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.

http://www.springer.com/978-3-319-55476-1

Data Science and Social Research
Epistemology, Methods, Technology and Applications
Lauro, N.C.; Amaturo, E.; Grassia, M.G.; Aragona, B.;
Marino, M. (Eds.)
2017, IX, 300 p. 77 illus., 51 illus. in color., Softcover
ISBN: 978-3-319-55476-1