Studies in Classification, Data Analysis, and Knowledge Organization

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia
Biagio Aragona · Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Studies in Classification, Data Analysis, and Knowledge Organization

More information about this series at http://www.springer.com/series/1564

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia · Biagio Aragona
Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Editors
N. Carlo Lauro
Department of Economy and Statistics
University of Naples Federico II
Naples
Italy

Enrica Amaturo
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Maria Gabriella Grassia
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Biagio Aragona
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Marina Marino
Department of Social Sciences
University of Naples Federico II
Naples
Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-319-55476-1 ISBN 978-3-319-55477-8 (eBook)
DOI 10.1007/978-3-319-55477-8

Library of Congress Control Number: 2017937698

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data Science is a multidisciplinary approach based mainly on the methods of statistics and computer science suitably supplemented by the knowledge of the different domains to meet the new challenges posed by the actual information society. Aim of Data Science is to develop appropriate methodologies for purposes of knowledge, forecasting, and decision-making in the face of an increasingly complex reality often characterized by large amounts of data (big data) of various types (numeric, ordinal, nominal, symbolic data, texts, images, data streams, multi-way data, networks, etc.), coming from disparate sources.

The main novelty in the Data Science is played by the role of the KNOWLEDGE. Its encoding in the form of logical rules or hierarchies, graphs, metadata, and ontologies, will represent a new and more effective perspective to data analysis and interpretation of results if properly integrated in the methods of Data Science. It is in this sense that the Data Science can be understood as a discipline whose methods, result of the intersection between statistics, computer science, and a knowledge domain, have as their purpose to give meaning to the data. Thus, from this point of view, it would be preferable to speak about DATA SCIENCES.

The Data Science and Social Research Conference has represented an interdisciplinary event, where scientists of different areas, focusing on social sciences, had the opportunity to meet and discuss about the epistemological, methodological, and computational developments brought about by the availability of new data (big data, big corpora, open data, linked data, etc.). Such a new environment offers to social research great opportunities to enhance knowledge on some key research areas (i.e. development, social inequalities, public health, governance, marketing, communication).

Along, the conference has been a crucial issue to discuss critical questions about what all this data means, who gets access to what data, and how data are analysed and to what extent.

Therefore, aim of the conference, and of the present volume, has been to depict the challenges and the opportunities that the “data revolution” poses to Social Research in the framework of Data Science, this in view of building a SOCIAL DATA SCIENCE … Let us own data science!

Naples, Italy

N. Carlo Lauro
Professor Emeritus of Statistics

Contents

Introduction

Enrica Amaturo and Biagio Aragona

One of the fundamental features of modern societies is the never-ending quantity of data they produce as direct and indirect effects of business and administrative activities, as well as the result of volunteer accumulation of information on the Internet by individuals who use the Web for social relationships and knowledge construction. Changes in social networking and the pervasive use of the Web in daily life, as well as improvements in computational power and data storage, are having impressive effects on data production and consumption. Social networks, sensors, and data infrastructure are generating a massive amount of new data (big data, big corpora, linked data, open data, etc.) that are readily available for the analysis of societies. That is why some talked about a "data deluge" (The Economist 2010) able to radically change both the individual and the social behaviours, and others (Kitchin 2014) have labelled the present time as "the data revolution era". Data revolution is the sum of disruptive social and technological changes that are transforming the routines of construction, management, and analysis of data once consolidated within the different scientific disciplines.

Digital technology for scanning, processing, storing, and releasing data has already had an impact on the quality of information available to social researchers. Other opportunities are opened by computational changes that have a radical effect on the nature of dissemination by allowing to deal with large data even for small areas (Uprichard et al. 2008) and to make data storage possible also to individuals and small businesses. The wide availability of software for analysis also has to be considered when drawing the picture of the new possibilities offered by technological changes that affect the production and consumption of data (Baffour et al. 2013).

E. Amaturo · B. Aragona (✉)
Department of Social Sciences, University of Naples Federico II,
Vico Monte di Pietà, 1, Naples, Italy
e-mail: aragona@unina.it

E. Amaturo
e-mail: amaturo@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_1

A first epistemological consequence of the passage from data scarce to data-intensive societies has been the re-emergence of data-driven science, which is opposed to hypotheses-driven science that is typical of post-positivist social science. The main argument of those who proclaim the "data first" model of science is that, being able to track human behaviour with unprecedented fidelity and precision, exploring existing data may be more useful than building models of why people behave the way they do. More specifically, in 2009, Lazer—in a paper that had great success within the scientific community (more than 1.527 citations)—identified Big Data as the core of a new field of social science which makes intensive use of computer science (computational social science (CSS)). For him, these vast data sets on how people interact were offering new perspectives on collective human behaviour.

As the availability of big quantities of data has grown, the main traditional empirical basis of quantitative social sciences (surveys and experiments) is being dismantled in favour of new data analysis. Market research, for example, widely employs studies on network communities instead of traditional survey's campaign and network, and sentiment analysis is substituting for traditional election pools, which proved to be less effective than they were in the past. Not to mention how documents' analysis has really changed with the advent of the Web and of social media (Amaturo and Aragona 2016). Because nearly all of our activities from birth until death leave digital traces in large databases, social scientist, who had to rely on account of actions for their research (through questionnaire), using new data can be in the action without asking questions or being seen.

After the early enthusiasm about the data deluge, in the past years critical data studies have been carried out to more deeply understand what is the context of validity of new data. Special attention has been paid, for example, to voluntarily generated contents on social network and Websites. They represent a massive quantity of data, but they need to be contextualized; otherwise, it becomes difficult to make sense of them. Moreover, despite the often made claim that Big Data provides total populations, ending our reliance on samples, this is rarely the case for social media data (Highfield et al. 2013). Boyd and Crawford, for example, have noted that working with Twitter data has: "serious methodological challenges that are rarely addressed by those who embrace it" (2012: 13) and that "Twitter does not represent people and it is an error to assume people and Twitter users as synonymous: they are a very particular sub-set" (2012: 12). When using data coming from the Web, researchers must recognize that part of the population is not accessible because does not have access to the Internet and that many are passive consumers of Internet information rather than active participants in the Web 2.0. Access may also be segmented according to socio-demographic characteristics (nationality, age, gender, education, income, etc.), systematically excluding some strata of the population from research. Surveys in the USA, for instance, show that Twitter has a disproportionate number of young, male black, and Hispanic users compared to the national population (Duggat et al. 2015).

These methodological concerns about validity and coverage biases rise also more deep sociological and political questions about to what extent these data may be used for the analysis of society, what aspects of the social reality they capture, and how they can be customized for designing, implementing, monitoring, and evaluating social policies. *La gouvernance par les nombres* (Supiot 2016) may be crucial to understand what will be the future of new data within both social sciences and society. Indeed, new digital data have been "normalized" within administrations, as showed by proliferating database-related technologies of governance. They are complementing existing uses of data with methods of digital governance, whereby digital technologies, software packages and their underlying standards, code, and algorithmic procedures are increasingly being inserted into the administrative infrastructure of our societies.

One interesting example is the fact that administrations, through local statistical offices, are giving access to their micro-data and have started to finance open-data initiatives with both a cognitive and a normative intent. On one side, open data help technicians, administrators, and politicians to redirect policies and, on the other side, allow citizens to check whether policies have had the desired impact. Opening data is therefore a consequence of the importance of transparency and accountability in our societies.

Another example concerns how Big Data have captured the interest of National Statistical Institutes (NSI) and related agencies such as Eurostat and the European Statistical System (ESS), who have formulated a Big Data roadmap. United Nations Economic Commission for Europe (UNECE) has established a High Level Group for the Modernization of Statistical Production and Services focused on Big Data with four "task teams": privacy, partnerships, sandbox and quality. Even the United Nations Statistical Division (UNSD) has organized a Global Working Group on Big Data and Official Statistics. The interest of official statistics is due to the fact that the developments in ICT help to handle these data sources and hence allow to drastically reduce the costs of statistics. However, a survey jointly conducted by UNSD and UNECE revealed that of the 32 NSIs that responded only a "few countries have developed a long-term vision for the use of Big Data", or "established internal laboratories, task teams, or working groups to carry out pilot projects to determine whether and how Big Data could be used as a source of Official Statistics" (EUESC 2015: 16).

A part from the efforts that NSI are doing in inserting Big Data in their statistical production, the use of Big Data, both structured and unstructured, can represent a valuable way to inspire decision-making at all level of public administration in a time of scarce resources. The technological revolution is in fact enabling governments to use a great variety of digital tools and data to manage all phases of the policy cycle's process more effectively, becoming a core element for e-governance applications and techniques. It has been widely claimed that this radical expansion of digital data is transforming the global evidence base and will lead to improved knowledge, understanding, and decision-making across the economy, in turn improving life chances and well-being for individuals and for the health and sustainability of economies and societies more broadly (Mayer-Schönberger and Cukier 2013; Margetts and Sutcliffe 2013). However, still more research is needed about what kind of analytics can be usefully managed, at what policy level they are really demanded, how they are collected, organized, integrated, and interrogated, by whom and for what purposes. Data are not useful in and of themselves. It is what is done with data that is important, and making sense of new data poses new analytical challenges.

First of all, new data usually require more attention to the processes where data have to be pre-prepared for analysis through data selection, curation, and reduction activities. Pre-analytical work can be extremely hard and time-consuming, so data scientists are devoting more research to seek the most productive, efficient and effective ways to undertake and especially, to automate, this work. Furthermore, the analysis of very large numbers of data records can be timely run only by computer algorithms, and then, much work is about developing automated processes that can assess and learn from the data and their analysis, the so-called machine learning. Machine learning seeks to iteratively evolve an understanding of datasets and is been used for data mining in order to detect, classify, and segment meaningful relationships, associations, and trends between variables. Data mining may employ a series of different techniques including natural language processing, neural networks, decision trees, and statistical (nonparametric as well as parametric) methods. The selection of techniques varies according to the type of data (structured, unstructured or semi-structured) and the objective of the analysis. Unstructured data in the form of natural languages raise particular data mining challenges; they need semantics and taxonomies to recognize patterns and extract information from documents. A typical application of such technique is sentiment analysis which seeks to determine the general nature and strength of opinions about an issue.

Another analytical challenge is about data visualization and visual analytics. Visual methods effectively communicate the structure, pattern, and trends of variables and their relations. Visualization created within the digital sphere can be used to navigate and query data, enabling users to gain an overview of their data. Visualization may also be used as a form of analytical tool, visual analytics, guided by a combination of algorithms and scientific reasoning which work to extract information, build visual models and explanations, and guide further statistical analysis (Keim et al. 2010). The last but not least challenge is about the stock of descriptive and inferential statistics that have traditionally been used to analyse traditional data. They are also being applied to new data though this is not always straightforward because many of these techniques were developed to draw insights form relatively scarce rather than exhaustive data. Further research is thus required to generate new methods or innovative combination of techniques that can make sense of and extract value from Big Data and data infrastructures.

These challenges are not simply technical, because analytics are the expression of a particular epistemology; therefore, both technical research and epistemological research are required to tackle the challenges of the data revolution. Data revolution is being a great opportunity of innovation of the social sciences. First of all, because it empowers the empirical base of social disciplines, furthermore, because it promotes interdisciplinarity between different areas of science, enhancing integration of data and methods. Only by mixing social theory and computation, data and modelling in an innovative way, social scientists can contribute to a clearer vision of social processes and to the quality of public choices, integrating the more traditional approaches already practiced in social research.

The volume aims to represent the complexity of the whole spectrum of epistemological, technical, and analytical challenges and opportunities that the datafication of society is posing to social sciences. The first section of the volume concentrates on the changes that new data have made to the core of the scientific method and to the theoretical and methodological assumptions that are behind these changes. Moreover, new theoretical reasoning is presented, also on the use of new data for the governance of public policies.

The second section is on methods, software, and data architectures to extract knowledge from data. All the chapters in this section concentrate with the difficult work of data preparation and data curation, focusing on how to manage different forms of data both in structured and in unstructured forms. More specifically, while some contributions deal with the construction of data matrixes ready for statistical analysis, others present softwares or techniques that can help in analysing the different kinds of new data.

The third and the forth parts of the volume present a series of applications. While section three focuses more specifically on the data of the Web (social network data, Web pages and so on), the contributions in section four are applications on data infrastructures' data or data produced by statistical offices. More specifically, in the third section, great relevance is devoted to the techniques such as sentiment analysis, lexical content analysis, and to some innovative efforts to combinate them with social network analysis. The fourth section deals more in depth with the issues of access, integration, and visualization of big databases concentrating both on the analytics required to make sense of them (visualization as well as traditional statistics techniques) and on the techniques needed for their construction.

References

Amaturo, E., & Aragona, B. (2016). La "rivoluzione" dei nuovi dati: quale metodo per il futuro, quale futuro per il metodo?. In F. Corbisiero and E. Ruspini (Eds.), *Sociologia del futuro. Studiare la società del ventunesimo secolo* (pp. 25–50). Milano: CEDAM.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society*, *XV*(5), 662–679. doi:10.1080/1369118X.2012.678878.

Highfield, T., Harrington, S., & Bruns, A. (2013). Twitter as a technology for audiencing and fandom. *Information Communication and Society*, *16*(3), 315–339.

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age solving problems with visual analytics*. Eurographics Association.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *I*(1), 1–12. doi:10.1177/2053951714528481.

Lazer, D., et al. (2009). Life in the network: The coming age of computational social science. *Science*, *CCCXXIII*(5915), 721–723.

Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of "Big data". *Policy & Internet*, *5*(2), 139–146.

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

Supiot, A. (2016). *La Gouvernance par les nombres*. Paris: Fayard.

The Economist. (2010). *The data deluge: Businesses, governments and society are only starting to tap its vast potential*, print edition.

Uprichard, E., Burrows, R., & Parker, S. (2009, February 25). Geodemographic code and the production of space. *Environment and Planning A*, *41*(12), 2823–2835.

Part I
Epistemology

On Data, Big Data and Social Research. Is It a Real Revolution?

Federico Neresini

Abstract This chapter aims at discussing critically some epistemological assumptions underlying a data science for social research. For this purpose, it is discussed the general notion of big data and the meaning of key-concepts such as those of information and data, mainly considering contributions coming from the science and technology studies (STS) and the sociology of quantification. In particular, it is argued the necessary shift from a discrete and transportable definition of data to a processual one, also taking into account the fact that data are always a process both when they are produced and when they are used/analysed in order to have research's results. The notion of data-base is compared with that of infrastructure as defined in STS, so that it is clear that they cannot be considered as repositories from which it is possible to extract meanings or results like getting minerals from a mine. Data and data-base are processes which cannot begin without a research question. For these reasons the debate opposing hypothesis-driven versus data-driven research should be overtaken: in social research, as well as in hard sciences, data-driven research simply doesn't exist. The last paragraph is devoted to draw some conclusions from the previous discussion in the form of hopefully useful suggestions for developing a data science for social research.

Keywords Big data • Data-base • Infrastructures • Data-driven/hypothesis driven research • Quantification

Answering the question posed by the title of this contribution might seem easy and straightforward: yes.

In fact it is hard not to recognize that the fast growth of digital data and their increasing availability have opened a new season for social sciences. The unceasing expansion of "datification" or "quantification" (Espeland and Stevens 2008) makes it possible that, for the first time in its history, social research has available a huge amount of data, not only regarding a great variety of phenomena, but also directly

F. Neresini (✉)
FISPPA Department, University of Padua, Padua, Italy
e-mail: federico.neresini@unipd.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_2

and "naturally" generated for the most part by social actors producing those phenomena. The volume of this spontaneous generation of digital data is truly striking: according to some estimates, every minute Google performs 2 million searches and 72 h worth of video is uploaded to YouTube; at the same time there are 1.8 million likes on Facebook, 204 million emails sent and 278,000 tweets posted.[1]

It was hence quite easy to predict that this almost sudden abundance of digital data would attract the interest of many social scientists, as proved by the flourishing of research centres established to exploit this new opportunity and the array of articles in which "big data" are involved.[2]

As a counterbalance to this enthusiasm there have not been lacking—of course, and fortunately—critical reflections, calling attention to the limits of "data-driven" social research (see for example Boyd and Crawford 2012) and to the problems deriving from the quantification processes (see, among others, Espeland and Sauder 2007; Lampland and Star 2009), highlighting the implicit assumptions laying behind the production of digital data by the social media platforms (Gillespie 2014) and the methodological traps to which researchers using, without the necessary awareness, those data and the automatic tools required for handling large amount of digital data are exposed (Giardullo 2015).

It is interesting to note that the debate on big data and social research is proposing again, almost without differences, those arguments that developed at the beginning of the new millennium within the molecular biology research field, and that it is not yet concluded.

The two parties are deployed in two opposite lines: on one side those who are maintaining the so-called "data-first" approach (Golub 2010), and, on the other side, those who are instead affirming the supremacy of the research questions both strategically and operatively orienting the work in the laboratories (Weinberg 2010). This opposition between "data-driven" and "hypothesis-driven" research clearly recalls what was proposed, back in 2008, by Chris Anderson—in a provocative way—as "the end of theory": "With enough data, the numbers speak for themselves" (Anderson 2008).

Already in 2001, John Allen was wondering whether: "With the flood of information from genomics, proteomics, and microarrays, what we really need now is the computer software to tell us what it all means. Or do we?" (Allen 2001). The same question could represent what we are now debating in the case of social sciences; it is enough to substitute the data source: with the flood of information from the web, the official statistics and the record of a huge amount of social activities, what we really need now is the computer software to tell us what it all means. Or do we?

But, this way of addressing the problem, as well as the opposition of data-driven versus hypothesis-driven research and the almost exclusive focus on how to handle

---

[1]See for example http://blog.qmee.com/qmee-online-in-60-seconds/ (05.06.2016).

[2]Between 2000 and 2015 there were published 2630 articles related to "big data" in the field of social sciences, 1087 only in 2014 and 2015 (Scopus).

data, produce the effect of leaving in the background the fact that data do not exist by themselves, being rather the outcome of a very complex process in which producing and using data are so deeply intertwined that they cannot be considered separately.

Nevertheless, we are inclined to treat distinctly the production of data—i.e. their collection—and the use of them—i.e. their analysis; and this distinction not only induces to paying more attention on the side of data-analysis, but it implicitly suggests also the idea that data simply are there, and that the only problem is how we can use them and with what consequences.

But, first of all, we should not forget that data—no matter how big or small they are—are always the result of a construction process, as should be obvious for social sciences and as is very clear also in the case of the so-called "hard sciences", at least in the wake of science and technology studies (STS).

Second, using and producing data cannot be considered separately because, on the one hand, the production process affects the possibilities of using data, and, on the other hand, the need to have data to be utilized affects the way they are produced. At the same time, focusing on both sides of producing and using data allows us to pay due attention to what data are, instead of taking them for granted.

Data which populate data-bases available for social sciences today are, in fact, the result of a long and complex process of manufacturing; moreover, the fact that social scientists increasingly seek to use those data for producing new knowledge—together with the fact that these attempts imply a range of problems regarding their accessibility, how to perform queries, the quantity and quality of meta-data, statistical techniques for reducing the complexity associated with their quantity, and the certainly not trivial interpretative work required for making sense of the outputs obtained from data-bases—all of these aspects testify that data entered in a data-base do not live by themselves, but depend on the fact that someone is utilizing them. This is a key point, even if it is very easy to think about "data" as "what remains at the end of these processes", while, on the contrary, at the end of these processes, nothing remains, because data *are* the process.

1 Data-Bases Are not a Repository

In order to justify the last statement and to explore what it actually implies with regards to the development of a data science for social research, it can be useful to focus our attention on what we still think of as—and therefore still treat as—"bags of data", i.e. "data-bases".

The reflection on data-bases has been developed by STS in the field of hard sciences, so that some interesting conclusions they reached can be regarded here as very interesting. It is not by chance that what is going on in the hard sciences can be observed also in the case of the social sciences.

As a starting point, we can refer to this passage by von Foerster, which fits perfectly with the aim of looking at big-data in a critical perspective and, in this case, specifically addressing the relationship between the intrinsic characteristics of what we are used to calling "data" or "information" and their supposed deposits (data-bases):

> Calling these collections of documents 'information storage and retrieval systems' is tantamount to calling a garage a 'transportation storage and retrieval'. By confusing *vehicles* for potential information with information, one puts again the problem of cognition nicely into one's blind spot of intellectual vision, and the problem conveniently disappears (von Foerster, 1981, p. 237).

So a data-base does not contain data or information, exactly as a garage does not contain transportation, because data, as well as data-bases, are nothing but processes, as has been made very clear by Shannon and Weaver as long ago as 1949:

> Information in communication theory relates not so much to what you do say, as to what you could say. That is, information is a measure of one's freedom of choice when one selects a message. If one is confronted with a very elementary situation where he has to choose one of two alternative messages, then it is arbitrarily said that the information, associated with this situation, is unity. Note that it is misleading (although often convenient) to say that one or the other message conveys unit information. The concept of information applies not to the individual messages (as the concept of meaning would), but rather to the situation as a whole, the unit information indicating that in this situation one has an amount of freedom of choice, in selecting a message, which it is convenient to regard as a standard or unit amount (Shannon and Weaver, 1949, p. 5).

Hence, precisely as in the case of information, data are not discrete entities, which can be treated as "packages" that can be transmitted from one point to another, or which can be collected and stocked in a deposit, or which can be extracted like precious minerals from a mine. Nevertheless still we substitute the unit of measurement, i.e. a quantity (bit), for what is measured, i.e. the process which, in the case of information, corresponds to reducing uncertainty.

As everybody knows, dimensions actually matter for big-data, supported by a long strain of increasing measures: giga-byte, peta-byte, exa-byte … but very few seem interested in the fact that the unit on which all these measures are based is a process, as clearly stated by Shannon and Weaver. It is possible to find the same conclusion within STS where there is a long array of studies showing the eminently "processual" character of data and data-bases in scientific research and therefore the necessity of not treating data-bases as mere repositories of information. Not only because "raw data is an oxymoron" (Gitelman 2013), but also because data as fixed entities, available for being transferred, transformed or simply used, do not exist. Information—or data—are not discrete elements, well established in time and space, but seamless processes of production and use; outside this process there are no data—nor information.

Being aware of this might lead us to avoid the risk of imperceptible, but—exactly for this reason—very insidious, meaning inversions like that we can see in this passage within a interview by Viktor Mayer-Schoenberg.[3] He maintains that for defining big data we should think about it as follows:

> it's like taking millions of fixed images and mounting them in a movie. The individual fragments, gathered together, take different forms and meanings. This is what happens with the data: the ability to work with a huge amount of numbers allows us to obtain billions of points of view on the world and then to understand it better. Until some time ago it was very expensive and difficult, but new technologies have made these procedures within the reach of many.[4]

We can see here a clear example of the inversion that is the basis on which big data are approached uncritically and naively: data allow us to obtain the points of view, instead of it being the points of view that allow us to generate the data. But a "point of view" is the inescapable starting point of the process which gives rise to data; at the same time, data are not the ending point: they are the process, and therefore we cannot split the expression "processing data" into "processing" and "data" without losing both data and process.

Looking at data which are at stake in doing social research when the data are a huge amount, suggests thinking about a data science for social research as an expression of what has been referred to as "virtual knowledge" and analyzing its relationship to "infrastructure":

> Virtual knowledge is strongly related to the notion that knowledge is embedded in and performed by infrastructures. (…) The infrastructures that are now taking shape are not developed to support well-defined research projects as to the generations of streams of yet undefined research. Most of the data infrastructures that have been built so far have promised the discovery of new patterns and the formation of new-data-driven research. (…) Increasingly, infrastructures and their component network technologies try to support possibility rather than actuality (Wouters, Beaulieu, Scharnhorst, Wyatt, 2013, p. 12).

The concept of "infrastructure", a notion which is clearly and strictly bound up with that of data-base (Mongili and Pellegrino 2015), was introduced into the STS field by Star and Ruhleder almost 20 years ago as follows:

> an infrastructure occurs when local practices are afforded by a larger-scale technology, which can then be used in a natural, ready-to-hand fashion. It becomes transparent as local variations are folded into organizational changes, and becomes an unambiguous home - for somebody. This is not a physical location nor a permanent one, but a working relation - since no home is universal (Star and Ruhleder, 1996, p. 114).

So data-bases, together with all their outfit of standards and routines, are exemplary cases of scientific infrastructures, and also they—as well as data produced, gathered and utilized by them—can exist until they are "in-action".

Moreover, the processual character of data-bases depends also on some aspects intrinsically pertaining to all "things" which can be categorised. It has been pointed

---

[3]He is the co-author with Cukier of the recently published book "Big-Data: A Revolution That Will Transform How We Live, Work and Think" (Mayer-Schoenberger and Cukier 2012).

[4]La Lettura, Il Corriere della Sera, 01.09.2013, p.14 (our translation).

out by Bowker that many "things" are hard to classify, others do not get classified (i.e. data-bases are selective), others get classified in multiple ways (Bowker 2000).

The data of big data are hence discrete representations of fluid realities—which are actually processes of interaction within a network of heterogeneous actors—they are frames of a film which cannot live outside the film; they appear static and this apparent "staticity" is what makes them exchangeable and transportable, in one word mobile, because they seem detached from the context of their production. For this reason, we should not conceive of data-bases as information's repositories, not only because data are always generated along a process in which many heterogeneous actors are involved and during which many "translations" occur (Latour 1987, 2005), but also because or, better, mainly because they only exist as processes, and the same goes for the informational infrastructures called data-bases.

2 Some Consequences for Building a Data Science for Social Research

The previous reflections regarding data and data-bases provide the opportunity for pointing out some consequences in order to develop a data science for social research upon fruitful assumptions.

First of all, it is important to stress again the centrality of research questions, and not for abstract reasons related to a supposed supremacy of theory, but mainly because questions play a strategic role in generating data: they create the conditions for facing an uncertainty to be reduced, i.e. for triggering the process through which data are produced and utilized, in both cases through a long array of tools.

Second, we should not forget that those tools—which in the case of big data become digital, as with search engines and their algorithms—are not neutral devices we can decide to use or not. On the one hand we simply cannot have data without this kind of tools; on the other hand, it is not true that "the Internet has no curriculum, no moral values, and no philosophy. It has no religion, no ethnicity, or nationality. It just brings on the data, railroad cars of it, data by the ton" (Sterling 2002, p.51). The Internet only "eclipses intermediaries" (Pariser 2011, p.53). Search engines—Google *in primis*—and other digital tools are not neutral devices, they always offer a selection of the world's complexity, a selection which is constructed at least for answering in a personalized way needs they ascribe to us as profiled users.

As a third point, the processual character of the digital data with which social research would like to work as well as the un-neutrality of the tools required for retrieving, collecting and processing them make clear what social sciences knew from the very beginning, even if they seem sometimes to forget it: the instruments used for processing data are intrinsically implied in the process of their construction. Put another way, there are not first data, then tools for collecting them, then those for analysing them and finally the results; on the contrary, data which we trust in order to obtain our results depend not only on the questions from which we start, but also on the tools we use for processing them. Exactly as data collected through a questionnaire and analysed with dedicated software are produced both by the questionnaire and by the software, in the same way data pertaining to the social media are produced by the algorithms of the digital platforms on which we "find" them, by the digital tools utilized for processing them and by our research questions, as well as this last depending on the availability of data shaped by the platform and by the tools used for processing them. So yes, questions first, even if questions are not independent from how data are produced and from the tools that can be used.

Furthermore, the heterogeneity of the actors involved in the processes of data construction and data utilization puts forward a very strong argument in favour of the fact that a data science for social sciences cannot be bounded within a single disciplinary domain. As a consequence, an interdisciplinary perspective cannot be avoided, maybe even less so than in the past. But interdisciplinarity is a time-consuming enterprise because it requires a great investment of resources in terms of intermediation among various actors, interests, points of view. It could seem a paradox that in the age of real-time interconnectivity, of fast and easy access to so many digitalized data through the web, of computational power, in short in the era of "speed data", we are requested to be aware of the fact that doing research is a matter of time. It is not by chance that in 2010 many scientists signed the "slow science manifesto": "We do need time to think. We do need time to digest. We do need time to misunderstand each other, especially when fostering lost dialogue between humanities and natural sciences", as is the case of a data science for social research.[5]

It should be clear, therefore, that the necessary interdisciplinarity for a data science even in the field of social research cannot be realized simply by putting together researchers with different training, or proposing training opportunities just as "one near another" classes of different disciplines in a curriculum. Also interdisciplinarity is, in fact, a process which requires time for building it; researchers have to find a new common domain in which they can actually "work together". This process, like any other process, must be fed by motivated actors and must be supported by favourable structural conditions. It means that, for example, it is important to invest in training opportunities for raising a new generation of researchers who have deeply experienced interdisciplinarity, i.e. not offering them just a patchwork of contributions coming from different fields. Moreover, and again as an example, articles published in journals outside the main field of their authors should be recognized institutionally as a valuable contribution and therefore should be considered as relevant in the evaluation exercises devoted to measuring scientific productivity.

---

[5]http://slow-science.org/ (accessed 16.06.2016).

In other words: building a data science for social research needs not only data and methodological solutions, but also resources, strategically allocated in a long-term strategy of scientific policy which cannot rely only on the goodwill of some social scientists.

References

Allen, J.F. (2001). Bioinformatics and discovery: induction beckons again. *Bioessay*, *23*(1), 104–107.

Anderson, C. (2008). The end of Theory. *WIRED, 16,* 07.

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

Bowker, G. (2000). Biodiversity datadiversity. *Social Studies of Science, 30*(5), 643–684.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: how public measures recreate social worlds. *American Journal of Sociology, 113*(1), 1–4.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology, 49*(3), 401–437.

Giardullo, P. (2015). Does 'bigger' mean 'better'? Pitfalls and shortcuts associated with big data for social research. *Quality & Quantity*. doi: 10.1007/s11135-015-0162-8.

Gillespie, T. L. (2014) The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, K.A. Foot (Eds.), *Media technologies* (pp. 168–193). Cambridge (MA): MIT Press.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge (MA): MIT Press.

Golub, T. (2010). Counterpoint: Data first. *Nature, 464,* 679.

Lampland, M., & Leigh Star, S. (Eds.). (2009). *Standards and their stories. How quantifying, classifying and formalizing practices shape everyday life*. Ithaca: Cornell University Press.

Latour, B. (1987). *Science in action*. Cambridge (MA): Harvard University Press.

Latour, B. (2005). *Re-assembling the social*. Oxford: Oxford University Press.

Mongili, A., & Pellegrino, G. (Eds.). (2015). *Information infrastructure(s): Boundaries, ecologies, multiplicity*. Newcastle: Cambridge Scholars Publishing.

Pariser, E. (2011). *The filter bubble. What the internet is hiding from you*. New York: Penguin Books.

Mayer-Schoenberger, V. & Cukier K. (2012). *Big data: A revolution that will transform how we live, work and think*. New York: Houghton Mifflin.

Shannon, C. H., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111–134.

Sterling, B. (2002). *Tomorrow now. Envisioning the next fifty years*. New York: Random House Inc.

von Foerster, H. (1981). *Observing systems*. Seaside: Intersystems Publications.

Weinberg, R. A. (2010). Point: Hypothesis first. *Nature, 464,* 678.

Wouters, P., Beaulieu, A., Scharnhorst, A., & Wyatt, S. (Eds.). (2013). *Virtual knowledge. Experimenting in the humanities and the social sciences*. Cambridge (MA): MIT Press.

New Data Science: The Sociological Point of View

Biagio Aragona

Abstract The objective of this chapter is to introduce the contribution that, apart from post-positivism, other sociological paradigms, such as interpretivism and social constructionism, may give to the development of research and thinking about data science in social research. These two paradigms have theoretical and methodological beliefs that seem unfitted to interpret the data revolution era because they are focused on individuals, *verstehen* and sense of action and have been usually associated with qualitative research. But they may be of great help in addressing the future of new data research in our discipline, especially on two important aspects: first, on how objective new data are and, furthermore, on the role of knowledge in new data use and construction.

Keywords New data • Data culture • Data assemblage • Social constructionism • Post-positivism

1 The Reframing of Social Sciences

Changes in social networking and the pervasive and ubiquitous Web use in daily life, as well as improvements in computational power and data storage, are having impressive effects on data production and usage. Social networks, sensors and data infrastructure are generating a massive amount of new data (big data, big corpora, linked data, open data, etc.) that are readily available for the analysis of societies.

A first consequence of the passage from data scarce to data-intensive societies has been the re-emergence of data-driven science, which is opposed to theory-laden science that is typical of post-positivist social sciences. This form of neo-empiricism implies both an epistemological and an ontological assumption. First of all, it sustains the adherence to inductivism, where a proposition is scientific

B. Aragona (✉)
Department of Social Sciences, University of Naples Federico II, Vico Monte Di Pietà 1, Naples, Italy
e-mail: aragona@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_3

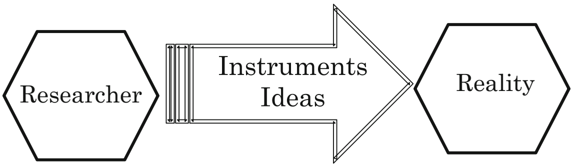

Fig. 1 Data collection in post-positivist view

if proven from facts, so research is data-driven. Furthermore, it supports data objectivity, where data are considered as neutral observation of reality; they are reality.

Neo-positivism and post-positivism, already in the mid of last century, have criticized both inductivism and data objectivity. Popper (1967, 1972) believed that no proposition can be proven from facts and that hypotheses are always theory-laden. Lackatos (1976) regarded empirical support as a three-place relationship between theory, evidence and background knowledge, where the latter is represented by the whole set of facts and parameters used in the construction of any given theory and, it could be added, of any given data.

According to post-positivist thinking, data are not objective and neutral. What we have is the researcher, the reality and, between these two, instruments (methods and techniques) and ideas (theories and empirical hypotheses) (Fig. 1).

As well as neo- and post-positivism started to criticize the ingénue epistemic and methodological view of first positivism, critical data studies started to criticize the neo-empiricism brought about by the present data revolution. Dalton and Tatcher (2014) have called critical data studies those studies that apply critical social theory to data to explore the ways in which they are never neutral, objective, raw representation of the world but are situated, contingent, relational and contextual. In this context, data are considered as complex socio-technical systems that are embedded within a larger institutional landscape of researchers, institutions and corporations (Ruppert 2013). And the inductive method as in the Anderson vision of the “end of theory” (2008) is considered an unsupportable fantasy. For Kitchin, for example, more common is the use of abduction, which enables to fit unexpected findings into an interpretative framework (Kitchin 2014). Abduction is also called the inference of best explanation, and it is for example applied through the Bayesian method of explanation called maximum likelihood. Peirce (1883) gave its first formulation, but for him abduction was typical of the context of discovery, while modern epistemology attributes it to the context of justification.

A further epistemological consequence of data revolution has been a redefinition of the boundaries of disciplines and the foundation of new interdisciplinary fields where technology plays a greater role, such as computational social science and digital humanities. In sociology, data revolution is undermining the already weak boundaries between quantitative and qualitative research. The big data realm in fact blends differences between textual and numerical data. Often numbers, texts and even images merge into the same database. Moreover, user-generated contents on social networks and Websites, which constitute a massive amount of data, are classified and analysed through techniques of sense making and meaning construction that have the features of deep data analytics, which are typical of qualitative research (Boccia Artieri 2015). That is why *mixed methods* (Hesse-Biber and Johnson 2013) researchers, who have always been keen to integration between quality and quantity, are paying great attention to the developments of new data use in sociology.

2 Data as Signification Acts

One difference between the epistemic position of Max Weber and that of neo-positivists and post-positivists is in the role that the point of view has in the connection between data, reality and knowledge. For Lackatos (1976), facts are just facts (data are just data), while the interpretation of facts (data) depends upon economic interests and points of view. For Weber, not only the interpretation of facts depends upon point of views, but also facts depend on point of views. As he has firstly stated (1904), the role of researchers in the construction of data is high and the distinguishing criteria used in their capture have consequences on the results.

Positivist and interpretivist paradigms have been put together inside social constructionism by Berger and Luckmann, referring to the work of Alfred Schutz (Schutz 1960; Berger and Luckmann 1966). It is interesting how Berger and Luckmann support the objectivity of data by defining them as signification acts. Signification is an objectivation through the definition of a sign: "able to communicate meanings that are not directs manifestations of *hic et nunc* subjectivity" (Berger and Luckmann 1966/1969, 58). The most important signification acts in history are language and numeration, actually also the two main forms of data. In a context of high data production as the contemporary societies that definition is meaningful, because it points out the symbolic dimension of both social life and data construction.

It is quite surprising how a mathematician as Tobias Dantzig shared the same notion of data as signification acts (1930/1954). He writes:

> Signification allows to transcend the subjective reality and pass to the objective reality. That reality is not a collection of frozen images, but a living, growing organism(…) an individual without a milieu, deprived of language, deprived of all opportunity to exchange impressions with his peers, could **not construct a science of number (a data science**[1]**)**. To his perceptual world data would have no reality, no meaning. (Dantzig 1930/1954, 253).

Both Dantzig and the constructionists hence recall the cognitive and symbolic aspects of data and believe that simple facts do not exist, but "facts are always interpreted" (Schutz 1962, 5). What Schutz and their fellows have proposed is a methodological constructionism where there is a suspension of ontology. Objectivity of data does not arise from a supposed reality but from the agreement of all

[1]Bold text and text in brackets are added by the author.

observers and through procedures which are intersubjectively defined. Scientific knowledge is therefore grounded on the agreement between observers, and it is not independent from the questions whose answers have been given.

3 Data Culture, Data Assemblage and Data Science

Two are the main reasons why social constructionism seems fitted to cross the data revolution era: first of all, because it is focused on daily life and new data bring in many ways the daily life of individuals in the hearth of social knowledge; furthermore, because it believes that the empirical process and the construction of reality are based upon the agreement of observers. And this perspective is useful to address the future of social sciences and data science into an interdisciplinary context.

Ubiquitous and pervasive technology brings daily life in data production. Social networks data, smartphone data and user-generated contents on the Internet are windows upon the subjectivity of individuals. These new data are able to track, trace, record and sense our complex interactions with the social world. For example, one point about the supremacy of big data on traditional data is that researchers may be in the action instead of collecting account of actions (through survey) (Savage and Burrows 2014). But the fact that there is no contact between researcher and subjects causes advantages as well as disadvantages. For example, automated and volunteer data are just founded data produced in an unobtrusive way (Webb et al. 1966). Unobtrusive methods collect data on research units that are not aware of being studied. There is not active participation (feedback) from those being researched; they are not-reactive. The most important advantage of using not-reactive methods is that the researcher does not perturb the behaviours of the subjects he/she is studying. There is no reaction to questions posed or observations made. The experimental “Hawthorne effect” (Mayo 1933) and the answers’ reliability problems (social desirability, memory effects, acquiescence) just vanish.

On the contrary, it is more important to study what is the context of validity of these data. Special attention must be paid to user-generated contents on social network and Websites. They represent a massive quantity of data, but their contents, and the ways of making sense of them through classifications, have the characteristics of deep data, those that are used in the realm of qualitative sociology. (Boccia Artieri 2015).

From a sociological perspective, changes in data construction are also changing the role of social actors in the production of data and their data culture. Data culture refers to the connection between the moment when data are constructed and the moment when they are used to produce knowledge in a specific domain (Sgritta 1988; Aragona 2008). Data culture is defined by two elements: the organizational and methodological changes which constitute the production of data in a specific time and the quantity of social data. The actual data culture era sees an impressive amount of social data produced daily, where the use of technology is more than ever. Producers and users may be very distant, and data that are generated by someone may be shared, sold, combined, merged and then analysed to produce knowledge on some specific domains. In this context, where many actors are involved in the production and use of data, empirical process truly becomes a cultural process which needs to be understood as such. What all this view on data suggests is that to put meaning into data and to understand what piece of reality that data is representing, we need to have a close look on what Kitchin and Lauriault (2014) call data assemblages.

Data assemblages are complex socio-technical system composed of many apparatuses and elements that are thoroughly entwined, whose central concern is the production of data. Data assemblages are made of two main activities:

- a technical process (operational definitions, data selection, data curation) which shapes the data as they are;
- a cultural process, which shapes the background knowledge (believes, instruments and others things that are shared in a scientific community) which enables the sharing of meanings.

Data science in social research requires therefore interdisciplinary and cross-cutting approaches, combining skills and viewpoints that cut across disciplines. The expertises needed are domain expertise, data expertise and analytical expertise. A dialogue on different technical and cultural processes is required to blend methodologies and disciplinary matrixes and shape what Lackatos (1976) called background knowledge (the whole set of facts and parameters used in the construction of any given theory and of any given data).

Methodological constructionism may be the right approach for addressing the future of new data research in social sciences.

4 Sense Making, Small Decisions and Social Actors

Some examples drawn from empirical research may clarify the contribution that constructivism may give to the development of new data research in social sciences.

A first example is the construction of meanings that is used to perform data curation in the analysis of *big corpora*, when researcher transforms unstructured texts in databases which can be analyzed through computational techniques. Di Maggio (2015) notes that topic-modelling programs require lots of decisions that most social scientists are ill-equipped to make. For example, it is essential to document the crawler's specification in detail to meet the standards of peer review and replicability, but not all social scientists are able to understand crawler language. Most textual analysis run on Web data requires close reading that has traditionally been conducted by hermeneutically oriented scholars who find not one simple uncontested communication, but multiple, contradictory and overlapping meanings.