2015 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
ARTS, HUMANITIES, SOCIAL SCIENCES & EDUCATION
JANUARY 03 - 06, 2015
ALA MOANA HOTEL, HONOLULU, HAWAII

CHALLENGES & OPPORTUNITIES OF BIG DATA FOR THE DIGITAL SOCIETY

PARK, JOOYEUN
HANKUK UNIVERSITY OF FOREIGN STUDIES
SOUTH KOREA

Prof. Dr. Jooyeun Park
Hankuk University of Foreign Studies
South Korea.

Challenges & Opportunities of Big Data for the Digital Society

Big data is at the core of today's information and digital society, representing a key engine for social, political, and economic relations. The potential benefits of big data are numerous. At the same time, the use of big data raises myriad issues, such as those related to privacy, data security, and ethics. In this paper will be discussed benefits, questions and concerns surround the big data phenomenon for the digital society.

Challenges & Opportunities of Big Data for the Digital Society

Jooyeun Park

Associate Professor

Hankuk University of Foreign Studies

Seoul (Korea)

1. Introduction

Big data is at the core of today's information and digital society, representing a key engine for social, political, and economic relations. The quantity of data is exploding worldwide, and the ability to analyze large datasets, Big Data is a central factor for competitiveness that is underpinning new waves of productivity, growth, and innovation (Kitchin, 2014). Advancements in Big Data analysis offer cost-effective opportunities for improvements in critical decision-making development areas such as health care, employment, economic productivity, crime, security, natural disasters, and resource management (Tinati et al., 2014). Big data technology is drastically revolutionizing commerce and society. The unlimited potential of a data-driven economy is widely recognized, and there is increasing enthusiasm for the notion of Big Data.

Although Big Data technology has the potential to provide powerful competitive advantages, governments and companies are struggling to establish effective governance and privacy in

connection with Big Data initiatives. While the potential of Big Data technology is real, the realization is lagging (Eynon, 2013). For example, fundamental concerns exist concerning Big Data development despite high expectations and exorbitant financial investment. As Boyd and Crawford (2012) critically note, contemporary discussions concerning Big Data have been technologically biased and industry-oriented, leaning toward the technical aspects of its design. Until now, most Big Data development efforts have focused on the commercialization of data technologies and resources. Existing concerns with Big Data such as the invasion of privacy, imperfect security, and limited interoperability are rarely examined compared to other technology concerns. Such issues, including the social, cultural, and ethical impacts of how we develop and manage the evolution of Big Data will be critical to its success (Shin, 2014).

The potential uses for Big Data analytics raises crucial questions about whether our legal, ethical, and social norms are sufficient to protect privacy and other values in a Big Data world. What is the status of so-called 'public' data on social media sites? What data are collected and what are not, and why? What critical or fundamental factors must be considered for a true understanding of a particular phenomenon? Asking such questions points to the necessity of engaging a more critical approach to Big Data in terms of understanding both what it is and how it should be used. Benefits, questions and concerns surround the Big Data phenomenon, and in this paper, both sides of the coin will be discussed. The purpose of this paper is to examine the implications of Big Data, with an eye to both the challenges and opportunities presented by the phenomenon.

2. Definition of Big Data

What is Big Data? There is no one definition of Big Data. Big data involves datasets that are far larger than those traditionally examined in science. Yet there has always been considerable variation in the size of datasets, ranging from small experimental studies to large samples involving census or polling data. Size alone is therefore an insufficient descriptor. Big data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets. Kitchin(2014) details that Big Data is huge in volume, consisting of terabytes or petabytes of data; Big Data is high in velocity, being created in or near real-time; Big Data is diverse in variety, being structured and unstructured in nature. Big data relates to data creation, storage, retrieval and analysis that is monumental in terms of volume, velocity, and variety. These three *V*s are often used to characterize different aspects of Big Data (ITU, 2013).

Volume may be the most compelling attraction of Big Data analytics. It means how fast data is being produced and changed and the speed at which data is transformed into insight. The examples have demonstrated that volume can present an immediate challenge to conventional resources, and that volume calls for scalable storage and capacity for distributed processing and querying. The amount of data generated or data intensity that must be ingested, analyzed and managed to make decisions based on complete data analysis. The speed of decision making – the time taken from data input to decision output – is a critical factor in the Big Data discussion. Emerging technologies are capable of processing vast volumes of data in real or near real time, increasing the flexibility with which organizations can respond to changes in the market, shifting customer preferences or evidence of fraud. Big data systems also need to be capable of handling and linking data flows entering at different frequencies.

Big data includes any type and structure of data for example, text, sensor data, call records, maps, audio, image, video, click streams, log files and more. Source data can be diverse, and it may require time and effort to shape it into a form fit for processing and analysis.

Table 1 Three Big Data characteristics

Characteristic	Description	Attributes	Drivers
Volume	The amount of data generated or data intensity that must be ingested analyzed and managed to make decisions based on complete data analysis.	- Exabyte, zettabyte, yottabyte, etc.	- Increase in data sources - Higher resolution sensors - Scalable infrastructure
Velocity	How fast data is being produced and changed and the speed at which data is transformed into insight	- Batch - Near real-time - Real time - Streams -rapid feedback loop	- Improved throughput connectivity - Competitive advantage - Precomputed information
Variety	The degree of diversity of data from sources both inside and outside an organization	- Degree of structure - Complexity	- Mobile - Social media - Video - Genomics - M2M/IoT

Sources: ITU-T Technology watch report (2013)

3. Benefits of Big Data

In a digitized world, when we are going out our day, communicating, browsing, buying, sharing, searching, we create our own enormous trails of data. Each time we use our smart phones, social media or search engine, digital traces of our activities are being stored. Emerging online apps will not only enable users to upload videos via mobile social networking, but will also incorporate wearable devices in the form of a digital watch or glasses to allow for continuous audiovisual capture. Social media platform with more than millions of users worldwide has been called the world's largest focus group, providing a platform for unfiltered expression. Digital devices that record our movements and communications, and digital sensors that record the behavior of inanimate objects and systems have become widespread and are proliferating wildly. This upsurge in data will greatly accelerate as we embark on the "Internet of Things," when millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters and automobiles that sense, create, and communicate data.

There has been a dramatic increase in the amount of data. Technological advances have driven down the cost of creating, capturing, managing, and storing information. According to the report of McKinsey (2011) and Kitchin (2014) more than 500 million photos are uploaded and shared every day. YouTube has 72 hours of video uploaded every minute. Facebook is processing about 2.5 billion pieces of content, 2.*7 billion* "like" actions, overall, terabytes of data *every day*. Volumes of data that were once unthinkably expensive to preserve are now easy and affordable to store on a chip the size of a grain of rice. It's no surprise that data sets are expanding exponentially.

Today we already benefit from Amazon, e-Bay, Netflix, and many other online merchants' use of Big Data to generate customized user recommendations. Big data is used to aggregate millions of GPS signals to predict commute times, to identify potential causes of disease, and to detect and prevent credit card fraud. Scientists are using massive data sets and powerful analytic tools to make progress on many of the most difficult problems in the health sciences and hard sciences. This change in tools and data sources has great potential to make our lives better.

1) Netflix

Netflix is the largest video streaming service player with a global streaming subscriber based of around 33 million. The firm's most recent success is a TV series, House of Cards. House of Cards is currently the most streamed piece of content in 41 countries and it has been viewed by millions of Netflix subscribers. Netflix used Big Data analytics technology to determine that House of Cards would be successful well. Netflix knew that the British version of House of Cards was well received to their users and that those who watched it, also tended to watch Kevin Spacey films or films directed by David Fincher. Big data indicated that this TV series project combining all these three factors would almost certainly appeal to viewers.

House of Cards was the first time any company had ever used such data in the creative production process for a TV show. According to Netflix, Netflix has sufficient raw data from their subscribers including complete details of the viewing patterns when they hit the pause button and whether they switch off. Netflix has track lists of more than 25 million users, about 30 million plays per day (tracked every time users rewind, fast forward and pause a movie), about 4 million ratings, about 3 million searches per day, geo-location, time of day and week, where users are watching (zip code), what device they use to watch, when they

leave content (and if they ever come back) as well as browsing and scrolling behavior. As the success of House of Cards has demonstrated, Big Data analytics can yield major dividends.

2) Seoul night line bus

Seoul is seeking to satisfy growing demand for public transport. Previously, night bus routes were designed by reference to daytime bus timetables, but did not reflect population movements by night. The City of Seoul and telecommunication company KT have worked together to enhance the quality of public services using telecommunication company's big data and the city's public data (ITU, 2013). KT analyzed the movement of citizens around the city at night based on localized call data, and found the specific areas most frequented at night. In terms of volume, over 300 million Call Detail Records (CDR) data were analyzed for this project, combined with a variety of Seoul's public data. These results were then related to a heat map of the floating population, grouped by zones. This analysis established the optimal location of night bus stops that satisfy the most number of citizens. Based on the results, bus routes were changed to include popular new stops, avoid stops little used at night (ITU, 2013).

4. Big data analysis in social sciences

The technologies of collection and analysis that fuel Big Data are being used in every sector of society and the economy. The end result is a massive increase in the amount of intimate information compiled about individuals. The financial, commercial and public sectors are already extracting value from the data. The issues in the social sciences are more complex

because they involve human behavior and people are slow to change.

Much of the current Big Data work in the social sciences, including communication, is still at first stage. But Big Data methods and sources will become increasingly important because they offer data and insights that could not be obtained in other ways. The methods of Big Data opens researcher to work with involving datasets of previously unimagined size (Parks, 2014).

There is little doubt that the development of Big Data and new data analytics offers the possibility of reframing the epistemology of science, social science and humanities, and such a reframing is already actively taking place across disciplines (Kitchin, 2013). As Kitchin (2013) and Ruppert (2013) argue, Big Data presents a number of opportunities for social scientists and humanities scholars, not least of which are massive quantities of very rich social, cultural, economic, political and historical data. Big Data and new data analytics enable new approaches to data generation and analyses to be implemented that make it possible to ask and answer questions in new ways. The computational tools of Big Data enhance researchers' ability to bring together multiple datasets—datasets of different times, from different places, or gathered at different times. This ability has always existed on a small scale, but new data management and analytic capabilities make it possible to conduct research of unprecedented complexity and scope (Parks, 2014).

Big data poses a number of challenges, including a skills deficit for analyzing and making sense of data. Big Data provides the problem of handling and analyzing enormous, dynamic, and varied datasets.

Twitter provides an example in the context of a statistical analysis. Because it is easy to

obtain – or scrape – Twitter data, scholars have used Twitter to examine a wide variety of patterns (e.g. mood rhythms (Golder & Macy 2011), media event engagement (Shamma et al. 2010), political uprisings (Lotan et al. 2011), and conversational interactions (Wu et al. 2011)). While many scholars are conscientious about discussing the limitations of Twitter data in their publications, the public discourse around such research tends to focus on the raw number of tweets available. Even news coverage of scholarship tends to focus on how many millions of 'people' were studied (Wang 2011).

In light of the predictive power of Big Data analytics, Big Data analysis often reflects deeply sensitive information about individuals. For example, a recent paper shows that merely using Facebook "likes" is sufficient to model and accurately predict a striking number of personal attributes including "sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender" (Kosinski, Stillwell, and Graepel, 2013).

In 2006, a Harvard-based research group started gathering the profiles of 1,700 college-based Facebook users to study how their interests and friendships changed over time (Lewis et al. 2008). These supposedly anonymous data were released to the world, allowing other researchers to explore and analyze them. What other researchers quickly discovered was that it was possible to deanonymize parts of the data set: compromising the privacy of students, none of whom were aware their data were being collected (Zimmer 2008). The case made headlines and raised difficult issues for scholars.

As mentioned by Boyd & Crawford (2012), there are significant questions of control and power in Big Data studies: researchers have the tools and the access, while social media users as a whole do not. Their data were created in highly context-sensitive spaces, and it is entirely

possible that some users would not give permission for their data to be used elsewhere. Many are not aware of the multiplicity of agents and algorithms currently gathering and storing their data for future use. Big Data researchers should acknowledge that there is a considerable difference between being in public (i.e. sitting in a park) and being public (i.e. actively courting attention) (Boyd & Crawford, 2012).

The process of evaluating the research ethics cannot be ignored simply because the data are seemingly public. The difficulty and expense of gaining access to Big Data produce a restricted culture of research findings. The chilling effects on the kinds of research questions that can be asked – in public or private – are something we all need to consider when assessing the future of Big Data (Boyd & Crawford, 2012).

5. Conclusion

We live in a world where data collection will be increasingly ubiquitous, multidimensional, and permanent. The resulting explosion of data is changing our world. Data is more deeply woven into the fabric of our lives than ever before. We aspire to use data to solve problems, improve well being, and generate economic prosperity. It is enabling important discoveries and innovations in health care, medicine, energy use, agriculture, and a host of other areas. While large, high-dimensional data sets have long been fundamental to research in such physical and life sciences, the use of Big Data beyond these disciplinary such as social science bounds has been much more limited.

Mobile web services including social media are producing an amount of data that are rich in detail concerning human and societal behavior and related contextual factors and dynamics,

including the attitudes, preferences, and sentiment of different individuals.

Amazon uses customer data to give us recommendations based on our previous purchases. Google uses our search data and other information it collects to sell ads and to fuel a host of other services and products. McKinsey (2011), the global management consulting firm, recently cautioned its business clients that privacy has become the "third rail in the public discussion of Big Data," noting the media attention paid to those who disregard consumer interests in collecting and using consumer information. According to the reports of Federal Trade Commission (2013) in the U.S., there are a series of cases against individual apps that have engaged in deceptive privacy practices. These include cases against a popular flashlight app and a social networking app that are sharing their location data with advertising networks.

Whether born analog or digital, data is being reused and combined with other data in ways never before thought possible, including for uses that go beyond the intent motivating initial collection. As a consequence, data, once created, is in many cases effectively permanent. Furthermore, digital data often concerns multiple people, making personal control impractical. The spread of these new technologies are changing the relationship between a person and the data about him or her.

Big data stands in high contrast to data avoidance and data minimization, two basic principles of data protection. Big data facilitates the tracking of people's movements, behaviors and preferences and, in turn, helps to predict an individual's behavior with unprecedented accuracy, often without the individual's consent. Large sets of mobile call records, even when anonymized and stripped of all personal information, can be used to create highly unique fingerprints of users, which in combination with other data such as geo-located tweets or "check-ins" may help to reveal the individual (ITU, 2013).

Communications metadata can be useful for telecom network management and billing, but, simply put, exploiting communications metadata on people is a form of surveillance. It not only reveals fine-grained details about people, but it also exposes the relationship between interacting entities. As the amount of personal data and global digital information grows, so does the number of actors accessing and using this information. Assurances must be given that personal data will be used appropriately, in the context of the intended uses and abiding by the relevant laws (ITU, 2013).

Another related concern of Big Data is security. A range of technical solutions (e.g. encryption, VPNs, firewalls, threat monitoring and auditing) can help in managing data privacy and mitigating security risks. Threats and risks need to be reassessed in view of Big Data, adapting technical solutions in response. The time is ripe to review information security policies, privacy guidelines, and data protection acts (ITU, 2013).

In the years to come, Big Data will increasingly become "part of the solution to pressing global problems like addressing climate change, eradicating disease and fostering good governance and economic development. Big Data methods and sources will become more and more important because they offer data and insights that could not be obtained in other ways.

Most sources of Big Data are related to Privacy of individuals. The wide variety of potential uses for Big Data about people raises some crucial questions about whether our legal, ethical, and social norms are sufficient to protect privacy and other values in our society. Asking such questions points to the necessity of engaging a more critical approach to Big Data in terms of understanding both what it is and how it should be used. There are serious issues involved in the ethics of online data collection and analysis (Ess, 2002). There are some significant and

insightful studies currently being done that involve Big Data, but it is still necessary to ask critical questions about who gets access to what data, how data analysis is deployed, and to what ends.

References

Bollier, D. & Firestone, C. M. (2010). *The promise and peril of Big Data*. Washington, DC, USA: Aspen Institute, Communications and Society Program.

Boyd, D. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Crawford, K. & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review, 55* (1), 93-128.

Cukier, K. & Mayer-Schonberger, V. (2013). *Big data: A revolution that will transform how we live, work and think*. John Murray.

Data & society research institute (2014). *The Social, Cultural, & Ethical Dimensions of "Big Data".* March 17, 2014 - New York, http://www.datasociety.net/initiatives/2014-0317/.

Ess, C. (2002) 'Ethical decision-making and Internet research: recommendations from the aoir ethics working committee', Association of Internet Researchers, [Online] Available at: http://aoir.org/reports/ethics.pdf (12 September 2011).

Eynon, R. (2013). The rise of Big Data. *Learning, Media and Technology, 38* (3), 1-20.

Federal Trade Commission (2014). The Power of Data. Paper presented at Georgetown University McCourt School of Public Policy and Georgetown Law Center Privacy Principles in the Era of Massive Data Washington, DC, April 22, 2014

Golder, S. & Macy, M. W. (2011) 'Diurnal and seasonal mood vary with work, sleep and day length across diverse cultures', *Science, 333, no. 6051,* 1878-1881, [Online] Available at: http://www.sciencemag.org/content/333/6051/1878.

ITU (2013). *Big data: big today, normal tomorrow.* ITU-T technology watch report, http://www.itu.int/ITU-T/techwatch.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage Publications.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & Society. Vol. 2,* 1-12.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. & Christakis, N. (2008) 'Tastes, ties, and time: a new social network dataset using Facebook.com'. *Social Networks, 30(4),* 330–342.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. & Boyd, D. (2011) 'The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions', *International Journal of Communications, vol. 5,* 1375–1405.

McCloskey, D. N. (ed.) (1985). *'From methodology to rhetoric'. The Rhetoric of Economics*. University of Wisconsin Press, Madison, pp. 20–35.

McKinsey global institute (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey & Company.

Shamma, D. A., Kennedy, L. & Churchill, E. F. (2010) 'Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events?,' Paper presented at the Computer-Supported Cooperative Work-2010, Association for Computing Machinery, February 6–10, Savannah, Georgia USA. Available at: http://research.yahoo.com/pub/3041.

Parks, M. R. (2014). Big Data in Communication Research: Its Contents and Discontents. *Journal of communication. vol. 64,* 355-360.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography*

3(3), 268–273.

Tinati, R., Halford, S., Carr, L. & Pope, C. (2014). Big data: Methodological challenges and approaches for sociological analysis. S*ociology, 48* (1), 23-39.

Wang, X. (2011) 'Twitter posts show workers worldwide are stressed out on the job', *Bloomberg Businessweek,* [Online] Available at: http://www.businessweek.com/news/2011-09-29/Twitter-posts-show-workers-worldwide-arestressed-out-on-the-job.html

Zimmer, M. (2008) 'More on the "Anonymity" of the Facebook dataset – it's Harvard College', MichaelZimmer.org Blog, [Online] Available at: http://www.michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/