# Looking into Big Data: The Case of the U. S. Federal Government

Sherry L. Xie
*School of Information Resource Management, Renmin University of China, 59 Zhongguancun, Beijing, China*

Keywords: Big Data, Data Feature, Existing Data, To-be-collected Data, Data Science, Data Analytics.

Abstract: This paper reports on a study that aimed to examine the term big data for its meaning in a particular setting. The study chose the U.S. Federal Government as its case and analysed all the big data projects and programs identified as representative of the U.S. Big Data Initiative. It constructed an analytical framework and generated findings in forms of statistic descriptions and narrative discussions. The study discovered that 1) not all the big data projects and programs possess in a collective manner the typical 3 Vs (i.e., volume, variety, and velocity), 2) variety appears to be the most valued characteristic, and 3) to-be-collected data lags largely behind existing data, indicating that technologies such as the Internet of Things are still at the stage of being developed. It also unrevealed that the U.S. Federal Government's current big data focus is heavily placed on IT and the term big data has made that focus hidden. It then suggests to sufficiently distinguish data and the technologies underlying the various features of data so that collaborations between the owners of data and technologies can be forged with easiness and big data benefits can be realized with efficiency and effectiveness.

## 1 INTRODUCTION

Big data has since its inception continued to receive focused attention. To use Web of Science Core Collection (WoS) as reference, more than seven thousand hits returned for the query of "big data" in Title, with the earliest one in 2004 as a conference paper and dramatic increases starting in 2013. Serving also as a strong indicator is the publishing of big data special issues in diverse fields such as Big Data by Nature (2008), Big Data by Significance (2012), Big Data in Political Science by Political Analysis (2013), Big Data and Organization Design, Journal of organization design (2014), Journalism in an Era of Big Data by Digital Journalism (2015), Big Social Data by Journal of Information Science (2015), The Value of Big Data in Agriculture by Journal of Business & Economic Statistics (2016) and Big Data in Psychology, Psychological Methods (2016).

Despite the rapidly increased volume of literature, the meaning of big data remains clouded: definitions abound, yet they do not necessarily agree with each other (Floridi, 2012; TechAmerica Foundation, 2012; Mayer-Schönberger, 2013; Ekbia, et al., 2014; Kitchin, 2014; Baro et al., 2015; NIST, 2016; Pirc, et al., 2016; Todman, 2016). Moreover, it is not just the wordings of the definitions that vary, so do the characteristics, i.e., the Vs. For example, Gartner (2016) proposes 3Vs (i.e., volume, variety, and velocity), IBM (2016) adds to the 3Vs with veracity, SAS (2016) with variability and Oracle (2016) with value. In addition, research dedicated to studying the meanings of big data appears to be sparse and the focuses of the handful articles (both journal and conference papers included) are on "big data analysis/processing" (e.g., Chebbi et al., 2015; Gandomi and Haider, 2015; Pashayev and Sabziev, 2016; Bendre and Thool, 2016) or the application of the concept to a specific area (e.g., Wielki, 2013; Elarabi et al. 2016; Miloslavskaya and Tolstoy, 2016; Drosio and Stanek, 2016). This situation continues to date, after the analysis of the present research was completed, as a search for dedicated research on big data concept after 2016 has showed (e.g., Venkatram and Geetha, 2017; Liu, et al., 2017; Gephart, et al., 2018).

In this reality, how to quickly and adequately gain understanding of big data has become a challenge for professionals whose fields are not in direct contact with the emergence of big data (e.g., computer science, data management, statistics etc.), yet, at the same time, are being constantly reminded that big data will ultimately impact their fields (e.g., Wong et



Xie, S.
Looking into Big Data: The Case of the U. S. Federal Government.
DOI: 10.5220/0006919001440152
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 3: KMIS*, pages 144-152
ISBN: 978-989-758-330-8

al, 2016 and the above listing of big data special issues). To examine existing big data definitions or to synthesize the varied numbers of Vs may be one way of gaining understanding (e.g. Baro et al, 2015; De Mauro et al, 2016), to examine big data projects in operation may be the other. This paper reports on a study that took the later method, which analysed operating big data projects and programs considered by the United States (U.S.) Federal Government as the highlights of its Big Data Initiative. This gives support to the selection of these projects and programs as a representative sample, which were coded against an analytical framework constructed by the study. The rest of this paper presents the sections of the U.S. Federal Government Big Data Context, the Analytical Framework, Analyses and Findings, Discussions, and Conclusion.

# 2 THE U.S. FEDERAL GOVERNMENT BIG DATA CONTEXT

## 2.1 Usage of the Term

The term big data (hereafter Big Data, or BD) appeared first in the U.S. Federal Government in December 2010, in a document entitled Report to the President and Congress Design a Digital Future: Federally Funded Research and Development in Networking and Information Technology. This report was produced by the President's Council of Advisors on Science and Technology (PCAST), an advisory group of leading scientists and engineers in the U.S. appointed by the White House for policy consultations on issues of science, technology, and innovation. The main purpose of this report is to present its assessment of the Federal Networking and Information Technology Research and Development (NITRD) Program, the primary source of advanced information technologies funded by the Federal Government in the contexts of the U.S. High-Performance Computing Act of 1991, the Next Generation Internet Research Act of 1998, and the America COMPETES (Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science) Act of 2007 (NITRD, 2016). For the 2010 assessment, the report focused on two aspects: the NITRD Program as 'the mechanism by which the Federal Government coordinates its unclassified research and development (R&D) investments in Networking and Information Technology (NIT)' and as 'the ensemble of unclassified R&D efforts in NIT supported by the Federal Government rather than the coordination effort' (PCAST, 2010, p. 1). Among the five major 'crosscutting themes' that the report identified, data volume was listed as the first one because of its exponential growth. A 'big data' strategy, therefore, was recommended for 'every agency' to have (PCAST, 2010, p. xvii). Since then, the term big data has appeared in many important Federal documents, first with quotation marks and all in lower letters (such as the one in the PCAST 2010 report), then with quotation marks being removed and first letters capitalized (such as the one in the 2016 Federal Big Data Research and Development Strategic Plan). Figure 2 depicts the occurrence of the term big data in some of the major documents of the U.S. Federal Government. MGI in the figure stands for McKinsey Global Institute, a research arm dedicated to business and economics established by McKinsey & Company in 1990. Although MGI is not a U.S. federal agency, its 2011 report, Big data: The Next Frontier for Innovation, Competition, and Productivity, is included due to the fact that the first author of the report, that is, James Manyika, was present at the Federal Government Big Data Rollout held on March 29, 2012 in the AAAS Auditorium in Washington, DC (NSF, 2012a). The Rollout consisted of agency announcements of Big Data projects and a panel discussion, of which Mr. Manyika was one of the panellists. As the Rollout was the immediate follow-up to the OSTP's announcement of the 2012 U.S. Federal Government Big Data Initiative, the influence of the MGI report seemingly should not be overlooked. OSTP in the figure refers to the Office of Science and Technology Policy of the Executive Office of the White House, which administers PCAST. The OSTP announcement of the Big Data Initiative and the companion Big Data Fact Sheet are the focus of the present study. The Big Data IWG refers to the Big Data Interagency Working Group, formerly known as the Big Data Senior Steering Group (BD SSG), formed in early 2011, with the tasks of identifying Big Data research activities across the Federal Government. In 2015, the NIST Big Data Public Working Group (NBD-PWG) published the first volume of its seven-volume series Big Data Interoperability Framework, which defines big data as "the inability of traditional data architectures to efficiently handle the new datasets" (NIST, 2016, p. 4). Same as the articles introduced above, this way of defining big data does not help professionals who traditionally do no possess knowledge of "traditional data architectures". The

most recent outcome of its work is the 2016 Big Data Research and Development Strategic Plan.

## 2.2 Big Data Initiative

By the recommendation of the 2010 PCAST report, the U.S. Federal Government formally announced its 'Big Data Research and Development Initiative' on the day of March 29, 2012, committing a more than 200 million USD to Big Data R&D to 'make the most of the fast-growing volume of digital data' and to 'help solve some the Nation's most pressing challenges'(OSTP, 2012a). Six Federal departments and agencies were identified in the announcement as major representatives, including the National Science Foundation (NSF), the National Institutes of Health (NIH), the Department of Defense (DOD), the Department of Energy (DOE), the Defense Advanced Research Projects Agency, and the US Geological Survey (USGS) of the Department of the Interior. The six departments and agencies then provided in the afternoon of the same day, more detailed information regarding their commitments on their respective websites. NSF (2012b), for example, listed 16 Big Data programs on its website with descriptions only for some projects and descriptions along with hotlinks for others. It needs to be pointed out that the announcing of the Big Data Initiative does not indicate the actual beginning of the 'big data action' in the Federal Government and what it did is to make the term 'big data' official. There were indeed many Big Data programs and projects already going on at the time of the announcement and the Big Data Fact Sheet clearly indicates it. Published by OSTP (2012b) on the same day of the announcement, the Fact Sheet identified in total 89 'ongoing' Big Data programs and projects across the Federal Government, including NASA's Earth Science Data and Information System (ESDIS), which started in 1997, and DOE's The Next Generation Networking program, which started in 2001.

A definition for big data was not provided by either the OSTP 2012 announcement or the PCAST 2010 report, both, however, provided explanatory information such as 'large and complex collections of digital data' and 'data volumes' in exponential growth. Definitions were found in less 'authoritative' documents such as the MGI 2011 big data report and the NSF's Big Data Solicitation. MGI considered big data as 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse) and the NSF Big Data Solicitation defines Big Data as 'large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future' (NSF, 2012c). In addition, another definition considered also relevant to the U.S. Federal Government setting was found in the report produced by the TechAmerica Foundation's Federal Big Data Commission, which read 'Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information' (TechAmerica Foundation, 2012). Like the situation outside, these definitions do not conform to each other in their wordings or intentions even though similarities do exist. As noted in the 2014 Big Data review report, Big Data definitions abound, and they vary depending on the party who defines it (The White House, 2014, p.2).

The 89 Big Data programs and projects listed in the OSPT Fact Sheet were considered 'highlights' of the Federal Government's Big Data action. They therefore were chosen to be data for analysis for the present study.

# 3 THE ANALYTIC FRAMEWORK

The analytical framework for the study consists of two parts, one addressing indicators of analysis and the other addressing the approach of coding.

Analysis indicators were identified relying on the various sources introduced above, including those in the Introduction section. The discussions in these sources collectively revealed the mostly recognized or referenced features of big data, which enabled the present study to form a set of indicators and weave them into a coherent analytical framework. Nine indicators were identified to form the framework, including Volume, Variety, Velocity, Data, Existing data, To-be-collected Data, NIT (Networking and Information Technology), Data Processing, and Data Collecting, assigned respectively to the categories of Data Presence (Data), Data Characteristics (Volume, Variety, Velocity; that is, the 3Vs), Data Type (Existing data, To-be-collected Data), NIT Presence, and NIT Function (Data Processing, Data Collecting).

The 3Vs in the present study has its study-specific meanings despite the fact that they now have gained a widespread familiarity with big data researchers and practitioners. Like big data, the 3Vs have so far not yet enjoyed the agreement of unambiguous meanings,

and consequently, different definitions and/or interpretations exist. For example, the 2010 PCAST report focused exclusively on volume, and accordingly, discussed variety and velocity as contributing factors to volume. Many others, however, treat the Vs as independent variables as in the cases of MGI and Gartner. Logically speaking, the PCAST view holds tight because it rightfully acknowledged the intertwined relationships among the 3Vs as revealed by existing narratives, and even if the 3Vs are defined with clear-cut conceptual boundaries, their connections to each other would still arise when big data is the backdrop. This kind of view, however, creates difficulties for studies that intend to understand a concept by decomposing it, like the present one. Volume in the present study, therefore, is used to refer strictly to size, and size only.

For Variety, discussions on it are observable of two types: one tends to be encompassing, referring to, for example, not only data format and/or data type but also data domain and/or data repository (NIST, 2016). The other narrows down the scope to only data format or data model and explains the concept by giving specific examples (NSF, 2012c). Common to both types is the understanding that variety implies the presence of both structured and unstructured data, which the present study adopts. The meaning of Velocity in the present study comes from the basic idea that was put forward in 2001 by now Gartner's Doug Laney, who was then with META Group. According to Laney, velocity is relevant to the 'increased point-of-interaction speed and, consequently, the pace data used to support interactions and generated by interactions'. While framed in the context of e-commerce, this idea can be easily transported into other settings including Internet-enabled governments or digital governments, where dynamic, real-time decision making is as well much envisioned and may even be more critical. Moreover, to understand Velocity as the focus on the interplay of interaction and speed aids technically the analysis as it excludes effectively the application of speed in stand-alone data actions such as collection, processing, or transmission.

Like big data, data too is a term that does not have a universally accepted definition and too is used with diversified meanings in different fields including those of the information technology, information management, and both the natural and social science. For the purpose of this study, the definition of data provided by the 2015 version of the ISO/IEC 2382 Information technology — Vocabulary was employed. Data in this context refers to 'reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing' (ISO, 2015).

Existing data refers to data that have participated, or are participating, in the operations of the departments and agencies of the Federal Government. Such data provides ready access for Big Data programs and project. To-be-collected Data, on the contrary, is considered requiring efforts additional to the activities that routinely generate Existing data in departments and agencies for its access. In other words, such data are determined to be collected specifically for a certain Big Data program or project despite the possibility that such collection may become routine in the future. Data Collecting as one of the two functions in the NIT Function category, refers exactly to the additional efforts that produce the data determined to be collected. Data processing, the other NIT function, refers mainly to the activities that yield results typically associated with Big Data programs and projects such as knowledge, insights, and/or actionable intelligence.

# 4 ANALYSES AND FINDINGS

Two rounds of coding were conducted, and discrepancies were sorted out during the comparing stage. Below are findings that emerged from the coding.

## 4.1 Overarching Tendency

Twelve departments and agencies collectively contributed to the 89 Big Data programs and projects, with varying numbers individually. These numbers were considered indicative of degree of focus, and when linked to the names of the departments or agencies, capable of revealing the overarching tendency of the Big Data Initiative.

It is rather clear that health, natural science, national security, and energy are the focused areas in the Federal Big Data movement health. It should be noted that given the fact that NITRD does not in its scope cover classified networking and information technology, the investment on national security Big Data programs and projects cannot be inferred as less than those on health and natural science. Also worth noting is the complete absence of education as an area. Education was pointed out in the OSTP 2012 announcement as one of the areas on which the Big Data Initiative would have a major impact, none of the 89 programs and projects, however, identified itself with it.

## 4.2 Observations by Indicators

### 4.2.1 Data Presence

Data is present in all of the 89 programs and projects, with 70 being explicit and 19 implicit. Examples of being explicit include the DOD Data to Decision project, the DOE High Performance Storage System (indeed, all of the 12 DOE projects), the NIH Internet Based Network for Patient-Controlled Medical Image Sharing project, and the NSF Data Citation project. Examples of being implicit include the DOD Cyber-Insider Threat project, the NIH National Centers for Biomedical Computing program, the NSF Expeditions in Computing project, and the NSA Combining Cybersecurity and Big Data program. Among the 70 explicit presences, 5 appeared stronger than the NIT presence in the same programs or projects (see also 4.2.4 NIT Presence and NIT Function) and they are the Cancer Imaging Archive project and the WorldWide Protein Data Bank project from NIH, and the Planetary Data System project, the Multimission Archive at the Space Telescope Science Institute program, and the Earth System Grid Federation program from NASA.

### 4.2.2 Data Characteristics

Variety was present in almost all programs and projects, with only one exception, the VA Informatics and Computing Infrastructure project. Among the 3Vs, Variety appeared also as the one that was mostly capable of being a stand-alone characteristic. Examples of projects and programs that had only Variety as their Big Data indicator include DOD's Machine Reading, Programming Computation on Encrypted Data, VA's Genomic Information System for Integrated Science, and HHS's Using Administrative Claims Data (Medicare) to Improve Decision-Making. For Volume, although it has a high percentage, close to 90% indeed, there are still programs and projects that did not consider it a necessity for their Big Data tasks. Velocity appeared to be unable to occur independently in any of the programs or projects as it was only discernable when Volume and Variety were both present. Moreover, its explicit presence is less than its implicit presence, making the certainty of its presence less than 20% against all the programs and projects. From a collective standpoint, the occurrence of the 3Vs all together does not account for a majority, and even the occurrence of 2Vs did not pass the 50% bar.

### 4.2.3 Data Type

Existing Data of the Data Type indicator appeared to be present in all programs and project, with 69 being explicit and 20 implicit. This 69 explicit: 20 implicit ratio is almost identical to that of Data presence, which is 70 explicit: 19 implicit. This suggests that the programs and projects that included in their descriptions a clear presence of data are those who also had a focus on Existing Data. The one program that is explicit in Data presence but implicit in Existing Data is the Office of Advanced Scientific Computing Research program at DOE, which focused primarily on technologies. The To-be-collected Data indicator has in total 36 presence, with 17 being explicit and 19 being implicit. Making use of Existing Data, therefore, clearly outweigh the collection of new data (78% vs. 19% explicitly). In other words, to collect new data was either not a priority or at most, a co-focus with Existing Data.

### 4.2.4 NIT Presence and NIT Function

NIT was present in all programs and projects, with 88 explicit and 1 implicit. The only exception is the Data and Communications in Basic Energy Sciences Workshop program administered by the Office of Basic Energy Sciences at DOE, which stated explicitly about data but nothing about NIT. The needing of NIT was thus inferred from the goal of the program, which aimed at handling scientific experimental data. When compared with Data presence, 21 among all the programs and projects (that is, 24%) had an equal focus on Data and NIT, 43 had a stronger focus on NIT (that is, 48%), and 20 had a much stronger focus on NIT (that is, 22%).

Data Processing of the NIT Function occurred alone in 53 programs and projects, and jointly occurred with Data Collecting in the rest. With 83 being explicit and 6 being implicit, Data Processing was thus at presence 100%. Not occurring as a stand-alone indicator in any of the programs or projects, Data Collecting occurred 36 times, with 31 being explicit and 5 being implicit. This finding corresponds to that on Data Type, where the type of To-be-collected Data did not occur at all as a stand-alone indicator. Among the 36 programs and projects in which both Data Processing and Data Collecting occurred, 29 treated these two equally, 6 focused more on Data Processing, and 1 focused more on Data Collecting (i.e., the VA's Million Veteran Program).

# 5 DISCUSSIONS

## 5.1 Data in Big Data

The above analyses suggest that BD can be understood in two parts, 'data' in BD and 'big' in BD.

In the analytical framework, data is represented by the categories of Data Presence (Data), Data Characteristics (Volume, Variety, Velocity), and Data Type (Existing data, To-be-collected Data). Data presence was found in all of the BD programs and projects, the commonly attributed BD 3V characteristics, however, were not. Variety appeared to be mostly visible, having an explicit presence of 64% (compared to Volume 52%, Velocity 17%), and when implicit presence is considered, Variety has a presence percentage close to 100%, higher than Volume (89%). This deviates from PCAST's original assessment of the BD situation in the Federal Government, which located data Volume as the focus. It is however in agreement with the viewpoint of Gartner. According to Gartner (Buytendijk, 2013), among the 3Vs, 'the new variety in the data' is where most of the value of the big data phenomenon lies. Gartner thinks also getting the most out of variety is more challenging than handling Volume and Velocity. The latter point, however, differs from the inference that the present study has drawn from its findings. Velocity in this study has only 17% explicit presence and this small percentage were only found in departments and agencies that had big budgets such as DOD and DOE or in research projects that were equipped with large funds awarded by NSF or NIH. Within these settings, the presence of Velocity did not appear to dominate either. For example, there were only 33% Velocity related projects and programs in DOE among all of its Big Data projects and programs, and similarly, only 30% in DOD, 17% in NIH, and 13% by NSF. These findings seem to suggest that either Velocity is not a priority compared to Variety or Volume, or it is challenging to many projects and programs in particular those who had to be budget-conscious. Nonetheless, the findings on the presence of the 3Vs were largely in consistence with Gartner's characterization of big data. As discussed in the Introduction section, Gartner's big data definition employs the expression 'and/or', indicating that any single V or any combinations of the 3Vs can be utilized to depict big data.

The difference between Existing Data and To-be-collected Data (78% vs. 19% explicitly) confirms the above finding that Velocity is not currently the dominating feature of the U.S. Big Data Initiative. While Existing Data does not constitute the entirety of Data presence, the 100% presence of Data Processing indicates sufficiently the reliance on Existing Data of the BD projects and programs. This focus on Existing Data reflects the reality that the U.S. Federal Government had accumulated, at the time of announcing its Big Data Initiative, a substantial storage of data, rich in both Volume and Variety. This reality gave rise also to the Obama Administration's Open Data initiative, which started in 2013 and has continued on since then. On May 9, 2013, an Executive Order was signed to boost 'openness in government' (The White House, 2013) and its companion government-wide policy – that is, the Open Data Policy – made it clear that the goal was to 'ensure that the Federal Government is taking full advantage of its information resources' (OMB, 2013). Building on the 2013 Open Data Policy, the OMB (Office of Management and Budget) issued on February 14 of 2014 its memo on administrative data, providing guidance to departments and agencies on leveraging 'existing data' for both the efficiency of their programmatic work and the benefits of the American public (OMB, 2014).

The much smaller presence of To-be-collected Data, as well as Velocity, seems to be attributable to the slow taking up of cloud computing and the Internet of Things, two enablers for real-time data collecting, processing, and presenting. The U.S. Government's Federal Cloud Computing Strategy, also known as the 'Cloud First Policy' (Kundra, 2011) , was issued in 2011, however, due to strong concerns with security (Corbin, 2015; Kapko, 2015), it is not until 2016 when the U.S. Federal Government as a whole 'finally loves cloud computing' (Darrow, 2016). The story with the Internet of Things is different because it arrived later than cloud computing and was found out with different causes than security for its slow adoption. The Federal Government was in 2016 still 'examining opportunities and challenges' of the Internet of Things (Bruce, Correa & Subramanyam, 2016), and a list of challenges were identified as causes, including 'a lack of leadership, skills, and funding, as well as cumbersome procurement policies and a risk-averse culture' (Castro, et al, 2016). A greater presence of Velocity, therefore, will have to come in the future. For the time being, data in BD means much less on the combined Vs, but more on 1) accumulated information, 2) from diverse sources and in a variety of formats, and 3) with the potential to be used as well as to be continuously accumulated for the same and/or a different use.

### 5.2 Big in Big Data

Unlike data, 'big' in BD was not identified as in association with any specific categories in the analytical framework. If 'data' can be considered as having a basic meaning in existing literature, yet the situation is different with 'big'. Its meaning, therefore, was expected to emerge in the analysis process of the study, which would also take into consideration any accompanying findings that could shed light on it. As it emerged, 'big' in BD appeared to be all about NIT, which was explicitly present in 88 of the 89 BD projects and programs, and 19 of them simply omitted mentioning data. Although the study categorized NIT Function into only two types, Data Processing and Data Collecting, the technological abilities connected with these BD projects and programs are evident in producing data, storing data, transmitting data, processing data, presenting data, and producing more data – the full circle from initiating a BD project to delivering outcomes. This finding largely corresponds to the development history of the term Big Data, which started as a 'problem' for computing capability because of the huge size of datasets (Cox and Ellsworth, 1997), then 'the result of recent and unprecedented advancements in data recording and storage technology' (Diebold, 2003), then 'the ability to gather information' (NIST, 2016), and then the 'information assets that demand … processing for enhanced insight and decision making' (Gartner, 2016).

This finding may appear insignificant because today's NIT, or information technology in general, underlies almost every type of human endeavour. It is this study's contention that the understanding of this underlying feature is not clear or intuitive to all the fields that hope to understand big data: for many, "big" describes the V-characterized features of data, not the whole set of digital technologies and devices. The implications may be instructive to organizations who are contemplating on initiating big data projects or academics who wish to join big data research. For data users, especially the general ones, the understanding of the Vs of big data is not important – that of the technologies needed for tackling their data analysis needs is.

## 6 CONCLUSIONS

Relying on the representative BD projects and programs identified by the U.S. Federal Government, the present study examined the meaning of the term big data in this particular setting. Although the examination relied primarily on descriptions of projects and programs, not their internal operations, the key features of these projects and programs were sufficiently clear for the intended analysis. The term big data started out in the U.S. Federal Government BD Initiative without a formal definition，but a focus on data volume. The subsequent NIST's effort to define big data struggled with bringing sufficient clarity to its conceptual construction (NIST, 2016). As the present study has demonstrated, the term is a simple combination of 'data' and 'big', with 'data' mostly referring to existing data for the time being and 'big' to digital technologies for now and for the future. While claimed to be a "movement", "transformation" or "revolution", big data is indeed nothing more than a newer – or the most recent – phase of digital data, exemplified by unstructured data, which is a long existing concept. The Vs, no matter how many, are unable to specialize or make unique enough the term as data can always be described by size, formats and/or forms, all existing concepts. It may be argued that, as a term, big data is convenient for usages in popular writings and news reporting, it should not be ignored, however, to use the term as a new, specialized one may create confusions for the professions that have legitimate interest in data. The conceptual identification of terms such as data, information, content, records, knowledge, intelligence, etc. do not always appear to be clear-cut and they typically vary in distinct contexts. For those professions, data and the technologies underlying the various kinds of data features are better to be distinguished. If data science is used to describe the whole treatment of data, including both the aspects of management and technology, and data analytics is used for the technological aspect, then, they would appear much clearer as specialized domain knowledge to those who owns certain type of data but may not own the whole set of data analytical technologies. As such, collaborations may be easier to be forged and the promising benefits of big data may be realized with both efficiency and effectiveness.

## ACKNOWLEDGEMENT

This study is supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNL032).

# REFERENCES


Baro, E. et al. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed Research International*, 2015, 1-9, doi: 10.1155/2015/63902.

Bruce, A, Correa, D, and Subramanyam, S. (2016*). Internet of things: examining opportunities and challenges*. https://obamawhitehouse.archives.gov/blog/2016/08/30/internet-things-examining-opportunities-and-challenges (accessed 5 July 2016).

Buytendijk, F. (2013). *The Art of big data innovation*. http://im.ft-static.com/content/images/e91a32d0-2bac-11e3-bfe2-00144feab7de.pdf (accessed 4 July 2016).

Castro, D., New, J., and McQuinn, A. (2016). *How is the federal government using the Internet of things?* http://www2.datainnovation.org/2016-federal-iot.pdf (accessed 6 July 2016).

Chebbi, I., Boulila, W. and Farah, I. R. (2015). Big data: concepts, challenges and applications. In Nunez, M. et al. (Eds.). *ICCCI 2015, Part II, LNCS 9330.* Switzerland: Springer International Publishing, pp. 638–647.

Corbin, K. (2015). *Government cloud adoption efforts lag as security concerns persist*. http://www.cio.com/article/2988495/cloud-computing/government-cloud-adoption-efforts-lag-as-security-concerns-persist.html (accessed 4 July 2016).

Cox, M. and Ellsworth. D. (1997). *Application-controlled demand paging for out-of-core visualization*. ww.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf (accessed 6 July 2016).

Darrow, B. (2016). *Why the U.S. government finally loves cloud computing*. http://fortune.com/2016/09/02/us-government-embraces-cloud/ (accessed 6 July 2016).

De Mauro, A., Greco, M., and Grimaldi, M. (2016). A Formal definition of big data based on its essential features. *Library Review*, 65 (3), 122-135.

Diebold, F.X. (2003). Big data dynamic factor models for macroeconomic measurement and forecasting. In M. Dewatripont, L.P. Hansen, and S. Turnovsky, (Eds). *Advances in economics and econometrics: theory and applications, Eighth World Congress of the Econometric Society*. Edinburgh: Cambridge University Press, pp. 115-122.

Drosio, S., and Stanek, S. (2016). The Big data concept as a contributor of added value to crisis decision support systems. *Journal of Decision Systems*, 25 (Sup1), 228-239.

Ekbia, H. et al. (2014). Big data, bigger dilemmas: a critical review. *Journal of the Association for Information Science and Technology*, 66 (8), 1523–1545.

Elarabi, T. et al. (2016). Big data analystics concepts and management techniques. In *Big data: concept, applications, & challenges, 2016 international conference on information management and technology (ICIMTech), IEEE, Coimbatore, India*, pp. 307-310.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy and Technology*, 25 (4), 435–437.

Gandomi, A., and Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management,* 35 (2), 137-144.

Gartner (2016). *IT Glossary*. https://www.gartner.com/it-glossary/big-data (accessed 22 January 2017).

Gephart, S. M., Davis, M., and Shea, K. (2018). Perspectives on policy and the value of nursing science in a big data era. *Nursing Science Quarterly*, 31 (1), 78-81.

ISO (2015). *ISO/IEC 2382:2015 Information technology: vocabulary*. https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en (accessed 5 July 2015).

Kapko, M. (2015). *U.S. CIO tells IT leaders to trust the cloud.* http://www.cio.com/article/2996268/cloud-computing/us-cio-tells-it-leaders-to-trust-the-cloud.html (accessed 6 July 2016).

Kitchin, R. (2014). *The Data revolution: big data, open data, data infrastructures & their consequences*. Thousand Oaks, California: SAGE Publications Ltd., 67-79.

Kundra, V. (2011). *Federal cloud computing strategy*. https://www.dhs.gov/sites/default/files/publications/digital-strategy/federal-cloud-computing-strategy.pdf (accessed 23 November 2017).

Liu, C. H., Wang, J. S., and Lin, C. W. (2017). The Concepts of big data applied in personal knowledge management. *Journal of Knowledge Management*, 21 (1), 213-230.

Mayer-Schönberger, V., and Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Miloslavskaya, N., and Tolstoy, A. (2016). Application of big data, fast data and data lake concepts to information security issues. In *2016 4th International conference on future Internet of things and cloud workshops, IEEE Computer Society, Vienna, Austria*, pp. 148-153.

NIST (2016). *NIST big data interoperability framework: volume 1, definitions*. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf (accessed 4 July 2016).

NSF (2012a). *Broadcast of OSTP-led federal government big data rollout*. www.nsf.gov/news/news_videos.jsp?cntn_id=123607&media_id=72174&org=NSF (accessed 3 July 2016).

NSF (2012b). *NSF leads federal efforts in big data*. www.nsf.gov/news/news_summ.jsp?cntn_id=123607 (accessed 3 July 2016).

NSF (2012c). *Solicitation 12-499: core techniques and technologies for advancing big data science & engineering.* www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf (accessed 5 July 2016).

NITRD (2016). *Federal big data research and development strategic plan*. www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf (accessed 22 January 2017).

OMB (2013). *Open data policy.* https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf (accessed 4 July 2016).

OMB (2014). *Guidance for providing and using administrative data for statistical purposes*.

www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf (accessed 5 July 2016).

Pashayev, A. B., and Sabziev, E. N. (2016). On the concept of big data analysis. In L.A. Zadeh et al. (eds.). *Recent developments and new direction in soft-computing foundations and applications, studies in fuzziness and soft computing.* Switzerland: Springer International Publishing, pp. 269-277.

Pirc, J. et al. (2016). *Threat forecasting: leveraging big data for predictive analysis*. Cambridge, MA: Syngress.

PCAST (2010). *Report to the president and congress designing a digital future: federally funded research and development in networking and information technology*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf (accessed 15 January 2016).

TechAmerica Foundation (2012). *Demystifying big data: a practical guide to transforming the business of government.* http://breakinggov.sites.breakingmedia.com/wp-content/uploads/sites/4/2012/10/TechAmericaBigDataReport.pdf (accessed 18 June 2016).

Todman, A. (2016). Privacy in the age of big data: recognizing threats, defending your rights and protecting your family. *Archives and Records*, 37 (1), 113-115.

Venkatram, K., and Geetha, M. A. (2017). Review on big data & analytics – concepts, philosophy, process and applications. *Cybernetics and Information Technologies,* 17 (2), 3-27.

OSTP (2012a). *Obama administration unveils 'big data' initiative: announces $200 million in new R&D investments.* www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf (accessed 2 July 2016).

OSTP (2012b). *Fact sheet: big data across the federal government.* www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf (accessed 3 July 2016).

White House (2013). *Executive order -- making open and machine readable the new default for government information.*www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government- (accessed 3 July 2016).

Wielki, J. (2013). Implementation of the big data concept in organizations – possibilities, impediments and challenges. In *Proceedings of the 2013 federated conference on computer science and information systems, Polish Information Processing Society, Krakow, Poland*, pp. 985–989.

Wong, H. T. et al. (2016). The Need for a definition of big data for nursing science: a case study of disaster preparedness. *International Journal of Environment of Research and Public Health*, 13 (1015), 1-13.

Xie, S. L. (2011)/ Building foundations for digital records forensics: a comparative study of the concept of reproduction in digital records management and digital forensics. *The American Archivist*, 74 (2), 576-599.