What is big data? A consensual definition and a review of key research topics

Andrea De Mauro, Marco Greco, and Michele Grimaldi

Citation: AIP Conference Proceedings **1644**, 97 (2015); doi: 10.1063/1.4907823
View online: http://dx.doi.org/10.1063/1.4907823
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1644?ver=pdfcov
Published by the AIP Publishing

What is Big Data? A Consensual Definition and a Review of Key Research Topics

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*

Abstract. Although Big Data is a trending buzzword in both academia and the industry, its meaning is still shrouded by much conceptual vagueness. The term is used to describe a wide range of concepts: from the technological ability to store, aggregate, and process data, to the cultural shift that is pervasively invading business and society, both drowning in information overload. The lack of a formal definition has led research to evolve into multiple and inconsistent paths. Furthermore, the existing ambiguity among researchers and practitioners undermines an efficient development of the subject. In this paper we have reviewed the existing literature on Big Data and analyzed its previous definitions in order to pursue two results: first, to provide a summary of the key research areas related to the phenomenon, identifying emerging trends and suggesting opportunities for future development; second, to provide a consensual definition for Big Data, by synthesizing common themes of existing works and patterns in previous definitions.

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

INTRODUCTION

Big Data[1] has now become a ubiquitous term in many parts of industry and academia. As often happens in these cases, the frequent utilization of the same words in different contexts poses a threat towards the structured evolution of its meaning. For this reason it is necessary to invest time and effort in the proposition and the acceptance of a standard definition of Big Data that would pave the way to its systemic evolution and minimize the confusion related to its usage. In order to describe Big Data we have decided to start from an "as is" analysis of the contexts in which the term most frequently appears. Given its remarkable success and its hectic evolution, Big Data possesses multiple and diverse nuances of meaning, all of which have the right to exist. By analyzing the most significant occurrences of this term in both academic and business literature we have identified four key themes to which Big Data refers: Information, Technologies, Methods and Impact. We can reasonably assert that the vast majority of references to Big Data encompass one of the four themes listed above. Understanding how these themes have been dealt with in existing literature and how they are mutually interconnected is the objective of the first section of this paper and is propaedeutic to the attempt of proposing a thorough definition, which is what the second section aims to provide. We believe that having such a definition will enable a more conscious usage of the term Big Data and a more coherent development of research on this subject.

[1] We have chosen to capitalize the term 'Big Data' throughout this article to clarify that it is the specific subject we are discussing.

International Conference on Integrated Information (IC-ININFO 2014)
AIP Conf. Proc. 1644, 97-104 (2015); doi: 10.1063/1.4907823

REVIEW OF MAIN RESEARCH TOPICS

This section represents a comprehensive but non-exhaustive review of research topics in the area of Big Data. We have examined a large number of abstracts of peer-reviewed conference and journal papers and identified recurring topics by looking at the appearance frequency of top keywords and making an educated guess on their interrelation. We needed to apply this heuristic approach in order to produce a depiction of the ample range of concepts related to Big Data while using a relatively small number of topic categories. A systematic literature review is beyond the scope of this paper and left as an opportunity for future work. The input list of documents was obtained from Elsevier's Scopus, a citation database containing more than 50 million records from around 5,000 publishers. On the 3rd of May 2014 we exported a list of 1,581 conference papers and articles that contained the full term "Big Data" in either the title or within the author-provided keywords[2]. We have removed those entries where the abstract text was not available and this left us with a corpus of 1,437 documents. By counting the appearance frequency of words included in the abstracts we have identified the most recurring items. Figure 1 shows a static tag cloud visualization (also known as "word cloud") of the most popular words in the abstracts we analyzed, obtained through the online tool ManyEyes [1].

By analyzing the most frequent keywords included in Big Data-related abstracts and considering their mutual relationships we have identified four top research themes in current literature, namely: 1. Information, 2. Technology, 3. Methods, 4. Impact. We believe that the great majority of papers written on Big Data touch upon one or more of these four topics. For each of them we will now describe content, trends and enlist a number of relevant works.

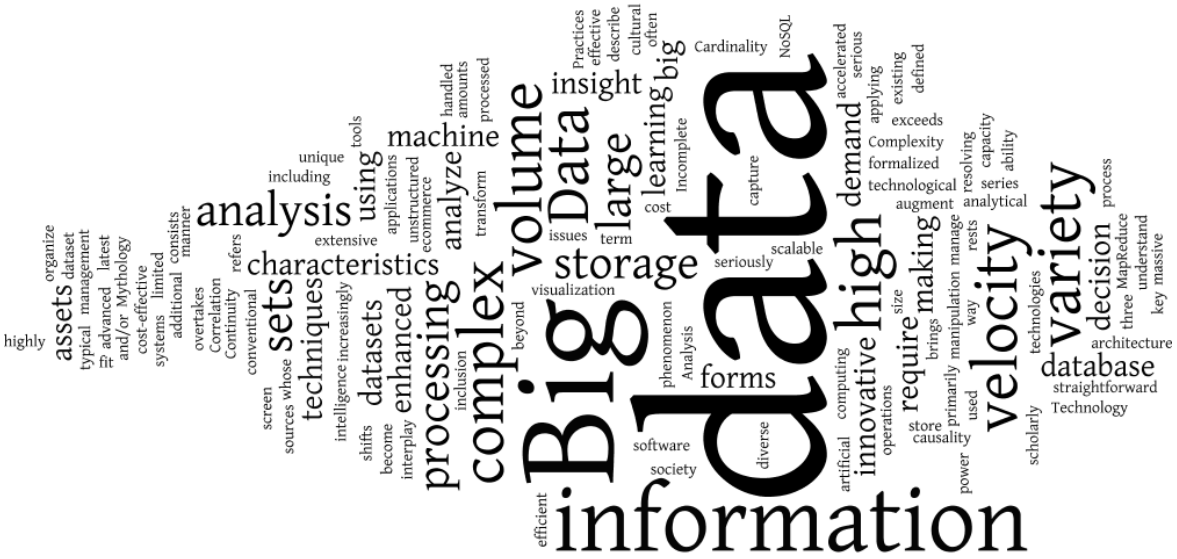

FIGURE 1. Static tag cloud visualization (word cloud) of key terms appearing in abstracts of Big Data-related papers.

The Fuel of Big Data: Information

One of the fundamental reasons for Big Data phenomenon to exist is the current extent to which information can be generated and made available. Digitization, i.e. the process of converting continuous, analog information into discrete, digital and machine-readable format, reached broad popularity with the first "mass digitization" projects. Mass digitization is the attempt to convert entire printed book libraries into digital collections by leveraging optical character recognition (OCR) software in order to minimize human intervention [2]. One of the most popular attempts of mass digitization was the Google Print Library Project[3], started in 2004, that aimed at digitizing more than 15 million volumes held in multiple university libraries, including Harvard, Stanford and Oxford. More recently it has been proposed a subtle differentiation between digitization and its next step, datafication, i.e. putting a phenomenon in a quantified format so that it can be tabulated and analyzed [3]. The fundamental difference is that

[2] We have used the following search query in Scopus: "AUTHKEY("Big data") OR TITLE("big data") AND (LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "ip"))"

[3] For more information you can visit the Google Books History page, available at http://www.google.com/googlebooks/about/history.html.

digitization enables analog information to be transferred and stored in a more convenient digital format while datafication aims at organizing digitized version of analog signals in order to generate insights that would have not been inferred while signals were in their original form. In the case of the previously cited Google mass digitization effort, the value of datafication came when researchers showed they were able to provide insights on lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology by using Google Books' data [4].

Digitization and datafication have become pervasive phenomena thanks to the broad availability of devices that are both connected and provided with digital sensors. Digital sensors enable digitization while connection lets data be aggregated and, thus, permits datafication. Cisco estimated that between 2008 and 2009 the number of connected devices overtook the number of living people [5] and, according to Gartner [6], by 2020 there will be 26 billion devices on earth, more than 3 devices on average per person. The pervasive presence of a variety of objects (including mobile phones, sensors, Radio-Frequency Identification - RFID - tags, actuators), which are able to interact with each other and cooperate with their neighbors to reach common goals, goes under the name of the Internet of Things, IoT [7,8]. This increasing availability of sensor-enabled, connected devices is equipping companies with extensive information assets from which it is possible to create new business models, improve business processes and reduce costs and risks [9]. In other words, IoT is one of the most promising fuels of Big Data expansion.

Another characteristic of the data generated today is its increasing variety in type. Structured data (traditional text/numeric information) is now joined by unstructured data (audio, video, images, text and human language) and semistructured data, such as XML and RSS feeds [10]. The diversity of data types is one of the challenges that organizations need to tackle in order to make value out of the extensive informational assets available today [11].

Equipment for Working with Big Data: Technology

The term Big Data is frequently associated with the specific technology that enables its utilization. The extent of the dataset size and the complexity of operations needed for its processing entail stringent memory storage and computational performance requirements. According to Google Trends, the most related query to "Big Data" is "Hadoop" that indeed is the most prominent technology associated with this topic. Hadoop is an open source framework that enables the distributed processing of big quantities of data by using a group of dispersed machines and specific computer programming models. The main components of Hadoop are: 1. its file system HDFS, that allows access to data scattered over multiple machines without having to cope with the complexity inherent to their dispersed nature; 2. MapReduce, a programming model designed to implement distributed and parallel algorithms in an efficient way. Both HDFS [12] and MapReduce [13] are the evolution of concepts that were originally proposed by Google [14] and that were then developed as open-source projects within Apache's framework. This proves the centrality of Google in the initiation of the current thinking about Big Data. The Hadoop framework contains multiple modules and libraries compatible with HDFS and MapReduce that enable the extension of its applicability to the various needs of coordination, analysis, performance management and workflow design that normally occur in Big Data applications.

The distributed nature of information requires a specific technological effort for transmitting big quantities of data and for monitoring the overall system performance using special benchmarking techniques [15].

Another fundamental technological element is the ability to store a bigger quantity of data on smaller physical devices. Although Moore's law suggests that storing capacity increases over time in an exponential manner [16], still it is required a continuous and expensive research and development effort to keep up with the pace at which data size increases [17] especially with the growing share of byte-hungry data types such as images, sounds and videos.

Transforming Big Data in Value: Methods

The analysis of extensive quantities of data and the need to grasp value out of individual behaviors require processing methods that go beyond the traditional statistical techniques. The knowledge of such methods, of their potential and, above all, of their limitations requires specific skills that are hard to find in today's job marketplace.

Both Manyika et al. [11] and Chen [18] propose a list of Big Data Analytical Methods, that include (in alphabetical order): A/B testing, Association rule learning, Classification, Cluster analysis, Data fusion and data integration, Ensemble learning, Genetic algorithms, Machine learning, Natural Language Processing, Neural

networks, Network analysis, Pattern recognition, Predictive modelling, Regression, Sentiment Analysis, Signal Processing, Spatial analysis, Statistics, Supervised and Unsupervised learning, Simulation, Time series analysis and Visualization.

Chen et al. [18] evoke the need for companies to invest in Business Intelligence and Analytics education that would be "*interdisciplinary and cover critical analytical and IT skills, business and domain knowledge, and communication skills required in a complex data-centric business environment*". The investment in analytical knowledge should be accompanied by a cultural change that would span across all employees and urge them to "*efficiently manage data properly and incorporate them into decision making processes*" [19]. Mayer-Schönberger and Cukier [3] envision the rise of new specific professional entities, called algorithmists, that would master the areas of computer science, mathematics and statistics and act as "*impartial auditors to review the accuracy or validity of Big Data predictions*". Also Davenport and Patil [20] describe data scientist as a hybrid of "*data hacker, analyst, communicator, and trusted adviser*", having also the fundamental abilities to write code and conduct, when needed, academic-style research. These skills are not sufficiently available to meet the increasing demand: according to Manyika et al. [11], by the year 2018 there will be a potential shortfall of 1.5 million data-savvy managers and analysts, in the US only. The analysis of competency gaps and the creation of effective teaching methods to fill them for both future and current managers and practitioners is a promising research area that has still much opportunity to grow.

Also the ability of making informed decisions is changing with the expansion of Big Data as the latter implies the shift from logical, causality-based reasoning to the acknowledgment of correlation links between events. The utilization of insights generated through Big Data Analytics in companies, universities and institutions provides for an adaptation to a new culture of decision making [21] and an evolution of the scientific method [22], both of which are still to be built and provide opportunities for future research.

Being aware of the limitations of Big Data Methods and potential methodological issues is a fundamental resource for organizations who want to drive data-based decision making: for example, predictions should always be accompanied by valid confidence intervals in order to avoid the false sense of precision that the apparent sophistication of some Big Data applications can suggest. Analysts should also be capable of avoiding models' overfitting that would facilitate apophenia, i.e. the tendency of humans to "*see patterns where none actually exist simply because enormous quantities of data can offer connections that radiate in all directions*", [23].

In a summary, Big Data requires the mastery of specific techniques, awareness of their strengths and limitations, and a spread cultural tendency to informed decision making that in most cases has still to be built.

How Big Data Changes our Lives: Impact

The extent to which Big Data is impacting our society and our companies is often depicted through anecdotes and success stories of methods and technology implementations. When these stories are accompanied by proposals of new principles and methodological improvements they represent a valuable contribution to the creation of knowledge on the subject. The pervasive nature of the current information production and availability leads to many applications spanning in numerous scientific fields and industry sectors that can be very distant from each other. Sometimes, the same techniques and data have been applied to solve problems in distant domains. For example, correlation analysis was leveraged to use logs of Google searches to forecast influenza epidemics [24] as well as unemployment [25] and inflation [26]. The existing Big Data applications are many and expected to grow: hence, their systematic description constitutes a promising development area for those willing to contribute in the scientific progress in this field.

Big Data can also impact society adversely. In fact, there are multiple concerns arising from the quick advancement of Big Data [23] first being privacy. Although large data sets would normally proceed from actions done by a multitude of individuals, it is not always true that consequences of using that data will not impact a single individual in an invasive and/or unexpected way. The identifiability of the individual person can be avoided through a thorough anonymization of the data set, although it is hard to be fully guaranteed as the reverse process of de-anonymization can be potentially attempted [27]. The predictability of future actions, made possible by the analysis of behavioral patterns, poses also the ethical issue of protecting free will in the future, on top of freedom in the present.

Other issues to be considered are related to the accessibility of information: the exclusive control over data sources can become an abuse of dominant position and restrict competition by posing unfair entrance barriers to the marketplace. For example, as Manovich notices [28], "*only social media companies have access to really large*

social data – especially transactional data" and they have full control over who can access what information. The split between information-rich and data-lacking companies can create a new digital divide [23] that can slow down innovation in the sector. Specific policies will have to be promoted and data is likely to become a new dimension to consider within antitrust regulations.

Not only society but also companies are heavily impacted by the rise of Big Data: the call to arms for acquiring vital skills and technology to be competitive in a data-driven market implies a serious reconsideration of the firm organization and the full realm of business processes [29]. The transformation of data into competitive advantage [21] is what makes "Big Data" such an impactful revolution in today's business world.

A DEFINITION FOR BIG DATA

A convincing definition of a concept is an enabler of its scientific development. As Ronda-Pupo and Guerras-Martin [30] suggest, the level of consensus shown by a scientific community on a definition of a concept can be used as a measure of progress of a discipline. Big Data has instead evolved so quickly and disorderly that such a universally accepted formal statement denoting its meaning does not exist. There have been many attempts of definition for Big Data, more or less popular in terms of utilization and citation. However, none of these proposals has prevented authors of Big Data-related works to extend, renovate or even ignore previous definitions and propose new ones. Although Big Data is still a relatively young concept, it certainly deserves an accepted vocabulary of reference that enables the proper development of the discipline among cognoscenti and practitioners.

In the first part of this paper we have identified the four main themes of Big Data and we have observed that they are the prevalent topics in the existing literature. In the next paragraphs we will review a non-exhaustive list of previously proposed Big Data definitions and we will conceptually tie them to the aforementioned four themes of research. After considering the existing definitions and analyzing their commonalities we will propose a consensual definition of Big Data. Consensus in this case comes from the acknowledgement of centrality of some recurring attributes associated to Big Data, and from the assumption that they define the essence of what Big Data means to scholars and practitioners today. We expect that such a definition would be less prone to attack from previous definitions' authors and users as it is based on the most central aspects associated until now to Big Data.

A thorough consensus analysis based on Cohen's K coefficient [31] and co-word analysis, as in [30], goes beyond the scope of this work and is left for future study.

Survey of Existing Definitions

Big Data has been often described "implicitly" through success stories or anecdotes, characteristics, technological features, emerging trends or its impact to society, organizations and business processes. In the existing attempts of explicit definitions for Big Data there is not even an agreement on what entity this term can be associated with. We have found that Big Data is used when referring to a variety of different entities including - but not limited to - social phenomenon, information assets, data sets, analytical techniques, storage technologies, processes and infrastructures. We have surveyed multiple definitions that have been proposed to date and listed them in Tab. 1: in this paragraph we will go through the most notable ones.

A first group of Big Data definitions focuses on enlisting its characteristics. What is probably the most popular definition falls within this group. When presenting the Data Management challenges that companies had to face in response to the rise of e-commerce in the early 2000's, Laney introduces a framework expressing the 3-dimensional increase in data Volume, Velocity and Variety and invokes the need for new formal practices that will imply "*tradeoffs and architectural solutions that involve/impact application portfolios and business strategy decisions*" [32]. Although this work did not mention Big Data explicitly, the model, later nicknamed as "the 3 V's", was associated to the concept of Big Data and used as its definition [33–35]. Many other authors extended the "3 V's" model and, as a result, multiple features of Big Data, like Value [36], Veracity [37], Complexity and Unstructuredness [38, 39], were added to the list.

A second group of definitions emphasizes the technological needs behind the processing of large amounts of data. According to Microsoft, Big Data is about applying "*serious computing power*" to massive sets of information [40] and also the National Institute of Standards and Technology (NIST) highlights the need for a "*scalable architecture for efficient storage, manipulation, and analysis*" when defining Big Data [41].

A few definitions associate Big Data to the crossing of some sort of threshold: for instance Dumbill [42] asserts that data is Big when it "*exceeds the processing capacity of conventional database systems*" and requires the choice

TABLE 1. Existing definitions of Big Data, adapted from the articles referenced in the first column. The last four columns indicate whether the definition alludes to each of the four Big Data themes identified in the first section of the paper, through the following legend: I - Information, T - Technology, M - Methods, P - Impact.

Source	Definition	I	T	M	P
[33]	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
[36]	The four characteristics defining big data are Volume, Velocity, Variety and Value.	x			x
[38]	Complex, unstructured, or large amounts of data.	x			
[39]	Can be defined using three data characteristics: Cardinality, Continuity and Complexity.	x			
[37]	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
[41]	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		
[44]	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	
[40]	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	
[42]	Data that exceeds the processing capacity of conventional database systems.	x	x		
[43]	Data that cannot be handled and processed in a straightforward manner.	x		x	
[45]	A dataset that is too big to fit on a screen.	x			
[11]	Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.	x	x	x	
[18]	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	
[23]	A cultural, technological, and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology.		x	x	x
[3]	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

of "*an alternative way to process it*". Fisher [43] acknowledges that the size that constitutes "big" has grown according to Moore's Law and links the absolute level of this threshold to the capacity of commercial storing solutions: Big Data "*is so large as to not fit on a single hard drive*" and, hence "*will be stored on several different disks*".

A last group of definitions highlights the impact of Big Data advancement on society. Boyd and Crawford [23] notice that "*Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets*". They define Big Data as "*a cultural, technological, and scholarly phenomenon*" that rests on the interplay of Technology (maximizing computation power and algorithmic accuracy), Analysis (to identify patterns on large data sets) and Mythology (meaning the belief that large data sets offer a higher form of intelligence with an aura of truth, objectivity and accuracy). Mayer-Schönberger and Cukier [3] describe Big Data by enlisting

the three key "*shifts in the way we analyze information that transform how we understand and organize society*": 1. "*More data*", in terms of "completeness" of the data set, using all of available data instead of a sample of it; 2. "*More messy*", meaning that we can loosen up on our desire for exactitude and use also incomplete or less accurate input data; 3. "*Correlation*" becomes more important and overtakes "causality" as a way to make sense of trends and finally make decisions.

Consensual Definition

By looking at both the existing definitions of Big Data and at the main research topics associated to it, we can affirm that the nucleus of the concept of Big Data can be expressed by:

- 'Volume', 'Velocity' and 'Variety', to describe the characteristics of Information involved;
- Specific 'Technology' and 'Analytical Methods', to clarify the unique requirements strictly needed to make use of such Information;
- Transformation into insights and consequent creation of economic 'Value', as the principal way Big Data is impacting companies and society.

We believe that the "object" to which Big Data should refer to in its definition is 'Information assets', as this entity is clearly identifiable and is not dependent on the field of application.

Therefore, we propose the following formal definition:

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

Such a definition of Big Data is compatible with the existence of terms like "Big Data Technology" and "Big Data Methods" that should be used when referring directly to the specific technology and methods mentioned in the main definition.

CONCLUSION

Big Data has recently become a voguish term among researchers and IT professionals. Its success is propelled by a frequent utilization in a broad range of contexts and with several, and often incongruous, acceptations. As a result, its meaning is still nebulous and this hinders an organized evolution of the subject.

We have conducted an analysis of the usage of this term in literature and concluded that the top four themes associated to Big Data are: Information, Technology, Methods and Impact. We have then have suggested a definition that is coherent with the current "as is" utilization of the term and consensual with the most prominent definitions that have been so far proposed. We suggest using Big Data as a standalone term when referring to those "Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" and as an attribute when denoting its peculiar requisites, e.g. "Big Data Technology" or "Big Data Analytical Methods". We believe that using this definition from now on will allow a more efficient scientific development of the matter.

Possible extensions to the present work include:

- A systematic literature review of "Big Data" by means of quantitative methods, such as co-word, cluster and frequency analysis. The review should also identify a more granular list of research topics through systemic methods like topic modeling.
- Study of how Big Data is systematically impacting on the creation of economic value in companies and a proposal of guidelines for a coherent development of system and processes related to Business Intelligence and Analytics. We can presume that the value creation chain would go through the four themes of Big Data and that maximizing the value each component brings would generate higher returns on BI&A investments.

REFERENCES

[1] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, IEEE Trans. Vis. Comput. Graph. **13**, 1121 (2007).

[2] K. Coyle, J. Acad. Librariansh. **32**, 641 (2006).

[3] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live* (2013).
[4] J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, Science **331**, 176 (2011).
[5] D. Evans, *The Internet of Things - How the Next Evolution of the Internet Is Changing Everything* (2011), pp. 1–11.
[6] Gartner, (2014), available at http://www.gartner.com/newsroom/id/2684616.
[7] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, IEEE Pervasive Comput. **1**, 59 (2002).
[8] L. Atzori, A. Iera, and G. Morabito, Comput. Networks **54**, 2787 (2010).
[9] M. Chui, M. Löffler, and R. Roberts, McKinsey Q. **291**, 10 (2010).
[10] P. Russom, TDWI Best Pract. Report, Fourth Quarter (2011).
[11] J. Manyika, M. Chui, B. Brown, and J. Bughin, *Big Data: The next Frontier for Innovation, Competition, and Productivity* (2011).
[12] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, in *2010 IEEE 26th Symp. Mass Storage Syst. Technol. MSST2010* (2010).
[13] J. Dean and S. Ghemawat, Commun. ACM **51**, 1 (2008).
[14] S. Ghemawat, H. Gobioff, and S.-T. Leung, ACM SIGOPS Oper. Syst. Rev. **37**, 29 (2003).
[15] W. Xiong, Z. Yu, Z. Bei, J. Zhao, F. Zhang, Y. Zou, X. Bai, Y. Li, and C. Xu, in *Big Data, 2013 IEEE Int. Conf.* (2013), pp. 118–125.
[16] G.E. Moore, IEEE Solid-State Circuits Newsl. **20**, (2006).
[17] M. Hilbert and P. López, Science **332**, 60 (2011).
[18] H. Chen, R. Chiang, and V. Storey, MIS Q. **36**, 1165 (2012).
[19] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, Bus. Inf. Syst. Eng. **5**, 65 (2013).
[20] T. H. Davenport and D.J. Patil, Harv. Bus. Rev. **90**, 70 (2012).
[21] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, Harv. Bus. Rev. **90**, 61 (2012).
[22] C. Anderson, Wired 3 (2007).
[23] D. Boyd and K. Crawford, Information, Commun. Soc. **15**, 662 (2012).
[24] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, Nature **457**, 1012 (2009).
[25] N. Askitas and K. F. Zimmermann, Appl. Econ. Q. **55**, 107 (2009).
[26] G. Guzman, J. Econ. Soc. Meas. **36**, 119 (2011).
[27] A. Narayanan and V. Shmatikov, in *Proc. - IEEE Symp. Secur. Priv.* (2008), pp. 111–125.
[28] L. Manovich, Debates Digit. Humanit. 1 (2011).
[29] T. Pearson and R. Wegener, Big Data: The Organizational Challenge, Bain & Company report (2013).
[30] G. A. Ronda-Pupo and L. Á. Guerras-Martin, Strateg. Manag. J. **33**, 162 (2012).
[31] J. Cohen, Educ. Psychol. Meas. **20**, 37 (1960).
[32] D. Laney, META Gr. Res. Note **6**, (2001).
[33] M. A. Beyer and D. Laney, The Importance of "Big Data": A Definition, Gartner report (2012), pp. 1–9.
[34] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. Zikopoulos, *Understanding Big Data* (McGraw-Hill Companies, 2012).
[35] A. Zaslavsky, C. Perera, and D. Georgakopoulos, preprint arXiv:1301.0159 (2013).
[36] J. Dijcks, Big Data for the Enterprise, Oracle report (2012).
[37] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, Analytics: The Real-World Use of Big Data, IBM report (2012), pp. 1–20.
[38] Intel, Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data, Intel report (2012).
[39] S. Suthaharan, ACM SIGMETRICS Perform. Eval. Rev. **41**, 70 (2014).
[40] Microsoft, (2013), available at https://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx.
[41] NIST Big Data Public Working Group, *Big Data Interoperability Framework: Definitions (draft)* (2014).
[42] E. Dumbill, Big Data **1**, 1 (2013).
[43] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, Interactions (2012).
[44] J. Ward and A. Barker, preprint arXiv:1309.5821 (2013).
[45] B. Shneiderman, in *Proc. 2008 ACM SIGMOD Int. Conf. Manag. Data* (2008), pp. 3–12.