A Formal Definition of Big Data Based on its Essential Features

Andrea De Mauro[1, a)], Marco Greco[2, b)] and Michele Grimaldi[2, c)]

[1]*Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Roma, Italy*
[2]*Department of Civil and Mechanical Engineering, University of Cassino and Southern Lazio, Via Di Biasio 43, 03043 Cassino (FR), Italy*

[a)] Corresponding author: andrea.de.mauro@uniroma2.it
[b)] m.greco@unicas.it
[c)] m.grimaldi@unicas.it

Structured Abstract

Purpose – This article identifies and describes the most prominent research areas connected with 'Big Data' and proposes a thorough definition of the term.
Design/Methodology/Approach – We have analyzed a conspicuous corpus of industry and academia articles linked with Big Data in order to find commonalities among the topics they treated. We have also compiled a survey of existing definitions with a view of generating a more solid one that encompasses most of the work happening in the field.
Findings – We've found that the main themes of Big Data are: Information, Technology, Methods and Impact. We propose a new definition for the term that reads as follows: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."
Practical implications – The formal definition we propose can enable a more coherent development of the concept of Big Data, as it solely relies on the essential strands of current state-of-art and is coherent with the most popular definitions currently used.
Originality/value – This is among the first structured attempts of building a convincing definition of Big Data. It also contains an original exploration of the topic in connection with library management.

Keywords: Big Data; Analytics; Information Management; Data Processing; Business Intelligence.

Article Classification: Literature Review

Introduction

At the time of writing, the term 'Big Data' is nearly omnipresent within articles and reports issued by Information Technology practitioners and researchers. The pervasive nature of digital technologies and the broad range of data-reliant applications have made this expression widespread also across other disciplines including sociology, medicine, biology, economics, management and information science. However, the degree of popularity of this phenomenon has not been accompanied by a rational development of an accepted vocabulary. In fact, the term 'Big Data' itself has been used with several and inconsistent acceptations and lacks a formal definition.

With this article we want to drive clarity upon the concept of Big Data so that the main connected "themes" are identified and a formal and widely acceptable definition is proposed. As we do so, we also explore the connections within this topic and Library Science. In order to pursue the first objective we analyse the most significant occurrences of the term 'Big Data" in both academic and business literature, intercepting the most debated topics. Then we propose a thorough definition of Big Data by discussing existing proposals, identifying their key features and synthesizing an expression that formally represents the essence of the phenomenon. We believe that doing so enables a more structured development of literature in this field.

Theoretical background: the Four Themes of Big Data

The aim of this section is to describe the essential features of Big Data by means of a broad yet non-comprehensive review of related literature. We have analyzed the abstracts of a large set of scientific papers and assessed the nature of the most recurring words, looking for usage patterns and making reasonable assumptions on their mutual interrelation. By doing so we recognized a subset of key topics that can be used to depict the multifaceted nature of the phenomenon under examination. A thorough, systematic literature review goes beyond the scope of this work and is left for future work.

We decided to use Elsevier's Scopus, a major reference database holding more than 50 million literature entries from around 5,000 publishers. In the month of May 2014 we exported a list of 1,581 conference papers and journal articles that contained the full term 'Big Data' in the title or within the list of keywords provided by the authors. We have cleaned up this initial list by removing those entries for which the full abstract was not available, leaving us with a corpus of 1,437 records. By counting the number of times each word was appearing in those abstracts we have identified the most recurring ones. Figure 1 contains the "word cloud" of the most prevalent words in the corpus we considered in our analysis: in this visualization, the font size of each term is directly proportional to its relative presence within the input text. Afterwards, we have collectively reviewed the list of the most frequent words included in the Big Data-related literature and made assumptions on their mutual interconnection. We have then grouped together the words that were more clearly connected with each other, from a conceptual viewpoint. By doing so we were able to recognize the existence of four fundamental "Themes", i.e. prevalent concepts that represent the essential components of the subject. We have described the four themes with the following titles: 1. Information, 2. Technology, 3. Methods, 4. Impact. As a last step, we have verified that the majority of papers on Big Data included in the list we retrieved deal with one or more of the themes we identified. In the following paragraphs we will proceed with a description of each of the four themes and report a list of the most representative related works.

Fig. 1 Static tag cloud visualization (or "word cloud") of key words appearing in the abstracts of papers related to Big Data, created through the online tool ManyEyes (Viegas et al., 2007), as appeared in (De Mauro et al., 2015).

The fuel of Big Data: Information

The first reason behind the quick expansion of Big Data is the extensive degree at which data is created, shared and utilized in recent times. Digitization, i.e. the transformation of analog signals into digital ones, had reached massive popularity in the early 1990s. At that time the first commercial Optical Character Recognition (OCR) tools were launched and paved the way for the kickoff of the first "mass digitization" projects, i.e. the conversions of entire traditional libraries into machine-readable files, (Coyle, 2006). A noteworthy example of mass digitization was the Google Books Library Project (2015), begun in 2004, whose objective was to fully digitize more than 15 million printed books held in several college libraries, including Harvard, Stanford and Oxford.

Once signals have been converted into a digital format they could be organized into more structured datasets. This further step, that Mayer-Schönberger and Cukier called "datafication" (2013), is capable of offering a unique, macro-level view-point for studying relevant trends and patterns that would have been impossible if all the data stayed in an analog format. In case of the

aforementioned Google mass digitization project, datafication started when the massive amount of textual strings was converted into sequences of contiguous words (n-grams) for which it was possible to track the level of occurrence over the centuries. In this way, researchers were able to find insights on disparate fields, such as linguistics, etymology, sociology and historical epidemiology by utilizing Google Books' datasets (Michel et al., 2011).

The Data-Information-Knowledge-Wisdom hierarchy offers an alternative view according to which information appears as data that is structured in a way to be useful and relevant to a specific purpose (Rowley, 2007). The identification of such purpose is a common element across all Big Data applications we have reviewed. In this perspective, information becomes a knowledge asset that can create value for firms (Cricelli & Grimaldi, 2008). Hence, we can conclude that information, not data, is the fundamental fuel of the current Big Data phenomenon.

According to Prescott (2013) *"library catalogues may be seen as representing an early encounter with Big Data"*. In fact, also library catalogues are characterized by a certain level of "heterogeneity" due to human mistakes and to the development of different standards for cataloguing over time. Big Data methods can be used to identify the various processes of cataloguing library assets over time and to find new inconsistencies in the data. For instance, different layers of data can be identified in the British Library catalogue because of its progressive reshaping: Big Data techniques can enable something like an "archaeology" of data in library catalogues.

We notice a strong connection between Big Data repositories and Digital Libraries (DLs). According to Candela *et al.* (2008) DLs are *"organizations which might be virtual, that comprehensively collect, manage and preserve for the long term rich digital content, and offer to its user communities specialized functionality on that content, of measurable quality and according to codified policies*". As noticed by Jansen (2013) the heterogeneity of content that can be found in a digital library, ranging from digitized versions of printed books to born-digital content, requires advanced data management technology.

A peculiar element of complexity comes from the fact that digital content can lie on different levels of syntactic and semantic abstraction. Big Data techniques offer enough flexibility to cope with such intrinsically heterogeneous information assets. When the magnitude of Digital Library in terms of Volume, Velocity and Variety of content, user base or any other aspect requires "specialized technologies or approaches" we enter the domain of Very Large DLs (VLDLs), (Candela et al., 2012). It is interesting to notice how the definition of VLDLs is coherent with the one we propose for Big Data: this suggests how Big Data technology and methods are strongly needed for a continued development of library and information management as a discipline.

Another prominent reason for the growing availability of information is the proliferation of personal devices connected to the Internet and equipped with digital sensors (such as cameras, audio recorders and GPS locators). Such sensors make digitization possible while network connection lets data be collected, transformed and, ultimately, organized as information. It was estimated that at some point between 2008 and 2009 the quantity of connected devices surpassed the number of human beings (Evans, 2011) and, Gartner forecasted that by 2020 there will be 26 billion devices on earth, more than three per living person (2014). The scenario in which artificial objects, equipped with unique identifiers interact with each other to achieve common goals, without any human interaction, goes under the name of Internet of Things, IoT (Atzori et al., 2010; Estrin et al., 2002) and represents a promising source of Information in the age of Big Data. An increasing amount of Data will also be generated by the collaboration among firms by means of internet-based tools (Michelino et al., 2008).

An important feature of the data that gets produced and utilized nowadays is its expanding variety in form. The traditional alphanumerical tables are being overtaken in presence by the growing availability of less structured data sources, like video, pictures and text produced by humans, (Russom, 2011). The multiplicity of data types and their co-existence is one of the major challenges associated with the handling of Big Data today (Manyika et al., 2011).

A necessary prerequisite for using Big Data: Technology

Another theme we found in Big Data literature relates to the specific technological issues that come hand in hand with the utilization of extensive amounts of data. Dealing with Big Data at the right speed implies computational and storage requirements that an average Information Technology system might not be able to grant.

Hadoop is an open source framework that was specifically designed to deal with Big Data in a satisfactory manner. The primary components of Hadoop are HDFS and MapReduce: both were originally developed by Google (Ghemawat et al., 2003), before becoming an Apache standalone project. HDFS (Hadoop Distributed File System) enables multiple, remotely located machines to co-operate seamlessly towards a common computational goal (Shvachko et al., 2010). MapReduce is a programming model intended to efficiently split operations across separate logical units (Dean & Ghemawat, 2008).

Hadoop and Map Reduce have proven to be very effective when exploring and managing metadata of large libraries: Powell *et al.* (2012) propose an implementation for extraction and matching of author names based on Big Data technology. Another example of how of a Big Data implementation designed to serve very large digital libraries is PuntStore (Wang et al., 2013): this

plug-in system supports the integration of several storage and index engines in order to maximize efficiency in making use of large collections of digital items.

Beside the complexity arising from the processing of Big Data, another fundamental technological issue, due to the dispersed nature of machines, is its transmission. Communication networks need to sustain bigger and faster data transfers and systems require specific benchmarking techniques for evaluating their overall performance (Xiong et al., 2013).

An additional technological requirement linked to the usage of Big Data is the capacity to store a greater extent of data on smaller devices. Moore's law states that the number of transistors that can be placed on a silicon chip tends to double every 18 to 24 months and this implies that memory storage capacity grows exponentially (2006). However, data also grows exponentially (Hilbert & López, 2011), and the issue of storing extensive amounts of data persists as a critical technological challenge in the age of Big Data.

Techniques for processing Big Data: Methods

Huge amounts of data need to be processed by means of more complex methods than the usual statistical procedures. Unfortunately, a specific competence about the potentiality and limitations of these techniques is not readily accessible in the job market at present.

Big Data Analytical Methods have been singled out by Manyika et al. (2011) and Chen (2012). They have obtained a list of the most usual procedures that includes: Cluster analysis, Genetic algorithms, Natural Language Processing, Machine learning, Neural Networks, Predictive modelling, Regression Models, Social Network Analysis, Sentiment Analysis, Signal Processing and Data Visualization.

According to Chen *et al.* (2012)., given the current, structured data-centric business environment, companies should invest in interdisciplinary Business Intelligence and Analytics education, so as to cover "*critical analytical and IT skills, business and domain knowledge, and communication skills*". At the same time, a cultural change should accompany this process, involving the company's entire population, urging its members to "*efficiently manage data properly and incorporate them into decision making processes*", (Buhl et al., 2013).

New professional skills could derive from such innovative education, that would help in training experts to assimilate various disciplines (Mayer-Schönberger & Cukier, 2013). These data scientists can be seen as hybrid specialists, able to manage both technologic knowledge and academic research (Davenport & Patil, 2012). There is a gap in the education for this professional profile (Manyika et al., 2011) and new productive learning subjects and methods are required for teaching future data specialists.

In addition, it must be noted that Big Data development has turned the method of decision making from a static process into a dynamic one; indeed, the analysis of the relationships among the many events derived from information data has replaced the pursue of traditional, logical connections. It is reasonable to presume that the consequence of the application of Big Data to companies, research and university institutions could modify, both the decision making rules (McAfee et al., 2012) and the scientific method (Anderson, 2007).

Learning about the strengths and weaknesses of Big Data Methods' application represents an undeniable resource for public and private institutions when carrying out strategic decision making processes(Boyd & Crawford, 2012). Clearly, the insight into the realm of future possibilities advanced by Big Data applications should be carefully verified, by ruling their high degree of complexity with the utmost cognizance.

Big Data touches our lives pervasively: Impact

Utilization and management of Big Data are impacting many fields of activities of our society. Big Data applications have shown a consistent level of adaptability to the different requirements arising from disparate scientific domains and industrial organizations. Problems originating in very distant areas were sometimes solved by making use of the same techniques and data types. An example of this is the application of correlation analysis to Google Search logs that produced insights applied to a range of domains, from epidemiology to economics (Ginsberg et al., 2009; Askitas and Zimmermann, 2009; Guzman, 2011).

Big Data is also a source of concern as its rapid growth preceded the establishment of exhaustive guidelines to protect private information(Boyd & Crawford, 2012). For example, it is necessary to prevent any possible identification of personal data by means of anonymization algorithms aimed to defend individual privacy.

Furthermore, the accessibility of information should be properly and impartially regulated in order to avoid anticompetitive business practices (Manovich, 2012) that could reinforce dominant positions in the market. The creation of a new digital divide among companies, driven by their different level of access to data is a potential impediment to the progress of innovation (Boyd & Crawford, 2012).

Big Data also impacts companies in depth, as they are forced to reconsider their organization and all of their business processes in light of the availability of new information that could be

transformed into a competitive advantage in a data-driven market (Pearson and Wegener, 2013; McAfee et al., 2012).

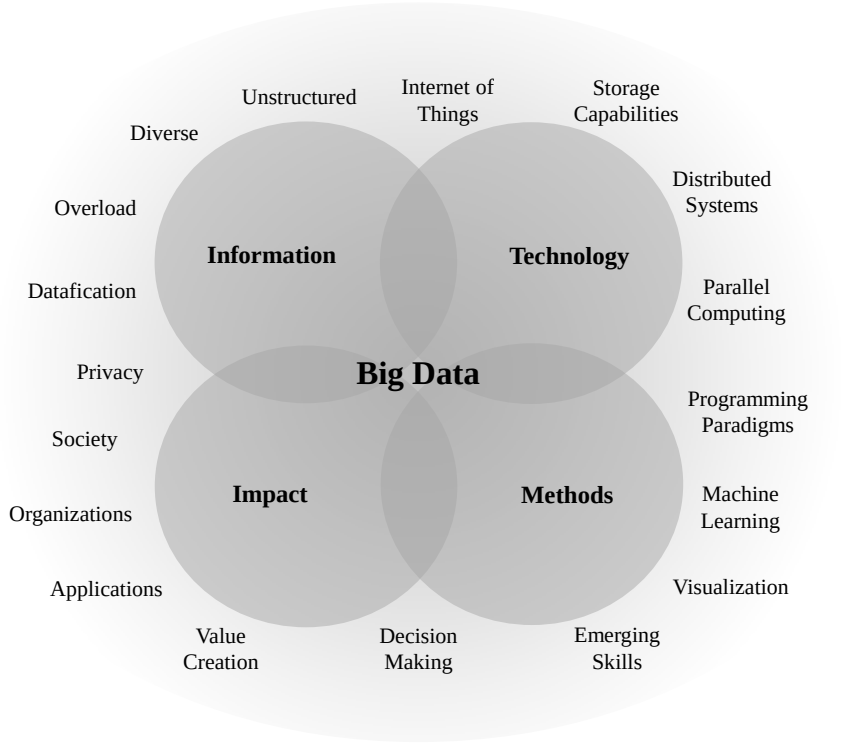


Fig. 2 Big Data themes and related topics in existing literature.

A thorough definition for Big Data

Emerging disciplines often experience a lack of agreement regarding the definition of core concepts. Indeed, the level of consensus shown by a scientific community when defining a concept is a proxy of the development of a discipline (Ronda-Pupo & Guerras-Martin, 2012). The quick and chaotic evolution of Big Data literature has impeded the development of a universally and formally accepted definition for Big Data. In fact, although several authors proposed their own definitions for Big Data (Beyer & Laney, 2012; Dijcks, 2013; Dumbill, 2013; Mayer-Schönberger & Cukier, 2013; Schroeck et al., 2012; Shneiderman, 2008; Suthaharan, 2014; Ward & Barker, 2013), none of such proposals has prevented subsequent works to modify or ignore previous definitions and suggest new ones (Ward & Barker, 2013). Such lack of agreement and homogeneity, although justified by the relative youngness of Big Data as a concept, limits the proper development of the discipline.

The next sub-section reviews a non-exhaustive list of existing definitions of Big Data, and ties them to the four themes of research previously identified in the theoretical background. The in-depth analysis of such definitions and of their shared characteristics is meant to recognize the critical factors that should be considered in a consensual definition of Big Data. Therefore, such definition should be less exposed to critiques than existing ones, being based on the most crucial aspects that have been associated so far to Big Data.

Existing definition: a survey

The absence of a consensual definition of Big Data often brought scholars to adopt "implicit" definitions through anecdotes, success stories, characteristics, technological features, trends or its impact on society, firms and business processes. The existing definitions for Big Data provide very different perspectives, denoting the chaotic state of the art. Big Data is considered in turn as a term describing a social phenomenon, information assets, data sets, storage technologies, analytical techniques, processes and infrastructures.

We observed that the definitions provided so far might be classified according to four groups, depending on where the focus has been put in describing the phenomenon: I. Attributes of Data, II. Technology needs, III. Overcoming of Thresholds, IV Social Impact. Table I shows the reviewed definitions, and describes their focus according to the four themes discussed in the theoretical background.

I group: Attributes of Data
The Big Data definitions pertaining to the first group enlist its characteristics. One of the most popular definitions for Big Data falls within this group (Laney, 2001). Noticeably, Laney's "3 Vs", underpinning the expected 3-dimensional increase in data Volume, Velocity and Variety, did not mention Big Data explicitly. His contribution was associated with Big Data several years later (Beyer & Laney 2012; Zikopoulos and Eaton, 2011; Zaslavsky et al. 2013). Several authors extended the "3 Vs" model, adding other features of Big Data, such as Veracity (Schroeck et al., 2012), Value (Dijcks, 2013), Complexity and Unstructuredness (Intel 2012; Suthaharan 2013).

II group: Technological Needs
The definitions falling within the second group emphasize the technological needs behind the processing of large amounts of data. Indeed, according to Microsoft, Big Data describes a process in which "*serious computing power"* is applied to *"seriously massive and often highly complex sets of information"* (Microsoft Research, 2013). Similarly, the National Institute of Standards and Technology emphasizes the need for a "*scalable architecture for efficient storage, manipulation, and analysis*" when defining Big Data (NBD-PWG NIST, 2014).

III group: Thresholds
Some definitions consider Big Data in terms of crossing thresholds. Dumbill (2013) proposes that data is Big when the processing capacity of conventional database systems is exceeded and alternative approaches are needed to process it. Fisher (2012) suggests that the concept of "big" in terms of size is linked with Moore's Law, and consequently with the capacity of commercial storage solutions.

IV group: Social Impact
Finally, several definitions highlight the effect of Big Data advancement on society. Boyd and Crawford (2012, p. 663) define big data as "*a cultural, technological, and scholarly phenomenon*" based on three elements: Technology (i.e. the maximization of computational power and algorithmic accuracy), Analysis (i.e. the identification of patterns on large data sets) and Mythology (i.e. the belief that large data sets offer a superior form of intelligence, carrying an aura of truth, accuracy and objectivity). Mayer-Schönberger and Cukier (2013) describe Big Data in terms of three main shifts in the way of analysing information that improve our understanding of society and our organization of it. Such shifts include: More data (i.e. all available data are used instead of a sample), More messy (i.e. even incomplete or inaccurate data may be used instead of limiting to complete ones), Correlation (i.e. correlation becomes more important, overtaking causality as privileged mean to make decisions).

< TABLE I GOES ABOUT HERE >

A Proposed Definition
The conjoint analysis of the existing definitions for Big Data and of the main research themes in literature allows us to conclude that the nucleus of the concept of Big Data includes the following aspects:

- 'Volume', 'Velocity' and 'Variety', to describe the characteristics of Information;
- 'Technology' and 'Analytical Methods', to describe the requirements needed to make proper use of such Information;
- 'Value', to describe the transformation of information into insights that may create economic value for companies and society.

On the whole, we argue that the definition for Big Data should refer to its nature of 'Information asset', a clearly identifiable entity not dependent on the field of application. Therefore, the following consensual definition is proposed:

"Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."

Such a definition is compatible with the usage of terms such as "Big Data Technology" and "Big Data Methods" when referring directly to the specific technology and methods cited in the main definition.

Conclusions

The concept of Big Data is as popular as its meaning is nebulous. With this article we have clarified the essential characteristics of Big Data, namely: Information, Technology, Methods and Impact. For each of them we have offered an exploration of the main research areas, bringing meaningful examples across multiple domains.

We have provided special focus on the impact of Big Data on information and library science: we noticed how the collection and organization of information resources have changed approach and methodology with the advent of Big Data, and how this is affecting the way libraries, especially digital

ones, are being managed. As noticed already in other sectors, researchers and professionals operating in the domain of libraries will have to upgrade their own digital and analytical skills in order to stay relevant with the upcoming data-driven innovations in the field.

We have also surveyed the most popular existing definitions of Big Data and suggested a new one that is congruent with its most prominent features: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value". A consistent utilization of this definition will contribute to the creation of an accepted terminology that will support a coherent scientific development of the topic.

Finally, some limitations regarding the present study should be considered: first, the amount of papers included in the initial list is limited in comparison with the extensive size of related literature currently available to researchers. Second, the approach we have adopted for identifying themes might have missed emerging or less evident topics as it is mainly based on human judgement. Third, our survey of definitions includes 15 popular entries while many more attempts have been presented to date and might have included other important elements to consider.

Future work includes: a more granular literature review based on quantitative methods that is capable to identify sub-topics for each of the themes we discussed; a summary of current best practises and trends by industry across the four themes of Big Data identified in this article.

References

Anderson, C. (2007). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, *16*(7).

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, *55*(2), 107–120. http://doi.org/10.3790/aeq.55.2.107

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, *54*(15), 2787–2805.

Beyer, M. A., & Laney, D. (2012). *The Importance of "Big Data": A Definition*. Stamford, CT.

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. http://doi.org/10.1080/1369118X.2012.678878

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big Data. *Business & Information Systems Engineering*, *5*(2), 65–69. http://doi.org/10.1007/s12599-013-0249-5

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., … Schuldt, H. (2007). *The DELOS Digital Library Reference Model. Foundations for Digital Libraries*.

Candela, L., Manghi, P., & Ioannidis, Y. (2012). Fourth workshop on very large digital libraries: on the marriage between very large digital libraries and very large data archives. *ACM SIGMOD Record*, *40*(4), 61–64. http://doi.org/10.1145/2094114.2094130

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Coyle, K. (2006). Mass Digitization of Books. *Journal of Academic Librarianship*, *32*(6), 641–645.

Cricelli, L., & Grimaldi, M. (2008). A dynamic view of knowledge and information: a stock and flow based methodology. *International Journal of Management and Decision Making*, *9*(6), 686–698. http://doi.org/10.1504/IJMDM.2008.021221

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job Of the 21st Century. *Harvard Business Review*, *90*(10), 70–76.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *International Conference on Integrated Information (IC-ININFO 2014) AIP Conf. Proc. 1644* (pp. 97–104). Madrid, Spain: AIP Publishing LLC. http://doi.org/10.1063/1.4907823

Dean, J., & Ghemawat, S. (2008). MapReduce : Simplified Data Processing on Large Clusters. *Communications of the ACM*, *51*(1), 1–13.

Dijcks, J. (2013). *Oracle: Big data for the enterprise*. *Oracle White Paper*. Redwood Shores, CA: Oracle Corporation.

Dumbill, E. (2013). Making Sense of Big Data. *Big Data*, *1*(1), 1–2.

Estrin, D., Culler, D., Pister, K., & Sukhatme, G. (2002). Connecting the physical world with pervasive networks. *IEEE Pervasive Computing*, *1*(1), 59–69. http://doi.org/10.1109/MPRV.2002.993145

Evans, D. (2011). *The Internet of Things - How the Next Evolution of the Internet is Changing Everything*. San Jose, CA: Cisco Systems.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with Big Data Analytics. *Interactions*.

Gartner. (2014). Gartner Says the Internet of Things Will Transform the Data Center.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, *37*(5), 29–43.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. http://doi.org/10.1038/nature07634

Google. (2015). Google Books Library Project – An enhanced card catalog of the world's books. Retrieved July 20, 2015, from https://www.google.com/googlebooks/library/

Guzman, G. (2011). Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations. *Journal of Economic and Social Measurement*, *36*(3), 119–167.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65. http://doi.org/10.1126/science.1200970

Intel IT Center. (2012). *Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data. Intel IT Center*. Santa Clara, CA: Intel Corporation.

Jansen, W., Barbera, R., Drescher, M., Fresa, A., Hemmje, M., Ioannidis, Y., … Stanchev, P. (2013). e-Infrastructures for digital libraries… the future. *Lecture Notes in Computer Science*, *8092*, 480–481. http://doi.org/10.1007/978-3-642-40501-3

Laney, D. (2001). 3-D Data Management:Controlling Data Volume, Velocity and Variety. *META Group Research Note*, (February), 1–4.

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). Minneapolis, MN: University of Minnesota Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey Global Institute.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 61–67.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. http://doi.org/10.1126/science.1199644

Michelino, F., Bianco, F., & Caputo, M. (2008). Internet and supply chain management: adoption modalities for Italian firms. *Management Research News*, *31*(5), 359–374.

Microsoft Research. (2013). The Big Bang: How the Big Data Explosion Is Changing the World.

Moore, G. E. (2006). Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Newsletter*, *11*(5), 33–35. http://doi.org/10.1109/N-SSC.2006.4785860

NBD-PWG NIST. (2014). *NIST Big Data Public Working Group. Draft of Big Data Definition*.

Pearson, T., & Wegener, R. (2013). *Big Data: The organizational challenge*, Bain & Company.

Powell, J. (2012). "At scale" author name matching with Hadoop/MapReduce. *Library Hi Tech News*, *29*(4), 6–12. http://doi.org/10.1108/07419051211249455

Prescott, A. (2013). Bibliographic records as humanities big data. In *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013* (pp. 55–58). Ieee. http://doi.org/10.1109/BigData.2013.6691670

Ronda-Pupo, G. A., & Guerras-Martin, L. A. (2012). Dynamics of the evolution of the strategy concept 1962-2008: A co-word analysis. *Strategic Management Journal*, *33*(2), 162–188. http://doi.org/10.1002/smj.948

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180.

Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter, pp 1-35*.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data.* New York, NY: IBM Institute for Business Value, Said Business School.

Shneiderman, B. (2008). Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 3–12). http://doi.org/10.1145/1376616.1376618

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In *IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST2010* (pp. 1–10). IEEE.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*. http://doi.org/10.1145/2627534.2627557

Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J. and McKeon, M. (2007), "Many Eyes: A site for visualization at internet scale", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13 No. 6, pp. 1121–1128.

Wang, J., Zhang, Y., Gao, Y., & Xing, C. (2013). A New Plug-in System Supporting Very Large Digital Library. In S. R. Urs, J.-C. Na, & G. Buchanan (Eds.), *15th International Conference on Asia-Pacific Digital Libraries, ICADL 2013, Bangalore, India, December 9-11, 2013* (Lecture No, Vol. 8279, pp. 45–52). Bangalore: Springer International Publishing. http://doi.org/10.1007/978-3-319-03599-4

Ward, J. S., & Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions*. *arXiv:1309.5821 [cs.DB]*.

Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y., … Xu, C. (2013). A characterization of big data benchmarks. In *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013* (pp. 118–125). http://doi.org/10.1109/BigData.2013.6691707

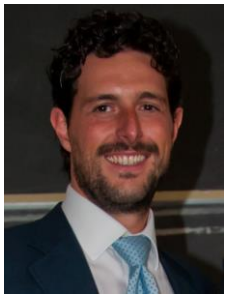
Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). *Sensing as a service and big data*. *arXiv:1301.0159 [cs.CY]*.

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media.

About the authors

Andrea De Mauro has over nine years of international experience in Data Analytics and Project Management, working for a leading FMCG multinational company. He holds a Master's Degree in Electrical and Computer Engineering from the University of Illinois at Chicago, a Master's Degree in ICT with honours from the Polytechnic of Turin and a diploma in Innovation from Alta Scuola Politecnica at Milan. He is currently pursuing his PhD in Business and Economic Engineering at "Tor Vergata" University of Rome, researching on Analytics, Data Visualisation and Decision Making. Andrea is the corresponding author and can be contacted at andrea.de.mauro@uniroma2.it.

Marco Greco is Assistant Professor at the Department of Civil and Mechanical Engineering at the University of Cassino and Southern Lazio. Graduated with honors in Business Engineering at the University of Rome "Tor Vergata", he gained the Ph.D. degree in Business and Economic Engineering at the "Tor Vergata" University of Rome. His research interests cover three disciplines: open innovation, strategic management and negotiation models.



Michele Grimaldi is an Assistant Professor at the Department of Civil and Mechanical Engineering at the University of Cassino and Southern Lazio. Graduated with honours in Business Engineering at the University of Rome "Tor Vergata", he gained the PhD degree in Business and Economic Engineering at the "Tor Vergata" University of Rome. He is a Tenured Professor of "Knowledge Management" in the second-level MS in Business Engineering organized by the "Tor Vergata" University of Rome and of "Intangible Assets" in the Executive Master Business Administration organized by the "Tor Vergata" University of Rome. He has published more than 80 papers in international journals and conference proceedings.

Preproof version – Published on Library Review, Vol. 65 Iss: 3, pp.122 – 135, DOI: 10.1108/LR-06-2015-0061

TABLE I. Survey of existing definitions of Big Data. The first column indicates the conceptual focus of the definition, namely: I. Attributes of Data, II. Technological Needs, III. Exceeding of Thresholds, IV. Social Impact. The last four columns flag whether the definition alludes to any of the four Big Data themes identified in this study, that are: I - Information, T - Technology, M - Methods, P – Impact.

Group	Source	Definition	I	T	M	P
I	(Dijcks, 2012)	The four characteristics defining big data are Volume, Velocity, Variety and Value.	x			x
	(Beyer and Laney, 2012)	High volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.	x		x	x
	(Intel, 2012)	Complex, unstructured, or large amounts of data.	x			
	(Schroeck et al., 2012)	Big data is a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace.	x			x
	(Suthaharan, 2014)	Can be defined using three data characteristics: Cardinality, Continuity and Complexity.	x			
II	(Ward and Barker, 2013)	The storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.	x	x	x	
	(Microsoft Research, 2013)	The process of applying serious computing power, the latest in machine learning and artificial intelligence, to seriously massive and often highly complex sets of information.	x	x	x	
	(NIST Big Data Public Working Group, 2014)	Extensive datasets, primarily in the characteristics of volume, velocity and/or variety, that require a scalable architecture for efficient storage, manipulation, and analysis.	x	x		
III	(Shneiderman, 2008)	A dataset that is too big to fit on a screen.	x			
	(Manyika et al., 2011)	Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.	x	x	x	
	(Fisher et al., 2012)	Data that cannot be handled and processed in a straightforward manner.	x		x	
	(Chen et al., 2012)	The data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies.	x	x	x	
	(Dumbill, 2013)	Data that exceeds the processing capacity of conventional database systems.	x	x		
IV	(Boyd and Crawford, 2012)	A cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology.		x	x	x
	(Mayer-Schönberger and Cukier, 2013)	Phenomenon that brings three key shifts in the way we analyze information that transform how we understand and organize society: 1. More data, 2. Messier (incomplete) data, 3. Correlation overtakes causality.	x		x	x