Tobias Wolbring

The Digital Revolution in the Social Sciences: Five Theses about Big Data and Other Recent Methodological Innovations from an Analytical Sociologist

Abstract: In recent years, both scholars and policy-makers place big hopes in the emerging fields of data science and computational social science to solve not only academic puzzles, but also to cure many "real-world" problems in a wide range of areas such as health, crime, and poverty. In this paper, we re-evaluate these claims, highlight current problems of these nascent fields, and show what sociology has to contribute to and can gain from the digital revolution in the social sciences. We thereby focus on analytical sociology – a field at the intersection of classical sociology and modern computational social science, which places a strong emphasis on mechanism-based explanations and rigorous empirical analyses. In a nutshell, we argue that sociology has to bring a lot to the table with important contributions concerning not only substantive research questions, but also theoretical insights and methodological skills. Both sides – not only sociology, but also data science – could thus substantially profit from a closer exchange, while some problems still remain that hinder an even more fruitful collaboration.

1 Introduction

This paper considers recent developments towards a digital social science in the light of promises, pitfalls, and challenges from the perspective of a quantitative, analytical sociologists. In a well-known paper Savage and Burrows (2007) predicted "The Coming Crisis of Empirical Sociology" due to the wide availability of large-scale data and related methodological innovations. Now a decade later, it appears worthwhile to critically assess whether sociology really lost ground as compared to other scientific disciplines, but also to re-evaluate what sociology has to contribute to and can gain from the digital revolution in the social sciences. The perspective of analytical sociology – understood in a rather broad sense as theory-driven social research aiming for mechanism-based explanations and rigorous empirical analyses (see Hedström and Bearman 2009) – thereby is particularly interesting as the field is located at the intersection of classical sociology and modern computational social science. On the one hand, analytical sociology is strongly inspired by classical sociological studies and methodological viewpoints, while at the same time it strongly builds on recent technical and methodological innovations such as agent-based modeling, web data scrapping, and social network analysis. In a nutshell, we argue that analytical sociology has to bring a lot to the table when discussing the role of empirical sociology in the age of digitization (see also Keuschnigg et al. 2018). We will develop the argument along the following five theses:

1) Not only the wide availability of big data, but also other methodological innovations in the digital age have fundamentally transformed the social sciences and will further do so in the coming years.

2) In contrast to current claims, big data do not replace theory, but highlight the need of thorough theoretical reasoning.

3) The nature and origin of big data often substantially limits the validity of empirical results.

4) A stronger interlinkage between classical tools of social research and computational social science is sorely needed.

5) The sheer amount of digital information forces researchers to reconsider established statistical approaches.

We conclude by shortly highlighting a number of additional issues ranging from methodological (e.g., reproducibility) over ethical/legal (e.g., data linkage, privacy, informed consent) to practical considerations (e.g., data access).

2 What is that revolution all about?

Not only the wide availability of big data, but also other methodological innovations in the digital age have fundamentally transformed the social sciences and will further do so in the coming years. Big data – characterized by the increased volume, velocity, and variety of available information – are usually at the forefront in discussions about the digital revolution and they are certainly one driving force for current developments. However, the digital revolution is more than just big data. It also encompasses fundamental methodological innovations such as the development of new techniques of data collection (e.g., web crawling), statistical tools to analyze large-scale, often time-stamped and geo-referenced "digital" data (e.g., analysis of longitudinal cell-phone data), and the purposeful intervention into online contexts (e.g., online field experiments). While the increase in data volume, velocity, and variety is just the logical sequel of a long known process of quantification and digitization, these new tools for social research have a much greater potential to cause a fundamental and lasting change in the social sciences. For example, relying on one or more of these methodological innovations major progress has been recently made in diverse fields such as research on the self-enforcing nature of status (van de Rijt et al. 2014), the role of reputation in online markets (Diekmann et al. 2014), ethnoracial neighborhood conflicts (Legewie and Schaeffer 2015), the effectiveness of crime prevention measures (O'Brien et al. 2015), and the cultural history of the world (Schich et al. 2014).

Recent advances, however, are most obvious for the area of social network analysis (e.g., Centola 2018; Giles 2012). For a long time social network analysis was mostly limited to research on small groups or ego-centric networks. Thereby, three unresol-

ved problems, from which the field suffered, were questions of (1) how to collect both complete and fine-grained social network data over longer periods of time, (2) how to investigate the structure of large networks, and (3) how to empirically separate selection from social influence (see also Golder and Macy 2014). The investigation of large-scale online networks has offered remedies to solve or, at least, attenuate these problems. For example, information on online networks can be – at least from a technical point of view – easily collected on the basis of web crawling. For sure many of the problems of classical network studies such as sample selectivity currently arise in new disguise, e.g. due to application programming interfaces (API) which allow only limited and selective access to data about online activities (see González-Bailón et al. 2014). However, it is likely that those problems will be cured in the future and, hence, the structural characteristics of complete online networks can be determined without strong auxiliary assumptions.

Repeated snapshots at different points in time additionally allow the investigation of the evolution and dynamics of social networks. Thereby, unobtrusive and objective measures of actual instead of self-reported relationships help to avoid reactivity, recall errors, and other well-known problems of survey research (for these and further advantages see Lewis et al. 2008). Of course, this doesn't mean that all online data are free from these or other problems. For example, social media analyses often depict a heavily distorted picture of actual human relations due to the fact that most websites have specialized on meeting certain demands (Ruths and Pfeffer 2014). Taking into account such limitations and pitfalls, the digital revolution offers new and powerful ways to deepen our understanding of human interactions. As a case in point, jointly with the development of sophisticated statistical models, advances have been made in disentangling contagion from homophily (Lewis et al. 2012) – even though only reliance on experimental intervention by researchers or natural exogenous shocks seems to allow the clean identification of both processes (Manski 1993). Although the majority of studies in this and the other above mentioned fields have applied the scientific method of formulating testable theoretical explanations and confronting them with empirical evidence, some have proclaimed that the digital revolution has brought about a fundamental change in the scientific method itself.

3 The End of Theory and the Age of Prediction

In a short, but widely read and heatedly debated essay "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" Anderson (2008) proposed that it is not required to have a good theoretical understanding of causal processes and relationships to make good predictions. Moreover, he argues, clinging to the well-established and widely applied scientific model of formulating informative theories and testing them empirically might even hinder progress of knowledge. According to Anderson – due to the sheer amount of available information and the

massive advances in computational power – it is sufficient to explore the data by means of statistical data mining and let the numbers speak for themselves: "We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." The successes of correlational consumer research and online recommendations (Couldry and Turow 2014), health care (Murdoch and Detsky 2013), early warning systems for crowd-related disasters (Haase et al. 2016), and crime hotspot detection (O'Brien et al. 2015) seem to corroborate Anderson's provocative conclusion that "correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." It appears that we have left the age of explanation and entered an age of prediction.

*In contrast to these claims, we argue that big data do not replace theory, but clearly highlight the need of thorough theoretical reasonin*g. Certainly, pattern recognition, machine learning, and other correlational techniques of data analysis are extremely powerful tools to explore large amounts of information. And certainly, these methods can give researchers many helpful hints on aspects and variables that might have been overlooked so far. Without any doubt application of these techniques can contribute to scientific progress. Even more in the short run practical impact of a correlational approach might be more effective than searching for the underlying causes of a phenomenon. For example, studying individual mobility profiles on the basis mobile phone data, Eagle and colleagues were able to predict cholera outbreaks and identify future cholera hotspots in Rwanda (Eagle and Greene 2014, p. 151 f.). From a practical point of view it is actually the duty of responsible policy makers to use this information for specifically targeting both the most affected geographical regions and transmission routes in order to prevent further infections and the spreading of the disease.

However, as helpful as such a correlational, purely data-driven approach might be in practice, it also bears several potential pitfalls to take data as given and mine them on purely statistical grounds. We want to discuss four issues in more detail here. First, understanding data as "raw data is an oxymoron" (Gitelman 2013). There simply is no such thing as raw data. Data have to be collected somehow, by somebody, and for some purpose; they are not independent from this context of data collection (see also Borgman 2014). In addition, in the course of an analysis plenty of (explicit or implicit) decisions have to be made ranging from data preparation issues over choice of statistical procedures to presentation of results. Thereby the meaning of variables and results time and again is a matter of interpretation and only becomes clear in the light of theoretical considerations.

Second, the sheer number of potential correlations one is often able to analyze entails a substantially increased risk of chance findings. Hence, as Bayesian statisticians have already emphasized for quite a while it is demanded to put stronger emphasis

on out of data predictions to avoid overfitting and fishing for significance (see recently Watts 2014). As for sociological theories, the proof of the pudding of statistical models lies in out-of-sample predictions and their empirical test on the basis of data which were not used for calibration of model parameters. Thereby, it is not sufficient to show that data collected shortly before an event of interest allow accurate prediction. For example, it is well-known that masses begin to move in a certain pattern shortly before the occurrence of a crowd disaster (Moussaid et al. 2011). Hence, observing such a pattern of crowd turbulence is an excellent predictor for subsequent injuries and deaths but it seems a matter of definition whether this is not already part of a crowd panic and it is clear that this approach does not provide an adequate answer to the question "what drives crowd disasters?".

Third, as is long-known in the philosophy of science, the fact that assumptions are consistent with empirical evidence does neither imply that underlying premises are correct nor that they will yield correct forecasts in the future (Nagel 1963). Leaving aside general problems of correct predictions in complex social systems (Martin et al. 2016) extrapolation of results becomes particularly problematic if the context-specific boundary conditions change, e.g., if we want to transfer the results from Rwanda to Haiti. Hence, to assess the generalizability of prediction models it is recommendable to additionally check the adequacy of forecasts for different populations, social contexts, and points in time. Theory can here offer a helpful bridge to transport local results to other contexts (see Deaton 2010).

Fourth, although prediction models such as the one on cholera hotspots and travel routes are extraordinary useful in practice, it would be even more valuable to have an in-depth understanding of the generative causal processes that bring about cholera infections and their spreading. Knowledge of such causal pathways or mechanisms, as called for by analytical sociologists (Hedström and Bearman 2009), is not only helpful out of academic curiosity, but also allows to uncover decisive points in the causal chain and hence enables practitioners to develop and implement more effective interventions. For example, for the above given example Eagle acknowledged in an interview that he had to realize that: "The model was not predicting cholera outbreaks, but pinpointing floods." (see Shaw 2014; p. 33). Since these catastrophes both significantly impair individual commuting by blocking roads and considerably increase the risk of a cholera outbreak, mobility data can be used for an early warning system. However, in the long run it seems wise to focus on the actual mechanisms and take measures against floods instead of curing the symptoms.

4 The Age of Messy Data

The nature and origin of big data often substantially limits the validity of empirical results. Big data are frequently "found" and the by-product of "real-world" processes rather than the result of a clearly designed study. For example, companies like

mobile- and internet-providers collect data for their own internal purposes (e.g., marketing, developing infrastructure plans) but do not have in mind scientific interests such as generating a valid measurement of a theoretical construct, isolating the causal effect of X on Y, or generalizing results to a certain population. Many advertisers of the digital revolution in the social sciences gladly neglect the reality that big data are frequently pretty messy, sometimes only insufficiently map theoretical constructs, suffer from sample restrictions as regards the general population, and lack context (see also Boyd and Crawford 2012). As Lazer et al. (2014: 1203) highlight: "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data".

Data errors are one issue. Since data collection is mostly automatized and relevant information like users' queries, self-reported characteristics and comments on online platforms is not or not fully standardized, big data do usually not only contain informative signals but also plenty of noise (Silver 2012). Another reason for this is that some indicators are pretty accurate in most cases but can dramatically fail in others. For example, the geolocation of cell-phones over time might describe individual mobility patterns for many users extremely well, but provides useless and misleading information if a person has forgotten the smartphone at home or lends it to someone else. The sheer amount of data prevents researchers from manually searching for such inconsistencies and errors making it unlikely to detect and fix them. At best researchers can try to validate samples of observations and run some consistency checks. Although specific tools are currently in development to cure big data from such obvious illnesses (e.g., artificial intelligence is making fast progress), at least presently data errors are still more problematic than in regular studies.

Murky indicators and measurement errors are other important issues. For example, research on social networks is sometimes based on only extremely rough approximations, such as online friendships, for actual social ties (Ruths and Pfeffer 2014). It goes without saying that findings based on such measures are not necessarily transportable to "real-world" networks. Large-scale research on scientific collaborations provides a further example of murky indicators. The main interest of this literature focuses on how the composition of a research team affects innovation and success. However, failed research often leaves no or less digital traces in the web, while successful collaborations tend to be continued (Guimerà et al. 2005). Thus, this strand of research is restricted to "successful" research projects in the form published articles or registered patents. As is generally well-known among methodologists (e.g., King et al. 1994; Chap: 4) selection of observations on the basis of a variable that is itself the outcome of interest in an investigation is fatal for the validity of empirical results. Selection on the outcome, in the given example about success of scientific collaborations, introduces systematic biases, irrespective of how big and rich the analyzed data are.

5 A Revival of Classical Tools of Social Research

We conclude from these observations in the previous section that a stronger interlinkage between classical tools of social research and computational social science is sorely needed. For many decades now objectivity, reliability, and validity have been core issues in empirical social science research. This expertise should certainly not be neglected in emerging disciplines like data science and the computational social sciences and instead should be utilized to produce valid and robust empirical results. Classical methods of social research, both quantitative and qualitative, have much to add to these nascent fields. For example, instead of proclaiming an end of surveys it appears much more promising to bring survey research and big data tools to a healthy marriage. As Callegaro (2016) argued in a keynote at the University of Mannheim big data appear to be particularly suited to answer what-questions about actual behavior whereas surveys seem to be especially useful to tackle why-questions of driving forces behind human actions, such as values, attitudes, and opinions. The variety of information sources is one of the defining features of big data and their specific strengths should be utilized to improve answers on both classical and new research questions.

Against this background we argue that two tools of data handling will become especially important in the near future. On the one hand, it is necessary to bring different data sources together to generate a richer, more informative dataset. For example, the combination of administrative and survey data promises major advances in various fields such health and labor market research as well as public administration (Connelly et al. 2016). Techniques for data linkage will thus substantially gain in relevance (see Harron et al. 2016 for an overview). While the ideal typical case would be that each dataset contains unique identifiers for units (individuals, firms etc.) and one simply has to merge the files on the basis of this ID variable, it is much more common that the data only contain information that allows probabilistic inferences about which observation in one data frame belongs to another observation in a second data base. Statistical analyses on the basis of such linked dataset should reflect this uncertainty related to the linkage procedure.

On the other hand, statistical techniques to handle missing data (Rubin and Little 2002), in particular multiple imputation and weighting, also become increasingly important in the course of the digital revolution. Since imputation tools provide statistical predictions for variables of interest based on available information, they can help to identify and correct inconsistencies in the data. However, the importance of these techniques is not only related to the frequent messiness of big data but also to their incompleteness. Since big data are at the same time rich in information, it suggests itself to draw – again probabilistic – conclusions about missing values by means of multiple imputation or weighting. For example, survey methodologists already exploit paradata from surveys such as call records and response

latency measures to improve coverage, reduce nonresponse bias, and attenuate measurement error (Kreuter 2013).

Not only as regards statistical techniques of data linkage and multiple imputation a stronger interlinkage between classical tools of social research and computational social science is sorely needed. A design-based approach to causal inference constitutes a particularly promising route for a quantitative digital sociology. In the social sciences such a turn from sophisticated statistical analysis to clever research design can be observed in recent years (Morgan and Winship 2015). Scholars more and more begin to acknowledge that „you can't fix by analysis what you bungled by design." (Light et al. 1990: i). With observational data alone it is typically quite difficult to isolate the effects of interest, while experimental designs, at least in theory, allow much cleaner identification of causal impacts (Rosenbaum 2010).

The interlinkage of an experimental approach with a reliance on digital traces helps to overcome fundamental concerns brought forward against each of the two approaches. On the one hand, experimental research frequently suffers from small, non-generalizable student samples, reactivity, and artificiality. On the other hand, online research typically lacks field control and exogenous stimuli resulting in limitations regarding causal inference. While a computational social science brings large, easy to collect and rather cost-efficient information and mostly unobtrusive measures of actual behavior to the table, a design-based approach strengthens the internal validity of causal inference.

A seminal study by Salganik, Dodds, and Watts (2006) on social dynamics in cultural markets strikingly illustrates the value added by staging unobtrusive field experiments in an online context. The scholars wanted to investigate the causal effect of social influence on further success of cultural products. More specifically, the theoretical prediction was that knowledge about consumption decisions of other users in the form of a chart table or download statistic causes a self-enforcing process of imitation and increases social inequality in the market. Unfortunately, it is nearly impossible to clearly answer this research question with observational data for an obvious reason: being a bestseller might simply signal the high quality of a song, film, or book – a thing that is inherently hard to measure and, hence, almost impossible to sufficiently control for in statistical models. Faced with this identification problem Salganik and colleagues decided to upload songs of nonfamous bands on a website, create multiple worlds by experimentally varying the reported download statistics, and collect information on user behavior on the homepage as regards listening, downloads, and ratings. Empirically, the multiple experimental worlds were characterized by hard to predict social dynamics and a high degree of social inequality. Product success was only weakly influenced by product quality, depended strongly on initial conditions, and was the result of path-dependent processes of herding behavior.

6 Statistical Inference, Pattern Recognition, and Data Visualization

The sheer amount of digital information forces researchers to reconsider established statistical approaches. With millions of observations even tiny effects and differences become statistically significant. Nonetheless they are probably irrelevant in substantive and practical terms. Additional concerns about the adequacy of statistical inference arise from full population coverage and an increasing number of potential correlations. The former raises general questions about the meaning and adequacy of the concept "statistical significance", while the latter can substantially inflate the danger of finding spurious associations due to chance findings.We want to shortly highlight three areas that deserve increased attention in the coming years. First, tools from Bayesian statistics (Gelman et al. 2013; Gill 2014; Jackman 2009) are helpful in the context of big data. This is not only the case because Bayesians give up the concept of significance in favor of an emphasis on perceived credibility, but also due to the idea of learning from a stream of data, updating one's subjective beliefs, and testing predictions with new data. A second core competence in the digital age is the ability to collect, store, and munge such massive data masses in an efficient way. This often requires familiarity with high-performance computing and data banks. If one aims to collect existing data from the internet or stage online experiments, additional skills like HTML programming and web crawling are indispensable. Third, data analysis needs to be reconsidered as well. Since it becomes increasingly difficult to fully exploit the potential of data which contain more and more information, tools of pattern recognition and data visualization gain in relevance. The same holds for techniques for the analysis of textual, graphical, and real-time data which rely on some sort of artificial intelligence such as machine learning.

7 Conclusion

Let us come back at the end of this paper to the initially mentioned claim of Savage and Burrows (2007: 895) that "the repertoires of empirical sociology need to be rethought" in order to avoid the next crisis of the discipline. In general, their diagnosis is certainly correct that sociology is endangered of falling behind and has already lost ground (see also Burrows and Savage 2014), since once innovative research tools like surveys and regression analysis do not secure the unique selling position of sociology among the different disciplines anymore. As a matter of fact, a lot of interesting and important "sociological" research is conducted outside of the classical realm of the discipline and published elsewhere. Fast progress is made in fields like data science and computational social science both on methodological and substantive grounds. Sociology must not blind itself from these insights and must actively participate in these scientific discourses.

However, it is not only in the interest of sociology to become part of this interdisciplinary community, but it is also desirable for nascent enterprises such as data science and computational social science (see also Salganik 2017). As shown in this

paper, sociology can make valuable contributions to these endeavors both on theoretical and methodological grounds. On a theoretical level, sociology can be an important corrective to purely data driven approaches to prediction tasks and can help to attenuate the dangers of extrapolation and out-of-sample predictions with its emphasis on explanations and causal mechanisms (see also González-Bailón 2013, 2017). On a methodological level, sociology has to offer many insights into how to sample populations, how to develop valid and reliable measure, and how to design social research allowing rigorous causal inference (see also Grimmer 2015). Recent methodological innovations share many problems and pitfalls with classical social research and can hence profit from the lessons learned in the past. It is not necessary to reinvent the wheel and to repeat mistakes of the past.

To reap this potential of the big data revolution in sociology, firm knowledge of methodological innovations is essential. Hence, to make a substantial and enduring contribution to the emerging field of computational social science, sociologist will need to reconsider the structure of their study programs putting a stronger emphasis on programming skills, management of data banks, and the large-scale analysis of spatial and temporal process. Besides these technical skills a reformed curriculum should certainly contain further topics which we couldn't discuss in as much detail as necessary in the course of this short contribution (for further reading Helbing et al. 2016; Mayer-Schonberger and Cukier 2013). In particular, students should understand methodological consequences of the use of big data such as reproducibility restrictions and changes of the typical research process, should learn about ethical and legal issues related to data linkage, privacy, and informed consent, and should be made aware of potential misuses like the manipulation of opinion, social disintegration, and authoritarian oppression.

Literatur

Anderson, Chris (2008): The end of theory: The data deluge makes the scientific method obsolete. Wired; 23.8.2008, Link: http://www.wired.com/2008/06/pb-theory/ (retrieved: 9.2.2016).

Borgman, Christine L. (2014): Big Data, Little Data and Beyond. Cambridge, MA.

Boyd, Danah/Crawford, Kate (2012): Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon. Information, Communication and Society 15(5): 662-79.

Burrows, Roger/Savage, Mike (2014): After the crisis? Big Data and the methodological challenges of empirical sociology. Big Data & Society 1(1): 1-6.

Callegaro, Mario (2016): Keynote "Importance of surveys in the era of big data", Kickoff Meeting International Program in Survey and Data Science, University of Mannheim, 20.2.2016.

Centola, Damon (2018): How Behavior Spreads. The Science of Complex Contagions. Princeton.

Connelly, Roxanne/Playford, Christopher J./Gayle, Vernon/Dibben, Chris (2016): The role of administrative data in the big data revolution in social science research. Social Science Research 59: 1-12.

Couldry, Nick/Turow, Joseph (2014): Advertising, big data and the clearance of the public realm: marketers' new approaches to the content subsidy. International Journal of Communication 8: 1710-1726.

Deaton, Angus (2010): Instruments, randomization, and learning about development. Journal of Economic Literature 48(2): 424-55.

Diekmann, Andreas/Jann, Ben/Przepiorka, Wojtek/Wehrli, Stefan (2014): Reputation formation and the evolution of cooperation in anonymous online markets. American Sociological Review 79(1): 65-85.

Eagle, Nathan/Greene, Kate (2014): Reality Mining: Using Big Data to Engineer a Better World. Cambridge, MA.

Gelman, Andrew/Carlin, John B./Stern, Hal S./Dunson, David B./Vehtari, Aki/Rubin, Donald B. (2013): Bayesian Data Analysis (3rd edition). London.

Giles, Jim (2012): Making the links: From e-mails to social networks, the digital traces left life in the modern world are transforming social science. Nature 488: 448-50.

Gill, Jeff (2014): Bayesian Methods: A Social and Behavioral Sciences Approach (3rd edition). London.

Gitelman, Lisa (2013): Raw Data is an Oxymoron. Cambridge, MA.

Golder, Scott/Macy, Michael (2014): Digital footprints: Opportunities and challenges for online social research. Annual Review of Sociology. 40:6.1–6.24

González-Bailón (2013): Social science in the era of big data. Policy & Internet 5(2): 147- 160.

González-Bailón (2017): Decoding the Social World: Data Science and the Unintended Consequences of Communication. Cambridge, MA.

González-Bailón, Sandra/Wang, Ning/Rivero, Alejandro/Borge-Holthoefer, Javier/Moreno, Yamir (2014): Assessing the bias in samples of large online networks. Social Networks 38:16-27.

Grimmer, Justin (2015): We are all social scientists now: How big data, machine learning, and causal inference work together. Political Science & Politics 48(1): 80-83.

Guimera, Roger/Uzzi, Brian/Spiro, Jarrett/Amaral, Luis A.Nunes (2005): Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722): 697-702.

Haase, Knut/Al Abideen, Habib Zain/Al-Bosta, Salim/Kasper, Mathias/Koch, Matthes/Müller, Sven/Helbing, Dirk (2016): Improving pilgrim safety during the Hajj: An analytical and operational research approach. Interfaces 46(1): 74-90.

Harron, Katie/Goldstein, Harvey/Dibben, Chris (eds.) (2016): Methodological Developments in Data Linkage. Chichester.

Hedström, Peter/Bearman, Peter (eds.) (2009): The Oxford Handbook of Analytical Sociology, Oxford.

Helbing, Dirk/Frey, Bruno S./Gigerenzer, Gerd/Hafen, Ernst/Hagner, Michael/Hofstetter, Yvonne/van den Hoven, Jeroen/Zicari, Roberto V./Zwitter, Andrej (2016): Digitale Demokratie statt Datendiktatur. Spektrum der Wissenschaft 2016.1: 50-58.

Jackman, Simon (2009): Bayesian Analysis for the Social Sciences. New York.

Keuschnigg, Marc/Lovsjö, Niclas/Hedström, Peter (2018): Analytical sociology and computational social science. Journal of Computational Social Science 1(1):3-14.

King, Gary/Keohane, Robert O./Verba, Sidney (1994): Designing Social Inquiry. Scientific Inference in Qualitative Research. Princeton.

Kreuter, Frauke (ed.) (2013): Improving Surveys with Paradata: Analytic Uses of Process Information. Hoboken, NJ.

Lazer, David/Kennedy, Ryan/King, Gary/Vespignani, Alessandro (2014): The parable of Google Flu: Traps in big data analysis. Science 343(6176): 1203-1205.

Legewie, Joscha/Schaeffer, Merlin (2016): Contested boundaries: Explaining where and when ethno-racial diversity provokes neighborhood conflict. American Journal of Sociology 122(1): 125-161.

Lewis, Kevin/Kaufman, Jason/Gonzalez, Marco/Wimmer, Andreas/Christakis, Nicholas (2008): Tastes, ties, and time: A new social network dataset using facebook.com. Social Networks 30(4): 330-342.

Lewis, Kevin/Gonzalez, Marco/Kaufman, Jason (2012): Social selection and peer influence in an online social network. PNAS 109: 68-72.

Light, Richard J./Singer, Judith D./Willett, John B. (1990): By Design. Planning Research on Higher Education. Cambridge, MA.

Manski, Charles F. (1993): Identification of endogenous social effects: the reflection problem. Review of Economic Studies 60(3): 531-42

Martin, Travis/Hofman, Jake M./Sharma, Amit/Anderson, Ashton/Watts, Duncan J. (2016): Exploring limits to prediction in complex social systems: Predicting cascade size on Twitter. Proceedings of the 25th International Conference on World Wide Web.

Mayer-Schonberger, Viktor/Cukier, Kenneth (2013): Big Data: A Revolution That Will Transform How We Live, Work and Think. London.

Morgan, Stephen L./Christopher Winship (2015), Counterfactuals and Causal Inference: Methods and Principles for Social Research (2nd edition). Cambridge, MA.

Moussaid Mehdi/Helbing, Dirk/Guy, Theraulaz (2011): How simple rules determine pedestrian behavior and crowd disasters. PNAS 108: 6884–6888.

Murdoch, Travis B./Detsky, Allan S. (2013): The inevitable application of big data to health care. Journal of the American Medical Association 309(13): 1351-1352.

Nagel, Ernst (1963): Assumptions in economic theory. American Economic Review: Papers and Proceedings 53(2): 211-19.

O'Brien, Daniel/Sampson, Robert J./Winship, Christopher (2015): Ecometrics in the Age of Big Data: Measuring and Assessing 'Broken Windows' Using Large-scale Administrative Records. Sociological Methodology 45: 101-147.

Rosenbaum, Paul R. (2010): Design of Observational Studies. München.

Rubin, Donald B./Little, Roderick J.A. (2002): Statistical Analysis with Missing Data (2nd edition). New York.

Ruths, Derek/Pfeffer, Jürgen (2014): Social media for large studies of behavior. Science 346(6213): 1063-1064.

Salganik, Matthew J. (2017): Bit by Bit: Social Research in the Digital Age. Princeton.

Salganik, Matthew J./Dodds, Peter Sheridan/Watts, Duncan J. (2006): Experimental study of inequality and unpredictability in an artificial cultural market. Science 311(5762): 854-856.

Savage, Mike/Burrows, Roger (2007): The coming crisis of empirical sociology. Sociology 41(5): 885-899.

Schich, Maximilian/Song, Chaoming/Ahn, Yong-Yeol/Mirsky, Alexander/Martino, Mauro/Barabási, Albert-László/Helbing, Dirk (2014): A network framework of cultural history. Science 345(6196): 558-562.

Shaw, Jonathan (2014): Why "big data". Harvard Business Magazine March/April: 30-35 and 74-75.

Silver, Nate (2012): The Signal and the Noise: Why So Many Predictions Fail – But Some Don't. London.

Van de Rijt, Arnout/Kang, Soong Moon/Restivo, Michael/Patil, Akshay (2014): Field experiments of success-breeds-success dynamics. PNAS 111: 6934-6939.

Watts, Duncan J. (2014): Common sense and sociological explanations. American Journal of Sociology 120(2): 313-351.

Prof. Dr. Tobias Wolbring
Chair of Empirical Economic Sociology
School of Business and Economics
Friedrich-Alexander University Erlangen-Nürnberg
Findelgasse 7/9
90402 Nürnberg
tobias.wolbring@fau.de