JOURNAL OF DIGITAL SOCIAL RESEARCH

WWW.JDSR.IO ISSN: 2003-1998

VOL. 1, NO. 1, 2019, 1–9

HACKING SOCIAL SCIENCE FOR THE AGE OF DATAFICATION

Simon Lindgren*

ABSTRACT

The ongoing and intensifying datafication of our societies poses huge challenges as well as opportunities for social science to rethink core elements of its research enterprise. Prominently, there is a pressing need to move beyond the long-standing qualitative/quantitative divide. This paper is an argument towards developing a critical science of data, by bringing together the interpretive theoretical and ethical sensibilities of social science with the predictive and prognostic powers of data science and computational methods. I argue that the renegotiation of theories and research methods that must be made in order for them to be more relevant and useful, can be fruitfully understood through the metaphor of *hacking social science*: developing creative ways of exploiting existing tools in alternative and unexpected ways to solve problems

* Umeå University, Sweden.

The launch of the *Journal of Digital Social Research (JDSR)* happens in the midst of the age of datafication, which poses challenges and opportunities for social science to rethink itself in methodological, and other, terms. The broad proliferation of big data and data science risks that its one-sided data-drivenness spills over into scholarly social research, and that the skill and role of theoretical interpretation gets lost among bits, bytes and shiny infographics. JDSR can hopefully be a forum for discussions about how we can think and operate as theoretically sensitive and methodologically flexible social scientists in the age of datafication. There lies great scholarly potential in bringing data-driven computational approaches into closer contact with social theory, but only if such approaches are given a critical framing.

1 DATAFICATION

Sociologist Deborah Lupton (2014, p. 101) argues that the hype that surrounds the new technological possibilities afforded by big data analysis contributes to the belief that such data are "raw materials" for information — that they contain the untarnished truth about society and sociality. But as we know, in reality, each step of the process in the generation of big data relies on a number of human decisions relating to selection, judgement, interpretation, and action. Therefore, the data that we will have at hand are always configured via beliefs, values, and choices that "'cook' the data from the very beginning so that they are never in a 'raw' state". So, there is no such thing as raw data, even though the orderliness of neatly harvested and stored big data sets can create a mirage to the contrary.

Sociologist David Beer (2016, p. 149) argues that we now live in "a culture that is shaped and populated with numbers", where trust and interest in anything that cannot be quantified diminishes. Furthermore, in the age of big data, there is an obsession with causation. As Crawford and boyd (2012, p. 665) argue, the mirage and mythology of big data demand that a number of critical questions are raised with regards to "what all this data means, who gets access to what data, how data analysis is deployed, and to what ends". There is a risk that the lure of big data will sideline other forms of analysis, and that other methods with which to analyse the beliefs, choices, expressions, and strategies of people are pushed aside by the sheer volume of numbers.

All the things that people do online in the context of social media generate vast volumes of interesting data from the perspective of social science. Such data have been approached in highly data-driven ways within the field of data science — an interdisciplinary specialisation at the intersection of statistics and computer science, focusing on machine learning and other forms of algorithmic processing of large datasets to "liberate and create meaning from raw data" (Efron and Hastie 2016, p. 471). Data science projects tend to be strongly data-driven, often with the aim of getting a general picture of some particular social pattern or process. Being data-driven is not a bad thing, but there must always be a balance between data and theory – between information and its interpretation. This is where social theories

come into the picture, as they offer a wide range of conceptual frameworks that can aid in the analysis and understanding of the large amounts and many forms of social data that are proliferated in today's world.

But in those cases where we see Big Data being analysed, there is far too often a disconnect between the data and the theory. One explanation for this may be that the popularity and impact of data science makes its data-driven ethos spill over also into the academic fields that try to learn from it. This means that we risk forgetting about the theoretical analysis that may fade in the light of sparkling interactive visualisations.

But social research that relies heavily on the computational amassing and processing of data must still have a theoretical sensitivity to it. While pure data science methods are extremely helpful when wrangling the units of information, the meanings behind the messy social data which are generated in this age of datafication can be better untangled if we also make use of the rich interpretive toolkit provided by sociological theories. The data do not speak for themselves, even though some Big Data evangelists have claimed that to be the case (Anderson 2008).

Big Data and data science are partly technological phenomena, which are about using computing power and algorithms to collect and analyse comparatively large datasets of largely unstructured information. But prominently they are also a cultural phenomenon that comes with a mythological belief that huge unstructured datasets, often based on social media interactions and other digital traces left by people, when paired with methods like machine learning and natural language processing, can offer a higher form of truth which can be computationally distilled rather than interpretively achieved.

Such mythological beliefs are not new however, as there has long been, if not a hierarchy, so at least a strict division of research methods within the cultural and social sciences, where some methods – those that have come to be labelled "quantitative", and that include survey methods analysed with statistical tools – have been vested with an "aura of truth, objectivity, and accuracy" (boyd and Crawford 2012, p. 663). Other methods – those commonly named "qualitative", and involving so-called close-readings of textual data from interviews, observations and documents – are seen as more interpretive and subjective. This distinction is not only annoying, but also wrong. We can get at approximations of "the truth" by analysing social and cultural patterns, and those analyses are by definition interpretive no matter the chosen methodological strategy. Especially in this day and age where data, the bigger the better, are fetishised it is high time to move on from the unproductive dichotomy of qualitative versus quantitative.

2 DATA-DRIVEN

As argued above, pure data science is most of the time too data-driven from the perspective of social science. It tends very strongly to focus simply on what is

researchable. It goes for the issues for which there is data, no matter if those issues have any real-life urgency or not. The last decade has seen parts of the field of data science and parts of the social sciences become entangled in ways that risk a loss of theoretical grounding. In a seminal paper outlining the emerging discipline of "computational social science", David Lazer and colleagues wrote that:

> We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies (Lazer et al. 2009, p. 721 , emphasis added)

Furthermore, they argued that there was an inherent risk to the fact that existing social theories were "built mostly on a foundation of one-time 'snapshot' data" and that they therefore may not be fit to explain the "qualitatively new perspectives" on human behaviour offered by the "vast, emerging data sets on how people interact" (Lazer et al. 2009, p. 723). While I agree that social analysis must be re-thought in light of these developments, I am not so sure that it is simply about "compiling" the data, and then being prepared that existing theories may no longer work. Rather, I argue, we should trust a bit more that even though the size and dynamics of the data may be previously unseen, the social patterns that they can lay bare – if adequately analysed – can still largely be interpreted with the help of "old" theories. After all, the theories are not designed to understand particular forms of data, but instead the sociality that they bear witness to. My point is that social theory is needed for considering both the data, the methods, the ethics, and the results of the research. By extension, still, theories may always need to be updated, revised, or even discarded — but that has always been true.

Noortje Marres and Caroline Gerlitz (2016) suggest that we should go beyond previous divisions of methods by thinking in terms of "interface methods". This means highlighting that digital methods are dynamic and under-determined, and that multiple methodologies are intersecting in digital research. By recognising "the unstable identity of digital social research techniques", we can "activate our methodological imagination" (Marres 2017, p. 106). Marres continues to say that:

> Rather than seeing the instability of digital data instruments and practices primarily as a methodological deficiency, i.e. as a threat to the robustness of sociological data, methods and findings, the dynamic nature of digital social life may also be understood as an enabling condition for social enquiry (Marres 2017, p. 107).

I would like to advocate for a general stance by which more integrated methodologies can be developed and propagated. In important respects, the data-drivenness of big data science is not different from the data-drivenness of ethnography and anthropology. There is a need to formulate approaches by which

theoretical interpretation and a "qualitative" approach to data is integrated with "quantitative" analysis and data science techniques. As a sociologist, I am particularly interested in what interpretive sociology can bring to the table here. With this concept I refer to the classic notion of sociology as "a science concerning itself with the interpretive understanding of social action (…) its course and consequences" (Weber 1922/1978, p. 4). This kind of sociology is about the understanding (Verstehen) of social life and has a focus on processes of how meaning is created through social activities. In other words, it is not a positivist and objectivist science. As Max Weber put it, "Meaning" never refers:

> to an objectively "correct" meaning or one which is "true" in some metaphysical sense. It is this which distinguishes the empirical sciences of action, such as sociology and history, from the dogmatic disciplines in that area (…) which seek to ascertain the "true" and "valid" meanings associated with the objects of their investigation (Weber 1922/1978, p. 4).

Still, he continued, interpretive sociology "like all scientific observations, strives for clarity and verifiable accuracy of insight and comprehension (Evidenz)" (Weber 1922/1978, p. 4). The interpretive stance should entail moving back and forth between such evidence – data – and their iterative and cumulative interpretation – theory. It is important that that we remind ourselves that also (or maybe especially) in the age of datafication, data (still) needs theory, and theory (still) needs data. One vision for JDSR is that it can enable discussions about how we can conceptualise and do research that aligns with that insight.

3 DATA/THEORY

Social theories, and often such theories that were developed more than a hundred years ago, can in fact contribute immensely to our understanding of things that we are now in the process of, maybe unnecessarily, inventing new names for: "viral communication", "user-generated content", "the blogosphere", "online hate", "cyber bullying", and so on. I do not mean that such words, at least not all of them, are merely superfluous synonyms for things that we already had adequate names for. Nor do I claim that any old theory is always better than a new one, or that such old theories can be applied unproblematically to 21st century phenomena without modification. But in many cases, we run the infamous risk of throwing the baby out with the bath water. When researching the peculiarities and novelties of interaction and communication in the datafied society, we risk mistaking theories about general patterns of social life for being obsolete just because they were developed in non-digital contexts.

The already established theories are useful because even though settings change, we may oftentimes be dealing with the same underlying social forms as before. Georg Simmel (1895, p. 54) argued that the most important task for the sociologist is to separate the form and content of social life, which are in reality inseparably united. The aim of the analysis must be to detach the forms from their

contents and to bring them together systematically: "For it is evident that the same form (…) can arise in connection with the most varied elements". Simmel continued to explain that:

> We find, for example, the same forms of authority and subordination, of competition, imitation, opposition, division of labor, in social groups which are the most different possible (Simmel 1895, p. 55).

Data scientists Rachel Schutt and Cathy O'Neil (2013, p. 9) argue that data scientists have much to benefit from collaborating with social scientists. This, they write, is because social scientists "do tend to be good question askers and have other good investigative qualities". They write about the hyped and still emerging speciality of data science that "it's not math people ruling the word". Rather, they argue that when different "domain practices", such as sociology, intersect with data science, each such practice is "learning differently" (Schutt and O'Neil 2013, p. 219). Taking my cue from Schutt and O'Neil, I would like to ask what type of such different learning – which methodological developments – can follow when social science meets data science.

This is obviously a vastly open question with a multitude of potential answers. Therefore, my suggestion, which draws to a great extent on my personal methodological and theoretical preferences as an interpretive sociologist, is but one possibility. The main idea that I am pushing is that the data-drivenness of interpretive sociology, as formulated as a hands-on framework by methodologists such as Glaser and Strauss (1967), and particularly Glaser's (1978) notion of theoretical sensitivity, can be dusted off and fruitfully brought together with the data-drivenness of data science practices.

Many would say that the respective general views on science and methodology between big data and grounded theory research are too divergent, to the point that they are even incompatible. I do not believe that to be the case. Still, experimentally merging methods that are labelled "qualitative" and "quantitative" is not a good idea if you want everyone to agree with you. In both camps (because sadly, that is still what they are), it is equally easy to find people who are dogmatic. So, to find productive ways across, there is definitely a need to think outside the box.

There are new types of data today, that demand new types of methods, while there are also new types of research questions arising that call for developing new approaches. This demands for advancing our perspective on both theory and methods in parallel. In other words, developing a data/theory approach.

4 HACKING SOCIAL SCIENCE

In light of the developments towards a datafication of society, there is a need to hack theories and research methods in order to make them more relevant and useful. Hacking is a fitting metaphor for the creative and somewhat anarchistic approach to existing theories and methods that I want to advocate, as it points to the idea that finding good solutions – rather than adhering to rules – should be the

end goal of any analytical strategy. This draws on Feyerabend's idea that anarchism in science, rather than "law-and-order science", is what will help achieve progress. And, as for the risk that such an approach will lead to an unproductive situation where anything goes, we must simply trust in our own ability to think in structured ways even without following rigid rules dogmatically:

> There is no need to fear that the diminished concern for law and order in science and society that characterizes an anarchism of this kind will lead to chaos. The human nervous system is too well organized for that (Feyerabend 1975, p. 13).

In spite of its popular reputation to the contrary, hacking is not (only) about breaking the law through forms of electronic vandalism. As argued by cryptologist Jon Erickson (2008), hacking can in fact be more about adhering to rules than about breaking them. Its goal, however, is to come up with ways of using, or exploiting, the structures and resources that are in operation in any given situation in ways that may be overlooked or unintended. Hacking is about applying existing tools in smart and innovative ways to solve problems. Erickson writes that:

> hacked solutions follow the rules of the system, but they use those rules in counterintuitive ways. This gives hackers their edge, allowing them to solve problems in ways unimaginable for those confined to conventional thinking and methodologies (Erickson 2008, p. 16).

The same can go for research, where we must allow ourselves to not think so much about which theoretical perspectives have been conventionally agreed to be compatible with one another, or about whether certain methods can be mixed together or not. A common conviction is that one cannot do "qualitative" and "quantitative" in the same breath, as they are based on different epistemologies. But this can in fact be debated. As argued by Bryman (1984), the difference may in practice not lie so much in different philosophical views on how knowledge about social reality is achieved, but simply in the path-dependent choices that are made by individual researchers who get stuck with one paradigm or the other. While it has become an eternal truth, reiterated by researchers and methods teachers alike, that "the problem under investigation properly dictates the methods of investigation" (Trow 1957, p. 33), very few of us adhere to this in practice.

But datafication presents us with a new data environment – with data traces, data fragments, and unsolicited data – that offers the opportunity to think in new ways about research in the "spirit of hacking", aiming to surmount "conventional boundaries and restrictions" for the goal of "better understanding the world" (Erickson 2008, pp. 16–18). What I describe here as anarchistic, and as hacking, may sound radical and dangerous – or maybe just plain stupid. But as a matter of fact, this approach is not very far from how science, as conceived by Bruno Latour, in general comes into being. Science and research happen in action. It is not ready-made. Interest should not be focused on any alleged intrinsic qualities of approaches, but at the transformations that they undergo in their practical use. Methods do not have any "special qualities", as their effects come from the many

ways through which they are "gathered, combined, tied together, and sent back" (Latour 1987, p. 258). Thus, "we are never confronted with science, technology and society, but with a gamut of weaker and stronger associations" (Latour 1987, p. 259). Knowledge about society is produced through more or less messy sets of practical contingencies.

The obvious connection between social science and computational methods may be through social science's quantitative specialisations, but there is much to gain from bringing data science methods in contact with the, maybe less expected, qualitative framework. The data-drivenness of data science can be more fruitfully construed as a form of digital fieldwork, rather than in terms of positivistic hypothesis testing.

In sum, hacking social science is first about *social science hacking*. That is, "qualitative" social scientists daring to play around with computational methods and techniques, more as participant observers than as faux computer scientists, second-rate statisticians, or bad mathematicians. This means engaging in a form of hacking, finding unexpected and creative solutions, but still as social researchers. Second, it is about *hacking social science*, that is, moving social research into new domains without fear of it losing its identity. In fact, many of today's computational opportunities may bring us closer than ever to realising the vision of the classic sociologists who wanted to study social processes in terms of systems, and often on a macro-level.

REFERENCES

Anderson, C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *WIRED*. Retrieved from https://www.wired.com/2008/06/pb-theory/

Beer, D. (2016) *Metric power*. London: Palgrave Macmillan, https://doi.org/10.1057/978-1-137-55649-3.

boyd, d., and Crawford, K. (2012) 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), pp. 662–679, https://doi.org/10.1080/1369118X.2012.678878

Bryman, A. (1984) 'The Debate about Quantitative and Qualitative Research: A Question of Method or Epistemology?', *The British Journal of Sociology*, 35(1), pp. 75–92, https://doi.org/10.2307/590553

Efron, B., and Hastie, T. (2016) *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Erickson, J. (2008) *Hacking: The art of exploitation*. San Francisco, Calif.: No Starch Press.

Feyerabend, P. (1975) *Against method: Outline of an anarchistic theory of knowledge*. London: NLB.

Glaser, B. G. (1978) *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, Calif.: The Sociology Press.

Glaser, B. G., and Strauss, A. L. (1967) *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine de Gruyter.

Latour, B. (1987) *Science in action: How to follow scientists and engineers through society*. Cambridge, Mass: Harvard University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Alstyne, M. V. (2009) 'Computational Social Science', *Science 323*, pp. 721–723.

Lupton, D. (2014) *Digital sociology*. Abingdon: Routledge.

Marres, N. (2017) *Digital sociology: The reinvention of social research*. Cambridge: Polity Press.

Marres, N., and Gerlitz, C. (2016) 'Interface methods: Renegotiating relations between digital social research, STS and sociology', *The Sociological Review*, 64(1), pp. 21–46, https://doi.org/10.1111/1467-954x.12314.

Schutt, R., and O'Neil, C. (2013) *Doing data science*. Beijing: O'Reilly Media.

Simmel, G. (1895) 'The Problem of Sociology', *Annals of the American Academy of Political and Social Science*, 6, pp. 52–63.

Trow, M. (1957). Comment on 'Participant observation and interviewing: A comparison'. *Human Organization*, *16*(3), 33.

Weber, M. (1922/1978). *Economy and Society: An outline of interpretive sociology. Vol. 1*. Berkeley, CA: University of California Press.