics businesses.

It is expected by 2018 customer will interact with data based on cognitive computing (Information Management, 2015). All business analytics software will incorporate prescriptive analytics that is based on cognitive computing

technology (IDC, 2016). Big data analytics will be incorporated into applications and shall become accessible to businesses and also those that are non-coders (Davis, 2014). This will make it easier for people to use data and discover new insights about organizations or their personal lives. As big data grows, accessibility to the data increases especially if the delivery of the big data software changes to a cognitive technology (Needham, 2013). The more people connect using various devices the more data is generated and shared. The compound annual growth rate (CAGR) of collaborative software applications is expected to increase 9% in five years (Tene & Polonetsky, 2013).

How Big Data Affects Industry

Big data affects the industry because it is part of disruptive technology. A disruptive technology is "one that displaces established technology and shakes up the industry or a ground breaking product that creates a completely new industry" (Wessell, 2016). However, big data encroaches across many industries and businesses (Disruptive Technology Reconsidered: A Critique and Research Agenda, 2004).

Many industries are using big data to enhance profits, customer relations and determine new emerging trends. Education uses big data to review student performance, adjust curricula, and identify the various types of students especially high-rusk or first time student from certain populations. The TRI0 program identifies students who are the first in their family to attend college. The program helps students adjust to college life by teaching study skills, provide counseling and academic support. The government uses big data to evaluate various programs, manage agencies, monitor utilities and populations. Adjustments to budgets for government programs are made by analyzing big data. Retailers use big data to determine buying trends or use customer purchases as suggestions for other purchases. The attributes of the individual purchaser are used to determine what he/she might be purchased the next time and send appropriate advertisements to encourage additional shopping. This information is also used to determine offline purchase behavior based on their online behavior. Manufacturing uses big data to monitor quality and efficiency of manufacturing processes. Changes to improve manufacturing processes can increase profits and efficiency.

Traditional data scientists are uncomfortable managing big data. However as new methods and software are introduced the comfort level increases. New tools and technology make data scientist become open minded and change the way they approach big data. The cost of technology will decrease which will open the door for data to be used in good or bad manner (Needham). "When asked if there should be an ethical framework for collecting and using data, 42% of JSM survey respondents agreed that an industry standard should be in place, while 43% said that ethics already plays "a big part" in their research" (Joint Statistical Meetings, 2014). The level of comfort might come at a cost in relation to privacy, security, ownership, data quality and decision making.

Information Systems, Information Technology and Ethics

Ethics has always been a part of Information Systems (IS) but it was not necessarily defined besides "doing the right thing." That leaves the definition of ethics in relation to big data open and vague. In the past, ethics were used to enhance or fill in the gap regarding big data where there is no precedent.

Literature on ethics and information science and technology has existed since the 1940s. Norbert Wiener wrote about computing issues which became the foundation for ethics in information systems today. Subjects included artificial intelligence, robotics, and computing for people with disabilities (Weiner, 1950). Weiner's writings covered the potential problems that might arise from the introduction of computers into society. Later in 1976, Walter Maner created a branch of ethics called computer ethics (Maner, 1996). He developed a computer ethics starter kit that was distributed to colleges and universities to teach the importance of developing ethical principles centered on computers.

Technology has changed over the past 35 years which has increased the potential for ethical problems. In 1985 Deborah Johnson added to the computer ethics literature by updating the need for "applying ordinary moral norms in uncharted realms" (Johnson, 1985). The primary concern of ethics in computers during this period was the human

aspect. Luciano Floridi (1990) created a new focus for computer ethics called information ethics which expanded the focus from only human to creating a new environment for information systems and technology.

Changes in computing power, reduction in data storage costs, better data analysis, and improvement in networks and the internet have increased the potential for ethical violations (Olumoye, 2013). As time moved forward, ethics became a part of the normal daily framework and developed into an ideology.

Big Data and Ethics

There is a rapidly growing volume of literature that combines big data and ethics. Between 2001 to 2016 big data usage has increased along with the instances of actual and perceived ethical violations. Major industries affected by ethics and by data included healthcare, education and information technology. Within these industries, four major themes were found within the reviewed articles (See Table 1). These themes include privacy, security, ownership and decision making. According to Zwitter (2014), industry is moving towards "changes in how ethics has to be perceived away from individual decisions with specific and knowledgeable outcomes towards actions by many unaware that they may have taken actions with unintended consequences for anyone". Each of these four themes is described next.

Table 1. Major Ethical Themes in Literature

Theme	Ethical Challenge	Example
Privacy	Sharing of personal information without permission – de-identifying information	Data used to determine Ebola outbreak in 2015 Facebook study in 2012 to test user's emotions without their consent
Security	Protection of data from outside threats	Hospital data ransomed in 2016 due to lax security
Ownership	The rightful ownership of the data used for analytics	Research in illegal behaviors where the courts want the data to build a case against a person
Evidence Based Decision Making	The use of data to make decisions about a population based solely on quantitative information	States make decisions about welfare guidelines using income as the sole factor

Privacy

Privacy is the ability to be free from disturbance or observation. In many cases, big data encompasses a variety of data which includes personal information. Privacy fears will not stop the collection of big data (Davis, 2014). Many individuals are affected when their personal information is shared with others, especially without their consent. Organizations, in general, do not provide the details regarding the source of big data unless it is specifically requested (Tene & Polonetsky, 2013). There are state and federal guidelines for patient privacy. The Health Information Portability and Accountability Act of 1996, which protects individual health information, are primarily designed to protect data that has identifying information. The Genetic Information Nondisclosure Act of 2008 (GINA) prevents health insurance carriers from making health related decisions based on a person's genetic makeup (Thorpe & Gray, 2015). Doctors, nurses and other medical practitioners take the Hippocratic Oath promising no harm to patients. There are ethical codes related to patient information and confidentiality that must be respected due to the law (McHale, 2012). States normally develop their own privacy requirements which echo federal laws (Thorpe & Gray, 2015).

However, there are times that the specific individual named information is removed from records and then the data is shared. Since so many people are connected through a variety of data networks, the ability to generate and share data increases daily (Tene & Polonetsky, 2013). There are situations where consent is not possible – is this an ethical issue? Is the greater good more important than a person's individual right to privacy? Confidentiality and privacy have to be weighed against the greater good (McHale, 2012). One example that uses big data analytics is the identification of drug trends. The increasing use of a particular drug can inform public health officials and try to prevent abusive drug use (Thorpe & Gray, 2015). In this case, the use of big data is beneficial.

There are situations where removal of private information can be reversed causing the identification of individuals. "Protecting privacy will become harder as information is multiplied and shared ever more widely among multiple parties around the world. As more information regarding individuals' health, financials, location, electricity use and online activity percolates, concerns arise regarding profiling, tracking, discrimination, exclusion, government surveillance and loss of control" (Tene & Polonetsky, 2013).

Security

Security is the protection of data to ensure that others who do not have permission do not access the data. As more data is produced, there are more opportunities for data breaches. Not long ago software was kept in-house but now with the cloud, web based platforms and mobile technology software can be used from any location. These platforms are vulnerable to hacking, especially those that are interconnected. Data is collected from a variety of locations before analysis (Lyon, 2014). Many processors are also located outside the jurisdiction of the company that requested the information (Thorpe & Gray, 2015). Often existing security is not sufficient to protect data, which can lead to breaches and hacking. Since big data is vast, it is hard to protect. The lack of protection for data is a violation of ethics because it affects more than just security. Other ethical categories such as privacy and ownership are affected. The public wants to feel confident their data is protected and secure.

Ownership

Individual control over information is important to most people. When data is used, informed consent should be given by the subject (Olumoye, 2013). Big data is collected from multiple platforms and this makes it harder to track and contain the data. When subjects give consent it is assumed they understand the disclosure statement contained in the form. More often the disclosure is for liability and not retention and protection of the (Tene & Polonetsky, 2013). Subjects believe they still own their data – but that is far from the truth. Most people do not understand or use the right to access their data. The option to opt-in or opt-out of the data collection process gives subjects the perceived level of ownership of their own data (McHale, 2012). However, is this really applicable? Does giving full control to the subject become a detriment for researchers and the need to use big data? It depends on why big data is needed. If the need for big data is important the owner of the data will be overridden in the name of the greater good (Johnson, 1985). In the end, the data is not owned by the individual. It is just a mere pit stop before hitting the super highway of data transmission known as big data.

Evidence Based Decision Making

The use of algorithms to make decisions using big data can create discriminatory situations. "Predictive analysis may have a stifling effect on individuals and society, perpetuating old prejudices" (Tene & Polonetsky, 2013, p. 253). Removing personal aspects of data "compartmentalizes" individuals or groups, ignoring personal characteristics that might force a different decision. Data scientist use predictive analysis to make decisions about individuals or populations, but when sensitive topics are included, it can create discriminatory situations. Whether the data is sensitive or not the effect can be detrimental. "The wealthy and well-educated will get the fast track; the poor and underprivileged will have the deck stacked against them even more so than before" (Tene & Polonetsky, 2013, p. 254). Target used predictive analysis to determine if customer purchases revealed a pattern that indicated they were pregnant. Coupons for baby goods were sent to customers based on the results of the predictive analysis However, what if a woman did not want to reveal her pregnancy? Was it ethical to send the coupons based solely on the data and not include the personal situation? That is one of the problems with predictive analysis to make decisions based on numbers only.

Each of the ethical themes affect the entirety of big data from collection to reporting. There is still room for the development of specified ethical guidelines related to big data. At this time, it is left up to the analyst to do the right thing per existing laws or norms. The themes in the literature review can be applied across a data analytics conceptual framework to provide a view of ethics and how it affects big data.

Data Analytics and How Ethics Affects the Conceptual Framework

The conceptual framework for big data analytics processes regarding big data includes two major concepts with subsets: Big Data (procurement and storage; mining; editing; and representation of big data), Analytics (analysis/review and explanation). Within the conceptual framework management of data is important especially when sensitive data sets are used. All points in this conceptual framework are vulnerable to ethics concerns which include the major points found in the literature: privacy, security, ownership, and evidence-based problems. Data analysts have to be careful when using big data. It cannot be assumed that big data is always free from controversy. There is an assumption that analyst will do the right thing when it comes to big data. Table 2 explains the conceptual framework, concerns (in the form of questions) and where the ethical themes are prevalent.

Table 2. Big Data Analytics Processes - Conceptual Framework, Questions and Ethical Concerns

Big Data	**Concerns**	**Ethical Theme**
Procurement and Storage of Big Data	Is the use appropriate? Is the source appropriate? Who owns the data – do they have access after collection?	Privacy Security Ownership Evidence-Based Decision Making
Mining	Is the source reputable?	
Representation of big data	Is the data truly representative of the population or subjects?	
Analytics	**Concerns**	**Ethical Themes**
Analysis/Review	Are mistakes made and who is responsible?	Evidence-Based Decision Making Ownership
Explanation	Has the data significantly changed? What are the consequences of reporting the findings?	

SUMMARY

Big data and ethics is a newer area that still needs specified guidelines. Within the next couple of years, it is recommended that specific ethical standards are developed to manage and care for big data. Creating laws based one this will be difficult. Big data is gathered from so many locations. One way to potentially to create laws based on ethics is to separate big data into specific categories and then design laws tailored to the category. It is critical to protect data to ensure the collection, analysis and reporting will always fall within the ethical guidelines. As data collection continues at an exponential pace, greater care must be given to the ethics involved in the manipulation and the use of big data for organizational decision making.

REFERENCES

AT Kearns Consulting. (n.d.).

Bearden, R. (2014). Hadoop Summit. *CEO Keynote Conference Speech.* San Jose, CA: Hadoop Summit.

Bienkowskie, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Washington, D.C.: U.S> Department of Education, Office of Educational Technology.

Boardman, J. K. (2011). Addressing the "research gap' in special eduacation through mixed methods. *Learning Disability Quarterly*, 208-218.

(2007). Business intelligence software. In *Network Dictionary* (p. 79).

Davis, J. (2016, January 4). *Big data predictions for 2016*. Retrieved May 5, 2016, from Information Week: http://www.informationweek.com/big-data/big-data-analytics/big-data-predictions-for-2016/d/d-id/1323671

Disruptive technology rconsidered: A critique and research agenda. (2004). *Journal of Product Innovation, 21*, 246-258.

Endley, K. (2009). Business intelligence: turning knowledge into power. *School Business Affairs*, 27-28.

Floridi, L. (2008). Foundations of computer ethics. In *The Handbook of Computes Ethics* (pp. 3-24). Hoboken, NJ: John Wiley and Sons.

Gunter, P. (2001). Data-based decision-making to ensure positive outcomes for children/youth with challenging behaviors. In L. B. (Eds.), *Addressing social, academic, and behavioral needs within inclusive and alternative settings* (pp. 49-52). Reston, Virginia: Council for Exceptional Children.

Gunter, P. L., Callicott, K., Denny, R. K., & Gerber, B. L. (2003). Finding a place for data collection in classrooms for stuents with emotional/behavioral disorders. *Preventing School Failure: Alternative Education for Children and Youth*, 4-8.

IDC. (2016). *IDC*. Retrieved May 5, 2016, from IDC predicts.
(2015). *Information Management.*

Innovation Enterprise. (n.d.). Retrieved May 5, 2016

Johnson, D. (1985). *Computer ethics.* Englewood Cliffs, NJ: Prentice-Hall.

(2014). *Joint Statistical Meetings.*

Jonaganti, T. K. (2008, December 30). *Data warehouse modeling.* Retrieved November 8, 2015, from SlideShare: http://www.slideshare.net/vivekjv/data-warehouse-modeling-presentation/13

Kavale K. A. & Forness, S. R. (2000). Policy decisions in special education: The role of meta-analysis. In E. P. R. Gersten, *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues* (pp. 281-326). Mahway: Lawrence Erlbaum.

Levine, E. (2002). Building a data warehouse. *American School Board Journal*, 1.

Lyon, D. (2014, July - December). Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data and Society*, 1-13.

Maner, W. (1996). Unique ethiccal problems in information technology. *Science and Engineering Ethics, 2*(2), 137-154.

McHale, J. V. (2012). Using anonymized NHS data without consent: A step too far? *British Journal of Nursing, 21*(1), 54-55.

National Survey of Childhood Health Data. (2009/10). *Survey results CSHCN with EBD issues*. Retrieved November 8, 2015, from Nation Survey of Childhood Health Data: http://www.nschdata.org/browse/survey/results?q=1820&r=1&t=1

Needham, J. (2013). *Disruptive possibilities: How big data changes everything.* Boston, MA: O'Reilly Media.

Olumoye, M. Y. (2013). Ethics and social impact of information systems in ethics and social impact of information systems. *International Journal of Science and Research (IJSR), 2*(11), 154-158.

Scientific Learning. (2015, January 05). *10 trends to watch in special education in 2015*. Retrieved Novemeber 8, 2015, from Scientific Learning: http://www.scilearn.com/blog/2015-special-education-trends

Sherlock, W. K. (2013). *Analytics tools and infrastructure.* Bolton, UK: CETIS for Jisc.

Tene, O., & Polonetsky, J. (2013, April). Big data for all: Privacy and user control in the age of analytics. *Northwesten Journal of Technology and Intellectual Propery, 11*(5).

Thorpe, J. H., & Gray, E. A. (2015, March-April). Big data and public health: Navigating privacy laws to maximize potential. *Law and the Public Health, 130*, 171-175.

Weathers, J., & Supovitz, J. (2004). *Dashboard lights: Monitoring implementation of district instructional reform strategies.* Philadephia: University of Pennsylvania Consortium for Policy Research in Education.

Weiner, N. (1950). *The human use of human beings: Cybernetics and society.* Boston, MA: Houghton Mifflin.

Wessell, M. (2016, January 27). *How big data is changing disruptive innovation*. Retrieved May 6, 2016, from Harvard Business Review: https://hbr.org/2016/01/how-big-data-is-changing-disruptive-innovation

West, D. M. (2012). *Big data for education: Data mining, data analytics, and web dashboards.* Washington, DC: The Brookings Institution.

Whitehurst, G. J. (2002). Evidenece-based education. *Student Achievement and Accountability Conference.*

Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. *Issues in Information Systems, 16*(II), 81-90.

Zwitter, A. (2014). Big data ethics. *Big Data and Society* (July - December), 1-6.