BENCE SÁGVÁRI *

The Computational Turn in Social Sciences. Challenges of the New Empiricism in the Age of Big Data

Intersections. EEJSP
3(1): 5-14.
DOI: 10.17356/ieejsp.v3i1.348
http://intersections.tk.mta.hu

* [sagvari.bence@tk.mta.hu] (Hungarian Academy of Sciences, Centre for Social Sciences, Hungary; International Business School (IBS), Hungary)

This work was supported by OTKA Grant No. K112713 ('Life cycle of an online social network. Big Data analysis').

Today large amounts of data are available to use for research on human behaviour: social media, data from online social networks, vast amounts of digital text, sensory information from personal hand-held and other devices, information from search engine usage and other online services, etc. The industry that relies on collecting, combining, selling and analysing digital footprints for all kinds of purposes ranging from simple targeted advertising to risk assessments and mass surveillance is developing with lightning speed (Van Es and Schäfer, 2017). However, such data could increasingly be used to address larger societal issues of social interactions and relations, inequality, education, healthcare, political participation, and more. The advances in the use of such data in social sciences offer the possibility to answer questions that were beyond research in the past, and this new generation of large-scale, complex, and usually unstructured data requires new forms of data analysis and scientific applications. Some also suggest that as a consequence of the data revolution that we are already living in, a major paradigm shift in science is expected with far-reaching consequences to how research is conducted and knowledge is produced (Mayer-Schönberger and Cukier, 2013; Meyer and Schroeder, 2015). While the course of development in the data-driven industries and research seems to be unambiguous for the future in terms of its expected impact on business and how societies function in general, today the possibilities are still frequently overestimated by some 'positivistic prophets' – coming mostly from outside academia. In addition to presenting the main arguments of the papers in this section, the purpose of this editorial is to highlight a few of those issues and challenges that may shape the future of social sciences and of those who pursue in it, in relation to the new data landscape. After briefly elaborating on the definitions of Big Data, the focus will move to the question of epistemology; the changing dynamics among various fields of sciences; the new divides in access to data; and the main ideas behind the critical approach that social sciences might follow to find their right place in the puzzle.

1. *The promise and the reality of Big Data*

The complex phenomenon described above is usually referred to as Big Data, however it might be misleading because of an inevitable limitation of the concept to a more mechanic and data-centred approach. Some authors argue that the phenomenon we are dealing with is rather the 'computational turn' in sciences and beyond (Van Es and Schäfer, 2017), where all aspects of life are transformed into quantifiable data, and it is used to predict human behaviour and automate human decision-making processes. Nevertheless, this editorial sticks to the use of Big Data as its key term, not just because of its history of nearly two decades, but also because of its general acceptance in multiple fields of science and beyond. Interestingly, the wider scientific and public consciousness of Big Data dates back only to a few years of active marketing activities by the largest IT companies in advertising and selling their analytical solutions (Gandomi and Haider, 2015). The literature on interpreting Big Data from a social science perspective is expanding fast, but we still miss a uniform definition (Borgman, 2015; Csepeli, 2015; Dessewffy and Láng, 2015; Kitchin, 2014; McFarland et al., 2015; Székely, 2015). It is by far no coincidence, since the evolution of Big Data has been too quick and disordered so far, characterised by rapid technological changes. There have been some attempts to create a comprehensive definition that considers the different perspectives of business and academia, but the results turned out to be overly complex and therefore hard to use routinely. Based on more than 1500 conference papers and articles, De Mauro (2015) and his co-authors defined four core areas that were found in most Big Data perspectives and definitions: (1) the nature of information; (2) technology, as the equipment for working with Big Data; (3) processing methods that go beyond the traditional statistical techniques; and finally (4) the impact that Big Data can have on our lives. Based on these premises they proposed the following formal (and fairly circuitous) definition: '*Big Data represents the Information assets characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value'* (De Mauro et al., 2015: 103). Kitchin (2014) identified the following seven general features of Big Data:

- huge in *volume* (i.e. gigabytes, terabytes or petabytes of data depending on their sources);
- high in *velocity* (created in or near real-time);
- diverse in *variety* (structured and/or unstructured in nature);
- *exhaustive* in scope (capturing entire populations or systems, as described by the popular *n=all* phrase);
- fine-grained in *resolution* and uniquely indexical in identification;
- *relational* in nature (with the ability of conjoining different data sets);
- *flexible*, extensional, and scalable.

Obviously, Big Data is not just about the data. It is the necessary first element, and it does not even have to be 'big'. Tera, or petabytes of meteorological data, gigabytes of social networking data, and only megabytes of processed data of literally anything can all qualify to be named Big Data. The question is rather how we access, collect, store, analyse, interpret and share it. If we believe the predictions on future developments, it might be accepted that currently we are still at the dawn of the new datafied world. From the perspective of social sciences it means that on the one hand, many old research questions could be approached anew from novel angles, but on the other hand, a whole new set of questions are also begging to be addressed (McFarland et al., 2015).

2. The end of theory in data-driven science?

The new empiricism that is frequently linked to Big Data rests upon the above traits, and it was first popularised by Chris Anderson, former editor-in-chief at Wired magazine (Anderson, 2008). He stated that '[...] *the data deluge makes the scientific method obsolete*', '[...] *with enough data, the numbers speak for themselves*', and '[...] *correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There is no reason to cling to our old ways.*' In other words, he projected a world of research and inquiry where there is no need for a priori theory, models of hypotheses, and it also implies a contradictory approach to deductive science. Besides this primarily inductive nature of Big Data, another often cited promise is the possibility of capturing the whole of a domain by providing full resolution, and obtaining a detailed and reliable picture of those niche groups that were previously beyond the reach of surveys and other sampling based techniques. Finally, based on the de-theorised approach to enquiry and the *n=all* nature of the data it might also be argued that information derived from Big Data can be interpreted and transformed into knowledge by anyone who is capable of decoding statistic and/or data visualisation. These premises could be valid in certain domains of (mostly) business analytics, where there might be no restrictions on access to data, and algorithm-based autonomous or semi-autonomous decisions are in the focus. However, from the perspective of social sciences and empirical research that attempt to rely on Big Data as raw material, some remarks need to be made here.

One of the most important unique characteristics of Big Data from an (academic) research perspective is that much of the data used for analysis are by-products of other (usually business related) activities, or they are owned by state organisations. In most cases, it also implies that they were generated before any kind of research question or hypothesis had been formulated (e.g. data from Twitter, Facebook, Google Trends, and other online service providers that offer usually restricted, but automated access to their data through APIs (application programming interface), or as the result of unique and occasional agreements.). To put it another way, on the one hand, we might see it as a new epistemological approach that differs from the traditional deductive design, where hypothesis and insights are born not from preliminary theories, but from the available data. On the other hand, it could also be revealed that data is never born in a scientific or cognitive vacuum: it is a product of human activity that is self-evidently framed by some conceptual framework. Therefore, the condition whether we have data or not related to a social phenomenon, and the exact elements of it are all constrained by external human, organisational, and technical factors that researchers need to adapt to. From a critical perspective, sometimes it is even more interesting to examine where and why there is no data available on certain phenomena.

As far as the data-driven, inductive approach of Big Data research is concerned, from a social scientific perspective, this is probably one of the most controversial statements. Big Data can truly open up new opportunities for discovery, that also requires innovativeness in both formulating questions and finding the right tools, thus hypotheses and insights might certainly be born directly from the data. Still, interpretation without context and domain-specific knowledge can hardly be deemed suitable. It is interesting to see how some studies that tend to neglect the theoretical and empirical research traditions of social sciences arrive at conclusions based on Big Data that overlook decades of scholarship established in these fields (Borgman, 2015), or simply prove relationships that have already been known for a long time. Thus, Big Data offers plenty of possibilities for theory-driven social scientists who cross the traditional borders of their disciplines. However, the mere theoretical knowledge of social scientists needs to be complemented by a solid understanding of how to process data to get information from them.

3. Re-defining roles between fields of sciences

Due to the advances in Big Data and social network analysis, it seems that social sciences have lost their former privilege to investigate the functioning of societies. The 'good old' historical lines of demarcation between disciplines in terms of general epistemology, research subjects and questions, dominant methods seem to be vanishing. Probably for the first time in the history of science the field of engineering, the Internet industry, the disciplines of natural and social sciences are all focusing on similar types of data and similar types of questions (McFarland et al., 2015). This process of convergence holds great potential for all players who take an active part in this exciting transformation. However, it is also evident that the former status quo between diverse fields of science is about to change. In other words, it is still unclear whether the new division of labour will be more symmetrical or asymmetrical between social sciences, and natural/computer sciences. It is far from impossible that social sciences, and particularly sociology may witness the surreptitious course of colonisation where their traditions would increasingly subvert to other fields. In the eyes of the 'outside world' social sciences are often seen as an 'easy prey' due to their confusingly high degree of fragmentation that is manifested in countless competing theories and fundamentally different methods (Balietti et al., 2015; Whitehouse et al., 2012).

A remarkable development of the past years is that such 'soft' rivalry between the fields of traditional sciences seems to be relocating to the area of the emerging field of Computational Social Sciences (CSS) (Lazer et al., 2009). By definition, CSS is much more than just 'pure' Big Data, since it comprises social network analysis, social simulation models, as well as other areas of scientific inquiry and methods. By way of illustration, the *Manifesto of computational social science* written by Rosaria Conte and her mostly non-social scientist co-authors in 2012 clearly demonstrates notable transformation of the dominant approaches.

> '[...] sociology in particular and the social sciences in general would undergo a dramatic paradigm shift, arising from the incorporation of the scientific method of physical sciences. Thus, the combination of the computational approach with a sensible use of experiment will bring the social sciences closer to establishing a well-grounded link between theory and empirical facts and research. Such links should inform all sciences in which human behaviour is the main object of research or interest, and should solve incompatibilities such

> as economics relying on the rational actor picture and sociology and social psychology outright rejecting it; on the other hand, the latter rely much more on facts (identified from experiments, surveys, etc.) than traditional economics, based on the strength of purely abstract analytical approaches. Computational social science would be a major factor toward this paradigm change in the social sciences.' (Conte et al., 2012: 341)

Fields in social sciences and humanities are increasingly facing the demand that they justify their activities by employing computer-aided methods and sophisticated quantitative analysis (Van Es and Schäfer, 2017). For this reason, there is growing motivation among social scientist (that is at least partly based on external pressures and the 'fear of missing out') to acquire new data analytic skills and somehow immerse themselves in Big Data research or in the broader field of computational social sciences.

4. Uneven access to data. The new division between data-rich and data-poor

As outlined above, there have been unprecedented opportunities in the collection and analysis of data about social phenomena. For example, using social media data, interactions among individuals can be measured in a precise and extensive way to understand behaviour (Felt, 2016); analysing Twitter data the use of language can predict certain health risk factors (Eichstaedt et al., 2015); and mapping the friendship ties in physical space at macro level can detect the structure of administrative regions in a given country (Lengyel et al., 2015), etc. Not surprisingly, in data-driven social science the key to success is to have access to good data – both in terms of quantity and quality. Therefore, the widely-held promise and simple statement of the Big Data era, that due to the data deluge limits of scientific discovery are fading away, evidently needs some clarification. It seems obvious that new divisions between the data-rich and the data-poor are emerging (Boyd and Crawford, 2012). So, it is not just the asymmetric relationship between the owners of the data (those who collect, store, mine, and analyse) and those whom data collection targets (i.e. the users) (Andrejevic, 2014), but the new kind of digital divide that becomes apparent between individual researchers or research groups; between industry and the academic world; or even between countries physically located 'closer to' or 'farther from' the original source of data. As an example, the few dozen data scientists who work at the research lab of Facebook[1] (in addition to being part of academia) are currently probably one of the most privileged researchers in the world. Outside this privileged social laboratory, independent data-collection from Facebook is rather limited using official APIs or by web crawling techniques. At the same time, establishing bilateral organisational relations between major Internet companies (such as Facebook) and research institutions (such as a university) from remote countries seems almost impossible. Local collaborations with major commercial data owners (e.g. telecom service providers, online media companies, or governmental

[1] https://research.fb.com/people/ Accessed: 26-03-2017.

organisations) are more likely to happen, but it also requires efficient negotiation skills, and experience in corporate or bureaucratic languages and cultures. The difficulties in getting access to data was also mentioned as the major obstacle for Big Data research in a (non-representative) international survey of social scientists conducted by SAGE Publishing (Metzler et al., 2016). In short, social scientists need to be able to acquire and utilise new research skills and methods that fit the new paradigm of datafied science. In many cases, it also includes the non-technical ability to acquire (big) data by being open to the demands and interests of other non-academic fields.

5. Challenging the positivistic notion of Big Data: the ground for Critical Data Studies

From a bird's eye view, research using Big Data is largely built around the principles of positivistic science (Kitchin, 2014). In this sense, Big Data research is fundamentally considered a neutral phenomenon, where social scientists could play a leading role is the emerging field of Critical Data Studies (CDS). Here the core idea is to tint the overly functionalist and result-oriented approach, and its initial assumption is that data are under any circumstances a form of power (Iliadis and Russo, 2016). The massive amount of user information collected from individuals constitute a unique form of capital. With this resource, accompanied by complex algorithms and powerful data processing tools, organisations are capable of influencing emotions and culture. This was most spectacularly reflected in the media by the current activities of the data mining and data analysis firm Cambridge Analytica during Ted Cruz's and Donald Trump's presidential campaign in the US, and the pro-Brexit campaign in the United Kingdom. Contrary to the high-sounding promises that mere utilisation of Big Data was enough to win the elections or the Brexit campaign, these analytic tools were only able to model the personalities of voters in unprecedented detail and thus identify target voters in a new and innovative form. Obviously, the long-term effects of these tools' capabilities should not be underestimated, and it also suggests that data are never raw, but always 'cooked', and it is extracted behind the user's back and might be seized to serve the interests of companies, political parties and other organisations. In short, the critical approach to Big Data challenges the ground upon which positivistic Big Data science stands.

As data are increasingly considered to be at the heart of the knowledge economies, data-savvy scholars from the humanities and arts (often in collaboration with information and computer scientists) have ignited critical public debates. Their perspectives on data science are important in that they bring the question of responsibility to the fore (Mann, 2017; McDermott, 2017; Tene and Polonetsky, 2012). Questions of responsible data production and use, ethics, privacy, data power and transparency of data handling form the core topics of this new paradigm (Schäfer and Es, 2017). The main objective of CDS is therefore to construct critical frameworks to exploit power structures related to the creation, curation and utilisation of (big) data.

6. Strategies for social sciences

Based on the trends explained above, the question then arises as to what role social scientists could fulfil in these changing circumstances. Or course, there can be many individual answers to this question, therefore no universal recipe exists. But if we accept that the current structural changes are similar in their effects to what happened in the second half of the 20th century with the statistical and survey turn particularly in the field of sociology, there is obviously a high demand for social scientists who are prepared both in their methodological skills and theoretical knowledge. Computational ethnography, computational linguistics, network science, machine learning, Big Data based experiments (McFarland et al., 2015), etc. are all streams of research that require new analytic techniques and hold great potential for interdisciplinary collaborations. While the multifaceted trading zone of computational social sciences may not rest upon the egalitarian principle, there is a precious place for social sciences to provide synthesis of information and narratives that enable us to understand the findings in a wider social context. Using new kinds of data and tools in a positivistic manner on the one hand, and being a critical, sometimes sceptic, theory-driven data practitioner on the other hand, seem to be the 'winning combo' for social scientists.

The articles in this special section demonstrate how the previously mentioned computational turn can be utilised to examine 'classic' matters of social research in a non-traditional way, and how social researchers can adapt to the new circumstances.

In the first article Mette My Madsen shows how in a data research project questioning the understanding of data itself could become a reality. The author demonstrates, through describing an 'experiment' conducted at the Danish Technical University (DTU) that was a collaborate work of different domains of science with heterogeneous types of data, how the emergence of new types, quantities, qualities and combinations of data, has potential to review our understanding of data anew. The author argues that instead of looking at data in the classic way as either 'raw' or 'shaped' we might gain a different perspective if we regarded it as 'monadic', which here stands for duality as simultaneously both unit and composition.

Dessewffy and Váry present an empirical case study that demonstrates how social media data can be used to address specific questions of cultural and media studies. They examine the relationship between the Hungarian celebrity sphere and social media fandom. Their approach is in line with the primary promise of computational social sciences: how can we ask and answer questions that could not have been asked or answered before. The article provides a network-based analysis of the most well-known Hungarian celebrities on Facebook.

The article written by Kmetty, Koltai, Bokányi, and Bozsonyi goes back to the earliest theoretical traditions of sociology by analysing the seasonality patterns of suicides in the US simultaneously using Twitter data and 'hard data'. They attempted to find grounds for the general negative social climate in the number of suicides committed, and in the aggregated content of tweets posted. Although they did not manage to find a straightforward link between the two, nevertheless the data used for this analysis, the applied methods, and the combination of 'new' and 'old' sources of data show an innovative approach, that also sheds light on the possible future directions of social science research.

Finally, the analysis of the hyperlink network of Hungarian websites from Romania by Boróka Pápay and Bálint Kubik is a remarkable work because of its efforts to bring together the crawled network data and the classic sociological phenomenon of minority societies. In their analysis, the authors were able to demonstrate that the network of Hungarian websites from Romania is strongly interconnected, forming a community with a separate reality. This article is another example of how sociological research can build on the tools of network science, and interpreting the results inside the 'good old' theoretical frames of social sciences.

References

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7). Available at: https://www.wired.com/2008/06/pb-theory/ Accessed: 26-03-2017.

Andrejevic, M. (2014) The Big Data Divide. *International Journal of Communication*, 8(8): 1673–1689.

Balietti, S., Mas, M. and D. Helbing (2015) On disciplinary fragmentation and scientific progress. *PLoS One*, 10(3): e0118747. DOI: https://doi.org/10.1371/journal.pone.0118747

Borgman, C. L. (2015) *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press.

Boyd, D. and K. Crawford (2012) Critical Questions for Big Data. *Information, Communication & Society*, 15(5): 662-679. DOI: https://doi.org/10.1080/1369118x.2012.678878

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., . . . Helbing, D. (2012) Manifesto of Computational Social Science. *The European Physical Journal Special Topics*, 214(1): 325-346. DOI: https://doi.org/10.1140/epjst/e2012-01697-8

Csepeli, G. (2015) A szociológia és a Big Data (Sociology and Big Data). *Replika*, (92-93): 171-176.

De Mauro, A., Greco, M. and M. Grimaldi (2015) *What is Big Data? A Consensual Definition and a Review of Key Research Topics*. Paper presented at the 4th International Conference on Integrated Information. September 4-5, 2014, Madrid. DOI: https://doi.org/10.1063/1.4907823

Dessewffy, T. and L. Láng (2015) Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon (The Chance Meeting on a Dissecting-Table of Big Data and Social Sciences). *Replika*, (92-93): 157-170.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. (2015) Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, 26(2): 159-169. DOI: https://doi.org/10.1177/0956797614557867

Felt, M. (2016) Social Media and the Social Sciences: How Researchers Employ Big Data Analytics. *Big Data & Society*, 3(1): 205395171664582. DOI: https://doi.org/10.1177/2053951716645828

Gandomi, A. and M. Haider (2015) Beyond the Hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, 35(2): 137-144. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Iliadis, A. and F. Russo (2016) Critical Data Studies: An Introduction. *Big Data & Society*, 3(2): 2053951716674238. DOI: https://doi.org/10.1177/2053951716674238

Kitchin, R. (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, 1(1): 205395171452848. DOI: http://doi.org/10.1177/2053951714528481

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Van Alstyne, M. (2009) Life in the Network: The Coming Age of Computational Social Science. *Science*, 323(5915): 721-723. DOI: http://doi.org/10.1126/science.1167742

Lengyel, B., Varga, A., Ságvári, B., Jakobi, A. and J. Kertész (2015) Geographies of an Online Social Network. *PLoS One*, 10(9): e0137248. DOI: http://doi.org/10.1371/journal.pone.0137248

Mann, S. (2017) Big Data is a Big Lie Without Little Data: Humanistic Intelligence as a Human Right. *Big Data & Society*, 4(1): 205395171769155. DOI: http://doi.org/10.1177/2053951717691550

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

McDermott, Y. (2017) Conceptualising the Right to Data Protection in an Era of Big Data. *Big Data & Society*, 4(1): 2053951716686994. DOI: http://doi.org/10.1177/2053951716686994

McFarland, D. A., Lewis, K. and A. Goldberg (2015) Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *The American Sociologist*, 47(1): 12-35. DOI: http://doi.org/10.1007/s12108-015-9291-8

Metzler, K., Kim, D. A., Allum, N. and A. Denman (2016) *Who Is Doing Computational Social Science? Trends in Big Data Research*. SAGE White Paper. London: SAGE. Available at: https://us.sagepub.com/sites/default/files/CompSocSci.pdf Accessed: 26-03-2017.

Meyer, E. T. and R. Schroeder (2015) *Knowledge Machines: Digital Transformations of the Sciences and Humanities*. Cambridge, MA: MIT Press.

Schäfer, M. T. and K. v. Es (2017) (Eds.) *The Datafied Society. Studying Culture through Data.* Amsterdam: Amsterdam University Press. DOI: https://doi.org/10.5117/9789462981362

Székely, I. (2015) Az adatmentes zónák szükségessége és esélye (The Need for and the Chances of Data-Free Zones). *Replika,* (92-93): 209-225.

Tene, O. and J. Polonetsky (2012) Big Data for All: Privacy and User Control in the Age of Analytics. *Northwestern Journal of Technology and Intellectual Property,* 11(5). Available at: http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1 Accessed: 26-03-2017.

Van Es, K. and M. T. Schäfer, M. T. (2017) Introduction. New Brave World. In M. T. Schäfer and K. Van Es (Eds.) *The Datafied Society.* Amsterdam: Amsterdam University Press. 13-23. DOI: https://doi.org/10.5117/9789462981362

Whitehouse, H., Kahn, K., Hochberg, M. E. and J. J. Bryson (2012) The Role for Simulations in Theory Construction for the Social Sciences: Case Studies Concerning Divergent Modes of Religiosity. *Religion, Brain & Behavior,* 2(3): 182-201. DOI: https://doi.org/10.1080/2153599x.2012.691033