ISR Italian Sociological Review
University of Verona, Italy
ISSN 2239-8589 | Editor in Chief Paola Di Nicola
Managing Editor Debora Viviani

The Use of Big Data: Some Epistemological and Methodological Considerations

Sonia Stefanizzi

How to cite
Stefanizzi, S. (2021). The Use of Big Data: Some Epistemological and Methodological Considerations. [Italian Sociological Review, 11 (4S), 193-205]
Retrieved from [http://dx.doi.org/10.13136/isr.v11i4S.431]

[DOI: 10.13136/isr.v11i4S.431]

1. Author information
Sonia Stefanizzi
Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy

2. Author e-mail address
Sonia Stefanizzi
E-mail: sonia.stefanizzi@unimib.it

3. Article accepted for publication
Date: January 2021

The Use of Big Data: Some Epistemological and Methodological Considerations

Sonia Stefanizzi*

Corresponding author:
Sonia Stefanizzi
E-mail: sonia.stefanizzi@unimib.it

Abstract

After presenting the main features of Big Data, the paper discusses its epistemological and methodological implications in social research. If the current interest in Big Data has generated the widespread impression that these methods are able to produce knowledge without the need to resort to conventional scientific methods, it remains an open question whether the possibility to work with a huge amount of data has really led to a new epistemological transition. A further aspect of critical reflection, present in the essay, concerns the quality and reliability of information, privacy issues and data ownership.

Keywords: Big Data, epistemological transition, data quality.

1. The Big Data phenomenon

The digital revolution of the last decades, with the development of computer science, has given the possibility to social scientists to have at their disposal a huge amount of data to share and a new type of research called eScience, based on the power of algorithms and computers. As several authors have well argued, in today's situation research would move within a data-driven science that incorporates theory, experiments and simulation and whose ambition would be to use the emerging correlations between these huge data sets to build powerful predictive models that can shed light on the complexity

* Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy.

of human behavior and offer insights to guide future behavior (Hendler et al. 2008; Dutton et al. 2010; Manovich, 2011).

The data currently available come from a variety of sources (companies, social media, web, data services, public institutions, etc.) and are constantly evolving and growing. In particular, when we talk about these data we refer to the so-called Open Data and Big Data. Without claiming to enter into the vast literature on this subject, here we limit ourselves to clarifying the general meaning of the terms. Open Data are data that can be accessed freely through the Internet (which is also the main channel of dissemination), without licenses or other forms of control that limit the reproduction and copyright restrictions with possibly only the obligation to cite the source. These data are related to the concept of Open Government, according to which the public administration should be open to citizens, both in terms of transparency and direct participation in the decision-making process. The distinctive elements of Open Data are the openness, the commitment of the public administration in questioning the relationship with citizens, providing them the opportunity to interact with institutions and put in place participatory behaviors based on a relationship of reciprocity, dissemination through information technology and free.

The term Big Data recalls an abstract concept, there is no exclusive definition and for this reason several conceptualizations have been proposed that differ in some elements, but share the idea that such data refer to a data set whose size goes beyond the capacity of a normal data base to capture, store, manage and analyze data (Manyika, 2011). Moreover, a common reference across all definitions are the so-called 5Vs that characterize them, namely volume, velocity, variety, veracity, and value (Opresnik, Taisch, 2015). Some scholars suggest adding two more V's to the definition of Big Data: variability and virality. Variability means that data must be contextualized, as its meaning can vary depending on context, while virality refers to the exponential growth of Big Data (Baker, 2014). Such peculiar characteristics require that, with respect to storage, the databases constituting Big Data, are both structured and unstructured, are expressed on different measurement scales, and/or are also qualitative in nature.

Moreover, some authors have pointed out that the expression Big Data is often misused, according to common sense, to refer to everything that is traceable through data. Its scientific meaning refers, instead, mainly to data sets that cannot be collected, stored, shared, analyzed, visualized without the help of appropriate IT tools. What characterizes Big Data is therefore not only the size, but the complexity of the data sets, as Hillard reminds us:

> ... A good definition of Big Data is to describe big in terms of the number of useful permutations of source making useful querying difficult (like the sensors in an aircraft) and complex interrelationships making purging difficult (as in the toll road example). Big then refers to big complexity rather than big volume. Of course, valuable and complex datasets of this sort naturally tend to grow rapidly ad so Big Data quickly becomes truly massive. Big Data can be very small and not all large datasets are big (Hillard, 2012: 120).

When we talk about Big Data we think, therefore, of a mass of complex data, extended in terms of volume, speed and variety to require technologies and analytical methods, complex for data extraction. For example, information inferred from websites (such as accesses, permanence, etc.), GPS data, sets of images, emails, information derived from social networks (Snijders et al. 2012). One of the key features of such data is the heterogeneity of the sources: they are dynamic streams of "metadata" from composite databases (Rezzani, 2013).

The challenges posed by Big Data to researchers are different. A first challenge concerns, as mentioned, the ability to analyze particularly complex databases derived from an increasing plurality of sources, from a mass of data-information at very high volume and with a plurality of formats and typologies without a common structure[1]. A second challenge is epistemological and refers to the criteria for the construction of scientific knowledge through the information produced by Big Data. In this regard, some authors argue that the digital revolution in science has created a new era in the production of scientific knowledge involving a sort of scientific revolution, or the transition to the so-called "fourth paradigm"[2]. The assumption at the basis of the so-called fourth paradigm is that the information collected in large quantities can be transformed

[1] In fact, they can be streams of communication collected from social media, audio, video, documents of various kinds, emails, web pages and posts.

[2] As is well known, it was Thomas Kuhn (1962) who introduced the concept of paradigm shift understood as the engine of science. Recall that the first and second paradigm in the history of scientific knowledge are based on rigorous theories and careful empirical verification and have concerned the description of natural phenomena and the discovery of the laws of nature. For three centuries these two paradigms defined science. In the second part of the twentieth century, with the birth and spread of computers, a third paradigm based on simulation has taken over. In many fields (as in the case of weather forecasting or in the study of climate change) scientific research has produced results that do not relate directly to reality, but only a more or less good approximation with the advantage of being able to build a series of probability scenarios. The possibility of working with a huge amount of data has led to a new epistemological transition. The conceptual and methodological model, the basis of the so-called fourth paradigm, is that to produce knowledge is no longer necessary theory, but only the production of algorithms.

automatically into a new way of producing scientific knowledge. On this aspect, scholars are divided between those who argue that Big Data can easily create a new form of knowledge and those, however, argue that the information itself cannot be considered knowledge (Hey, Tansley, Tolle, 2009). In particular, it would seem no longer necessary to follow the traditional model of scientific inquiry: the conceptual tools such as theory, hypotheses would become obsolete representing, for some authors, only a further mental complication having available such masses of data that, if properly analyzed with the appropriate mathematical techniques, can show interesting correlations between the data themselves (Anderson, 2008). In this regard Prensky (2009) argues that it is not necessary to build hypotheses or models to be tested empirically as it would seem sufficient to produce knowledge, the simple correlation between data without necessarily seek models of causation. Following this type of argumentation appears outdated the old design of scientific research based on the elaboration, from a theory, of hypotheses on the functioning of a social phenomenon and then proceed to empirical verification. With Big Data, therefore, the idea is affirmed that it is possible, in the absence of any research question, to process huge amounts of data to search for meaningful correlations from extremely complex data systems, regardless of the knowledge of their content (Baldassare, 2016).

While undoubtedly Big Data has expanded knowledge about individuals, increased the efficiency of production processes, and helped to improve the decision-making ability of policy makers, it cannot be considered alternative to theory. In the academic debate that has developed in recent years, although a consensus has been reached at the definitional level regarding the concept of Big Data and awareness of the risks and benefits associated with them, the question remains whether these types of data are able to explain social phenomena without the need to resort to conventional scientific methods.

We believe, as Etzioni argues, that "To make a significant improvement in the accuracy of predictions, we need together, basic knowledge, deductive reasoning, and more sophisticated semantic models" (Etzioni, 2016: 1519). The role of theory, therefore, even when using large amounts of data, seems important as it can help the researcher to decide which variables, in the immensity of the available data, to concentrate the analysis on. Moreover, it can give a relevant contribution to identify causal relationships between different phenomena. For example, if we want to see the effect of cultural capital on the average income from future work of individuals, the simple correlation tells us nothing about how some intervening variables can influence the relationship, which, however, thanks to a theoretical framework can be easily identified as, in the case under consideration, the individual's innate abilities and socio-family

background[3]. Big Data are prevalently used to search for a posteriori correlation, that is not a priori seen on the basis of a theoretical model, but simply identified in the data and to which we will try, posteriori, to give an explanation.

Big Data are preferably used to search for a posteriori correlation, that is, they are not seen a priori on the basis of a theoretical model, but simply identified in the data and to which we will try, posteriori, to give an explanation[4].

The use of Big Data refers, albeit with new tools (the algorithms) to a logic of inductive type. Inductive inferences undoubtedly lead to conclusions whose information content is greater than that of the premises, even if this step is not guaranteed by a logical condition.

As is well known, the limitations of inductivism have been discussed by many logicians and philosophers who have highlighted the problems of this type of reasoning (Russel, 1912; Popper, 1959). As Popper argues, data do not speak for themselves. Any observation presupposes prior knowledge of the reality being observed. One cannot observe anything without first knowing how the world works: observation only shows us that "things" are just as we expected them to be, or not. Another criticism that Popper addresses to the inductive method concerns the idea that science starts from observation to get to the formulation of hypotheses or laws. But if observing presupposes not only to know, but also to have hypotheses, the process of knowledge consists, in Popper's reasoning, in conceiving hypotheses, working with them until they are

[3] Becker (1994) had already identified in the innate abilities of individuals a variable difficult to measure that influences both the years of study and the income received. In this regard, Griliches and Mason (1972) and Griliches (1977) suggested the existence of an upward bias in the estimate of performance if ability is not considered. In particular, the effect of socio family background consists of the different cultural and income opportunities enjoyed by individuals from more affluent families that affect the eventual choice to pursue education and, therefore, future income leading to an upward bias in the rate of return to education (Griliches, 1977).

[4] Many examples can be given of cases in which, by analyzing a large amount of data, correlations between completely independent phenomena have been found. For example, it is known that in Italian cities both the number of churches and the number of homicides committed each year are proportional to the population, but this does not mean that increasing the number of churches increases the number of homicides and vice versa. This example serves to show that the presence of a correlation does not imply the existence of a causal link. In fact, in the huge amount of data available today, one can find correlations that are spurious and have no meaning because the two phenomena (as in our example, the number of churches and the number of murders) represent only two processes of penetration that arose and grew together in a completely random way.

confuted. From Popper onwards becomes an idea shared by many scientists and philosophers that the scientific method consists in trying to break down the theoretical statements from which knowledge itself starts. Peirce distinguishes another type of inference in addition to the two traditional forms of induction and deduction. The third type of logic is that of abduction or hypothetical reasoning which consists of formulating a causal hypothesis from a given effect (Peirce, 1960)[5]. According to Peirce (1960), abductive reasoning must be distinguished from both deductive and inductive reasoning since it has a specific structure; however, other scholars consider abduction a form of inductive reasoning since both the abductive and inductive conclusion have a probable character. Abduction is a backward process that is employed when the rules and conclusion are known and the premises are to be reconstructed. It considers a specific fact (the consequent), connects it to a hypothetical rule (implication relation) and derives an uncertain result, i.e. a hypothetical conclusion (the antecedent). It is used in diagnostic reasoning (e.g.: a doctor faced with a symptom, an electrician faced with a breakdown, etc.), in investigative reasoning (e.g.: a detective faced with a case), in scientific reasoning (a researcher faced with a hypothesis to be tested).

Without entering into the debate on the various forms of inference and the vast philosophical discussion on abduction, we can say that abductive logic is a method to generate hypothesis when we do not have them, it will not be as "reliable" as the deductive one, but not as fallacious as the inductive one and it can help to increase our knowledge about the world because it suggests a path to be verified. The question is whether the abductive approach is useful for analyzing Big Data. The answer is affirmative because the logic that guides the analysis of Big Data assumes, as in the case of abductive reasoning, a backward cognitive process that is used when rules and conclusion are known and we want to reconstruct the premises. In other words, having huge amounts of data at our disposal we can start from observed facts without having any particular theory in mind, we hypothesize causal connections between phenomena or facts and even if we can discover something new the conclusions we reach are exposed to the risk of error. In fact, following abductive reasoning the possible confirmation of a hypothesis does not prove the truth of that hypothesis and, as, Oldroyd argues, the experimental confirmation of a hypothesis does not prove the truth of that hypothesis and the assumption that it implicates such truth leads to fall into the "error of asserting the consequent" (Oldroyd, 1986). An abductive inference, therefore may present in its premises-facts that have

[5] Remember that the abductive inference presupposes a reasoning through which, starting from some facts that you want to explain (premises) you try to identify a possible hypothesis that explains them (Frixione, 2007).

some familiarity with the conclusion but that could be true without the conclusion being true. A further problem that arises when proposing models that are exclusively data-driven is that there is no guarantee that the same model can be derived from a different dataset.

In other words, Big Data cannot aspire to impose itself as the "fourth paradigm", however they can be a valid research tool because they allow to indicate a path but they cannot do without theory and experimental verification.

When we talk about Big Data, therefore, we should not only focus our attention on the huge amount of data available, but also on the analyst's ability to extract meaning from the enormous amount of data at stake[6]. Therefore, it becomes important to evaluate the quality level of the available information in order to be able to reuse the acquired information for interpretative and decision-making purposes. The role of sharing quality data is central, above all, for the subsequent production of information. Individually considered data are not particularly significant, but when analyzed in large volumes they can lead to the delineation of patterns and trends, for example behavioral, which when added to other data sources then produce knowledge.

If, on the one hand, Big Data represent a great opportunity for analysts as they enable them to identify consumer behavior patterns and shed light on their intentions, improve decision-making processes and increase company productivity, on the other hand, as we have seen, they pose a series of barriers to their use.

2. Big Data challenges

Big Data, differently from Small Data[7], are inhomogeneous data, coming from various structured and unstructured sources, they can be messy, disorderly and necessarily varied. In a certain sense, everything is collected and everything

[6] In this regard, the UN in the report "A world that counts", built to assess the opportunities arising from the innovation of technological advances and the explosion of the number of public and private data producers, in order to provide guidance on possible forms of evolution of conventional systems of statistical production, clearly indicates that the true meaning of the data revolution. This revolution is given by the awareness of the value that lies in such huge amounts of data that are generated daily and in being able to find new ways of using data in order to concretely improve the quality of people's lives (Independent Expert Advisory Group, 2014).

[7] We would recall that Small Data refers to data in a highly structured format and in an easily manageable re-ducted volume. The characteristics of this data, therefore, are: low volumes, fast response times, and decentralization of data, with the creation of autonomous repositories to be used for specific purposes.

can only be disorder. Hence their problematic nature, which mainly concerns the quality and reliability of information, privacy issues and data ownership. Focusing on the problems of quality and reliability of data, it is well known that to verify the adequacy of data is important to have information about the procedures of generation of these, in order to interpret them correctly. If in the primary analysis the researcher is able to control every phase of the construction of the data, in the secondary analysis, especially of heterogeneous data such as Big Data, often collected in different formats not directly accessible for analysis, it is necessary to have the relevant metadata. Metadata are undoubtedly a great resource, as if properly interrogated they contribute to an adequate interpretation of the phenomena being analyzed, but at the same time they are a problem for privacy promoters, who consider them, if not properly regulated, an invasion of the personal sphere of the individual. With the sudden development of technology and the consequent speed and amount of data collected, it is beginning to be considered that the existing rules on Small Data are no longer sufficient to ensure the privacy of individuals in a context in which any information can be transformed into data, or better into Big Data, through the process of data processing (Mayer-Schönberger, Cukier, 2013). The other critical aspect concerns, as it is well known, the protection of privacy as the different typology of available data coming from an increasing number of heterogeneous sources, have produced an extraordinary intrusion in everyone's life, a real surveillance, with important effects on individual and collective behaviors, on the same principles of our democracies. It is important, therefore, to focus on their preservation and correct use, clearly establishing for how long data can be retained and setting rules on how they can be used, whether they can be sold, and the level of consent to be obtained from the person providing them[8].

Today, therefore, scientific research is facing a new historical moment characterized by the production, analysis and sharing of an enormous amount of data. In this context, new prospects for social research are opening up and, above all, as we have said, new challenges. A further challenge certainly concerns the quality of data and ensuring that the large mass of data available today meets these requirements. Developments in information and communication technologies and their incorporation into social practices (social media data) potentially make available databases with new features that also derive from the computational capabilities of processing systems that make it possible to explore and analyze multiple and diverse databases. Knowledge derived from diverse data sources, such as data produced by social media, should be subjected to the most open scrutiny possible, should be cross-

[8] Please refer to Margo Seltzers talk at the 2015 Word Economic Forum in Davos.

subjectively vetted and validated methodologically and in light of the wealth of knowledge and theories produced to date by the Social Sciences.

Moving back to the issue of data quality, three dimensions have been identified (Karr et al. 2006) that within them are broken down into specific areas to focus on to ensure this requirement. These are:

1. The process (data creation, description, and management). There are three areas to consider here: reliability (which is the aspect on which the literature insists most, particularly in the data collection phase, i.e. wording, operational definitions, construction of collection tools); metadata, meta-dating would ensure that the content, the collection procedures, are documented in a clear and unambiguous way and in a form accessible to users; security and confidentiality, i.e. the need to find the right balance between protecting sensitive data and producing quality research data;
2. Data (this dimension is related to the actual data). Here we talk about accuracy (data have been reported correctly), completeness (no missing values), consistency (in the sense of the relationship between variables, i.e., that a value assumed by a case is consistent with that assumed by the same case in another variable, e.g., consistency between age and educational attainment), validity (a value is valid if it falls within the specific definition of that attribute);
3. Use (i.e. the use made of the data). We refer, first of all, to accessibility, that is, the possibility of being able to access and, therefore, use the data. This dimension also concerns the format in which the data are distributed (i.e., whether the data are in a format that is widely used within the relevant scientific community). In addition, integrability, the ability to integrate multiple data sets; interpretability, the use of clear, shared and stable definitions; the possibility of rectification, the presence of procedures to correct data in case of errors; relevance, the ability of data to respond to the interests of the community; timeliness, data made available as soon as possible.

In order to meet the demands of data quality, additional guidelines have been introduced, useful, above all, in the phase of sharing and, therefore, of secondary analysis. These guidelines are inspired by the principles defined in the acronym FAIR for a data management that makes information content effectively available, accessible, interoperable and reusable[9]. The actual

[9] Specifically, the acronym FAIR refers to the following properties of data: findable (i.e., existing, having DOI cataloging, being described with metadata), accessible, interoperable, and reusable. These guidelines have also been adopted by the European Commission for data management in projects funded under the Horizon 2020 Program.

possibilities of implementing the FAIR principles in practice are also affected by other factors, such as, for example, logistical constraints. In this regard, it is useful to recall the problem related to the impossibility of long-term preservation of all data used during a research project. The main reasons for this impossibility are of two types: the first one concerns the huge physical space needed to store the data; the second reason is related to the cost of managing these spaces. Therefore, it is necessary to choose which data to store and make them always accessible.

An important aspect of data quality, in addition to those previously mentioned, is the categorization of information: since it is not always possible to distinguish different meanings of the same word, it is important to ensure syntactic and semantic accuracy of data. In the case of syntactic accuracy, the interest does not lie in the evaluation of the value of the data with the real value, but the comparison is with the set of all domain values of the attribute. One has, instead, semantic accuracy when the value of the data corresponds with the real value. Consequently, semantic accuracy intrinsically expresses the concept of correctness of the data (Rezzani, 2013).

In addition to a data quality problem, the reuse of this kind of databases involves security and privacy issues. In fact, one of the most recurrent concerns, when using Big Data is the protection of the privacy of people. If in the Orwellian imaginary every individual lived in a condition of absolute lack of privacy because any noise and / or movement (that was not made in the dark) would be heard and / or seen by "Big Brother" in our society, increasingly connected, the increasing use of Big Data, generated by an increasing number of sensors, of various orders and sizes, located in the environment that surrounds us or in our possession makes almost trivial scenarios described by Orwell[10] (Palanza, 2016). Speaking of personal data protection, it is well known that one of the aspects on which we do not compromise is the so-called requirement of notice and consent to data collection, which is based on the principle that users must be warned that their data may be collected[11]. When we

[10] The British writer George Orwell, whose real name was Eric Arthur Blair, in 1948 produced his famous, and incredibly current, book: 1984 in which he described a society governed by an infallible and omniscient entity, that no one knew, called "Big Brother". Every house had a "tele-screen" capable of receiving and transmitting, and anyone in the field of vision of this instrument could be both seen and heard.

[11] In this regard, we point out that on May 25, 2018, the new European regulation relating to the protection of personal data of natural persons and their free movement came into effect. This new regulation has imposed stringent obligations and introduced new responsibilities aimed at ensuring greater security measures to the protection of personal data. In fact, the regulation has introduced clearer rules on information and

talk about Big Data, this requirement would lose its meaning for some jurists in the phase of data collection because, as it is known, Big Data, by their nature, come from a heterogeneity of sources (Focarelli, 2015). The problem would arise in the analysis and storage phase. An example of criticality arising from particular models of analysis of Big Data, essentially based on algorithms concerns a project of analysis of judgments issued by the Supreme Court of the United States[12]. On the basis of a series of statistical models, a group of researchers succeeded in predicting with accuracy 75% of the sentences of the judges by analyzing a series of databases that collected sentences and other legal information.

Another area of criticality regarding the protection/security of large databases concerns data coming from social networks. Again, in the United States, during the last elections some analysts showed how the "personalized" political messages created through digital traces found in Facebook profiles, e-mails, sites viewed, contributed to support candidates considered by pre-election polls to be at a disadvantage compared to their opponents (Lucchini, Matarazzo, 2014; Palanza, 2016).

Underlying the concept of Big Data are three epochal changes that must be considered when discussing whether or not such data can be reused. The first change involves the shift from the sample to the universe. With Small Data, given the impossibility of processing all possible information, representative samples of the population were selected whose analysis provided statistically significant results on the total estimated universe of reference. The second change concerns the passage from the so-called order to disorder: Small Data were necessarily ordered, as they were collected with method and rigor by the interviewer in tabular form.

The real challenge today, therefore, does not lie in the number of observations available, but in the ability to extract meaning and sense from the mass of data. The current situation of social research, characterized by a huge amount of available data, undoubtedly raises the problem of data quality and consequently the issues related to their sharing and reuse. Therefore, it is not the data that are revolutionary, but the ability to analyze them. Secondary analysis, carried out according to the appropriate and correct procedures, assumes, therefore, a fundamental role since it is the model that is elaborated that contributes to give value to the data. In other words, the challenge consists in the ability to attribute a sociological meaning to databases produced by

consent, defined the limits of automatic processing of personal data and also established strict criteria (and sanctions) in the event of personal data violations.

[12] The team was coordinated by Theodore W. Ruger, a professor at the University of Pennsylvania.

multiple sources, overcoming the self-referentiality typical of the single systems of reference from which these sources originate (Agodi, 2010).

References

Agodi, M. C. (2010), L'estrazione di dati dalla Rete: una nota introduttiva, *Quaderni di Sociologia,* 54, 11-21.

Anderson, C. (2008), The end of theory. Will the data deluge make the scientific method obsolete? *Wired Magazine*, Issue 16.07 (available at: http://edge.org/3rd_culture/anderson08/anderson08_index.html)

Baldassare, M. (2016), Think big: learning contexts, algorithms and data science, *Research on Education and Media*, 8, 2.

Baker, P. (2014), *Data Divination: Big Data Strategies*, Cengage Learning, Boston

Becker, G. (1994), *Human capital: a theoretical and empirical analysis with special reference to education*. Third Edition, The University of Chicago Press, Chicago.

Dutton, J., Roberts, L.M., Bednar, J. (2010), Pathways to Positive Identity Construction at Work: Four Type of Positive Identity at Building of Social Resources, *Academy of Management Review*, 35, 2, 265-293.

Etzioni, O. (2016), The Elephant in the room: getting value from Big Data. *ACM SIGMOD* Blog available at http://wp.sigmod.org/?p=1519.

European Data Protection Supervisor, (2015), Meeting the Challenges of Big Data. A Call for Transparency, User Control, Data Protection by Design and Accountability, *Opinion* 7/2015, 19 November 2015.

Focarelli, C. (2015), *La privacy. Proteggere i dati personali oggi*, Il Mulino, Bologna.

Frixione, M. (2007), *Come ragioniamo*, Laterza, Bari.

Griliches, Z. (1977), Estimating the returns to schooling: some econometric problems, *Econometrica,* 45, 1, 1-22.

Griliches, Z., Mason, W. (1972), Education, Income and Ability, *Journal of Political Economy,* 80, 3, S74-S103.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D. (2008), Web Science: An Interdisciplinary Approach to Understanding the Web, *Communications of the ACM,* 51, 7, 60-69.

Hey, A., Tansley, S., Tolle, K. (Eds.) (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery,* T. Redmond, VA: Microsoft Research, available at: http://fourthparadigm.org.

Hillard, R. (2012), *Information-driven business*, John Wiley e Sons disponibile a: www.infodrivenbusiness.com)

Independent Expert Advisory Group, (2014), *A World that Counts. Mobilising the Data Revolution for Sustainable Development*, available at: http://www.undatarevolution.org/report/.

Karr, A.F., Sanil, A.P., Banks, D.L. (2006), Data quality: A statistical perspective, *Statistical Methodology*, 3, 2, 137-173.

Kuhn, T. (1962), *The structure of scientific revolution*, Chicago University Press, Chicago.

Lucchini, S., Matarazzo, M. (2014), *La lezione di Obama. Come vincere le elezioni dell'era della politica 2.0.*, Baldini & Castoldi, Milano.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, A., Byers, H.l (2011), Big Data: the next frontier for innovation, competition and productivity, *Mckinsey Digital*, available at: http://www.mckinsey.com/big-data-the-next-frontier-for-innovation.

Manovich, L. (2011), Trending: The Promises and the Challenges of Big Social Data, *Debates in the Digital Humanities*, 2, 460-475.

Mayer-Schönberger, V., Cukier K. (2013), *Big Data. A Revolution That Will Transform How We Live, Work and Think*, Houghton Mifflin Harcourt, New York.

Oldroyd, D. (1986), *The arch of knowledge: an introductory study of the history of the philosophy and methodology of science*, Methuen, London.

Opresnik, D., Taisch, M. (2015), The value of Big Data in servitization, *International Journal of production economics*, 165, 174-184.

Palanza, S. (2016), Internet of things, Big Data e privacy: la triade del futuro, *Documenti IAI,* ottobre 2016.

Peirce, C.S. (1960), *Collected papers of Charles Sanders Peirce*, Harvard University Press, Cambridge.

Popper, K. (1959), *The logic of scientific discovery*, Routledge, London.

Prensky, M. (2009*),* H. Sapiens Digital: from digital immigrants and digital natives to digital wisdom*, Journal of online Educations*, 5, 3.

Rezzani, A. (2013), *Big Data. Architettura, tecnologie a metodi per l'utilizzo di grandi basi dati*, Maggioli, Santarcangelo di Romagna.

Russel, B. (1912), *The problems of philosophy*, Williams and Morgate, London.

Snijders, C., Matzat, U., Reips, U. D. (2012), Big Data: big gaps of knowledge in the field of internet science, *International Journal of Internet Science*, 7, 1, 1-5.

The Royal Society, (2012), Science as an open enterprise, *Science Policy Centre Report,* 2, 12.