BIG DATA & SOCIETY

Original research article

Big Data & Society
April–June: 1–15
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517251330181
journals.sagepub.com/home/bds
Sage

Making sense of "big data": Ten years of discourse around datafication

Charlotte Knorr[1,2] and Christian Pentzold[1]

Abstract

This article reconstructs the sociotechnical imaginaries of "Big Data" in Germany, South Africa, and the United States over 10 years. Our inquiry into the meaning-making undertaken on expansive datafication processes began from the observation that since its inception circa 2010, the buzz phrase "Big Data" has not only denoted a technology but has also gestured toward a vast array of ambitions and concerns that are reflective of values, economic perspectives, and cultural preoccupations. We use a frame analysis to investigate the unfolding journalistic discourse and discuss the sociotechnical imaginaries of Big Data in cross-national comparison. We found three dominant views that centered chiefly on rebuilding a datafied society, reviving datafied business, and retooling datafied surveillance. Despite substantial data scandals and whistleblower revelations, affirmative views prevailed. The trope of reviving datafied business was most often evoked from 2011 to 2013, but rebuilding a datafied society became the central perspective from 2014 onward. This order of prominence existed in the United States and Germany, whereas a business-oriented view predominated in South African media. Some publications in all three countries ran counter to the general trend but only exerted limited influence on the overall picture.

Keywords

Datafication, big data, media frames, discourse, sociotechnical imaginaries, frame analysis

In the past two decades, social processes and personal lives have increasingly been accompanied by the computation of large amounts of data, which are stored and analyzed for numerous purposes (Flensburg and Lomborg, 2023; Kitchin, 2014; van Dijck, 2014). This is an almost universal transformation characterized by unequal opportunities and potentially conflicting economic interests and legal contexts (Couldry and Mejias, 2019; Zuboff, 2019). Data scandals such as the revelations of Edward Snowden or the Cambridge Analytica exposures have highlighted the global dimension of datafication processes. At the same time, these events have demonstrated that the significance of data and data-based operations can be understood quite differently (Di Salvo and Negro, 2016).

Since its inception circa 2010, the term "Big Data" has been invoked to refer to a broad array of concepts, technologies, and metrics. It fueled excitement around large quantities of data and analytical procedures (Pentzold and Knorr, 2023) as it referred to the volume and variety of data, the speed with which they are gathered and analyzed, and the algorithmic operations, storage, and computing resources required to exploit them (Mayer-Schönberger and Cukier, 2013). The term entails expectations of precise calculations and reliable probabilistic prognoses. According to boyd and Crawford (2012), it thus engenders a peculiar "mythology"—namely, "the widespread belief that large datasets offer a higher form of intelligence and knowledge … with the aura of truth, objectivity, and accuracy" (p. 663).

At the same time, the usage of large datasets in an increasing number of sectors is accompanied by public debates about their social consequences as well as their political, economic, and cultural implications (Paganoni, 2019; Pentzold and Fischer, 2017). The issues at stake

[1] Department for Communication and Media Studies, Leipzig University, Leipzig, Germany
[2] Department of Media and Communication, LMU Munich, Munich, Germany

Corresponding author:
Christian Pentzold, Department for Communication and Media Studies, Leipzig University, Nikolaistrasse 27–29, D-04109 Leipzig, Germany.
Email: christian.pentzold@uni-leipzig.de

include questions of data ownership, privacy, security, and other fundamental human rights. In these discourses, datafication processes—which are usually deemed opaque—are captured through metaphors such as "data explosion," "data deluge," or "dataverse" that chime with figurative expressions like "data harvesting" and "data mining" (Couldry and Yu, 2018; Portmess and Tower, 2015; Puschmann and Burgess, 2014). For some, data are the "new oil" whereas others liken them to "Big Brother."

However, this "interpretative work" (Bowker, 2013: 170), in which the social meaning and significance of Big Data and data-based processes are articulated and contested, remains largely unexplored. Indeed, despite the eminent reality-making power of communication, deliberation, and imagination, our understanding of discourses around datafication is still quite limited. Little is known about the framing of datafication processes in cross-country comparison, particularly in non-Western countries (Milan and Treré, 2019). Moreover, given that the notion of Big Data has been around for a while, there is a dearth of knowledge about how the ideas used to make sense of Big Data may have changed over time.

To address these lacunae, we explore debates in cultural analysis and critical data studies that question the ontology and epistemology of data (Andrejevic, 2014; Boellstorff, 2015). Challenging data's facticity and neutrality, we stress that data presuppose interpretation. "Data need to be imagined as data to exist and to function as such," Gitelman and Jackson (2013) likewise argue, "and the imagination of data entails an interpretative base" (p. 3). Along these thoughts, the article examines the sociotechnical imaginaries of Big Data in Germany, South Africa, and the United States over the course of 10 years (2011–2020) with the help of a frame analysis.

Background: Big Data discourses

In general, the public understanding of technology, its diffusion and acceptance, are driven by public discourses and the imaginaries they evoke (Barbrook, 2007; Wyatt, 2021). Each technology is embedded in a web of cultural ideas, values, social aspirations, and political projects. These webs enable people to make sense of available tools and applications and are crucial for anticipating the utility and ramifications of forthcoming innovations (Natale, 2016). Multiple frameworks exist to analyze such patterns of thought as, for instance, the technological sublime (Marx, 2000; Nye, 1996), myths (Mosco, 2004), metaphors (Burgers, 2016), or imaginaries.

Concept: Sociotechnical imaginaries

Studying the discursive dimension of datafication is important because the ideas, affects, visions, and beliefs encapsulated in a discourse have tangible effects on the course of technological development, policymaking, and public expenditure. To express this conjunction terminologically, Jasanoff and Kim (2009) speak of sociotechnical imaginaries, which Jasanoff (2015) later defines as "collectively held, institutionally stabilized and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology" (p. 4). In this broad competence, sociotechnical imaginaries interlace semantics and materiality; they drive commercial investments; they are employed to back up political decisions; and they undergird cultural expectations and concerns.

On a similar note, Marcus (1995) has coined the notion of "technoscientific imaginaries" (p. 3) which seeks to reflect both the future possibilities that come with technological and scientific innovations and the constraints set by current conditions. On a more fundamental level, Taylor (2004) writes about a "social imaginary" (p. 23) that refers to the expectations and normative visions of how people imagine their social existence and togetherness. A prospective impulse is central to all these conceptions. Beyond capturing the status quo, imaginaries involve ambitions and normative stances about potential, desirable, or unwanted ways of living and of sociotechnical futures. "By guiding the making of things and services to come," as Mager and Katzenbach (2021) put it, "imaginations of the future are co-producing the very futures they envision" (p. 224).

In principle, this pertains to all kinds of technology, established or emergent. Yet it is particularly effective in the early roll-out stages of a pioneering technology with as-yet undetermined paths of development. Hence, imaginations of data-intense technologies and their related practices are what Mosco (2004) has identified as today's formidable "repository of the future" (p. 15). The repository's conceptions of the future are consequential and powerful elements in the construction of the new technologies, making imaginaries highly contested. In effect, imaginaries are a matter of semantics manifesting in discourse. They are more than mere talk or a kind of shared mindset: imaginaries are of strategic importance as they become employed to legitimize decisions, mobilize resources, and devise pathways towards investment and technological innovation. For instance, China's sociotechnical imaginary culminating in the "Digital Silk Road" initiative is reflected in legal documents regulating cross-border data flows and trade, in geopolitical relationships, in technological investments and commercial decisions, as well as in the corporate data management of businesses like Tencent and Alibaba (Zhang and Negro, 2019). Or take the sociotechnical imaginary that infuses "India Stack," a volunteer-built public system around open APIs that forms the base for a range of services clustered around Aadhaar, India's national identity program, like digital IDs, digital payments, and data exchange. These digital identity products are not alone technological enterprises but are undergirded by notions of empowerment and emancipation (Kanna et al., 2024).

Imaginaries are not only instrumental in character but can, when widely shared, become a taken-for-granted interpretative framework that guides people's understanding and public sensemaking. Other concepts such as myths or frames also aim at capturing the semantic dimension that supports technological choices and social development, yet they are of limited help in addressing the constitutive entanglement of materiality and discourse that is at the heart of the notion of sociotechnical imaginaries. As such, sociotechnical imaginaries may congeal in metaphors and other linguistic tropes yet they hardly merge into them. This is, for instance, the case with the persistent metaphorical talk about Big Data as the "new oil." For some, the metaphor is useful as it suggests comparing data to some natural resource readily available, an undue naturalization criticized by others using the same metaphor to challenge the extractivist venture and draw attention to the limited rights of those who become implicated in the generation and processing of data (Couldry and Yu, 2018; Taffel, 2023).

State-of-Research: Imagining data technologies

The internet is a prime example of the co-constitutive entanglement found in sociotechnical imaginaries, and many have attempted to shed light on the ideologies and cultural concepts encapsulated in an "internet imaginary" (Flichy, 2007; Mansell, 2012). These contributions show that the internet did not emerge out of technological necessity or from a clear-cut plan (Bory, 2020; Stefik, 1996). As there are not one but many ideas of what the internet is or should be, its history is neither unilinear nor teleological (Driscoll and Paloque-Berges, 2017). This multiplicity is evidenced by the different metaphors used to characterize aspects of the internet. Metaphors imagine a frontier, library, village square, superhighway, a galaxy or a place for web surfing, and these metaphorical framings have affected the accessibility and development of networked infrastructures (Markham and Tiidenberg, 2020).

Furthermore, sociotechnical imaginaries have animated a prolific strand of research dealing with the discursive dimension of data technologies and telecommunications. In this vein, Bucher (2017) and Schulz (2023) each explored a version of the "algorithmic imaginary": the former considered what users believe algorithms to be, and the latter considered how designers and algorithms anticipate user behavior. Based on statements made by Tim Berners-Lee, Lesage and Rinfret (2015) examined contrasting imaginaries of the web as a semantic web or the Web 2.0. In another example, Bareis and Katzenbach (2022) studied AI imaginaries from policy documents in China, the United States, France, and Germany. They found that the overall narratives of AI posing a massive and inevitable disruption are similar, but the imaginaries vary with, for instance, Chinese documents stressing the potential of AI for social control or a French commitment to help humans flourish through AI.

Beer (2019) focused on data and datafication in particular and maintained that the "way in which data are seen is crucial to the power that they afford and the possibilities that are available for the enlargement of data-led thinking, judgment, ordering and governance" (p. 5). In reference to the notion of sociotechnical imaginaries, he proposed a "data imaginary" which he saw as "powerful visions of what data can achieve, what they can solve, how they might help us to thrive, what they are able to reveal and how they are able to make us more informed, efficient or better at things" (p. 14). Beer traced this data imaginary in marketing material from the data analytics industry. There, data and data analytics were envisioned as making insights rapidly accessible to nonexperts. They were treated as being panoramic in scope and anticipatory—that is, they should provide reliable outlooks. In a similar vein, van Dijck (2014) argued that dataism predicates on beliefs and trust: "belief in the objective quantification and potential tracking of all kinds of human behavior and sociality through online media technologies … trust in the (institutional) agents that collect, interpret, and share (meta)data" (p. 198).

Studying datafication discourses

Inquiries into datafication discourses are important, yet a comparative perspective that looks beyond highly industrialized G7 countries is often missing. Sociotechnical imaginaries are not universal, and research into people's optimistic or pessimistic attitudes toward science and new technologies has shown that in addition to individual factors they hinge on a country's humanitarian, democratic, as well as scientific and technological development (Nisbet and Nisbet, 2019). Thus, insights from the United States or European countries may not apply to other contexts. A comparative approach is required because a "data divide" (Andrejevic, 2014) not only exists within and between Western states but also profoundly conditions how people and communities in other parts of the world become implicated in the vast infrastructures of the capitalist exploitation of data (Couldry and Mejias, 2019; Milan and Treré, 2019).

A diachronic examination of datafication featured in discourse has also been neglected. This is remarkable because datafication is a highly dynamic and very protean process that touches many different sectors and evolves on technological, methodical, social, and political dimensions simultaneously. When the term Big Data entered the lexicon circa 2010, datafication was very different from the datafication prevalent today, and the question of how these tremendous changes have been reflected in public discourses remains open (Koenen et al., 2021).

Finally, studies on the sensemaking around datafication have mainly looked at policy documents (Couldry and

Yu, 2018; Nolin, 2019; Paganoni, 2019; Rieder, 2018) and corporate statements (Beer, 2019; Couldry and Yu, 2018; Paganoni, 2019). This is unsurprising since politicians and businesses are prime actors in defining and implementing sociotechnical imaginaries. Yet next to policymakers and IT professionals, journalists are key in propagating and popularizing an imaginary—usually at the expense of others (Droog et al., 2020). Nevertheless, media reports and the press have only received scant attention (Paganoni, 2019; Pentzold and Fischer, 2017; Puschmann and Burgess, 2014).

So we ask: With which imaginaries do journalistic reports make sense of Big Data? (*RQ1*) How do these imaginaries evolve over time? (*RQ2*) To what extent are the imaginaries similar or different across countries? (*RQ3*).

First, we suppose that there is not one imaginary but many. In Jasanoff's (2015) conception, struggles are inevitable as imaginaries often do not present equally valid views but instead engender material consequences and political decisions that are difficult to undo. This holds true for datafication processes as they can follow alternative pathways of regulation, accountability, and investment. What is more, with pervasive datafication, the collection and exploitation of data necessarily becomes a topic of debate in almost any sector or area of concern. So no single universal dataist imaginary is championed, for example, by Silicon Valley conglomerates (Ferrari, 2020), welfare agencies (Dencik and Kaun, 2020), or data activists (Lehtiniemi and Ruckenstein, 2019) but instead there are many different and also conflicting versions. Journalists not only make people aware of the value and usage of data and its ramifications but they also create arenas for articulating different available imaginaries.

Secondly, it can be assumed that the discourse was punctuated by events, most notably "data scandals" where the massive misuse of consumer, telecommunication, health, personal, or payment data were revealed, such as the Snowden affair in 2013, the #GuptaLeaks in South Africa in 2017, or the 2018 Cambridge Analytica revelations. Hence, Russell and Waisbord (2017) referred to the Snowden disclosures as encompassing a number of "news flashpoints" (p. 858) while Haim et al. (2018) called it a "killer issue" (p. 282) that made headlines. In the discursive material we study, these phases should quantitatively result in a higher volume of reports; qualitatively, they may cause shifting, more critical interpretations.

Thirdly, we suggest that there are different imaginaries in different countries due to different political systems, cultural environments, and economic developments. Research has shown that people in countries with a high standard of living are more skeptical about science and technology than people in countries with a lower standard of living. Nisbet and Nisbet (2019) call this the "postindustrial paradox" (p. 13). At the same time, however, the level of technological development and innovation activity have positive effects on optimistic attitudes toward science and technology. The level of political and journalistic freedom has shown no significant effects so far (Allum et al., 2018; Nisbet and Nisbet, 2019).

Data and methods

Sample countries

We selected three countries—South Africa, Germany, and the United States—according to their level of humanitarian, political, and technological development (Nisbet and Nisbet, 2019). This is based on the assumption that the range of circulating frames and their acceptance or rejection can be attributed to sociostructural influences, the interplay of which has not yet been discussed systematically and comparatively at country level. Instead, studies on the framing of new technologies have primarily examined factors at the individual level. The country-specific level of human development, the level of democratic development, and the level of science and technology development can be used as differentiation criteria (Dalton and Welzel, 2014).

When we started our research in 2017, Germany ranked 5th out of 189 countries in the Human Development Index (HDI) compiled by the United Nations based on national statistics on life expectancy and health, school attendance, and gross national income per capita (2017); 25th out of 201 countries in the Freedom of the Press Index by Freedom House (FH), which assesses political, legal, and economic restrictions on press freedom (2017); and ninth out of 129 countries in the Knowledge Creation Index (KCI), which combines information on the H-index and the number of scientific articles and patent applications relative to the gross national income (2017). The United States ranked 13th in the HDI; 38th in the FH; and 3rd in the KCI. South Africa ranked 113th in the HDI; 76th in the FH, and 63rd in the KCI.

We opted for a variation of similar and different cases and chose a purposive sampling design. While Germany and the United States rank high in all three indices, South Africa, as a BRICS emerging market, is in the middle. Moreover, the United States is a pioneer of data-based practices: Technological infrastructures and economic activities are dominated by U.S. companies. In the United States, positive views on technologies have predominated (Allum et al., 2008), whereas in the German discourse on new technologies questions of data protection and the consequences of technology receive a great deal of attention—assumedly also because Germany is part of the European Union (EU) and reflects the data protection debates prompted by EU regulations (acatech and Körber-Stiftung, 2023). In South Africa, according to previous findings, there has been an ambivalent view of the consequences of new technologies that cut across socioeconomic, gender, and ethnic differences (Guenther and Weingart, 2018). As in other BRICS countries, an IT sector is emerging in partnership with international firms. Data-based surveillance was enacted as early as 2003 through the Regulation of Interception of Communications and Provision of Communications Related Information Act. These measures were the subject of heated debates and newly installed parliamentary oversight bodies such as the Matthews Commission (Nathan, 2017).

Sample publications

The material comes from 26 periodicals published between 2011 and 2020 in the three countries. The first article came out on 26 January 2011; the last article is from 29 December 2020. We looked at online editions of newspapers and weeklies to capture a range of relevant Big Data topics and viewpoints and we supplemented them with specialized magazines or their respective online formats in the fields of finance and technology because Big Data and data-based processes were expected to be of special interest here. The sampled articles had to contain the keywords "Big Data" or "dataf*" in the headline, subheadline, or first paragraph so that only texts with a clear topical focus were selected. Images were not considered. We chose "Big Data" because it served as an important keyword that attracted widespread attention—especially in the first years of our timeframe in the discourse around data analytics (Diebold, 2012; Mayer-Schönberger and Cukier, 2013; Pentzold and Knorr, 2023).

Legacy news is still a key arena for the framing of new technologies where societal contentions around sociotechnical imaginaries play out. They provide one avenue for investigation, though the public discourse is also carried out elsewhere, most notably in networked communication where a diversity of opinions and alternative frames may be ventilated. At the same time, research has shown that there are notable intermedia framing influences in-between media outlets and major platforms (Lo et al., 2021). In this discursive environment, which also includes material and statements issued by nonprofits and high-profile commentators, we use journalistic media as a gateway to grasp major sociotechnical imaginaries that are taken up by journalists when assuming societal significance and reverberate, among others, in political decisions, in innovation or in research and development.

We queried the LexisNexis database and the WISO database for German publications as well as individual newspaper archives. For South African newspapers, we chose *The Star*, *Sowetan*, *Financial Mail*, and *Business Day*; for magazines or weeklies: *Mail and Guardian*, *Sunday Times*, *NAG*, *Tech Central*, and *Brainstorm Magazine*. From German newspapers, we selected *Frankfurter Allgemeine Zeitung* (*FAZ*), *Süddeutsche Zeitung*, *taz*, *Die Welt*, and *Handelsblatt*; for magazines and weeklies: *Die Zeit*, *Der Spiegel*, *Wirtschaftswoche*, *Wired Deutschland*, and *c't*. For the United States, the newspapers were the *New York Times* (*NYT*), *Washington Post*, *Financial Times*, and *Wall Street Journal* (*WSJ*); weeklies and magazines: *Newsweek*, *The Atlantic*, *Forbes*, *Wired*, and *The Verge*. These outlets play a leading role in national discourse, and they enable a detailed comparative examination of the development of imaginaries over time. They vary in technological affinity, political orientation, and topic specialization yet the overall focus of our analysis is not on some paper's editorial stance but on a country's journalistic sphere and the discursive side of imaginaries that unfold in this environment.

After data cleaning, a total of $N = 1456$ texts out of 13,097 items (11%) remained in the sample, most of them featuring "Big Data" ($n = 1450$) and a minority only "dataf*" ($n = 6$). The major part of the unselected pieces did not contain the search terms in either headline, subheadline, or first paragraph ($n = 7055$); others were book or exhibition reviews ($n = 323$), noneditorial content like letters to the editors, interviews, or event notes ($n = 1926$), or duplicates and pages not found—especially on South African news websites ($n = 1904$). The sample ($N = 1456$) consists of 603 texts from Germany (42%), 731 from the United States (50%), and 122 from South Africa (8%). The South African publication with the most articles was *Business Day* ($n = 36$, 29%); in Germany it was *FAZ*, which published almost half of all texts ($n = 251$, 42%); and in the United States it was the *Financial Times* ($n = 275$, 37%) together with the *WSJ* ($n = 283$, 38%), which accounted for almost four-fifths of the entire sample. Thus, most of our material originated in the two highly industrialized countries, and there was also a noticeable emphasis in some periodicals that may be reflective of their editorial profile and issues of interest. About a third of the articles (30.4%) were in an editorial rubric; 51% of them in business or economics sections, 13% appeared in the society section, and 11% came out in the technology section.

Frame analysis

We used a frame analysis (Matthes and Kohring, 2008) to operationalize the reconstruction of Big Data imaginaries and analyze the text material. In our approach, examining relevant imaginaries meant considering them in terms of frame packages that are formed by a meaningful composition of reasoning and framing devices (Van Gorp, 2007, 2010). In Van Gorp's (2007) definition, reasoning devices are "explicit and implicit statements that deal with justifications, causes, and consequences in a temporal order" while framing devices are a matter of "word choice, metaphors, exemplars, descriptions, arguments, and visual images" (p. 64).

In public discourse, frames are not only communicated by legacy media. Instead, approaching public discourse through the lens of frame analysis underscores the strategic agendas of social elites, most notably in politics, business,

Table 1. Key variables for manual analysis and intercoder reliability.

Variable	Description	Krippendorff's α
Media outlet	Publication venue of the news item	1
Title	Headline of the article[a]	—
Author	Name(s) of authors of a news item	—
Format	Information-based format or opinion-based format	0.553
Length	Word count	1
Section	Editorial rubric	1
Keyword	Occurrence of keyword (headline, subheadline, first paragraph, body)	0.443
Country	References to countries	1
Event	Reference to events (from expected/scheduled to unexpected/unscheduled)	0.558
Event quality	Characterization of event (integrative, disruptive, ambiguous)	0.719
Actor	Actors quoted/mentioned in the news item	1
Role	Actor roles (hero, antihero, supporter, nonsupporter, third parties)	0.622
Addressed actors	Actors spoken about/held responsible	0.821
Cultural motif	Central organizing idea (e.g., innovation for progress, preventing wrongs, profit, empowerment)	0.810
Problem definition	Problem identified (e.g., lack of transparency, unused potential, resource requirements)	1
Causal attribution	Causes identified for the problem (e.g., deficient laws, technological innovations)	0.669
Treatment	Treatment in relation to the problem	0.896
Treatment evaluation	Evaluation of the treatment (positive/negative)	0.938
Implications	Implications or consequences of a problem	0.869
Temporal anchor	Temporal classification of implications (past/upcoming)	0.950
Moral evaluation (actor)	Moral evaluation through actors quoted/mentioned	0.936
Moral evaluation (journalist)	Moral evaluation through authors	0.486
Moral evaluation	Moral evaluation with no actor link	0.057
Comparison	Reference to connate events	0.283
Framing devices	Catchphrase, metaphors, tropes	0.841

Note. The full codebook can be downloaded from here: https://www.sozphil.uni-leipzig.de/fileadmin/user_upload/Pentzold_Knorr_2024___Public_Codebook_FBD_fin.pdf. All coders were tested for interrater reliability during the coding period and prior to it in the training phase, using a total of seven test items (four test items by five coders and three test items by four coders). Following Krippendorff (2004), values below 0.667 were considered unacceptable. In line with Matthes and Kohring (2008), fixed ideal measures should be evaluated in the context of categories, in this case, for each semantic unit at the frame level.
[a]Only main titles were coded.

the nonprofit sector, in culture and media, who have vital interests in establishing their views by sponsoring certain issues while marginalizing others (Vliegenthart and van Zoonen, 2011). When their interests overlap, they might form advocacy coalitions seeking hegemony. Legacy media take an important role here, following Entman's (2010) cascading activation model, since frames promoted by elites spread through media to public discourse thus informing the ways people talk and reason about an issue. This is not only a struggle over semantics. Rather, frames go along with sociotechnical imaginaries which they support or challenge, sometimes within a national sphere or in transnational perspective (Bareis and Katzenbach, 2022).

Following Entman (1993), frames can be treated as semantic patterns articulated in a text and formed by at least four elements: problem definition, causal attribution, treatment recommendation, and moral evaluation. Each frame is made up of a meaningful package of these components that makes some of them salient while demoting others. The packages are organized around a cultural motif, which Van Gorp (2010) argues can be a common idea or archetype. Moreover, in addition to reasoning devices that reside on the semantic level, a frame usually takes shape through framing devices which exist on the stylistic level, like catchphrases, metaphors, and tropes (Gamson and Modigliani, 1989).

The codebook for the manual analysis contained 32 variables (Table 1). Five coders went through two pretests between May and June 2022, and each variable was trained with 30 to 50 codes. Weekly meetings were held during the period of coding (July to December 2022) to discuss the analytical progress and address open questions. The coding was done on the level of propositional units—that is, the smallest meaningful string of words. This could mean anything from a single word up to cross-sentence phrases to account for the varying linguistic complexity of the framing and reason devices. A total of $N = 3338$ propositional units were coded (on average, 2.3 per article). In the texts from South African media, 252 of these units were coded; in texts from the United States 1673; and in texts from Germany 1413. Krippendorff's α was between 1 and .486. Coders had particular trouble distinguishing the agent of a moral evaluation (the author of the text or a quoted source), recognizing the addressed actors, and

Table 2. Frames, frame packages, and number of coded propositions.

Frame	Frame package	No of coded propositions
Rebuilding a datafied society	*Cultural motifs*	
	Preventing wrongs	556
	Profitable predictions	451
	Problem definitions	
	Potentials of Big Data	289
	Requirements of Big Data	284
	False understanding and misuse of Big Data	138
	Causal attributions	
	Advances in health and medicine	447
	Treatment recommendations	
	Use technologies to rebuild society	192
	Nonuse of technology	80
	Implications	
	Social benefit	241
	Economic and financial prosperity	204
	Moral evaluations	
	Negative	301
	Positive	616
	Ambivalent	217
	Framing device categories	
	Revolution (e.g., "rapid analysis and sharing")	204
	Data as force for good (e.g., "Imagine what is possible")	141
	Overwhelming abundance of data (e.g., "tsunami of data")	109
Reviving datafied business	*Cultural motifs*	
	Profitable predictions	888
	Shift in surveying and datafying society	72
	Problem definitions	
	(Unused) potentials of Big Data	473
	Requirements of Big Data	199
	Causal attributions	
	Data for profit	1166
	Treatment recommendations	
	Use technology to rebuild society	189
	More education	105
	Implications	
	Economic and financial prosperity	551
	Better decision-making	142
	Moral evaluations	
	Positive	618
	Ambivalent	338
	Negative	
	Framing device categories	
	Effectiveness/efficiency (e.g., "be more efficient")	176
	Revolution (e.g., "rapid analysis and sharing")	156
	Overwhelming abundance of data (e.g., "tsunami of data")	87
Retooling datafied surveillance	*Cultural motifs*	
	Negative consequences of Big Data	171
	Low profile surveillance	150
	Problem definitions	
	General suspicion	238
	Microtargeting	231
	Lack of transparency, lack of regulation	187
	Causal attributions	
	Governmental exploitation	129
	Treatment recommendations	
	Political regulation	99
	Data rights	80

(continued)

Table 2. Continued.

Frame	Frame package	No of coded propositions
	Implications	
	Mass surveillance	136
	Moral evaluations	
	Negative	400
	Framing devices categories	
	Mass surveillance (e.g., “use and abuse the data”)	49
	Data as force for good (e.g., “Imagine what is possible”)	35
	Data as raw material (e.g., “gold mine”)	22
		$N = 3338$

deciding if a certain event was the flashpoint of a news report. The three variables that fell below .667 were not considered in the analysis (Krippendorff, 2004; Matthes and Kohring, 2008).

To determine the frame packages, we employed a hierarchical and a subsequent partitioning cluster analysis (k-means) to see how framing and reasoning devices systematically group together and to identify these packages across the sample (Matthes and Kohring, 2008).[1]

Results

The cluster analysis yielded altogether three frame packages (*RQ1*; Table 2). Almost half of all propositional units (44%) contained devices that were clustered into a package that we named “rebuilding a datafied society.” Its central cultural motif is the overcoming of challenges in society by the use of technology. Big Data features here in terms of a revolution where the potentials of large-scale quantitative analytics return social benefits and make a difference in numerous other areas of application. On that note, *Newsweek*, for instance, postulated that “[f]rom lethal disease to murders, to deadly workplace accidents, suicides, fatal domestic violence incidents and natural disasters, researchers are now harnessing vast amounts of data to more specifically forecast mortality” (1 August 2014). Within this frame, the correlation between the framing devices “revolution” and “innovations for societal progress and/or personal advancement” are highly significant, χ^2 (60, $N = 740$) = 166.09, $p < .001$; Cramer’s V = .474, $p < .001$.

In the frame, economic motifs and implications exist but are less frequent than civil and individual aspects. Hence, the *Sunday Times* heralds

> [t]he coming together of the sensor networks known as the Internet of Things and the massive information flows we call big data have obvious benefits for industrial and consumer activities, ranging from managing factories to monitoring traffic patterns. Now, this combination could come to the rescue of threatened wildlife. (1 July 2018)

Table 3. Top-media outlets per country.

	Rebuilding a datafied society		Reviving datafied business		Retooling datafied surveillance	
Top 3 media outlets	*N*	%	*N*	%	*N*	%
United States	723	100.0	614	100.0	336	100.0
Financial Times	262	36.2	285	46.4	141	42.0
WSJ	282	39.0	229	37.3	117	34.8
NYT	80	11.1	51	8.3	26	7.7
Germany	649	100.0	431	100.0	333	100.0
FAZ	272	41.9	218	50.6	152	45.6
Handelsblatt	128	19.7	93	21.6	52	15.6
Süddeutsche Zeitung	108	16.6	31	7.2	46	13.8
South Africa	94	100.0	121	100.0	37	100.0
Business day	24	25.5	39	32.2	8	21.6
Brainstorm Magazine	24	25.5	25	20.7	1	2.7
Tech central	13	13.8	9	7.4	13	35.1
Total	1466	100.0	1166	100.0	706	100.0

There was a particular focus on the health sector and self-optimization through technological innovations. The causal attribution “advances in health and medicine, including self-optimization” was often mentioned ($n = 447$ in total), though its frequency differed significantly over the course of time (Welch’s-F(1; 805) = 4.069, $p < .044$). In the 10-year period of investigation, the variable peaked between 2014 ($n = 98$) and 2016 ($n = 59$). Furthermore, the variables “positive potential/social benefit” and “health sector” showed a highly significant correlation (χ^2 (36, $N = 528$) = 219.05, $p < .001$; Phi = .644, $p < .001$). For example, the *WSJ* wrote about a massive data-driven venture for “technology leading to a cure for cancer” (2 June 2016). The framing devices underscored the idea of Big Data’s revolutionary potential and the use of data for social good. In the United States, this frame was most strongly supported by the *WSJ* (with 282 propositional units), the *Financial Times* (262 units), and the *NYT* (80 units). In Germany, *FAZ* was on top (272 units), and in South Africa it was mainly sported by *Business Day* and *Brainstorm Magazine* (24 and 24 units, respectively; Table 3).

Table 4. Frames packages per outlet.

	Rebuilding a datafied society		Reviving datafied business		Retooling datafied surveillance		Total
Media outlets	*N*	%	*N*	%	*N*	%	*N*
United States	723	43.2	614	36.7	336	20.1	1673
Financial Times	262	38.1	285	41.4	141	20.5	688
WSJ	282	44.9	229	36.5	117	18.6	628
NYT	80	50.9	51	32.5	26	16.6	157
Newsweek	35	58.3	8	13.3	17	28.4	60
The Verge	26	44.1	14	23.7	19	32.2	59
WP	30	52.6	15	26.3	12	21.1	57
Forbes	6	30.0	10	50.0	4	20.0	20
Wired Magazine	2	50.0	2	50.0	0	0	4
Germany	649	45.9	431	30.5	333	23.6	1413
FAZ	272	42.4	218	34.0	152	23.6	642
Handelsblatt	128	46.8	93	34.1	52	19.1	273
Süddeutsche Zeitung	108	58.4	31	16.8	46	24.8	185
WirtschaftsWoche	48	46.2	41	39.4	15	14.4	104
c't	27	39.7	24	35.3	17	25.0	68
WELT online	26	44.1	18	30.5	15	25.3	59
Taz	19	48.7	4	10.3	16	41.0	39
ZEIT online	6	28.6	2	9.5	13	61.9	21
SPIEGEL online	9	69.2	0	0	4	30.8	13
Wired Germany	6	66.7	0	0	3	33.3	9
South Africa	94	37.3	121	48.0	37	14.7	252
Business Day	24	33.8	39	54.9	8	11.3	71
Brainstorm Magazine	24	48.0	25	50.0	1	2.0	50
Tech Central	13	37.1	9	25.7	13	37.2	35
The Star	17	51.5	15	45.5	1	3.0	33
Financial Mail	3	11.1	19	70.4	5	18.5	27
Sunday Times	7	38.9	8	44.4	3	16.7	18
Mail and Guardian	4	28.6	4	28.6	6	42.8	14
NAG	2	50.0	2	0	0	50.0	4
Total	1466		1166		706		3338

WP: Washington Post; WSJ: Wall Street Journal; NYT: New York Times; FAZ: Frankfurter Allgemeine Zeitung.

Interestingly, although the package contains strong positive connotations, there are also noticeable ambivalent and even negative evaluations that point to the potential misuse afforded by the available large datasets, the errors produced by biased data, and the many requirements necessary to exploit Big Data properly. Thus the possibility of not using Big Data technologies was also discussed.

We called the second most prominent frame package "reviving datafied business." A third of all propositional units contained devices that were clustered into this package (35%). It had a strong commercial orientation and sprang from the positive opportunities offered by Big Data analytics—in particular, for turning predictions into profits and capitalizing on surveying and datafying society. Hence there was a strong correlation between the problem definition that Big Data harbors potentials that are not yet realized and the cultural motif "profitable predictions" (χ^2 (48, $N = 843) = 322.01$, $p < .001$; Cramer's V = .681, $p < .001$). Accordingly, Big Data stood for technical progress and was part of a global transformation where the economic and financial sector were believed to benefit most (χ^2 (4, $N = 551) = 22.01$, $p < .001$; Cramer's V = .237, $p < .001$). There was also a strong correlation between the reasoning device "potential of Big Data" and the cultural motif "profitable predictions" (χ^2 (48, $N = 843) = 322.01$, $p < .001$; Cramer's V = .681, $p < .001$).

As could be expected, this perspective prevailed in outlets with a distinct business focus (Table 3). It was most apparent in the *Financial Times* (with 285 propositional units coded), *WSJ* (229 propositional units), and *FAZ* (218 propositional units). It was furthermore the most prevalent frame in South African media (Table 4). There are highly significant differences between the outlets in their annual distributions of coded items ($F(23, 25.431) = 5.82$, $p < .001$). The *WSJ* articles mainly advocate the view in 2013 and 2014, *FAZ* and the *Financial Times* peak in 2015. The framing devices also differ strongly over the years, with "revolution" spiking in 2012 and "effectiveness/efficiency" in 2015 ($F(9, 15.91) = 3.21$, $p < .001$). Thus, while a rhetoric of newness initially dominated, it gave way to expectations

Table 5. Distribution of frame packages across years.

	Rebuilding a datafied society		Reviving datafied business		Retooling datafied surveillance	
Year	*N*	%	*N*	%	*N*	%
2011	9	0.6	16	1.4	3	0.4
2012	99	6.8	92	7.9	21	3.0
2013	222	15.1	223	19.1	115	16.3
2014	270	18.4	210	18.0	128	18.1
2015	172	11.7	170	14.6	81	11.5
2016	202	13.8	162	13.9	108	15.3
2017	156	10.6	95	8.1	78	11.0
2018	137	9.3	91	7.8	77	10.9
2019	98	6.7	71	6.1	45	6.4
2020	101	6.9	36	3.1	50	7.1
Total	1466	100.0	1166	100.0	706	100.0

Note: There are highly significant differences between the frames over the years (Welch's $F(2, 1839.36) = 20.88$, $p < .001$). The statistically significant differences exist between Frames 1 and 3, as well as between Frames 1 and 2. No significant difference was found between Frames 2 and 3.

about the best strategy to exploit data. Some hurdles and challenging developments are mentioned though, as the enormous potentials often remained unused due to high requirements (such as large computing systems) and as datafication engenders a shift in power relations between enterprises. *Tech Central*, for instance, declared: "All businesses, big and small, need data and the insights that can be derived from it. When it comes to understanding customers' preferences and needs, big data plays a crucial role" (29 August 2019).

The third frame package was a cluster grouped around "retooling datafied surveillance." It was the least frequent view yet it highlighted some crucial issues. It gravitated around three problems: "general suspicion," "microtargeting," and a "lack of transparency." All three problems correlate significantly with the causal attribution that Big Data could be used for state-controlled purposes ("governmental exploitation"; χ^2 (14, $N = 487) = 213.787$, $p < .001$; Cramer's V = .663, $p < .001$). Laws and data rights were criticized as being obscure and the general suspicion directed against citizens was criticized (χ^2 (20, $N = 270) = 47.414$, $p < .001$; Phi = .419, $p < .001$). Among South African media, it appeared mostly in *Tech Central* (13 units); in German media *FAZ* (152 units); and in the United States the *Financial Times* (141 units; Table 3).

Within the frame package, the three problem definitions correlated with ambivalent to negative evaluations (χ^2 (4, $N = 606$) $= 23.539$, $p < .001$). Mass surveillance was discussed as an implication of Big Data exploitation (χ^2 (8, $N = 328) = 40.281$, $p < .001$; Phi V = .350, $p < .001$). Big Data's potential for mass surveillance was addressed as an implication ($n = 129$, χ^2 (14, $N = 487) = 213.787$, $p < .001$; Phi V = .663, $p < .001$). Political regulation and the strengthening of data rights were recommended. Both of these treatment recommendations correlated with regulation and transparency problems (χ^2 (8, $N = 280) = 25.31$, $p < .001$; Cramer's V = .301, $p < .001$). On this note, the *Financial Times* wrote: "it is important that there is greater and genuine transparency about the use of such techniques to ensure that people have control over their own data" (17 May 2017) and the *NYT* stated: "What we need is for an ethics of data to be engineered right into the information skyscrapers being built today. We need data ethics by design" (25 March 2018). Regarding framing devices, "retooling datafied surveillance" was associated with notions of mass surveillance and data as "raw material," but also linked to terms suggesting Big Data could nevertheless be "a force for good" (χ^2 (70, $N = 270) = 134.523$, $p < .001$; Phi = .786, $p < .001$).

Looking at the existence of frame package devices across time (*RQ2*), we found a highly significant correlation between the mix of framing devices and reasoning devices and the 10 consecutive years (Welch's $F(2, 1839.36) = 20.88$, $p < .001$). The ANOVA suggested that there was a statistically significant difference between "reviving datafied business" and "rebuilding a datafied society," as well as between "reviving datafied business" and "retooling datafied surveillance" in their annual occurrence, whereas the society-centered and critical perspectives had no significant peak.

Although each year was marked by a peculiar composition of framing devices and reasoning devices, there were clear dominances. In the first three years (2011–2013), the majority of items belonged to the business frame package; subsequently the emphasis shifted to the broadly conceived societal perspective (Table 5). The individual set of devices, their presence and rate, changed from year to year, yet this did not change the overall importance of the packages much. This was mainly due to the dominance of the handful of outlets that contributed most of the sampled news items. Many outlets only dedicated a few articles to the topic. Hence we found a shift in the microconfiguration and not the overall macro structure. Looking at the volume of articles per year, we found a noticeable peak in the period from 2013 to 2016 across all three frame packages.

There was furthermore a highly significant difference between the frame packages and media outlets (Welch's $F(2, 1814.98) = 6.38$, $p < .002$). In other words, each newspaper or magazine was characterized by a distinct frequency of framing and reasoning devices. Differences existed between the business and the societal views, and between the business and surveillance views, yet not between the societal perspective and the critical position (Table 4). The frame package around "retooling datafied surveillance" ranked second in five German publications, while the business-centered perspective loomed large in South African outlets.

Finally, when considering country-specific relations (*RQ3*), we again found highly significant differences between the three frame packages and the three countries. That is, each country could be identified by its signature mix of framing and reasoning devices (Welch's $F(2, 1819.31) = 6.77$, $p < .001$). As with the medium-specific

Table 6. Frame packages per year and country.

	Rebuilding a datafied society		Reviving datafied business		Retooling datafied surveillance	
Year	*N*	%	*N*	%	*N*	%
2011–2013	330	100.0	331	100.0	139	100.0
United States	205	62.1	198	59.8	68	48.9
Germany	94	28.5	84	25.4	65	46.8
South Africa	31	9.4	49	14.8	6	4.3
2014–2017	800	100.0	637	100.0	395	100.0
United States	378	47.3	313	49.1	170	43.0
Germany	393	49.1	287	45.1	213	53.9
South Africa	29	3.6	37	5.8	12	3.0
2018–2020	336	100.0	198	100.0	172	100.0
United States	140	41.7	103	52.0	98	57.0
Germany	162	48.2	60	30.3	55	32.0
South Africa	34	10.1	35	17.7	19	11.0
Total	1466		1166		706	

Note: The mean of all articles is given as 2015.41, with a standard deviation of 2.278 ($N = 3338$). The minimum value (min) is 2011 and the maximum value (max) is 2020. The 25th percentile (P25) is 2014 and the 75th percentile (P75) is 2017.

and year-specific proportions, differences could be confirmed for the package favoring a business view and the one stressing the societal dimension of Big Data, as well as the business-inclined position and the one underscoring the problematic issues. No such difference could be detected for the relationship between “retooling datafied surveillance” and “rebuilding a datafied society.”

We were able to identify three periods when considering the volume of coded propositions. In the first phase from 2011 to 2013, the U.S. publications were producing the most coverage yet they were outstripped by German media in the years from 2014 to 2017. Subsequently, a more complex pattern emerged. While almost all prolific outlets in the United States and Germany followed the general order of prominence, South African media most often chose a business-oriented framing, with the societal-inclined perspective only coming second. Given the much lower number of news texts and propositional units in the sample from South Africa, this variant focus is not reflected in the overall ranking (Table 6).

Discussion

The limited general evolution of the issues at stake is particularly striking when looking back over the 10 years of discourse on the large-scale collection and exploitation of data. Although each year is distinct and marked by its own composition of frame packages, there is still not much overall topical development. This discursive inertia of three sociotechnical imaginaries is somewhat surprising for two reasons.

Firstly, the practices of datafication have matured tremendously, but while technology and practices have changed considerably, the sociotechnical imaginaries seem little altered. Instead, the basic set of views around the social and commercial benefits and the critical perspective on the control and surveillance made possible by ubiquitous data analytics have neither ceased to attract nor have they been replaced by other, more congenial imaginaries. This set of imaginaries still prevails in public sensemaking. In a way, datafication does not stop to amaze journalists in all three countries, or rather they keep resorting to the well-known revolutionary rhetoric when writing about new applications and forms of use. As such, they perpetuate an excitement that resonates with techno-solutionist thinking in its discursive persistence and cultural allure (Morozov, 2014). Each novel service or application is thus again framed in terms of its enormous potential and far-reaching implications. Similar observations have been made for other recent innovations like AI that evoke imaginaries pivoting on a “technological fix” as a cure to social ills (Bareis and Katzenbach, 2022; Katzenbach, 2021). In this way, journalists risk circulating imaginaries championed by those invested in data analytics like Silicon Valley conglomerates and the global startup scene (Daub, 2020; Levina and Hasinoff, 2017).

Secondly, the preponderance of positively toned perspectives is startling because it does not mirror the series of data scandals that have been spawned by whistleblowers in the wake of Edward Snowden (Di Salvo, 2020). Next to the high-profile cases of Christopher Wylie or Brittany Kaiser, who were both embroiled in the Cambridge Analytica revelations, there are also more local cases like the #GuptaLeaks in South Africa (Dodd and Merwe, 2019) or the Handygate Affair in Germany (Pentzold and Fischer, 2017). The frame package “retooling datafied surveillance” certainly epitomizes this critical view, yet we found no evidence for its increasing significance. These events neither led to spikes in the volume of news reports nor had a lasting effect.

The overall presence of frame packages and thus sociotechnical imaginaries is mainly due to the unequal distribution of publication volume. There are many outlets whose editorial line seems to counter the general trend. Among the nine German publications, there are five prioritizing the frame package “retooling datafied surveillance” over the commercial view; in the United States and South Africa there are another two that emphasize the critical view more. At the same time, almost all remaining South African media and two U.S. business magazines place the economic perspective first. They thus underscore the editorial discretion of setting a framing agenda that highlights some aspects of Big Data. Moreover, the decision of German media to endorse critical positions resonates with the reservations found in the national public opinion and among policymakers in Germany (acatech and Körber-Stiftung, 2023). It may also be taken as confirming the “postindustrial paradox” (Nisbet and Nisbet, 2019: 13) of increasing standards of living going hand-in-hand with rising tech skepticism, especially because media from the BRICS country South Africa champion affirmative and business-friendly views. Because the limited absolute number of texts and propositional units of all the outlets does not support the general ranking, they do not leave their mark on the overall tendencies.

Arguably, the most obvious reason for the permanence of the three sociotechnical imaginaries and the limited echo of the data scandals lies with the buzz phrase "Big Data" itself. In the discourse, the term seems to have been occupied by overtly positive aspirations and associated with commercial ambitions. For example, the many books that popularized the expression during the same period were also brimming with expectations and often authored by journalists and IT commentators also present in the news media and magazines we studied (Pentzold and Knorr, 2023). Because the term strongly connotes profits, benefits, and prosperity at the expense of more critical views, it might have been eschewed or demoted by other writers and commentators. Indeed, looking back at ten years of discourse it seems that the term has increasingly fallen out of use (Pentzold and Knorr, 2023). In the first years of the period studied, it was evoked to gesture to datafication's revolutionary potential in all sorts of sectors. Arguably, this auspicious view was key in stirring interest among politicians, managers, engineers, and educators that has made massive datafication possible. Yet today, the term "Big Data" and the anticipations it entailed seem outdated and of little use to come to terms with the ever-increasing generation and exploitation of data.

A point in case is the large number of articles that were initially sampled yet not selected for analysis. There is a large proportion of items that did not contain the search terms in either headline, subheadline, or first paragraph ($n = 7055$). Notwithstanding the ancillary role of the notion in these articles, they still dealt with the use of large quantities of data but without using the tainted keyword. Following the keyword "Big Data" therefore afforded access to an outlier portion of a more widespread discourse. There, a range of applications from epidemiological forecasting to data-driven profiling are discussed. While these are clearly examples of pervasive datafication, the media reports dedicated to them refrain from using "Big Data" prominently. Hence, the large number of articles not selected for further analysis can be due to the problematic nature of "Big Data" as a keyword. The term may have been used to classify and index news stories in a news website's backend yet it was not taken up in the piece itself. Also note that our search for "dataf*" led to virtually no results, thus stressing the concept's mostly academic resonance.

This is not to say that the articles did not cover topics related to Big Data, but they either evaded the keyword and its connotations on purpose or the lack of explicit mentions is indicative of a kind of discursive naturalization: with datafication becoming omnipresent, it stops being a tech issue and becomes part and parcel of a broad range of topics that are discussed under more apt terms like predictive policing or microtargeting. So the absence of the buzz phrase may signal the missing need to employ this somewhat vacuous term anymore to indicate a topical focus on digital data and data analytics—with datafication threading through all walks of life, it ceases to be an IT issue and morphs into social affairs writ-large. It becomes the focus of general journalistic inquiry.

Conclusion

In our study, we looked at the *zeitgeist* of the ever-increasing gathering, appropriation, and analysis of digital data. This turn to the discursive dimension of Big Data is important, we argue, and does not detract from its concrete technological force. Quite the opposite: It orients our attention to discourses that shape the social construction of datafication and translate into practices, organizational forms, policies, and institutions. In a word, discourses are integral to how we engage with Big Data. As such, discourses allow us to appreciate and critically engage with what datafication is or should be for various speakers and how they discuss, criticize, or envision the collection and use of data at different places, speaking from different situations and at different times (Brown et al., 2000; Hajer, 1995).

To acknowledge the consequential nature of discourses and their material ramifications in envisioning technological choices, steering expenditures, and the orientation of public attention, we refer to sociotechnical imaginaries that are performative and forward looking (Jasanoff, 2015). To speak of sociotechnical imaginaries is to assume that ideas manifesting in the rhetoric associated with technology are not separate from material tinkering and engineering. Likewise, there is no technological precedence in which technology comes first and discourses second. To the contrary, conceptions of upcoming or nascent technologies and projections of their desirable or undesirable potential are intimately linked to design affordances, scientific discovery, commercial prospects, perceptions of customers, and policy decisions (Beckert, 2016).

In the light of these assumptions, we analyzed articles from publications in three countries: the United States, Germany, and South Africa. To reconstruct the sociotechnical imaginaries, we chose a frame analysis that yielded three frame packages. The most frequently evoked perspective suggested that datafication processes are key to rebuilding society. The second most prominent frame package highlighted the economic potential of large quantities of data, and the third frame package underscored the critical expansion and intensification of datafied surveillance. This set of imaginaries endured across the 10-year period studied in spite of all the profound developments in datafication practice. This could be read as a sign of semantic inertia or as proof of their persistence. Even the data scandals that punctuated this era had no radical echo. Instead, the discourse was led by a few prolific media outlets that formed the lion's share of the sample. Still, some publications accentuated critical views while others had a more pronounced commercial focus. There were also variations at the country level. For example, South African media and their affirmative, business-orientated inclination resonate with other insights into the assessment of new technologies among South Africans (Guenther and Weingart, 2018). Likewise, the critical views propagated by some German publications tie in with tech-skeptical positions held by the German public.

So in essence, the frames we found are limited and have a tendency to solidify in sociotechnical imaginaries that come with technological lock-ins and long-term decisions for programs of public expenditure and lawmaking. Looking back at 10 years of discourse as it is reflected in news media shows how little has changed over time. Nevertheless, discursive closure is a precarious state that can give way to alternative interpretations. In this regard, moving beyond the journalistic arena to networked communication and mainstreaming voices from the Global South can help to transform and manifold the interpretative repertoire which harbors opposing sociotechnical imaginaries.

Admittedly, our sample of major media in the United States, in Germany, and South Africa is only a starting point for a more extensive cross-country comparison that should include a diversity of non-Western contexts of data application and interpretation. Although of global significance and scope, perspectives on Big Data are examined almost exclusively for Western countries, if at all. This is critical because the data divide not only exists within and between Western countries, but also shapes the interpretation and handling of datafication processes in emerging and developing countries and reveals a disparity in data-based opportunities for influence and knowledge (Arora, 2019; Milan and Treré, 2019). The question of hegemonic interpretations and the power to influence framing processes also remains unresolved, not only between countries but also within different news media and public forums of a national public sphere (Vliegenthart and van Zoonen, 2011). In that regard, the present analysis needs to be extended so to include, for example, media outlets in other dominant languages in South Africa like Zulu and Afrikaans, and to take into account online platforms and networked communication that needs to be studied along a more in-depth interpretative engagement with the journalistic sources.

Although the term "Big Data" may have lost some of its allure in recent years and has never captured the broad variety of data-based practices, neither the term nor the sociotechnical imaginaries with which it is associated were obsolete from the outset. On the contrary: they were instrumental in drawing attention to the profound changes brought by the extensive accumulation and computation of data (Lohr, 2015; McAfee and Brynjolfsson, 2012). The notion provided a moniker for vast and abstract data practices and placed the issue on the public agenda. Yet rather than animating scrutiny, to some extent the unfolding discourse contributed to the hype and furthered the prospects of industries thriving on "surveillance capitalism" (Zuboff, 2019). In this capacity, "Big Data" surfaces in many other venues such as business reports and policy documents that give voice to a common concern around the exploitation of data and, at the same time, further encourage interests in their commodification and extended collection.

ORCID iDs

Charlotte Knorr https://orcid.org/0000-0001-6591-3781
Christian Pentzold https://orcid.org/0000-0002-6355-3150

Statement and declarations

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) (Grant No. 447465824/PE 2436/3-1).

Conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Note

1. For the cluster analysis, 34 binary-coded variables were used. All had a sufficient intercoder reliability with Krippendorff's α above 0.76, and they highly correlated with each other, as determined by a χ^2-test for each pair of variables. The variables that were difficult to recognize by the coders and did not strongly correlate with each other were not used in the cluster analysis. A hierarchical cluster analysis (Ward's method) was conducted, selecting variables for cultural motifs, reasoning devices, and framing devices that strongly correlated in the dataset. The dendrogram showed seven clusters (elbow criterion). A scree plot of a complementary exploratory factor analysis indicated seven factors. Yet, Kaiser–Mayer–Olkin test and Bartlett test showed low suitability (29%). We cross-tabulated seven clusters with framing and reasoning devices using χ^2-tests. That procedure served to narrow down the data set to the variables and codes that correlated strongly with each other and it served as another preparatory step for the second cluster analysis. In the end, we used the following variables in the cluster analysis based on intercoder reliability and correlation: eight framing devices (out of 12), six of 10 problem definitions, seven of eight causal attributions, three of five treatment recommendations, moral judgment, five of seven cultural motifs, where three codes had been merged before. We conducted a hierarchical cluster analysis based on a factor analysis with 34 variables that yielded seven clusters of which four were semantically coherent. Also, in the k-means cluster analysis, the

iteration values between three and four differed only marginally so we merged two clusters. Finally, using k-means cluster analysis, we then specified three clusters with distinct sets of cultural motifs, reasoning devices and framing devices. The three final clusters were cross-tabulated again. For statistical analysis, we used standard measures and descriptive correlations (mean comparisons, test for normal distribution, Shapiro–Wilk test), as well as the test for mean comparisons and ANOVA tests. We chose Welch's ANOVA tests to avoid missing homogeneity in framing and reasoning devices.

References

acatech und Körber-Stiftung (Eds.). (2023) TechnikRadar 2023. Was die Deutschen über Technik denken. Schwerpunkt: Bauen und Wohnen. https://www.acatech.de/publikation/technikradar-2023/download-pdf?lang=de.

Allum N, Besley N, Gomez L, et al. (2018) Disparities in science literacy. *Science* 360(6391): 861–862.

Allum N, Sturgis P, Tabourazi D, et al. (2008) Science knowledge and attitudes across cultures. *Public Understanding of Science* 17(1): 35–54.

Andrejevic M (2014) Big data, big questions: The big data divide. *International Journal of Communication* 8(1): 1673–1689. https://ijoc.org/index.php/ijoc/article/view/2161.

Arora P (2019) *The Next Billion Users*. Cambridge, MA: Harvard University Press.

Barbrook R (2007) *Imaginary Futures: From Thinking Machines to the Global Village*. London: Pluto Press.

Bareis J and Katzenbach C (2022) Talking AI into being. *Science, Technology, & Human Values* 47(5): 855–881.

Beckert J (2016) *Imagined Futures*. Cambridge, MA: Harvard University Press.

Beer D (2019) *The Data Gaze*. London: Sage.

Boellstorff T (2015) *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton, NJ: Princeton University Press.

Bory P (2020) *The Internet Myth: From the Internet Imaginary to Network Ideologies*. London: University of Westminster Press.

Bowker G (2013) Data flakes. In: Gitelman L (ed) *"Raw Data" is an Oxymoron*. Cambridge, MA: MIT Press, 167–172.

boyd d and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Brown N, Rappert B and Webster A (2000) *Contested Futures*. Farnham: Ashgate.

Bucher T (2017) The algorithmic imaginary. *Information, Communication & Society* 20(1): 30–44.

Burgers C (2016) Conceptualizing change in communication through metaphor. *Journal of Communication* 66(2): 250–265.

Couldry N and Mejias U (2019) Data colonialism. *Television & New Media* 20(4): 336–349.

Couldry N and Yu J (2018) Deconstructing datafication's brave new world. *New Media & Society* 20(12): 4473–4491.

Dalton R and Welzel C (2014) *The Civic Culture Transformed*. Cambridge: Cambridge University Press.

Daub A (2020) *What Tech Calls Thinking*. London: Macmillan.

Dencik L and Kaun A (2020) Datafication and the welfare state. *Global Perspectives* 1(1). https://doi.org/10.1525/gp.2020.12912.

Diebold FX (2012) On the Origin(s) and Development of the Term 'Big Data'. PIER Working Paper No. 12-037. http://dx.doi.org/10.2139/ssrn.2152421.

Di Salvo P (2020) *Digital Whistleblowing Platforms in Journalism*. Basingstoke: Palgrave.

Di Salvo P and Negro G (2016) Framing Edward Snowden: A comparative analysis of four newspapers in China, United Kingdom and United States. *Journalism* 17(7): 805–822.

Dodd N and Merwe J (2019) The Gupta leaks and intra-BRICS collusion. In: Merwe J, Dodd M and Bond P (eds) *BRICS and Resistance in Africa*. London: Bloomsbury, 152–169.

Driscoll K and Paloque-Berges C (2017) Searching for missing "net histories". *Internet Histories* 1(1): 47–59.

Droog E, Burgers C and Kee KF (2020) How journalists and experts metaphorically frame emerging information technologies. *Public Understanding of Science* 29(8): 819–834.

Entman RM (1993) Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4): 51–58.

Entman RM (2010) Cascading activation: Contesting the white house's frame after 9/11. *Political Communication* 20(4): 415–432.

Ferrari E (2020) The dominant technological imaginary of Silicon Valley. *Communication, Culture and Critique* 13(1): 121–124.

Flensburg S and Lomborg S (2023) Datafication research. *New Media & Society* 25(6): 1451–1469.

Flichy P (2007) *The Internet Imaginaire*. Cambridge, MA: MIT Press.

Gamson WA and Modigliani A (1989) Media discourse and public opinion on nuclear power. *American Journal of Sociology* 95(1): 1–37.

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed) *"Raw Data" is an Oxymoron*. Cambridge, MA: MIT Press, 1–14.

Guenther L and Weingart P (2018) Promises and reservations towards science and technology among South African publics. *Public Understanding of Science* 27(1): 47–58.

Haim M, Graefe A and Brosius H-B (2018) Burst of the filter bubble? *Digital Journalism* 6(3): 330–343.

Hajer MA (1995) *The Politics of Environmental Discourse*. Oxford: Oxford University Press.

Jasanoff S (2015) Future imperfect. In: Jasanoff S and Kim S (eds) *Dreamscapes of Modernity*. Chicago: University of Chicago Press, 1–33.

Jasanoff S and Kim S-H (2009) Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* 47(2): 119–146.

Kanna T, Raina A and Chawla R (2024) India Stack: Digital Public Infrastructure for All. HBS Case Collection. Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=64379.

Katzenbach C (2021) “AI will fix this” – The technical, discursive, and political turn to AI in governing communication. *Big Data & Society* 8(2): 1–8.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Koenen E, Schwarzenegger C, Kittler J (2021) Data(fication): “Understanding the world through data” as an everlasting revolution. In: Balbi G, Ribeiro N, Schafer V, et al. (eds) *Digital Roots*. Berlin: De Gruyter Oldenbourg, 137–156.

Krippendorff K (2004) Reliability in content analysis. *Communication Research* 30(3): 411–433.

Lehtiniemi T and Ruckenstein M (2019) The social imaginaries of data activism. *Big Data & Society* 6(1): 1–12.

Lesage F and Rinfret L (2015) Shifting media imaginaries of the web. *First Monday* 20(10). https://doi.org/10.5210/fm.v20i10.5519.

Levina M and Hasinoff AA (2017) The Silicon Valley ethos. *Television & New Media* 18(6): 489–495.

Lo WH, Lam BSY and Cheung MMF (2021) The dynamics of political elections. *Social Science Computer Review* 39(4): 627–647.

Lohr S (2015) *Data-ism: Inside the Big Data Revolution*. New York: Oneworld Publications.

Mager A and Katzenbach C (2021) Future imaginaries in the making and governing of digital technology. *New Media & Society* 23(2): 223–236.

Mansell R (2012) *Imagining the Internet*. Oxford: Oxford University Press.

Marcus GE (1995) *Technoscientific Imaginaries*. Chicago: University of Chicago Press.

Markham A and Tiidenberg K (2020) *Metaphors of Internet: Ways of Being in the Age of Ubiquity*. New York: Peter Lang.

Marx L (2000) *The Machine in the Garden*. Oxford: Oxford University Press.

Matthes J and Kohring M (2008) The content analysis of media frames. *Journal of Communication* 58(2): 258–279.

Mayer-Schönberger V and Cukier K (2013) *Big Data*. New York: Houghton Mifflin Harcourt.

McAfee A and Brynjolfsson E (2012) *Big Data: The Management Revolution*. Cambridge, MA: Harvard Business Review.

Milan S and Treré E (2019) Big data from the south(s): Beyond data universalism. *Television & New Media* 20(4): 319–335.

Morozov E (2014) *To Save Everything, Click Here*. New York: PublicAffairs.

Mosco V (2004) *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge, MA: MIT Press.

Natale S (2016) There are no old media. *Journal of Communication* 66(4): 585–603.

Nathan L (2017) Who is keeping an eye on South Africa’s spies? CNBC Africa. https://www.cnbcafrica.com/2017/who-is-keeping-eye-south-africas-spies-nobody-thats-problem/.

Nisbet MC and Nisbet EC (2019) *The Public Face of Science Across the World*. Cambridge: Cambridge University Press.

Nolin JM (2019) Data as oil, infrastructure or asset? *Journal of Information, Communication and Ethics in Society* 18(1): 28–43.

Nye DE (1996) *). American Technological Sublime*. Cambridge, MA: MIT Press.

Paganoni MC (2019) *Framing Big Data: A Linguistic and Discursive Approach*. Cham: Springer.

Pentzold C and Fischer C (2017) Framing Big Data: The discursive construction of a radio cell query in Germany. *Big Data & Society* 4(2): 1–11.

Pentzold C and Knorr C (2023) When data became big: Revisiting the rise of an obsolete keyword. *Information, Communication & Society* 27(3): 600–617.

Portmess L and Tower S (2015) Data barns, ambient intelligence and cloud computing. *Ethics and Information Technology* 17(1): –9.

Puschmann C and Burgess J (2014) Metaphors of Big Data. *International Journal of Communication* 8: 1690–1709.

Rieder G (2018) Tracing big data imaginaries through public policy. In: Sætnan A, Schneider I and Green N (eds) *The Politics of Big Data*. London: Routledge, 89–109.

Russell A and Waisbord S (2017) Digital citizenship and surveillance. *International Journal of Communication* 11: 858–878. https://ijoc.org/index.php/ijoc/article/view/5526.

Schulz C (2023) A new algorithmic imaginary. *Media, Culture & Society* 45(3): 646–655.

Stefik M (1996) *Internet Dreams. Archetypes, Myths, and Metaphors*. Cambridge, MA: MIT Press.

Taffel S (2023) Data and oil: Metaphor, materiality and metabolic rifts. *New Media & Society* 25(5): 980–998.

Taylor C (2004) *Modern Social Imaginaries*. Durham, NC: Duke University Press.

van Dijck J (2014) Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12(2): 197–208.

Van Gorp B (2007) The constructionist approach to framing. *Journal of Communication* 57(1): 60–78.

Van Gorp B (2010) Strategies to take subjectivity out of framing analysis. In: D’Angelo P and Kuypers JA (eds) *Doing News Framing Analysis*. London: Routledge, 84–109.

Vliegenthard R and van Zoonen L (2011) Power to the frame. *European Journal of Communication* 26(2): 101–115.

Wyatt S (2021) Metaphors in critical internet and digital media studies. *New Media & Society* 23(2): 406–416.

Zhang Z and Negro G (2019) Introduction: Exploring flows and counter-flows of information along the new silk road. *Communication and the Public* 4(4): 255–260.

Zuboff S (2019) *The Age of Surveillance Capitalism*. New York: Profile Books.