Datafication research: Mapping the field for a future agenda

Stine Lomborg (ORCID: 0000-0002-5672-7298)

Sofie Flensburg

Department of Communication, University of Copenhagen, Denmark

slomborg@hum.ku.dk (corresponding author)

sofie.flensburg@hum.ku.dk

Author bios:

Stine Lomborg is associate professor at the Department of Communication, University of Copenhagen, and the PI of the Datafied Living project, which studies infrastructures and experiences of tracking of the self across personal, work and institutional life in the welfare state.

Sofie Flensburg is postdoc in the Datafied Living project at the Department of Communication, University of Copenhagen. She researches digital infrastructures and their political, economic and regulatory implications in the context of the welfare state.

Funding: Research for this article was funded by the European Union (ERC, DATAFIED LIVING, grant no. 947735). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.
Datafication: A research review and future agenda

Abstract: A growing body of research centers around the concept of ‘datafication’ suggesting a buzz around data studies and, perhaps, the emergence of a research field. This article analyses and discusses the current state of datafication research. Our dataset comprises 463 publications on datafication identified through a systematic literature search in Web of Science and Scopus, an explorative network analysis of keyword co-occurrences and a content analysis of these publications. We map datafication research interests in various research fields, find that the majority of studies are theoretically oriented, whereas empirical analyses largely apply qualitative approaches and rarely make use of data-driven methods. We suggest studies on datafication can be devised into categories reflecting research interests in either user understandings and practices, or in infrastructure and technological processes of datafication. The latter strand is particularly sparse in empirical anchoring, and needs empirical and methodological attention. We conclude by outlining three paths for future datafication research to cross-pollinate infrastructural and user perspectives, highlighting the bridging role of communication research in such an endeavour.

Keywords: datafication; literature review; communication theory; user imaginaries; data infrastructures.
Datafication research: Mapping the field for a future agenda

Introduction

Datafication is commonly associated with the collection, databasing, quantification and analysis of people’s information, and the uses of these data as resources for knowledge production, service optimization and economic value generation (Mayer-Schönberger and Cukier, 2013). Despite current invocations of paradigmatic change, the rendering of people’s information into quantitative data in databases and the analytical operations performed on such data is in fact an old phenomenon. It is historically rooted in needs of population management and became widespread with the emergence of modern forms of bureaucratic organizing of states and companies (Porter, 1995), and historically assisted by old media technologies such as writing and print to register, analyze, archive, index and administer people and society (Mayer-Schönberger & Cukier, 2013). Citizens’ tax, health and educational records and social security benefits in the welfare state were historically administered in written documents (Kaun and Dencik, 2020); and so was productivity and performance measurement of workers (Skouvig, 2012), customer lists in retail (Turow, 2014) and so on. Datafication historically precedes digitalization. Yet, following the digitalization of the Global North through broad public and private development of digital infrastructures and online presence since the 1990s, such operations of collection and processing of information about people have become pervasive and unprecedented in scope and scale, and increasingly automated (Andrejevic, 2014). Most aspects of our daily lives are now rendered into data. Datafication, today, thus marks how especially digital
systems fuel, intensify and automate historical practices of databasing, analyzing and using information about people as a key resource for value-creation, as well as a propelling of these practices into everyday life as such.

In 2019 and 2020, 273 unique research publications in the databases Web of Science and Scopus used the term ‘datafication’ or ‘datafied’ in their titles, abstracts, or keywords. In comparison, these terms are mentioned only twice, and without definitional interest, in research publications up until 2013, where Mayer-Schonberger & Cukier (2013) coined the term. This development indicates a recent, significant growth in the scholarly attention to datafication as a phenomenon, but also suggests a kind of buzz around data studies and, perhaps, that a research field with a particular terminology is emerging. The rapid growth in research about datafication is evident across in particular humanistic and social sciences, but with outlooks to technical sciences, including computer science, although the term ‘datafication’ is less widely used there. Research about datafication seeks to understand the phenomenon in itself, its uses and implications for human existence and societies at large. Datafication research is timely and very important, but often conceptually unclear, empirically sparse and scattered across scholarly journals and fields with the result that research accumulates unevenly. Therefore, to devise paths for more systematic, integrative efforts, we believe it is time to take stock of datafication research.

This article analyses and discusses the key characteristics and current state of datafication research. Aspects of datafication have been the topic of a couple of literature reviews in recent years. For instance, Kennedy et al. (2020) map “original empirical research into public understanding and perceptions of, attitudes towards and feelings about data practices and related phenomena” (p. 3), a trajectory that has flourished lately according to the authors. From a topical angle, Ruckenstein and Schüll (2017) review datafication in the
context of health and identify two attitudes to data: on the one hand, “the so-called “big-data fundamentalists” promote the view that large data sets, properly mined for correlations and patterns, will render up previously elusive insights, predictions, and answers to long-standing challenges of individual and collective life, replacing the need for theory and science” (p. 262). On the other hand, they see researchers coming out of the social sciences and humanities, to promote “a more skeptical stance, emphasizing the cultural, political, economic, and rhetorical dimensions of the data paradigm shift, typically by focusing on particular cases of “datafication,” or the conversion of qualitative aspects of life into quantified data” (ibid.).

In this review, we cast a wider net to understand research on datafication as such; in relation to a range of topics, and with a view to both scholarship on the technological and infrastructural underpinnings and public understandings and uses of data. This move, we suggest, is pertinent, if we accept that technical and experiential aspects of datafication are distinct yet intertwined, equally important for understanding the nuts and bolts of data in digital systems, and, by extension, that analyses of these distinct aspects should inform one-another in future research. Some current research might be cautious to accept an ontological and analytical distinction between material technological infrastructure and social processes and uses of said infrastructure. We maintain that keeping technology and social uses as discrete components of datafication provides conceptual clarity and is analytically productive, because it helps us to better identify the theoretical, methodological and empirical trajectories of datafication research. Our aim in reviewing datafication research is thus also to build on the current state of the art to devise conceptual and empirical pathways for advancing our understanding of the phenomenon of datafication, its technological underpinnings and social implications.
The article proceeds in four sections. First, we visit key texts that have shaped scholarly understanding of datafication across domains to discuss definitional traits and tensions associated with ‘datafication’, and establish the fundamental questions datafication has raised. We then present the methodology and findings of a systematic inquiry into the datafication literature, analyzing this research along trajectories of disciplinary grounding, theoretical and empirical orientations and thematic foci. We conclude by outlining three paths for future datafication research, arguing in particular for the need of more holistic empirical inquiry and how communication research might offer conceptual fuel to overcoming some of the divisions we see in the current literature on datafication.

**Datafication research**

The concept of “datafication” was introduced by Mayer-Schönberger and Cukier in their 2013 book on “big data” processes. They define datafication as involving the combined processes of registration and quantification of various types of data, and the different kinds of value that may be derived from these processes. Datafication, then, is the cultivation of various data resources for value-generation purposes: standardization, typification, personalization, optimization, commodification etc. (Mayer-Schönberger & Cukier, 2013). The data in question is typically, but not exclusively about humans, as pointed out by Mejias and Couldry (2019). In the digital age, these data resources include aspects of human life that could not or at least not easily be rendered into data and made analysis-ready before, such as our online behavior and information (meta-data) associated with it (Kennedy et al., 2015). This overarching definition echoes Lycett’s (2013) conceptualization of datafication in
information science and business insights, where datafication is conceived as a technology-driven sense-making process of dematerialization (extraction of information), liquefication (data flow across contexts) and density (data enrichment through cross-referencing and reuse).

While Mayer-Schönberger and Cukier’s book continues to be a highly cited source for defining datafication, media scholars such as (van Dijck, 2014) have critiqued their conceptualization as reifying an ideology of data as a resource to be taken and molded; an ideology that underpins the practices of social media companies and, by extension, the whole industry of surveillance capitalism (Zuboff, 2019). Van Dijck (2014) instead advances an understanding of datafication that builds from boyd & Crawford’s seminal article (2012) on critical questions for big data studies, to suggest that datafication is also an expression of a specific mythology and logic of dataism that attends to data as if it were a raw material to be cultivated and that need to be deconstructed and contextualized. Van Dijck (2014) argues against seeing data as simply imprints of behavior that is neutrally curated by platforms, and for seeing platform interventions and interpretation as necessary conditions for cultivating data resources in the first place. For her, then, “datafication” also refers to a privately and publicly held imaginary of dataism; the widespread belief in the objective quantification and potential tracking of all kinds of human behavior through digital systems. This imaginary or mythology of data can further be traced as a central stream in current research (Andrejevic, 2020). And it is a key interest in the recently established field of critical data studies, which is focused on the politics and power of data (Andrejevic, 2014, Milan, 2018), and the relationship between data and knowledge production (Kitchin, 2014, Iliadis, 2018).
This brief review of key conceptualizations of datafication places datafication research at the intersection of different sciences: technological, business, social and humanistic sciences. It also suggests that datafication research covers a lot of ground: from infrastructural and data analytics perspectives, represented by Mayer-Schönberger & Cukier (2013), to more socio-culturally oriented approaches dealing with individuals’ and organizations’ imaginaries, practices and attitudes towards data, represented by van Dijck (2014) and others. Perhaps, as Dencik (2019) suggests, the two identified trajectories of conceptualizing datafication translate into two separate strands of research: namely studies concerned with the technological aspects of datafication and studies concerned with its social implications. This begs questions for an empirical analysis of the state of the art in datafication research: What are the dominant research fields and topical areas for studying datafication? What do they have in common and how might they differ? What kinds of theoretical perspective and research questions are pursued and with what methods and approaches? What are the most common analytical foci and units of analysis for understanding datafication? And how might our understanding of the impact of datafication on human existence and social life be advanced in future research by cross-pollinating insights in the field?

**Methods and data**

The findings presented in this article have been generated through a three-step process consisting of: 1) a systematic literature search for academic publications that explicitly mention the concept of datafication, 2) an explorative network analysis of the identified publications’ bibliographic information, and 3) a content analysis of the included
publications’ research fields, methodological approaches, scopes and analytical entry points.

We used the databases Web of Science and Scopus to search in titles, abstracts and keywords for the terms ‘datafication*’ or ‘datafied*’. This search strategy meant that we excluded publications that do not use these terms explicitly in their titles, abstracts, or keywords but nonetheless address big data collection, processing and analysis practices and their social consequences (e.g., Andrejevic (2014), who is frequently cited by the included references, uses the term ‘big data’ and not ‘datafication’ or ‘datafied’). A broader search for, for instance, ‘big data’ would result in an overwhelming number of results (88,572 references in Scopus and 53,421 in Web of Science). Moreover, we consider ‘big data’ a broad term rather than a conceptual anchor with potential of uniting a specific research area. Our goal is to explore the concept of datafication and studies explicitly designated as datafication research. We have therefore excluded publications that do not self-identify as dealing with datafication from the dataset. While necessary for untangling “datafication” in itself, this choice is also a limitation of the study in terms of scoping the broader research field concerned with the technological and social implications of data. We will return to this limitation in the discussion.

The searches were conducted in October 2020 and included all years and all the collections included in the two databases. However, we cannot assume that we have found all relevant references, since some research fields and journals, might not be fully represented in Scopus and Web of Science, especially when it comes to historical publications. The search returned a total of 415 references from Web of Science and 418 references from Scopus that were imported into Endnote and checked for duplicates. We
found 336 duplicates, bringing the total number of references down to 498 publications that formed the basis for the analyses.

An explorative network analysis was conducted using the VosViewer software (van Eck and Waltman, 2010) that enables various types of bibliographic analyses and mappings. We used it to map out the co-occurrence of officially assigned keywords to provide an overview of included publications and how they relate to each other in terms of terminology. As such, the keyword map sensitized us to both the different scholarly traditions and the key topics that inform datafication research. On par with the definitional texts of datafication previously discussed, the network analysis also served as a basis for designing the content analysis, as it suggests connections and clusters across the different publications in the sample that can beexplained by looking at the individual publications.

In preparing the content analysis, we exported the data to Excel. We deleted all non-English references leaving us with a total of 463 publications. Hence, 35 non-English references (corresponding to 7% of the dataset, and primarily in Chinese and Spanish) were left out as we could not properly code the content of these publications due to our lack of linguistic skill in the languages in question (mainly Chinese and Spanish). This is a limitation of the content analysis, specifically. In the network analysis, non-English language texts are still included as the bibliometric data on which the networks are based is provided in English. Based on the existing definitions of datafication research as well as the network analysis, we developed and tested a coding scheme that was used to code all included publications according to their research field, publication type, the methods and data sources used, the units of analysis, and the contexts and scopes of the study. The design of the coding scheme reflects our interest in determining the research fields that have addressed and adopted the concept of datafication, the relative distribution of theoretical
and empirical studies as well as their methodological approaches and analytical foci. The coding was conducted by four trained research assistants who were briefed and supervised by the authors and followed a detailed codebook. We controlled the quality of the coding through comparing a sub-sample (4.3%) of the dataset’s nominal categories that all the research assistants had initially coded before carrying out the full coding of the data set. To further ensure the quality and consistency of the coding, the authors re-coded the open-ended coding categories (research fields, specific methods, specific contexts, and data sources) together after the initial coding and reliability testing.

The intercoder reliability was high or acceptable for most categories (between 1 and 0.774 between all four coders, as calculated with Krippendorff's Alpha), apart from one category: unit of analysis (intercoder reliability at 0.47), that is the coding of publications as either offering infrastructural or user-centric analysis. This may be a result of much current scholarship subscribing to socio-material approaches which tend to collapse the material-infrastructural and the social or experiential to the extent that these entities become both ontologically and analytically blurred. While we recognize the mutual shaping of the material and the social, there are important differences between technological infrastructures of datafication (e.g. pipes and cables, computer software for data collection and storage, algorithmic data processing and data circulation) and the socio-cultural uses, understandings, and imaginaries of datafication. Keeping the material and infrastructural aspects of datafication ontologically and analytically distinct from imaginaries and experiences of said infrastructures - however difficult it might be - is necessary when seeking to understand how datafication is defined, approached, and studied. The units of analysis have important repercussions for the methodological approaches, the theoretical assumptions, and the empirical results. The conflation between material and socio-cultural
approaches can be seen as an interesting finding in itself when seeking to explain the somewhat blurry definitions of datafication. As a result of the initially low intercoder reliability, we clarified the category when further instructing the coders and later manually checked and validated the coding of this category together.

Findings
This section presents the main findings from the network and content analysis and provides examples of different types of datafication research. We first elaborate on how datafication has been addressed within various research fields reflecting different scholarly traditions, then examine the different methods and approaches that have been applied, and lastly identify analytical foci in studies of datafication relative to context.

Research disciplines and traditions
The network and content analyses indicate that studies of datafication are rooted in distinct research fields that influence the questions that are asked and the ways they are addressed and answered. The keyword map in Figure 1 gives an impression of the organization of datafication research through illustrating how the included publications cluster according to research fields and topics. The map is based on a network analysis of the co-occurrence of officially assigned keywords. Each node represents a keyword that appears at least five times across the 498 included references. The keywords are connected to each other if they appear in the same reference. We have removed the most dominating, incomplete, or misleading keywords (specifically ‘datafication’, ‘big data’, ‘social sciences - other topics’, ‘big’, ‘data’, ‘online’, ‘literature’) and the final map comprises a total of 115 keywords.
The network analysis identifies five clusters, each represented by a color. The most common keyword, Communication, is assigned to 104 of the references and co-occurs with 86 other keywords. Communication thereby connects several of the clusters but has the strongest links to the blue cluster where it is connected to keywords such as media, information, journalism, capitalism, social media, cultural studies, and platforms. The network analysis suggests that the field of media and communication studies is strongly represented and arguably a key driver in datafication research, which the content analysis confirms (see Table 1). Of the 463 coded publications, 134 (equivalent to 29%) are written by media or communication scholars (determined on the basis of either their educational
background or their institutional affiliation). These publications include a wide range of studies addressing, for instance, how algorithmic processes in social media and self-tracking have influenced social interactions, information streams and public opinion (e.g., Splichal, 2019), how digital technologies and devices enable self-tracking and transform everyday lives and call for new forms of literacy and agency (e.g., Yates and Carmi, 2020), and how digitization of legacy media has transformed audience measurements, content curation, and business models focusing in particular on the media industries (e.g., Andrew, 2019). These studies often tie datafication closely to the related concept of platformization arguing that “datafied user feedback” (Nieborg and Poell, 2018) make up a vital part of platform business models. These examples suggest that the cluster centered on communication includes both technological and business-oriented strands and socio-cultural and critical approaches to datafication outlined above.

<table>
<thead>
<tr>
<th>Research field (1st author)</th>
<th>Publications</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>134</td>
<td>29%</td>
</tr>
<tr>
<td>Education</td>
<td>56</td>
<td>12%</td>
</tr>
<tr>
<td>Sociology</td>
<td>48</td>
<td>10%</td>
</tr>
<tr>
<td>Arts &amp; cultural studies</td>
<td>30</td>
<td>6%</td>
</tr>
<tr>
<td>Computer Science</td>
<td>25</td>
<td>5%</td>
</tr>
<tr>
<td>Information science</td>
<td>23</td>
<td>5%</td>
</tr>
<tr>
<td>Law</td>
<td>23</td>
<td>5%</td>
</tr>
<tr>
<td>Health</td>
<td>22</td>
<td>5%</td>
</tr>
</tbody>
</table>
The green cluster in Figure 1 represents another dominating field within datafication research, namely Education and educational research that occurs as a keyword in 48 of the 498 publications. It co-occurs with 56 other keywords in the network including governance, politics, science, learning analytics, and smart cities that are also part of the green cluster. In the content analysis, 56 (equivalent to 12%) of the included publications are by authors from the field of educational research (see table 1). Studies within this field for instance ask questions relating to digital literacy (e.g., Grant and Rogers, 2018), the use of digital technologies in teaching (e.g., Manolev et al., 2019), and, especially, the use of data for performance monitoring, testing, benchmarking and so forth (e.g., Roberts-Holmes and Bradbury, 2016, Stevensen, 2017). The field of education is one where datafication studies have emerged as a natural continuation of existing research agendas concerned with

### Table 1. Primary research fields of first authors.

<table>
<thead>
<tr>
<th>Field</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political Science</td>
<td>21</td>
<td>5%</td>
</tr>
<tr>
<td>Anthropology</td>
<td>16</td>
<td>3%</td>
</tr>
<tr>
<td>Business &amp; organization</td>
<td>14</td>
<td>3%</td>
</tr>
<tr>
<td>Geography</td>
<td>13</td>
<td>3%</td>
</tr>
<tr>
<td>Philosophy</td>
<td>11</td>
<td>2%</td>
</tr>
<tr>
<td>Economy</td>
<td>8</td>
<td>2%</td>
</tr>
<tr>
<td>Engineering</td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>Psychology</td>
<td>4</td>
<td>1%</td>
</tr>
<tr>
<td>Architecture</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>1%</td>
</tr>
<tr>
<td>N=</td>
<td>463</td>
<td>100%</td>
</tr>
</tbody>
</table>
how to register, monitor and compare for instance students’ performances and grades. Digital data processing can, as such, be seen as a new solution to an old problem since digital tools and programs have extended teachers’ and school managers’ abilities to carry out existing tasks, while also raising new questions, concerns and dilemmas (Selwyn, 2020).

The three remaining clusters are less clearly tied to specific research disciplines than the Communication and Education clusters. The yellow cluster contains the third and fourth most common keywords across the 498 references: Surveillance (42 occurrences and 74 links to other keywords) and Sociology (38 occurrences and 60 links). Surveillance and Sociology co-occur with keywords such as Power, Dataveillance, Self-tracking, Quantified self, Life, Work, Health, Care, and Identity. The yellow cluster represents a social-science based and growing body of research investigating the implications of tracking and surveillance for optimization of individual lives, for instance through a focus on health, and with questions of power. As Table 1 shows, the content analysis finds that of 48 (equivalent to 10 %) the publications are from Sociology and 22 publications (5%) are from health and healthcare sciences. Similar to Education, the field of Health has a relatively long tradition for investigating datafication processes (cf. Ruckenstein & Schull, 2017). Going back centuries, registration, cataloging, journaling and so forth have been key components in the work of doctors and other health professionals who need to keep track of patients’ disease trajectories, understand the development and spread of, for instance, viruses, analyze test results, and so forth. As such, neither registration, quantification, nor the ethical issues related to privacy and protection from harm, are new to health studies, but the technological affordances promoting datafication are. The studies of health and datafication investigate, for instance, how the use of IT transforms the day-to-day practices of health care workers (e.g., Hoeyer and Wadmann, 2020), how improved data processing can
support cancer treatment (Binning et al., 2018), and how users track their health through mobile apps (e.g., Fiske et al., 2019, Esmonde 2020).

Closely related to the yellow cluster, and in particular to Surveillance, the red cluster centers around Privacy (34 occurrences and 66 links) and Ethics (19 occurrences, 47 links) as prominent keywords across the 498 references. The red cluster seems to cut across disciplines as it also contains keyword such as Information science & library, Government & Law, Public administration, Computer science, Data science, Business & economics, and Education. This cluster illustrates how legal and ethical concerns related to datafication appear across contexts and raise general questions and challenges related to for instance data processing, governance, and so forth. This cluster also covers computer-science contributions to datafication research. According to the content analysis, publications written by computer scientists make up 25 of the 463 publications (equivalent to 5%). These include, for instance, studies of cyber security (Lam and Chi, 2016), the Internet of Things (Jesse, 2016), and artificial intelligence (Bañeres et al., 2020).

Finally, the purple cluster in the network map, is made up by 13 keywords among which Film, radio & television (14 occurrences, 15 links) and Children (13 occurrences, 28 links) are the most prominent. The purple cluster clearly emphasizes user and citizen perspectives by linking keywords such as Agency, Open data, Data activism, Rights, Participation and Risk. Often originating in audience studies, research in this vein highlights “the interest of audiences” (Ytre-Arne & Das 2019) and “fears of audience gullibility, ignorance, and exploitation” (Livingstone 2019) in relation to datafication.

The analysis shows that a wide variety of research fields, including humanities, social sciences, health, business and technical sciences (see table 1), underlining the inherently interdisciplinarity of datafication research, as also reflected by the overlapping
clusters in the network map. Despite the scholarly differences, a range of topics (e.g., education, health, ethics) and concepts (e.g., surveillance, communication, privacy) also bridge the different research fields, suggesting that the different sub-fields can profit from cross-pollination in the study of datafication.

Approaches & methods in datafication research

As described above, the different strands of datafication research address the phenomenon in a variety of ways reflecting scholarly traditions and specific research questions. In order to further investigate how knowledge on datafication is generated, we looked closer at each of the included publications to assess their overall research approach. Figure 2 illustrates the relative distribution between different types of publications and their development over time. The numbers are accumulated, so that all studies published since 1994 figure under 2020. The graph is based on a multiple-choice coding of each reference where the coders were asked to classify the main contribution of the publication in question. Specifically, the coders chose between ‘Theoretical’, ‘Methodological’, ‘Empirical’, ‘Systematic literature review’, ‘Editorial’, ‘Book review’, and ‘Other’ and based their assessment on the abstract and the general structure of the publication. For instance, theoretical publications are defined as being concerned with conceptual discussions and non-empirical research questions, while empirical publications typically contain a methods section, present and describe data sources, and so forth.
Figure 2: Research publication identified through Scopus and Web of Science in October 2020 and coded according to publication type (N=463).

Figure 2 illustrates the significant increase in the number of publications that explicitly mention datafication in their titles, keywords or abstracts. Since the publication of Mayer-Schönberger & Cukier’s seminal book in 2013, datafication literature has undergone an exponential growth and the total number of publications have more than doubled since 2018. Across all years, the majority of publications addresses datafication from a mainly theoretical perspective, while fewer, despite a growth in recent years, address the phenomenon either from a methodological or empirical perspective. Examples of
theoretical contributions include van Dijck’s (2014) deconstruction “of the ideological grounds of datafication”, Galić et al.’s (2017) “overview of surveillance theories and concepts that can help to understand and debate surveillance in its many forms”, and Robinson’s (2018) exploration of “the power and efficacy of databases as part of the turn to intensive datafication in contemporary life”. To compare, the empirical strand of datafication research includes studies of, for instance, “how activists in the open data movement re-articulate notions of democracy, participation, and journalism” (Baack, 2015), how “datafication impacted the practice of student grouping and students’ experience of education” (Neumann, 2021) and “how nurses and patients collaboratively form and interact through digital representations” (Kempton and Grisot, 2019).

If we look at the relationship between disciplines and theoretical vis-à-vis empirical outlook, we find little indication that specific disciplines are the key drivers of either theoretical, methodological or empirical studies. That is so say, each discipline largely follows the same pattern of distribution as seen in the aggregate data in Figure 2, with the exception being educational research which has a relatively larger portion of empirical vis-à-vis theoretical publications (34% and 38%, respectively). We thus cannot infer from the analysis that specific fields are either under-theorized or in need of empirical inquiry, nor do any specific discipline drive methodological innovation from within datafication research. However, methodological innovation might be found elsewhere, in studies that are not necessarily indebted to the conceptual framing of data research in terms of datafication, for instance in studies seeking to unravel ecosystems of digital tracking, a point to which we will return in the discussion.

We coded the 115 empirical publications according to their general methodology as well as the specific methods that they apply. We found that eighty of the
empirical studies (equivalent to 70%) build on qualitative analyses, while only ten empirical studies (equivalent to 9%) use quantitative methods, and 18 studies (equivalent to 16%) are based on mixed-method designs (the methods of the last seven studies, equivalent to 6% are unclearly described). The predominance of qualitative studies reflects that especially interviews are a preferred method for datafication researchers. 66 (equivalent to 57%) of the empirical studies either exclusively use interview material or combine interviews with other methods (e.g., observations, document analysis, survey, etc.). Only twelve publications (equivalent to 10%) use so-called data-driven methods to collect and analyze, for instance, data harvesting activities, data flows, and so forth. These are furthermore scattered across disciplines. We find the lack of data-driven research quite surprising, given the new methodological opportunities that come with big data in research (Kitchin, 2014). One explanation for the widespread use of interviews is, as illustrated in the examples of empirical research mentioned above, that many of the studies focus on how people understand and articulate datafication processes and that analyses of data practices and processes are often based on descriptions from, for instance, programmers or other professionals working with data.

**Analytical foci and units of analysis**

As mentioned in the introduction, definitions of datafication emphasize technological and social aspects, respectively. This definitional debate further suggests a distinction in the research literature between studies concerned with the technological aspects of datafication, and studies concerned with its social implications. Inspired by this, we coded the units of analysis of all included publications to determine the relative distribution between studies that focus on, respectively, ‘infrastructure’ and ‘users’. While the first
refers to studies that focus on technological systems, platform design, computational processes, and so forth the latter refers to studies that investigate, for instance, user understandings and representations and practices with data. A number of studies (30 publications equivalent to 6.5% of the included studies) fall in-between or outside these two main variables. These include literature reviews of specific domains in relation to datafication and studies that focus or conceptual or theoretical issues such as “surveillance capitalism” (Zuboff, 2019). However, we find that the large majority of studies (326 of the 463 publications, equivalent to 70.5%) have a user-centered focus, emphasizing for instance “the relationship between digitally tracked work behaviors and employee attitudes” (Bertolotti et al., 2020) or “how individuals made sense of behavioral data and algorithmic recommendations” (Haapoja & Lampinen, 2018). This confirms previous observations of public perceptions of data as a flourishing research topic (Kennedy et al., 2020). In contrast, the infrastructurally oriented publications (that make up 107, equivalent to 23% of the publications), often focus on the technological compositions, architectures and protocols of specific software system or platforms (e.g., Marachi and Quill, 2020). When zooming in on empirical research specifically, it stands out that only 21 (equivalent to 18 percent) of the empirical studies identified are concerned with infrastructure, and these typically build on interviews with professionals or various types of document analysis, but also, sometimes, make use of data package inspection and other computational methods for analyzing data traffic. The overall lack of empirical research on infrastructure suggests that there is a vital research gap to be addressed.

We coded the analytical foci of the different studies to further understand the included publications’ research contexts and scope. The context of study was coded as a multiple-choice category distinguishing between six different spheres that are influenced by
datafication (‘Everyday life’, ‘Public institutions’, ‘Private organizations’, ‘NGOs’, ‘Society’, and ‘Other’). Since, for instance, the healthcare sector, the educational sector, and the media can be both privately and publicly governed, we recoded this category so that it distinguishes between everyday life, institutions and organizations, or societies in a more general sense. The analysis shows that 64 (equivalent to 14%) focus on a micro-level everyday life. 231 publications (50%) apply what could be interpreted as a meso-perspective focusing on particular sectors, institutions and organizations. This meso-perspective, to some extent, reflects the findings presented in the research fields section, since for instance media, educational, or healthcare studies naturally have a sector focus and often center around specific institutions or organizations. 150 studies (equivalent to 32%) apply a macro-perspective to address the societal consequences of datafication emphasizing the emergence of, for instance, surveillance capitalism as a new economic order (Zuboff 2019) or dataism as a new governing ideology (van Dijck 2014). Studies in this vein are often theoretical and macro-oriented in contrast to the more empirical organizational or user-oriented studies.

<table>
<thead>
<tr>
<th>Analytical foci</th>
<th>Infrastructure</th>
<th>Users</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>2</td>
<td>61</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>Meso</td>
<td>65</td>
<td>148</td>
<td>18</td>
<td>231</td>
</tr>
<tr>
<td>Macro</td>
<td>31</td>
<td>110</td>
<td>9</td>
<td>150</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 2. Analytical foci and units of analysis.

When looking at analytical foci relative to units of analysis (table 2), it stands out that infrastructural studies are skewed towards studies of organizations and sectors in society. These include research on how data infrastructures work in various organizations and
institutions (e.g. Andrew, 2019), how datafication processes transform industries and
business models (e.g. Nieborg and Poell, 2018). Studies focusing on users appear more
diverse. These include organizational-level studies of for example how datafication impacts
the day-to-day practices of professionals within different sectors (e.g. Hoyer and Wadmann,
2020), and societal-level studies of data rights and ethical issues pertaining to for instance
personal data protection and exploitation (Milan and Trëré, 2019). But it also contains a
body of studies that address users from an everyday life perspective, for instance analyses
of people’s knowledge on, awareness of and practices with data (e.g., Esmonde, 2020).
Given the relative prominence of the users as units of analysis, one might expect more
emphasis on the lived experiences of datafication in everyday life as such. It is not evident
that infrastructural and user-centred lines of inquiry intertwine in any significant and
systematic way across the datafication literature.

**Discussion**

The findings presented above contribute to ongoing efforts of understanding and defining
datafication as an increasingly important societal phenomenon and, thereby, as a significant
research object. Our analysis of datafication research confirms the alleged distinction
between studies focusing on technological processes (e.g. the implementation and use of
data-driven systems and data mining techniques in organizations) and more critical,
humanistic and social science-based, studies that scrutinize the uses, imaginaries and social
consequences of datafication. It suggests that datafication over the last decade has
developed into a research concept and has been adopted by a wide range of disciplines.
However, we also find that the definition and use of the concept – despite the many
theoretical contributions – continues to be somewhat unclear and fragmented. Datafication literature seldomly reflects critically on the use of the concept and how it relates to the chosen objects of analysis, methods, and empirical source materials. If we are to stick with datafication as a useful conceptual lens onto contemporary data developments and uses in society, our analysis raises questions for the ability of research to cumulate systematically to the mutual enrichment of the fields that have stakes in the topic of datafication. It is our hope that this article will encourage such critical reflections and serve as a foundation for further discussions, critical scrutiny, and evaluations of the growing body of scholarly work dealing with datafication.

The publications included in this review tend to focus on the theoretical aspects of datafication and empirical studies are usually based on interviews and other qualitative methods that typically generate knowledge on users’ practices, understandings and experiences of data. That said, the research publications that ground the analyses presented above do not represent the numerous publications that address questions related to datafication without using this particular term. As mentioned earlier, our review only includes publications that explicitly mention datafication in their titles, keywords, or abstracts and thus excludes important work that surely informs datafication research. Importantly, however, critical theories of big data that does not explicitly address the term ‘datafication’ (e.g., Andrejevic, 2014; Kitchin, 2014) are strongly referenced throughout the datafication literature and with impact on how especially the theoretical study of datafication unfolds. Building on the findings presented in the analysis, we might consider datafication research a theoretically rich source of scholarship for understanding (big) data. Yet, datafication research is not equally diverse and rich when speaking in methodological and empirical terms. Apart from a few recent exceptions (e.g., Pybus & Coté, 2021),
datafication research will need to look elsewhere for methodological innovation to advance both user and infrastructural analyses. For instance, the term is not particularly common in computer science, although a body of computer-science research on tracking in digital systems has clear overlaps with datafication research (e.g., Binns et al., 2018). Data-driven and quantitative methods seem to be more prevalent in this literature (whereas theoretical and conceptual aspects are perhaps given less attention in such research). Datafication research will miss an important opportunity for deepening scholarly understanding of especially infrastructures of datafication if we fail to engage with and learn from such empirical research. In this sense, the concept of datafication should not foreclose what counts as relevant research to inform the further investigation of data use in society, but it might offer theoretical fuel for data-driven research in the area.

By acknowledging that people’s experiences, perceptions, and imaginations of for instance tracking, data processing, algorithms and so forth are rooted in the material conditions that enable and constrain their action, user studies can gain deeper insights into the forces that shape human agency and capabilities. In parallel, infrastructural and technological approaches to studying data flows and digital ecosystems, can be strengthened by focusing more explicitly on the socio-cultural implications of datafication. We suggest that a clearer conceptualization of datafication in individual studies, combined with empirical experiments and methodological innovation can stimulate increased cross-pollination between disciplines, between empirical and theoretical approaches, and between studies focusing on the technological and infrastructural aspects of datafication and the more socio-culturally oriented approaches.
Conclusion: towards an interdisciplinary and empirical research agenda

Our mapping of datafication research has documented datafication as a multi-disciplinary research field that has grown immensely since its inception in 2013 and is nested across the humanistic and social sciences as well as technical sciences. It extends historical procedures, problems and practices of data collection, databasing, analysis and indexing of information and people for value-creation, knowledge production, management and governance. These procedures are amplified in the historically rapid, ongoing and ever-expanding process of digitalization of business, society and human life.

First, we found that to some degree, research on datafication is sector specific (e.g., focusing on tech giants and media industries) and topical (e.g., focusing on datafication in education or healthcare), but research also indicates that there are significant practical as well as conceptual challenges that cut across domains and scientific fields, for instance regarding communication, surveillance, and ethics. Future research may want to explore such cross-cutting concepts further and use them to compare how infrastructures of datafication work across contexts and what user experiences are general and specific to contexts. This would not only help us identify the fixed and variable components of datafication across sectors and spheres of life, but also to explore how different fields, via shared concepts, can learn from each other. In addition, a comparative focus should include studies of datafication across the globe, as we may see important variations in technological infrastructures, users’ experience and societal implications of datafication across, e.g., the global north and global south, as argued by Milan and Treré (2019).
Second, we found that datafication research is largely dominated by theoretical contributions, but empirical work to substantiate theoretical frameworks and assess implications of datafication in specific contexts is growing and much needed. Yet, the majority of empirical studies base their analyses on qualitative interviews and other qualitative methods to address very context-specific research questions, and now is arguably the time to expand the methodological toolbox to complement existing research and broaden the kinds of questions that can be asked. One key candidate for methodological development is the inclusion of data-driven methods, which are surprisingly absent in research on datafication. Such methods for data collection and analysis work through datafying the empirical domain of inquiry and thus may be said to constitute methods for datafication par excellence (Lomborg, Dencik & Moe, 2020; Kitchin, 2014). Data-driven methods may be particularly well suited for eliciting empirical material about the actual technological operations of the platforms and systems that datafy us (Rogers, 2020, Helles and Ørmen, 2020). By extension, mixing data-driven methods with ‘traditional’ ones might help us combine user perspectives and infrastructural investigations to reach a better, more comprehensive understanding of the interplay and mutual shaping of technological infrastructures and human practices and experiences. One implication in addressing this crucial research need in scholarship on datafication across disciplines is the importance of ramping up computational skills to be able to integrate data-driven methods and insights in research on datafication. This might also prove a useful effort for exploring and innovating towards new disciplinary frontiers.

Finally, in the analysis of existing scholarly work on datafication, we have identified two main lines of inquiry in scholarly literature on datafication, focusing on user perceptions, practices and imaginaries, and technological processes of datafication,
respectively. These can be traced from the early conceptual anchor texts through the growing literature and appear to foster a divide between infrastructure and user-centered studies of datafication. Further cross-pollination, we think, will have to rely not just on systematic, comparative and multi-method empirical research, but also on the development of a conceptual repertoire that bridges the infrastructural and user-centric dimensions. One key theoretical perspective to start with for this endeavor is communication theory. As suggested by the research clusters found in our network analysis, communication inhabits a bridging role in the current interdisciplinary research landscape and thus might offer conceptual fuel to overcome some of the divisions in conceptualizations and analytical foci we see in the current landscape of datafication research. Datafication, after all, is conditioned on communication: on people registering information in (analogue or digital) systems and databases by way of communication technologies, and on people and systems communicating about and acting on that information. The pervasive collection, processing and analysis of big data hinges on people’s intensive and extensive communication in digital systems, whereby they leave all kinds of digital traces behind. The cross-referencing and analysis of big data, and the resulting typification, personalization and optimization of user experiences, enable and shape future communications through open-ended feedback loops. A comprehensive communicative conceptualization of datafication, one that considers both the material technology, the meanings generated through data and the socio-cultural consequences of the ubiquitous datafication of our everyday lives, might offer a promising next step in integrating and advancing current research on datafication, as well as grounding datafication more systematically in a historical trajectory of the mutual shaping of communication technology and society.
References

ESMONDE, K. 2020. ‘There’s only so much data you can handle in your life’: accommodating and resisting self-surveillance in women’s running and fitness tracking practices. Qualitative Research in Sport, Exercise and Health, 12, 76-90.


