International Journal of Communication 16(2022), 1482–1505 1932–8036/20220005

Unboxing Computational Social Media Research From a Datahermeneutical Perspective: How Do Scholars Address the Tension Between Automation and Interpretation?

JAKOB JÜNGER
STEPHANIE GEISE
MARIA HÄNELT
University of Münster, Germany

Communication researchers have fruitfully applied computational methods in their analysis of communication processes. However, the automation of scientific data collection and analysis confronts scholars with fundamental epistemological and practical challenges. Particularly, automation implies that the processing of data is highly standardized for all cases. In the context of social science research, this contrasts with the expectation that meaning is always attributed in individual interaction processes. Based on a literature review of peer-reviewed journal articles, our study explores the resulting tension between automated and interpretive research. We first analyze the extent to which automated methods play a role in social media research. We then identify the challenges and limitations researchers addressed in their studies. On this basis, we propose steps for a data hermeneutical perspective that combines computational methods with interpretive approaches.

Keywords: computational communication science, computational social science, computational methods, automated data collection, process-generated data, data hermeneutics, interpretive paradigm

Automation From an Interpretive Perspective

A central characteristic of computational communication science is its focus on computer-based methods—such as machine learning, text and data mining, computer simulations, or generating data sets from social media platforms—to investigate social phenomena and processes associated with digitization in the widest sense (Cioffi-Revilla, 2017). In light of recent technological and social transformations, these methods are prominent in current scientific debates (e.g., Alvarez, 2016; Choi, 2020; Lazer, Pentland, Adamic, & Aral, 2009; Shah, Cappella, & Neuman, 2015). They make it easier for researchers to systematically collect and analyze large quantities of data (Hox, 2017). Accordingly, computational

Jakob Jünger: jakob.juenger@uni-muenster.de
Stephanie Geise: stephanie.geise@uni-muenster.de
Maria Hänelt: maria.haenelt@uni-muenster.de
Date submitted: 2021-02-26

Copyright © 2022 (Jakob Jünger, Stephanie Geise, and Maria Hänelt). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at http://ijoc.org.

communication science has been characterized as involving large and complex data sets that often consist of digital traces and other "naturally occurring" data, require algorithmic solutions to analyze, and allow for the study of human communication by applying and testing communication theory (Shah et al., 2015; van Atteveldt & Peng, 2018).

As scholars have begun to take advantage of computational approaches to answer fundamental questions about human behavior, interaction and communication, the field of computational communication science has recently emerged. The increasing availability of new types of data and computational methods makes it possible to explore and empirically test ideas that could not be tested with classical methods (González-Bailón, 2017), and some researchers have argued that communication science is therefore about to undergo an "unprecedented boost to progress" (van Atteveldt & Peng, 2018, p. 82).

At the same time, the application of computational methods has been criticized as primarily data driven and lacking theoretical positioning, both for understanding the social phenomena under study and for developing an analytical view of the research process. While data and interfaces used for research purposes were originally created to stimulate user behavior (Keyling & Jünger, 2016), the communication processes under study may not always mirror an independent social reality, or they may also mirror processes oriented toward private or governmental institutions' purposes (Couldry & Hepp, 2017). Valid criticism has been made about the reliability, validity, and reproducibility of computational methods applied in communication science (e.g., Alvarez, 2016; Hargittai, 2015, 2018; Lazer et al., 2009; Stockemer, Koehler, & Lentz, 2018; van Atteveldt, Strycharz, Trilling, & Webers, 2019; Wallach, 2016). In addition, in recent articles, scholars have increasingly argued that computational social research should consider theoretical grounding (e.g., Waldherr, Geise, & Katzenbach, 2019; Waldherr, Geise, Mahrt, Katzenbach, & Nuernbergk, 2021), ethical reasoning (e.g., Kosinski, Stillwell, & Graepel, 2013; van Atteveldt et al., 2019), and technical challenges (e.g., van Atteveldt & Peng, 2018) more carefully.

On the one hand, these debates reflect scientists' attempts to deal with the challenges of computational methods and their application. On the other hand, they refer to the *epistemological problem* of how computational methods relate to fundamental assumptions of social research. For example, scientific knowledge production, social action, and the organization of technical infrastructures are inherently intertwined (Marres & Weltevrede, 2013)—with research automation appearing to be at odds with basic methodological assumptions about *human interaction*. In this regard, Marres (2017), for example, pointedly states, "Yes, computational social science presents a problem, in the good sense, and this problem is as much an intellectual one as anything else: how is social inquiry possible under the conditions of interactivity?" (p. 190).

Although a growing number of studies have discussed the promising and limiting issues at the methodological level (e.g., Driscoll & Walter, 2014; Howison, Wiggins, & Crowston, 2011; Hox, 2017; Jünger, 2018; Keyling & Jünger, 2016; Mahrt & Scharkow, 2013; Shah et al., 2015; van Atteveldt & Peng, 2018; van Dijck, 2014), it is unclear how researchers regularly deal with challenges concerning the conditions of justified knowledge based on automated processes of data collection and data analysis. In this respect, the automation of data collection and analysis processes is a kind of black box whose inner structure, functions, and operations are potentially unclear in the research process. Borrowing from

theoretical frameworks such as systems theory (Ashby, 1957; Bunge, 1963; Luhmann, 1993, p. 156) or actor-network theory (Latour, 2002, p. 373), we thus use the black box metaphor to describe the challenge researchers face in relation to the automation of data collection and analysis. Ultimately, the intention to further open up this black box leads to our central question:

RQ: How do scholars address the tension between automation and interpretation?

Traditionally, aiming to reconstruct how humans make sense of social reality the social sciences have not only valued standardized measurement of phenomena under examination but have also systematically applied a dual hermeneutics, thus acknowledging interpretative practices in research processes. Because from a *representative perspective* empirical social research aims at systematically collecting data on social phenomena to which conclusions can be drawn by observation, surveys, interviews, or by collecting process-generated data, computational data collection and analysis can be valued as a way of increasing efficiency. Following an *interpretive paradigm*, this view, however, is challenged: Based on the basic theoretical assumption that all interaction is an interpretive process in which actors relate to each other through meaning-making interpretations, social reality is constituted by acts of interpretation (Giddens, 1984; Marres, 2017). The social contexts and phenomena studied by the researcher can therefore not be understood as objectively given and deductively explainable social facts, but as the result of an interpretive interaction process that largely defies automation—or at least raises fundamental epistemological problems (Jünger, forthcoming).

Aiming to more closely inspect and reflect on these challenges, we conducted a systematic literature review of scholarship on digital communication processes in social media to gain an overview of computer-based methodological approaches. In doing so, we follow a constructively critical approach: With our analysis we sound out how the basic scientific assumptions of traditional quantitative and qualitative methodology can accommodate recent computational methods. Our intention is that this consideration helps us to better understand and improve the application of recent computational methods. Contributing to a more nuanced picture of the computational black box resulting from the intertwined behavior of platform users and providers as well as from the applied research methods themselves, we ask the following:

RQ1: What computational methods have been applied in the field of social media research?

RQ2: What challenges do researchers address about data collection and analysis in computational communication science?

While we do not assume that every researcher shares the assumptions of the interpretive paradigm, we value this perspective and acknowledge it as fruitful background to highlight theoretical assumptions of computational methods. Building on our overview of scholarly applications and reflections, we are interested in ways in which interpretive and automated dimensions of computational methods are combined, asking how interpretability is related to computational methods in social media research:

RQ3: How can researchers address the tension between computational and interpretive approaches?

On the basis of our literature review, we develop a model demonstrating that the interactional and representative perspectives are not mutually exclusive in computational communication science but can complement each other usefully. Building on our examination, we draw the conclusion that computational methods can bridge the long-lasting differentiation between qualitative and quantitative research approaches if the interactivity of the research object is considered more closely.

Background: Computational Communication Research Methodology

Approaches of computational communication science touch on two key aspects: The first relates to the nature of the data being studied—process-generated, large-scale, complex data, such as digital traces produced on and by social media platforms—usually before the researcher intentionally entered the fray. The second aspect highlights the special circumstances of the scientific research process and concerns computer-based automation tools and processes. Both of these aspects—the data and the methods—have consequences for social research that are briefly summarized in the following sections.

Computational Communication Science as Automation of Research

In general, empirical research processes begin with data collection, in which the world under investigation is transformed into data (Schnell, Hill, & Esser, 2013). Communication researchers mostly use empirical methods such as surveys, interviews, or document collection to set up this transformation. The second research stage is the transformation of data into data sets through data preparation that includes different kinds of coding or statistical aggregation procedures. Finally, by analyzing the data in the last stage, the researcher transforms data sets into propositions about the world.

In the data sciences, the transformation of data is often extensively elaborated. For example, the widespread notion of the data pyramid places data at the base of the pyramid (indicating low meaning and value) and knowledge or wisdom at the top (Rowley, 2007). A similar conception of the research process from a more technical perspective involves knowledge discovery in databases (Cleve & Lämmel, 2014; Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Both conceptions assume that by transforming data into information, the data is enriched with meaning, which finally leads to knowledge.

While data collection and analysis are primarily carried out through *manual* collection and interpretation in the social sciences, in the field of computational communication science, at least parts of the research processes are *automated* by computer programs. As automation in general is defined as the procedure of making a process operate without manual control ("Automation," 2018), automated data collection procedures are practices in which data is not generated by manual coding using questionnaires but by using algorithms in the form of computer programs or scripts. In digital media environments, automated data collection techniques include Web scraping (extracting data from websites), access to application programming interfaces (APIs), and aggregation of tracking data, log files, or similar data sets (Jünger, 2018, p. 107; Keyling & Jünger, 2016). In contrast to manual data collection, the marginal costs for collecting additional data approach zero as the data volume increases (Monroe & Schrodt, 2008).

With regard to the stage of data analysis, computation has a long-standing tradition in statistics, and this makes it difficult to conceptionally differentiate between pure computational and noncomputational approaches. In the recent scientific discourse, the following techniques are usually considered as computational methods: automated text analysis, network analysis, simulation methods, and machine learning (e.g., Cioffi-Revilla, 2017, pp. 12–18). In that vein, computational data analysis encompasses all procedures using computer algorithms that go beyond standard statistics (Cioffi-Revilla, 2010).

The high degree of efficiency and innovativeness of computational methods, however, comes with its own price. While scientists have demonstrated that computational approaches make it possible to find meaning among digital data, providing unprecedented fine-grained and diverse information about human communication and interaction behavior (e.g., Mukerjee, Majó-Vázquez, & González-Bailón, 2018; Wettstein, 2020; Yarchi, Baden, & Kligler-Vilenchik, 2021), automation, ultimately, is a strong type of standardization in which all units are processed identically. Because computational automation therefore generally implies that the collection, processing, and analysis of data are unvarying, it bears the risk that particular cases and specific details may be overlooked—particularly when working with "big" computational approaches processing large volumes of data.

Yet, and probably more importantly, a high degree of standardization often contrasts with fundamental theoretical and methodological assumptions that underscore the central position of the researcher in understanding and interpreting the data collected. According to the interpretive paradigm, meaning results from individual attribution processes that depend strongly on the context, the situation, and the researchers' dispositions. This leads to two crucial assumptions about the research process. First, data do not have any meaning in themselves, but acquire it only through the interpretation of the individual researchers (Giddens, 1984; Wilson, 1973)—regardless of whether the data collection is carried out by automated or manual procedures or whether the analysis follows statistical or hermeneutic principles. Secondly, since the sense of social action is always negotiated in specific situational contexts, the everyday practices of human meaning attribution have to be embedded in the scientific knowledge process. Accordingly, by the terms *indexicality* and *reflexivity*, ethnomethodology has pointed to the context-bound nature of actions and understands social research as an investigation of "contingent ongoing accomplishments of organized artful practices of everyday life" (Garfinkel, 1967, p. 11). Building on these ideas, in the following analysis, we ask how researchers deal with the interplay of machines and humans in the stages of data collection and analysis.

Consequences of Process-Generated Data

Computational methods are closely related to so-called big data phenomena (Hox, 2017, p. 3). These data are defined as having a high volume, variety, and velocity (Laney, 2001). The amount and complexity of big data comes with challenges, particularly when data do not fit into one device, and thus the analysis needs to be distributed (Cox & Ellsworth, 1997). Nevertheless, epistemological consequences are related more to the source and content of big data. In contrast to data collected purposefully for research, computational social science often deals with "naturally occurring" (Shah et al., 2015, p. 7) or behavioral "trace" data (Lazer et al., 2009, p. 721; Welser, Smith, Gleave, & Fischer, 2008, p. 117) data. One example of communication environments in which such data are generated is *social media*, which

Kaplan and Haenlein (2010) define as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (p. 61). In these applications, the processes inscribed into the platforms incidentally generate data for potential further examination.

From the social science methodology viewpoint, data collected at an earlier time by someone other than the current researcher are known as *secondary data* (Johnson & Turner, 2003). While secondary data include data left behind by previous research projects, data on social media platforms appear as by-products of human interactions and social processes and thus are comparable with data in administrative bookkeeping systems (Bick & Müller, 1980), which social sciences call process-generated data (Baur, 2011). However, the circumstances of process-generated data have consequences for the stages of both data collection and analysis.

Issues about data collection begin before the data collection even starts. In process-generated data, communication automatically produced by bots as well as strategic communication by organizations mingle with individuals' messages (Pfaffenberger, 2018; Woolley, 2016). As long as such procedures and structures remain hidden behind process-generated data and are not explicitly analyzed and decoded, they remain black boxes for the researcher in the sense that they are a source of possible research distortions and resulting misinterpretations, which can obscure what the extracted information really stands for (Driscoll & Walter, 2014). Especially when working with data from social media platforms, different types of bias can occur (Ruths & Pfeffer, 2014). For example, users of social media platforms do not represent the whole society, nor do they necessarily represent specific subgroups of Internet users (Hargittai, 2018). Moreover, working with process-generated data may cause ethical issues because participants' informed consent is usually missing: "The process of evaluating the research ethics cannot be ignored simply because the data is seemingly public" (boyd & Crawford, 2012, p. 672).

Regarding data analysis, limitations of measured variables also have significant consequences. Given researchers' restricted control of the data-generation process, the operationalization of theoretical concepts is highly constrained (Bick & Müller, 1984). Frequently, researchers can only work with a predefined set of available metrics, and these are not necessarily the best proxies for their targeted constructs. The question of representation is also closely connected to established procedures of testing relationships by applying statistical procedures and measures. More precisely, in large data sets, even "insignificant findings seem meaningful because they achieve conventional thresholds of statistical significance" (Shah et al., 2015, p. 11).

Thus, while working with prearranged process-generated data can be convenient and promises new insights into previously hidden fields, it also challenges the application and evaluation of established quality criteria of empirical research, such as validity, reliability, objectivity, or ethics (Bryman, Becker, & Sempik, 2008; Lincoln, 1995). Such issues have been discussed in dedicated methodological studies (see Gerlitz & Rieder, 2013; Ho, 2020; Kwon, Priniski, & Chadha, 2018; Thelwall & Stuart, 2006; van Atteveldt & Peng, 2018). To supplement this perspective, our literature review examines how challenges of process-generated data are addressed in typical empirical studies.

Method: Literature Review of Social Media Studies

The aim of this study is to discuss the methodological challenges of communication studies applying computational methods. Aggregating researchers' experiences and considerations from a range of different studies not only provides a systematic overview of current scholarship it also helps to determine where research gaps and methodological challenges exist (Budgen & Brereton, 2006). Based on the evolving demarcation of computational communication science, we analytically focus on *data collection* and *data analysis* as two important stages of the research process in which computational methods are increasingly used. The field of *social media studies* seems particularly suited to answering our research questions since platforms such as Facebook and Twitter regularly produce process-generated data, allowing both automated data collection and computational analyses of communications and interactions. From the selected articles, we first identify the textual sections on methods and challenges and then code these with regard to aspects of our three research questions, particularly, applied *computational methods*, the *challenges* addressed, and the role of *interpretation* in the research process.

Sampling of Research Articles

Aiming to answer our research questions, we first identified communication research articles that applied different types of computational methods for data collection and analysis. We are not trying to provide a representative overview of the field, but rather focus on identifying typical challenges. For this reason, we selected a database frequently used in communication studies for conducting literature research and limit ourselves to studies that deal with established online platforms. We focus on prominent online platforms providing an application programming interface (API) to ensure that the selected studies can in principle make use of automated data collection.

The articles included in our review were selected from the full text database Communication and Mass Media Complete, which provides a wide range of articles on communication and media studies. We focused on peer-reviewed journal articles written in English and published in 2016 or 2017 that were available in the database in March 2018. Our review thus provides an overview of mostly high-quality scholarly work, with an average Scimago scientific journal rank of 1.6 (SJR).[1]

Since some years have elapsed since then, the selection does not mirror the most recent studies. Instead, we identify studies from a phase in which computational methods are becoming increasingly established in the discipline. We assume that it is primarily during this period that fundamental concerns with epistemological parameters can be found. The sample comprises standard communication research but not necessarily the most innovative studies, which are probably found in more specialized outlets or published as conference articles. However, the sample reflects an extract from the scientific community, which allows for assessment of early developments in the field of computational methods.

[1] See the appendix (https://doi.org/10.17605/OSF.IO/RYJ8G).

Our sampling followed the flow of the PRISMA standard that provides a systematic structure for documenting the steps involved in screening out studies (Moher, Liberati, Tetzlaff, & Altman, 2009).[2] In the first step, we manually screened all social media studies found in the database. To identify social media studies, we applied search terms referring to the most prominent social media platforms—namely "Facebook," "Twitter," and "YouTube" (Innes, Roberts, Preece, & Rogers, 2017; Stoycheff, Liu, Wibowo, & Nanni, 2017). In addition, to broaden the view to also include platforms such as Weibo, Instagram, or Snapchat, we used the phrase "social media" as a search term as well. This search strategy resulted in 497 distinct peer-reviewed articles written in English.

In the second step, we systematically screened out all studies that were not empirical. The screening was based on the abstracts and the methods sections of the articles and thus included only articles for which the full text was available, resulting in 260 original research articles. For further screening out all articles in which scholars did not apply computational methods, we made use of our definitions of computer-based approaches to data collection and analysis (see the Background section, above). All sampling decisions were documented with corresponding paragraphs and reviewed by two researchers. This filtering procedure resulted in 68 original research articles representing a sample of social media research that includes studies with a potentially high reach, providing a wide range of articles on communication and media studies.

Coding Procedure

In the next step, we conducted an in-depth analysis of the 68 identified studies, applying the methodological approach of qualitative content analysis (Kuckartz, 2012). To this end, we first read the full text and according to the three research questions identified all statements about computational methods, methodological challenges, and interpretive steps. Second, guided by the open and selective coding principles of grounded theory (Holton, 2010), we developed inductive subcategories and categorized the text passages accordingly. Every coding decision was documented with corresponding text passages. This resulted in a profile matrix (Kuckartz, 2012, p. 73) containing a row for each case, a column for each category, and text, paraphrases, and codes in each cell. This matrix allowed for both quantifying the categories and summarizing the text passages. The resulting category system was composed of three aspects.

1. *Computational methods:* We found five types of data collection methods: (a) using APIs with dedicated scripts, (b) working with specialized tools, (c) using data collection platforms, (d) buying data from providers, and (e) working with compiled databases. Methods of data analysis were also grouped into five categories: (a) corpus analysis, (b) content analyses based on dictionaries, (c) supervised machine learning, (d) unsupervised clustering techniques in the broadest sense, and (e) network analyses using, for example, community detection algorithms.

2. *Challenges and limitations:* We carefully selected every expression that referred to issues, challenges, and limitations of computational methods in data collection and data analysis in the

[2] The PRISM diagram can be found in the appendix (https://doi.org/10.17605/OSF.IO/RYJ8G).

broadest sense. The resulting list of issues regarding data collection includes (a) interactivity between platforms and users, (b) methodological issues (e.g., sampling decisions), and (c) technical or organizational challenges to data access. Issues related to analysis are concerned with (a) research designs, (b) problems with data processing, and (c) the interplay of scientific disciplines.

3. *Interpretation*: We identified which studies documented any interpretive steps. For the data collection and preparation methods, every method involving human reading of materials was considered. In addition to (a) hermeneutical approaches such as discourse analysis, we also included all forms of (b) classical content analysis. Moreover, human reading is involved in (c) training corpora used for automated classification tasks and (d) quality checks or (e) illustration of results. Some studies discussed (f) quantitative data from an interpretive instead of statistical perspective; we summarized them as data hermeneutics. Finally, one study conducted (g) qualitative interviews.

Based on the text passages, we evaluated how interpretive and automated procedures related to each other in each research project.

Computational Methods in Social Media Research

Our literature review provides two interconnected perspectives regarding the current debate on computational methods. First, we examine how communication scholars have analyzed social media by making use of different computational approaches and methods of data generation, collection, and analysis (RQ1). Then, we closely investigate the challenges of computational communication science (RQ2).

Methods of Data Collection and Analysis[3]

Scholars working with computational methods primarily examined microblogging services such as Twitter or Sina Weibo (65%), while others presented case studies examining multiple platforms (15%) or centering on Facebook (12%; Figure 1). Other sites such as YouTube or LiveJournal were less studied (9%). In all the studies in our sample, digital content was analyzed, demonstrating that this is a kind of precondition for many automated methods. Interestingly, only a few authors expanded their analytical angle by making use of multimethod approaches—for example, by combining content data with data extracted from surveys (9%) or panels (3%).

[3] The list of analyzed studies can be found in the appendix (https://doi.org/10.17605/OSF.IO/RYJ8G). We mark references to this list with an asterisk.

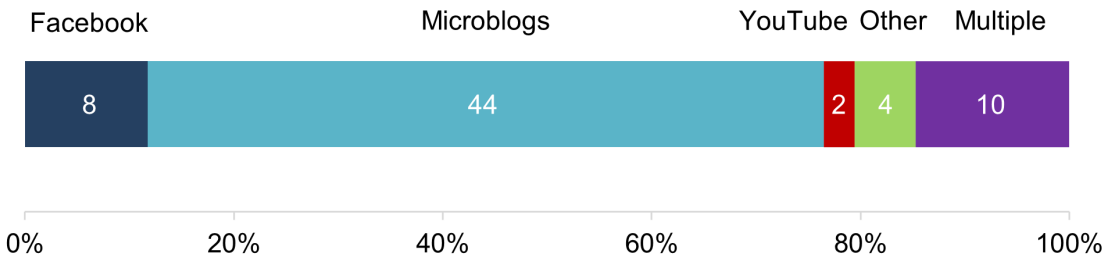

Figure 1. Platforms investigated in studies using automated methods. Basis: n = 68 social media studies with computational methods in 2016 and 2017. Absolute numbers inside boxes.

Regarding computerized data collection procedures, the largest share (34%) of studies relied on third-party platforms such as DiscoverText, Gnip, Crimson Hexagon, or Weiboscope (e.g., Rossi & Giglietto, 2016*; Zhao, 2017*; Figure 2). The mentioned services have very different relationships to their data sources. Gnip, for example, belonged to Twitter, giving paid access to data that were not accessible with the standard Twitter API. Meanwhile, their services are integrated into the Twitter API. As an academic project located at the university of Hong Kong, Weiboscope collects censored messages from Sina Weibo, the Chinese microblogging service. Crimson Hexagon is one of the bigger companies offering social media analytics; it relies on cooperation with platforms such as Twitter and Facebook. Some authors (10%) went even further, buying fully hydrated data sets from social media analytics companies (e.g., Akpinar & Berger, 2017*; Bulut & Yörük, 2017*). Thus, it is clear that researchers are dependent on cooperation with nonscientific organizations, which means that parts of the collection process take place outside scientific control.

Direct access to APIs (32%) was usually managed by self-written Python scripts (e.g., Bozdag & Smets, 2017*; Hayat & Samuel-Azran, 2017*). Some authors (10%) also reported using tools such as Facepager (e.g., Fenoll & Cano-Orón, 2017*; McKinnon, Semmens, Moon, Bamarasekara, & Bolliet, 2016*). While these methods allow for more control, they are also more demanding in terms of competences and skills. Using raw databases is uncommon in social media studies, as these would mostly be restricted to researchers inside the organizations. Rahman, See, and Ho (2017*) used an already publicly available compilation of videos to train machine learning algorithms. Overall, all forms of data access were subject to restrictions with uncertain effects on research results.

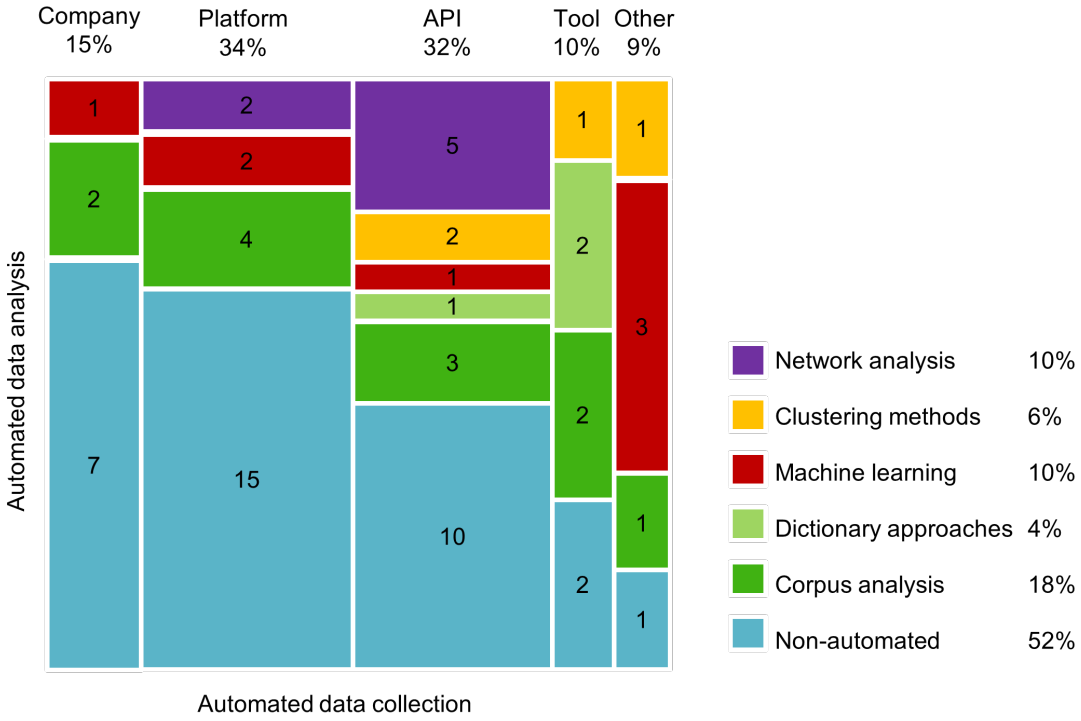

Figure 2. Automated data collection and analysis. Basis: n = 68 social media studies with computational methods in 2016 and 2017. Absolute numbers inside boxes.

In nearly half of the studies, computational procedures were implemented to analyze data (47%); this usually goes hand in hand with automated data collection but not necessarily vice versa (Figure 2). Only one study manually compiled a set of Facebook posts and comments, and then analyzed them with AntConc (Palacio & Gustilo, 2016*). Using computational tools for text analysis was one of the more prevalent methodological approaches (18%). While AntConc is freeware, commercial applications, such as QDA Miner (Provalis), or built-in modules of popular qualitative data analysis software, such as NVivo (QSR), were also used repeatedly (e.g., Al-Rawi, 2016*; Saura, Muñoz-Moreno, Luengo-Navas, & Martos-Ortega, 2017*). In addition, some authors made use of self-written scripts. In most cases, the implemented software supported the counting of word frequencies, collocation analysis or keywords in context analyses. This can be broadly conceived as automated data analysis, converting textual material to data (Krippendorff, 2013, p. 213; Shah et al., 2015, p. 13), even if the complexity and computing demands for such analytical procedures are usually low.

From a methodological standpoint, content analysis is based on the operationalization of theoretical constructs. In our sample, three studies followed this idea, implementing automated content analysis through a dictionary approach (Fenoll & Cano-Orón, 2017*; Melek, 2017*; Naidoo & Dulek, 2017*).

Automatic classification with machine learning approaches, such as support vector machines, were conducted in 10% of the studies (e.g., Li et al., 2016*), while unsupervised machine learning approaches, such as topic modelling or cluster analysis, were employed in 6% (e.g., Bodrunova, Koltsova, Koltcov, & Nikolenko, 2017*).

The seven studies (10%) in our sample that applied network analyses were exploratory or netnographic in character. The underlying construction of networks is usually based on who follows, retweets, or mentions whom on Twitter and who likes what on Facebook (e.g., García-Perdomo, 2017*; Hayat & Samuel-Azran, 2017*). Accordingly, network analysis is mainly used to identify certain subgroups or cumulation points. For example, the distance between Facebook pages is calculated from the number of users who co-liked the pages (Šisler, Švelch, & Šlerka, 2017*).

Although an ongoing further development and improvement has been observed in computational communication science (e.g., van Atteveldt & Peng, 2018), other, more computing-intensive approaches (e.g., artificial neural networks; agent-based modelling) were not applied by scholars in our sample.

Challenges and Limitations

In addition to an inventory of computational data collection and analysis techniques applied by communication scholars, our study paid particular attention to certain problems, pitfalls, and issues that go hand in hand with the implementation of computational approaches mentioned by the researchers. We manually extracted 209 propositions from 52 different social media research articles that referred to such issues in the broadest sense. Fourteen articles did not mention any challenges or limitations related to computational data collection and analysis. Three-quarters of the statements corresponded to data selection and collection (n = 156; 75%), and 53 statements (25%) corresponded to data analysis.

In half of all of propositions regarding the *data collection* stage, researchers tended to mention general methodological challenges, but did not further concretize the limitations arising from them (n = 80; 51%). For example, authors mentioned being limited by focusing on a single platform or topic, thus counteracting generalization (e.g., Blackstone, Cowart, & Saunders, 2017*; Liang & Fu, 2017*). Likewise, scholars felt that the platform under analysis was not representative of the whole community under study (e.g., Jiang, Leeman, & Fu, 2016*; Murthy, Gross, & Pensavalle, 2016*). Scholars regularly reflected that by focusing on specific platform users only, they could not generalize their results to the whole society and therefore suggested that "future research may also want to include other social media sites than Facebook like, for example, Instagram" (de Vries, Gensler, & Leeflang, 2017*, p. 27). Some researchers also stressed the importance of data preparation as part of the data collection stage, which could lead to incorrect conclusions if not adequately performed (e.g., Liu, Burns, & Hou, 2017*).

In one-quarter of the data collection challenges (n = 38; 24%), researchers discussed more precise issues arising from technological challenges relating to computational data gathering. For example, if page administrators were included as users during data collection, they could not later be conceptionally eliminated by the researchers when analyzing the data (Chen et al., 2016*). Data collection challenges also included specific technological requirements of data access and computational data "harvesting" set by the

platform operators. Typical examples mentioned were Facebook's or YouTube's API limitations (e.g., Hayat & Samuel-Azran, 2017*; Onyancha, 2017*), Instagram's data archiving restrictions (e.g., Beach, 2017*; Darwish, 2017*), or Twitter's platform access restrictions (e.g., Rodrigues & Niemann, 2017*; Vessey, 2016*). In some cases, the researchers also had difficulties with adapting programs (e.g., archiving tools, Web scrapers) to their specific needs. However, sensitivity to the limitations resulting from data access (e.g., the generalizability of the findings) was still relatively rare in our sample.

Besides technological challenges of data collection, scholars further reflected possible limitations grounded in psychological or social biases connected to individual user behavior mirrored in the data under study (n = 20; 13%). Some scholars considered, for example, issues linked to social platform dynamics, such as social navigation or echo chambers, selective exposure or homogeneity biases, or social inequalities limiting access to certain platforms (e.g., Murthy et al., 2016*; Verbeke, Berendt, d'Haenens, & Opgenhaffen, 2017*) that potentially affected their data collection.

Ethical considerations about computational data collection were also exceptional in our sample (e.g., Mercea & Bastos, 2016*). For example, Bozdag and Smets (2017*) noted that collecting "large amounts of data does not disburden researchers from . . . ethical responsibilities vis-à-vis marginalized subjects" (pp. 4051–4052).

With regard to the *data analysis* stage, 53 concerns were discussed. A further look at these statements indicated that half of them (n = 29; 55%) referred to challenges on a rather abstract level, thus leaving the scholarly reflection fairly general. Some researchers, for example, addressed the need to reflect on the structures and contexts of the data during data analysis (e.g., Fuchs, 2016*; Verbeke et al., 2017*). Others mentioned that domain knowledge and the understanding of different platforms and communication patterns proves particularly helpful at this stage (Verbeke et al., 2017*; Walker, Baines, Dimitriu, & Macdonald, 2017*).

In one-fifth of the statements about computational data analysis (n = 10; 19%), researchers discussed technological challenges in more specific terms. For example, some researchers critically reflected that the use of different social media analytic tools such as quintly, Keyhole, and Twitter reach produced different data and results (Darwish, 2017*; Jungherr, Schoen, & Jürgens, 2016*). Furthermore, the instability of algorithms in the context of topic modelling was discussed; Bodrunova and colleagues (2017*), for example, mentioned a custom C++ implementation of Gibbs sampling. In another 10 statements (19%) related to the computer-based analysis of data, the scholars reflected on challenges resulting from disciplinary boundaries and attempts at interdisciplinary cooperation to overcome such limitations (e.g., Bulut & Yörük, 2017*; Verbeke et al., 2017*). For example, exploring data-mining practices and their impact on societal discourse, Verbeke and associates (2017*) reported cooperation between social science and computer science scholars. While computer science seemed better at coping with the speed at which Twitter data come in and the various data formats, domain expertise from journalism and media studies was used to explain different styles when tweeting news content and users' corresponding retweeting behavior.

Automation and Interpretation—How to Open Up the Black Box?

Automating research generally means relying on algorithms for collection or analysis that are as standardized as possible. This makes sense from a representative research perspective as long as the samples adequately capture reality. In contrast, a key assumption from the interpretive paradigm is the need to analyze actions and meanings in specific situations. The overview about methods and challenges highlights that researchers are yet aware of digital and computational black boxes and their limited representativeness and generality, but rarely discuss issues connected to black boxing in light of constructive solutions. In the last step we, therefore, reflect possible strategies dealing with the tension between automation and interpretation as one approach to further open up the black boxes inherent in applying computational methods. More precisely, we argue that the loss of control and transparency due to blind spots embedded in the processes of data collection and analysis can be partially compensated by applying interpretive methods that further look at individual cases, because interpretation can put results in perspective and enrich them with meaning.

Most of the 68 studies in our sample not only executed automated data collection procedures but also documented at least minimal *interpretive steps* (*n* = 44; 64%; Figure 3). While manual content analysis of automatically collected messages was most prevalent (*n* = 26; 38%; e.g., Blackstone et al., 2017*; García-Perdomo, 2017*), manual coding was also used to train software (*n* = 4; 6%; e.g., Li et al., 2016*; Ordenes, Ludwig, Ruyter, Grewal, & Wetzels, 2017*) or to conduct quality checks by assessing the precision and recall of automated classification algorithms (*n* = 2; 3%; Abril, Szczypka, & Emery, 2017*; Jang & Park, 2017*). Three studies (4%) also used interpretive approaches for illustrative purposes, aiming to better explain and portray individual cases (Bulut & Yörük, 2017*; Knight, 2017*; López-García, 2016*). Automated data collection thus does not generally rule out interpretive processing of material; quite the opposite: As such examples illustrate, both perspectives can complement and enrich each other.

While most of the researchers working with automated data collection procedures applied standardized interpretive techniques, more open interpretive methods were also used, if only occasionally. For example, hermeneutic approaches appeared in our sample for ten case studies (7%) that were deepened through automatically collected material. Some further studies came from a hermeneutic tradition but looked at structural data like counts or network relations (*n* = 4; 6%; e.g., Arvidsson & Caliandro, 2016*; Vessey, 2016*). Likewise, one study used automated procedures to identify "prolific postees" among 2.5 million Twitter users, and these identified individuals were then interviewed (Mercea & Bastos, 2016*). In addition, some authors in the sample briefly discussed the relationship between small and large samples (e.g., Bozdag & Smets, 2017*, p. 4051) or quantitative and qualitative approaches (e.g., Saura et al., 2017*, p. 47). Further, some fundamental limitations of computational methods were critically examined. Jungherr et al. (2016*), for example, noted that "it is important to systematically analyze the various . . . user-based . . . processes involved in creating the data traces in the first place" (p. 65).

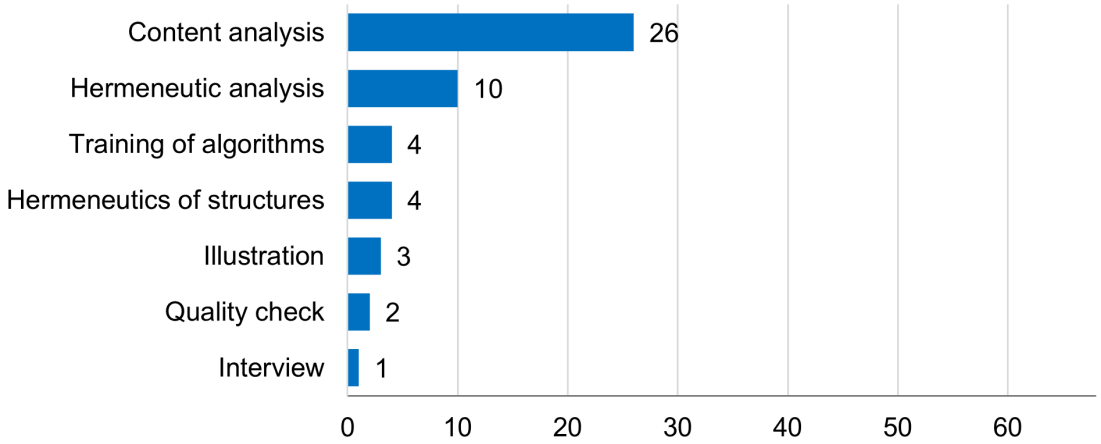

Figure 3. Types of interpretive steps. Basis: n = 68 social media studies using computational methods in 2016 and 2017. Absolute numbers, multiple assignments possible.

Such studies indicate diverse relationships between interpretation and automation. First, from a computational perspective, it is tempting to *maximize automation*. This strategy bears tremendous opportunities for social science research, as content can be compressed to standardized metrics that allow the handling of large-scale databases. In addition, by reducing manual coding, technically, reliability gains are achieved. Fundamental behavioral patterns hidden behind heterogeneous data can be distilled from the complexity arising out of large numbers of variables and cases. In contrast, hermeneutic *case studies* start with a single case, like one company or one typical politician or campaign, and then expand to time-consuming, large-scale analyses. The opportunities of such approaches come from stepwise following interesting cases and close reading of content in context.

The two strategies can easily be linked to either quantitative or qualitative research traditions. Extending this view, inspired by articles in our analysis, we propose a third approach balancing interpretive and standardized methods: *data hermeneutics*. We understand data hermeneutics as a systematized procedure for understanding and interpreting data in a reflective way. In the hermeneutical tradition of the social sciences (e.g., Soeffner & Hitzler, 1994), the methodically controlled understanding that characterizes a data hermeneutical approach takes place by adopting a theoretical attitude of principled doubt about self-evident "facts" provided by digital data. From the perspective of data hermeneutics, structured data, such as the output of machine learning models or network structures *have to be interpreted* hermeneutically.

The *data hermeneutical* approach combines the representative and interpretive perspective of social science research. On the basis of our analysis, we particularly propose four steps that should be included in such analyses (Figure 4):

1. ***Combine calculating metrics with close reading of content.*** As our review indicates, researchers include interpretive steps into their computational research designs, because they expect to be able to better explain, understand and portray individual cases. Metrics can be used for finding interesting cases in the data set. This is the very virtue of computational methods: They can bridge the different traditions of standardized quantitative and nonstandardized qualitative research.

2. ***Conduct statistical modeling, not for representing the world, but for inspired interpretation***. Such models can shed light onto otherwise unseen correlations. While statistical analyses are good at revealing general patterns on the basis of single test statistics or estimators, manual inspection can provide insights into the stories behind the data. From the combination of computational and interpretive approaches it becomes apparent that one's own role as a researcher and the steps of data collection and analysis are deeply intertwined. Researchers should be aware of this linkage—at least when it comes to discussing the limitations of computational methods.

3. ***Compile large data sets to zoom in and out phenomena of interest.*** Jumping between micro and macro level perspectives gives context to the data. On an aggregated level, structures and correlations may be distilled, but to understand mechanisms and causes behind those general patterns, interpreting single cases has proven to be important. Accordingly, computational social science "allows us to zoom from a particular 'data-point' out to the whole (data set), and back again" (Marres, 2017, p. 18).

4. ***Start with manually coded material for training and end with manual coded material for quality check***. At least some kind of interpretive quality check seems vital for gaining a deeper scholarly understanding. Particularly, when developing sophisticated machine learning techniques based on human-coded data, analyzing recall and precision are vital examples of how to deal with the black boxes of computational methods. Working with large-scale process-generated data using automated methods obviously demands these validation efforts. Otherwise, researchers risk losing the links to actual behavior. Therefore, when reporting study results, provide example cases to illustrate findings and limitations.

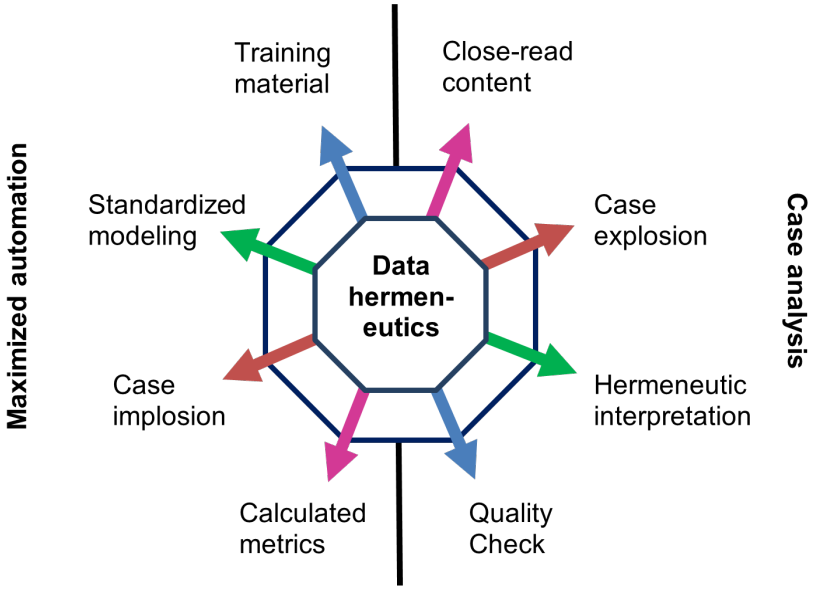

Figure 4. The data hermeneutic octagram.

Following this understanding, we argue that data hermeneutics help to open up the black box of automated methods. By hermeneutically exploring data structures, the peculiarities of specific cases, the data-generating processes, and the role of the researchers are brought into focus.

Conclusion

About one quarter of all empirical social media studies screened in the first step were based on computational methods. Nearly all of these social media studies in our sample made use of automated data collection methods. Most studies worked with Twitter data collected from third-party platforms or with scripts leveraging platform APIs. Automated analysis procedures were conducted in about half of these studies, often through counting words. The methods used were not necessarily technically or statistically demanding. Challenges arose rather from the characteristics of process-generated data. With the sociotechnical conditions managed by platform providers, user behavior was confounded. Thus, the processes behind process-generated data were often opaque, and findings about one platform could not be simply generalized to other communication settings. Such limitations of representativeness and validity were usually discussed, but rather on a general level. Aiming to cope with the arising challenges, researchers not only discussed how to adapt established requirements and quality criteria of social science research but also looked beyond the interdisciplinary boundaries of their subject. It seems likely that research in the field of computational communication science would benefit particularly from the methodological and conceptual openness suggested here.

For opening up the black box of computational methods (i.e., reflecting how automated systems work), we propose to view these methods from a data hermeneutical perspective. While maximizing automation achieves efficiency gains, case analyses risk overlooking general pattern. It is the combination of computational and interpretive approaches that not only fosters validation efforts, rather zooming in and out into the data sets gives context to metrics and action vice versa. Particularly, it brings to light how data-generating processes on online platforms and the steps of data collection and analysis are intertwined. As the epistemic operation of automated scientific research becomes more pervasive across academic disciplines, a hermeneutic perspective becomes increasingly important for making sense of convoluted data and opaque systems.

References

Alvarez, R. M. (2016). *Computational social science: Discovery and prediction.* New York, NY: Cambridge University Press.

Ashby, W. R. (1957). *An introduction to cybernetics.* London, UK: Chapman & Hall.

Automation. (2018). *Merriam-Webster's online dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/automation

Baur, N. (2011). Mixing process-generated data in market sociology. *Quality & Quantity, 45*(6), 1233–1251. doi:10.1007/s11135-009-9288-x

Bick, W., & Müller, P. J. (1980). The nature of process-produced data: Towards a social-scientific source criticism. In J. M. Clubb & E. K. Scheuch (Eds.), *Historical social research: The use of historical and process-produced data* (pp. 396–413). Stuttgart, Germany: Klett-Cotta.

Bick, W., & Müller, P. J. (1984). Sozialwissenschaftliche Datenkunde für prozeßproduzierte Daten: Entstehungsbedingungen und Indikatorenqualität [Social-scientific data literacy for process-produced data]. In W. Bick, R. Mann, & P. Müller (Eds.), *Sozialforschung und Verwaltungsdaten* [Social research and administration data] (pp. 123–159). Stuttgart, Germany: Klett-Cotta.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Bryman, A., Becker, S., & Sempik, J. (2008). Quality criteria for quantitative, qualitative and mixed methods research: A view from social policy. *International Journal of Social Research Methodology, 11*(4), 261–276. doi:10.1080/13645570701401644

Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. In L. Osterweil, D. Rombach, & M. L. Soffa (Eds.), *Proceedings of the 28th International Conference on Software Engineering* (pp. 1051–1052). New York, NY: ACM Press. doi:1-59593-085-X/06/0005

Bunge, M. (1963). A general black box theory. *Philosophy of Science, 30*(4), 346–358. doi:10.1086/287954

Choi, S. (2020). When digital trace data meet traditional communication theory: Theoretical/methodological directions. *Social Science Computer Review, 38*(1), 91–107. doi:10.1177/0894439318788618

Cioffi-Revilla, C. (2010). Computational social science. *WIREs Computational Statistics, 2*(3), 259–271. doi:10.1002/wics.95

Cioffi-Revilla, C. (2017). *Introduction to computational social science. Principles and applications* (2nd ed.). Cham, Switzerland: Springer.

Cleve, J., & Lämmel, U. (2014). *Data mining.* München, Germany: De Gruyter Oldenbourg.

Couldry, N., & Hepp, A. (2017). *The mediated construction of reality*. Cambridge, MA: Polity.

Cox, M., & Ellsworth, D. (1997, October). Application-controlled demand paging for out-of-core visualization. In R. Yagel (Ed.), *Proceedings: Visualization '97*, *Phoenic* (pp. 235–244). New York, NY: IEEE. doi:10.1109/VISUAL.1997.663888

Dijck, J. van. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197–208. doi:10.24908/ss.v12i2.4776

Driscoll, K., & Walker, S. (2014). Big data, big questions. Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication, 8*, 1745–1764.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54. doi:10.1609/aimag.v17i3.1230

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.

Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal, 2*(16), 1–18. doi:10.5204/mcj.620

Giddens, A. (1984). *Interpretative Soziologie: Eine kritische Einführung* [Interpretive sociology: A critical introduction]. Frankfurt, Germany: Campus-Verlag.

González-Bailón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. Cambridge, MA: MIT Press.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *Annals of the American Academy of Political and Social Science, 659*(1), 63–76. doi:10.1177/0002716215570866

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review, 38*(1), 10–24. doi:10.1177/0894439318788322

Ho, J. C.-T. (2020). How biased is the sample? Reverse engineering the ranking algorithm of Facebook's graph application programming interface. *Big Data & Society, 7*(1), 1–15. doi:10.1177/2053951720905874

Holton, J. (2010). The coding process and its challenges. *Grounded Theory Review, 9*(1). Retrieved from http://groundedtheoryreview.com/2010/04/02/the-coding-process-and-its-challenges/

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems, 12*(12), 767–797. doi:10.17705/1jais.00282

Hox, J. (2017). Computational social science methodology, anyone? *Methodology, 13*(Suppl.), 3–12. doi:10.1027/1614-2241/a000127

Innes, M., Roberts, C., Preece, A., & Rogers, D. (2017). Of instruments and data: Social media uses, abuses and analysis. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 108–124). Los Angeles, CA: SAGE Publications.

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 297–319). Thousand Oaks, CA: SAGE Publications.

Jünger, J. (2018). Mapping the field of automated data collection on the Web: Data types, collection approaches and their research logic. In C. Stützer, M. Welker, & M. Egger (Eds.), *Computational social science in the age of big data: Concepts, methodologies, tools, and applications* (pp. 104–130). Köln, Germany: Halem.

Jünger, J. (forthcoming). Verhaltens-, Forschungs- oder Datenschnittstellen? Drei Perspektiven auf die sozialwissenschaftliche Bedeutung von Application Programming Interfaces (APIs) [Behavioral, research, or data interfaces? Three perspectives on the significance of application programming interfaces for social science research]. In E. Koenen, T. Birkner, C. Pentzold, C. Katzenbach, & C. Schwarzenegger (Eds.), *Digitale Kommunikation und Kommunikationsgeschichte* [Digital communication and communication history] (pp. 2–31). Berlin, Germany: Digital Communication Research. doi:10.17174/dcr.v10.6

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. doi:10.1016/j.bushor.2009.09.003

Keyling, T., & Jünger, J. (2016). Observing online content. In G. Vowe & P. Henn (Eds.), *Political communication in the online world: Theoretical approaches and research designs* (pp. 183–200). New York, NY: Routledge.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Los Angeles, CA: SAGE Publications.

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung* [Qualitative content analysis: Methods, practices, computer-assistance]. Weinheim, Germany: Beltz Juventa.

Kwon, K. H., Priniski, J. H., & Chadha, M. (2018). Disentangling user samples: A supervised machine learning approach to proxy-population mismatch in Twitter research. *Communication Methods and Measures, 12*(2/3), 216–237. doi:10.1080/19312458.2018.1430755

Laney, D. (2001, February 6). 3D data management: Controlling data volume, velocity, and variety [Blog post]. META Group Inc. Retrieved from https://web.archive.org/web/20200731220550/https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Latour, B. (2002). *Die Hoffnung der Pandora: Untersuchungen zur Wirklichkeit der Wissenschaft* [Pandora's hope: Essays on the reality of science studies]. Frankfurt, Germany: Suhrkamp.

Lazer, D., Pentland, A. S., Adamic, L., & Aral, S. (2009). Life in the network: The coming age of computational social science. *Science, 323*(5915), 721–723. doi:10.1126/science.1167742

Lincoln, Y. S. (1995). Emerging criteria for quality in qualitative and interpretive research. *Qualitative Inquiry, 1*(3), 275–289. doi:10.1177/107780049500100301

Luhmann, N. (1993). *Soziale Systeme. Grundriß einer allgemeinen Theorie* [Social systems]. Frankfurt, Germany: Suhrkamp.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20–33. doi:10.1080/08838151.2012.761700

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

Marres, N., & Weltevrede, E. (2013). Scraping the social? *Journal of Cultural Economy, 6*(3), 313–335. doi:10.1080/17530350.2013.772070

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine, 6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Monroe, B. L., & Schrodt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis, 16*(4), 351–355. doi:10.1093/pan/mpn017

Mukerjee, S., Majó-Vázquez, S., & González-Bailón, S. (2018). Networks of audience overlap in the consumption of digital news. *Journal of Communication*, *68*(1), 26–50. doi:10.1093/joc/jqx007

Pfaffenberger, F. (2018). What you tweet is what we get? *Publizistik, 63*(1), 53–72. doi:10.1007/s11616-017-0400-2

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science, 33*(2), 163–180. doi:10.1177/0165551506070706

Ruths, D., & Pfeffer, J. (2014). Social sciences: Social media for large studies of behavior. *Science, 346*(6213), 1063–1064. doi:10.1126/science.346.6213.1063

Schnell, R., Hill, P. B., & Esser, E. (2013). *Methoden der Empirischen Sozialforschung* [Empirical methods in social research] (10th ed.). Munich, Germany: Oldenbourg.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *Annals of the American Academy of Political and Social Science, 659*(1), 6–13. doi:10.1177/0002716215572084

Soeffner, H.-G., & Hitzler, R. (1994). Hermeneutik als Haltung und Handlung: über methodisch kontrolliertes Verstehen [Hermenutics as attitude and action]. In N. Schröer (Ed.), *Interpretative Sozialforschung: auf dem Wege zu einer hermeneutischen Wissenssoziologie* [Interpretive social research] (pp. 28–54). Opladen, Germany: Westdeutscher Verlag.

Stockemer, D., Koehler, S., & Lentz, T. (2018). Data access, transparency, and replication: New insights from the political behavior literature–corrigendum. *Political Science & Politics, 51*(4), 977–803. doi:10.1017/S1049096518000926

Stoycheff, E., Liu, J., Wibowo, K. A., & Nanni, D. P. (2017). What have we learned about social media by studying Facebook? A decade in review. *New Media & Society, 19*(6), 968–980. doi:10.1177/14614448176957

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771–1779. doi:10.1002/asi.20388

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures, 12*(2/3), 81–92. doi:10.1080/19312458.2018.1458084

van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Toward open computational communication science: A practical road map for reusable data and code. *International Journal of Communication, 13*, 3935–3954.

Waldherr, A., Geise, S., & Katzenbach, C. (2019). Because technology matters: Theorizing interdependencies in computational communication science with actor–network theory. *International Journal of Communication, 13*, 3955–3975.

Waldherr, A., Geise, S., Mahrt, M., Katzenbach, C., & Nuernbergk, C. (2021). Toward a stronger theoretical grounding of computational communication science: How macro frameworks shape our research agendas. *Computational Communication Research*, *3*(2), 1–28. doi:10.5117/CCR2021.02.002.WALD

Wallach, H. (2016). Computational social science: Towards a collaborative future. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. 307–316). Cambridge, UK: Cambridge University Press.

Welser, H. T., Smith, M., Gleave, E., & Fischer, D. (2008). Distilling digital traces: Computational social science approaches to studying the Internet. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 116–140). Los Angeles, CA: SAGE Publications.

Wettstein, M. (2020). Simulating hidden dynamics: Introducing agent-based models as a tool for linkage analysis. *Computational Communication Research*, *2*(1), 1–33. doi:10.5117/CCR2020.1.001.WETT

Wilson, T. P. (1973). Theorien der Interaktion und Modelle soziologischer Erklärung [Theories of interaction and models of sociological explanation]. In Bielefeld Sociologists' Group (Ed.), *Alltagswissen, Interaktion und Gesellschaftliche Wirklichkeit, Band 1, Symbolischer Interaktionismus und Ethnomethodologie* [Everyday knowledge, interaction, and social Reality, Volume 1, Symbolic interactionism and ethnomethodology] (pp. 54–79). Reinbek, Germany: Rowohlt.

Woolley, S. C. (2016). Automating power: Social bot interference in global politics. *First Monday, 21*(4). doi:10.5210/fm.v21i4.6161

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *38*(1/2), 98–139. doi:10.1080/10584609.2020.1785067