Technische Universität München
TUM School of Social Sciences and Technology

# COVID-19 and Data Journalism. Insights from Computational Social Science

Benedict Ludwig Witzenberger

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

| | | |
|---|---|---|
| Vorsitz: | | Prof. Dr. Orestis Papakyriakopoulos |
| Prüfende der Dissertation: | 1. | Prof. Dr. Jürgen Pfeffer |
| | 2. | Prof. Dipl.-Journ. Christina Elmer |

Die Dissertation wurde am 19.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 02.05.2024 angenommen.

# Acknowledgments

I would like to express my heartfelt gratitude to several individuals who have played a pivotal role in my academic and personal journey:

I am deeply grateful to Prof. Dr. Jürgen Pfeffer for his invaluable supervision and guidance. His insightful perspectives and encouragement have been fundamental in shaping this research.

I am also immensely grateful to Vanessa Wormer, whose guidance has been crucial in my professional and academic development.

My parents deserve special acknowledgment for their support and encouragement throughout my life.

Lastly, I deeply appreciate my wife, Sabrina, for her enduring support and patience. Her presence has been a source of strength throughout the challenges and achievements of this academic endeavor.

# Abstract

This dissertation examines COVID-19's impact on data journalism through the lens of Computational Social Science methods. It reveals the increased use of infographics, discovers expanded collaborations between data and science journalists, and observes divided perceptions of predictive reporting. It underscores the importance of interdisciplinary collaboration and the application of mixed-method analysis in understanding newsroom dynamics and audience engagement.

# Zusammenfassung

Diese Dissertation untersucht die Auswirkungen von COVID-19 auf den Datenjournalismus mithilfe von Methoden der Computational Social Science. Sie zeigt die vermehrte Nutzung von Infografiken auf, untersucht verstärkte Kooperationen zwischen Datenjournalisten und Wissenschaftsjournalisten und beobachtet geteilte Einstellungen zu predictive journalism. Sie betont die Bedeutung interdisziplinärer Zusammenarbeit und die Anwendung von Mixed-Method-Analysen zum Verständnis der Dynamiken in Redaktionen und der Einbindung der Nutzer.

# Publications

Table 1: International, peer-reviewed publications in this dissertation

| No. | Year | Authors | Paper | Venue |
|---|---|---|---|---|
| 1 | 2023 | Benedict Witzenberger, Angelina Voggenreiter, Jürgen Pfeffer | Popular and on the Rise - But Not Everywhere: COVID-19-Infographics on Twitter | Peter Haber, Thomas J. Lampoltshammer, and Manfred Mayr, eds. (2024). *Data Science—Analytics and Applications. Proceedings of the 5th International Data Science Conference—iDSC2023.* Cham: Springer. 104 pp. ISBN: 9783031421716 |
| 2 | 2023 | Benedict Witzenberger, Nicholas Diakopoulos | Election predictions in the news: how users perceive and respond to visual election forecasts | Information, Communication & Society |
| 4 | 2024 | Benedict Witzenberger, Jürgen Pfeffer | More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany | Journalism and Media |

Table 2: Other publications in this dissertation

| No. | Year | Authors | Paper | Venue |
|---|---|---|---|---|
| 3 | 2023 | Benedict Witzenberger, Jürgen Pfeffer | Unleashing Data Journalism's Potential: COVID-19 as Catalyst for Newsroom Transformation | Pre-print: Arxiv.org, Under Review for: Journalism Practice |

Table 3: Talks and Presentations

| Date | Title and Venue |
|---|---|
| June 4th, 2021 | Bulletproofing Data, Seminar Critical Data Studies, Technical University of Munich SoSe 2021, Munich, Germany |
| November 12th, 2022 | Gender dynamics of German journalists on Twitter, The 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2022), Istanbul, Turkey |
| May 2nd, 2023 | Popular and on the Rise — But Not Everywhere: COVID-19-Infographics on Twitter, 5th International Data Science Conference (iDSC 2023), Krems, Austria |

# Contents

# Chapter 1

# Introduction

*"The plural of anecdote is data"*[1]

- Raymond Wolfinger, *Political Scientist*

"The plural of anecdote is data," a maxim attributed to American political scientist Raymond Wolfinger. It encapsulates the essence of this dissertation through the interplay of the COVID-19 pandemic and data journalism, using Computational Social Science methods. The COVID-19 pandemic, starting in 2020, marked an unprecedented and transformative moment in modern history. Beyond its immediate and profound impact on public health, the pandemic catalyzed societal changes and innovations, reshaping how we live, work, and communicate.

Also, journalism underwent a significant transformation due to the impact of the pandemic. As people increasingly relied on information to navigate the virus and their world, journalism evolved to meet their needs and provide clarity amidst confusion. In this rapidly changing landscape, data journalism emerged as a pivotal and resilient approach that uses the power of data to convey complex information, raise the perspectives from individual cases to a societal level, and turn anecdotes into data.

This dissertation explores the changes brought to data journalism during the COVID-19 pandemic, unveiling how the crisis catalyzed innovation while leveraging Computational Social Science methods to uncover these influences. Through four papers, this dissertation navigates the multifaceted dimensions of this relationship, offering a nuanced understanding of journalism's response to the crisis and the broader implications for media practices in the digital age.

## 1.1 Background and Motivation

The COVID-19 pandemic, which emerged in late 2019, has had profound and far-reaching impacts on various facets of society like health, mobility, and workplaces (Barouki et al.,

[1] According to his recount, Professor Wolfinger said this in a graduate seminar at Stanford during the 1969-1970 academic year after a student dismissed another student's factual statement as an anecdote. It has since been used numerous times in different, even twisted, contexts. See https://web.archive.org/web/20130628221221/http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0407a&L=ads-l&P=8874 and https://blog.revolutionanalytics.com/2011/04/the-plural-of-anecdote-is-data-after-all.html for further context.

2021; Kniffin et al., 2021; Daly, Sutin, and Robinson, 2022).

One of the many areas affected by this global crisis is journalism (E.-J. Lee, 2021). Although the industry is in a state of constant flux (Spyridou et al., 2013) with innovation management being a constant and critical element of media management (Mierzejewska, 2011), Covid-19 was a change on another scale (José A. García-Avilés, 2021; José A. García-Avilés et al., 2022). Journalists had to leave their newsrooms to work from home. Sources were only available via phone or video. On the other hand, the uncertainty of the situation led to enormous demand for journalism, rising digital subscriptions, and new skills for journalists (Quandt and Wahl-Jorgensen, 2022), offering new ideas for a revised societal role for journalism by calling for care and offering communities relevant information (Hess and Waller, 2020; Fernández-Pedemonte, Casillo, and Jorge-Artigau, 2020). In the wake of this crisis, data journalism has surfaced as a potent and essential method for conveying intricate information to the general public (Pentzold and D. Fechner, 2019; Danzon-Chambaud, 2021; José A. García-Avilés et al., 2022).

Understanding the relationship between COVID-19 and data journalism is important because it can shed light on how crises can drive innovation and transformation within the field of journalism. Furthermore, integrating Computational Social Science methods offers new avenues for studying these influences in a data-driven and systematic manner.

## 1.2 Research Questions & Objectives

The rise of data journalism has been studied continuously for some years now. Focus was placed on perspectives like the work with data as a source (Howard, 2014), adapted training requirements (Weiss and Retis, 2018; Burns and Matthews, 2018), or its divergence of diverse fields and practices (Coddington, 2015). Great attention was also paid to integrating data journalists within the newsrooms (Parasie and Dagiral, 2012; Hermida and M. L. Young, 2019).

Newer studies emphasized developments like the publications of predictive models and the visualizations of their implied uncertainties (Pentzold, D. J. Fechner, and Zuber, 2021; Pentzold and D. Fechner, 2021).

Most of these publications, however, are based on surveys or manual coding of publications, a common methodology in data journalism studies (Ojo and B. Heravi, 2017; Loosen, Reimer, and De Silva-Schmidt, 2017; Auväärt, 2023), which leads to the main contribution of the planned work: it aims to measure connections, prevalence, and propagation of data journalistic work with Computational Social Science methods. This enables comparable and repeatable measurements to track changes over time. It also allows quantifying these variances. Furthermore, using datasets bound to behavior adds another dimension to the validity of observations.

To guide this dissertation, we will focus on these broad research questions:

**RQ 1**: How has COVID-19 influenced the practice of data journalism?

**RQ 2**: How effective are Computational Social Science methods in systematically analyzing the impact of COVID-19 on data journalism practices in terms of accuracy and comprehensiveness?

To further concretize the questions, this work seeks to achieve the following research objectives:

**Objective 1: Quantitative Analysis of Data Journalism Outputs**
The first research objective is to quantitatively analyze data journalism outputs within German newsrooms during and after the COVID-19 pandemic. This analysis aims to provide insights into the extent of the increase in data-driven publications during this period.

**Objective 2: Utilization of Computational Social Science Methods**
The second research objective is to demonstrate the efficacy and utility of Computational Social Science methods in systematically analyzing the impact of COVID-19 on data journalism practices. This objective underscores the importance of employing advanced methodologies to study complex phenomena.

**Objective 3: Increased Understanding of COVID-19's Influence on Data Journalism**
The third and final research objective is to understand better how the COVID-19 crisis has reshaped data journalism and its role within the media landscape. This objective contributes to the broader discourse on journalism's response to crises in the digital age, enhancing our knowledge of journalism practice during pivotal historical moments.

To achieve these objectives, this dissertation is structured around four papers:

Paper 1 - "Election predictions in the news: how users perceive and respond to visual election forecasts" explores how users perceive and respond to visual election forecasts in news media, particularly in the context of election predictions, which evolved as a concurrent approach of predictive journalism next to models forecasting the spread of COVID-19.

Paper 2 - "Popular and on the Rise - But Not Everywhere: COVID-19-Infographics on Twitter" provides an in-depth analysis of the popularity and prevalence of COVID-19 infographics on Twitter, shedding light on how these visualizations were used to communicate pandemic-related information.

Paper 3 - "Unleashing Data Journalism's Potential: COVID-19 as Catalyst for Newsroom Transformation" delves into the transformational impact of COVID-19 on German newsrooms, measured by authorship outputs and cooperations.

Paper 4 - "More inclusive and on wider sources: A Comparative Analysis of Data and Political Journalists on Twitter in Germany" conducts a comparative analysis between sexes, and data and political journalists on Twitter in Germany, examining their practices and interactions — and aims to illustrate the use of Computational Social Science methods to understand social groups better.

## 1.3 Significance and Contributions

This dissertation aims to significantly contribute to journalism studies, Computational Social Science, and crisis communication by addressing these research objectives through a combination of four papers. It provides a nuanced understanding of how the COVID-19 pandemic has shaped data journalism practices, offers insights into user perceptions of visual forecasts, explores newsroom transformation, and analyzes the practices of journalists on social media. It seeks to increase practitioners' and researchers' understanding of newsroom innovation processes. By employing Computational Social Science methods, this dissertation further aims to prove the value of their application within communication studies.

# Chapter 2

# Literature Review

The following literature review sections aim to create a common basis for understanding the evolution of data journalism, its innovative potential for journalism, and the role of technology — starting with a definition of the object of the study.

## 2.1 Definitions and Structure of Data Journalism

### 2.1.1 Definitions and Characteristics

"Comment is free", wrote The Guardian's editor CP Scott in 1921, "but facts are sacred" (S. Rogers, 2013a). These words were laid down as an editorial manifest of the British The Guardian newspaper.

While facts were mostly conveyed in textual form for a long time, the early Twenty-first century was accompanied by a new way of publishing those "sacred facts": *data-driven journalism*, often shortened to *data journalism*.

One of its earliest descriptions was written in 2006 by Adrian Holovaty, an American journalist and web developer (Holovaty, 2006). He argued that for many forms of news and information, a typical journalistic story might not be the best-suited form of storytelling. News that was based on quantifiable information could get repurposed for further use. Quantifiable information that journalists would already collect, although not in a structured way, could be saved in a database, analyzed, and visualized to the public. Holovaty gave examples, like sport, crime (Holovaty, 2005; Holovaty, 2008), or fire statistics, of classic examples of what we would clearly regard as some form of data journalism today.

But Holovaty already gave a glimpse at ideas that were even more innovative at the time: analyzing obituaries or wedding announcements, agendas of parliaments or political advertisements in electoral races, analyzing the approximate age, birthplace, or formal charges of Guantanamo inmates, or researching the claims of politicians, which led to the creation of the fact-checking website PolitiFact (Waite, 2007).

Holovaty's blog post led to some discussion about whether or not the presentation of structured data had the right to be called journalism. In 2009, Holovaty commented on those arguments with a: "Who cares?" (Holovaty, 2009). In this post, he also labeled the combination of structured data and journalism as “data journalism”. Although this was framed in a question, it is one of the initial appearances of the term.

Over the years, data journalism has been defined in various ways, some very broad, some aiming to increase specificity.

While some journalists define data journalism as journalism using datasets, others focus more on the storytelling and presentation aspects (Hermida and M. L. Young, 2019, p. 35-36). For instance, Cushion et al. defined data journalism as the “use of data by journalists” (Cushion, J. Lewis, and Callaghan, 2016, p. 1200), very close to the description Holovaty gave in 2006. Anderton-Yang et al. (2012) referred to journalism professor Philip Meyer, who pointed out the value of processing information in an age where information is abundant. He argued that data should be analyzed to structure it and present the results in their importance and relevance to the audience. This idea was further emphasized by Antonopoulos and Karyotakis (2020), who described data journalism as a way of enhancing reporting with the use and examination of statistics to increase insights into a news story or to uncover previously hidden aspects in forms like interactive online content, or through data visualization tools that can create tables, graphs, or maps.

Berret and Phillips (2016, p. 9) also included using computational methods, algorithms, machine learning, or other emerging technologies in the journalistic toolbox as a data journalism feature. Furthermore, “The ability to use, understand, and critique data amounts to a crucial literacy that may be applied in nearly every area of journalistic practice.”

Throughout this dissertation, both perspectives are used to derive various working definitions to research data journalism. Some focus more on the output when analyzing visualizations, and some emphasize the data-driven workflow that data journalists employ in line with the primary definition.

While these definitions bring along various workflow changes, new forms of journalistic sources, and different ways of showing stories more than telling them, there are limited differences to ”traditional journalism” from a foundational perspective. Some parts of the data journalistic community use the long-standing prototype of the “newshound approach,” which describes the traditional investigative journalist serving as a watchdog over democracy, as a role model to align on (Hermida and M. L. Young, 2019, p. 24).

Rather than talking mostly to contacts on the phone or in interviews, they leverage automation to dig through vast amounts of data or documents to uncover misbehavior. This does not mean that data journalists do not seek clarification or further information from human sources to add contexts and make the data more understandable — but the initial and underlying information for the story comes from data (Kalatzi, Bratsas, and Veglis, 2018, p.37), promoting the image of a data-literate, computational-savvy journalist, who can make use from complex streams of information (Berret and Phillips, 2016, p. 5), while still generally adhering to traditional news values and formats such as objectivity (Tandoc and Oh, 2017, p. 997). Working with data is by some regarded as a core skill, comparable to interviewing or the ability to interpret charts (visual literacy) (Burns and Matthews, 2018, p. 94). Burns and Matthews (2018) described data as a 'live' source that can be questioned and might provide answers but is not - like human sources - free of error and value, as data is collected and operationalized by humans in the first place, which requires the same scrutiny as journalists would show towards human sources.

In short, journalism remains key, not data analysis (S. Rogers, 2011).

However, increasing attention to the topic has brought different characteristics that can help to arrive at a taxonomy of data journalism. Davies (2018, p. 108) categorized data journalism into four categories:

- **Data visualization**: ways of telling stories using maps, charts, or timelines — while the visualization remains the center of the story, not just an accompanying image to a long text.
- **Quantitative literacy**: the knowledge of statistical methods, descriptive and predictive statistics, or significance or error margins.
- **Data access**: methods and knowledge about how to find and potentially scrape information that might not be available as open access, machine-readable datasets.
- **Coding and data extraction**: skills in programming languages or easy-to-access tools to utilize datasets of all sizes.

This taxonomy already reveals the broad range of skills that data journalism requires. It is, therefore, not surprising that the field is a divergence between separate specializations and practices like statistical analysis, computer science, visualization, web design, and reporting (Coddington, 2015, p. 334). While it is closely aligned with the open-data movement, it does not specifically require open data.

### 2.1.2 Actors and Workflows

The global development of technology drives the evolution of data journalism. In addition, data journalism may provide some answers to emerging crises of journalism (Hermida and M. L. Young, 2019, p. 53), for instance, for trust issues (Toff et al., 2020). Data journalists understand themselves as an international, "independent community of practitioners" (Morini, Dörk, and Appelgren, 2022, p. 15) that shares knowledge and data (Appelgren, 2016; Porlezza and Splendore, 2019) - which is to some extent unusual for journalists that commonly keep their sources and approaches more secretive towards competitors.

One central function for data in journalism is the claim of empirical evidence (Engebretsen, Kennedy, and Weber, 2018, p. 9), driven through a clear story. Whether that is achieved by using the written word, tables or raw numbers, or visualization is something novel in journalism that is explained by data journalists' increased reflexity, which describes the intersection of developing new ways of telling stories while keeping existing journalistic values in place (Borges-Rey, 2017).

This image is, however, not complete without regard to the shifts of actors to an increased open source or open data perspective. Data journalism brought increased cooperation with non-journalistic actors, mainly from civic tech that complemented the journalistic work of holding the powerful accountable and working in the public interest by facilitating the data work by creating access to information and enabling ways to analyze and visualize data (Baack, 2018).

These technology-oriented strangers in journalism can be categorized as explicit and implicit interlopers and intralopers (Belair-Gagnon and Holton, 2018). Bloggers or citizen journalists are regarded as *Explicit Interlopers* who "frequently and overtly challenge journalistic norms, calling for improved practices (e.g., more transparency through linking in social media spaces; fact-checking that includes public input)" (Belair-Gagnon and Holton, 2018, p. 73).

Implicit Interlopers are not expected to be as critical as explicit interlopers, as they are mostly already part of news organizations and might adapt to changed requirements for news gathering. Examples include programmers or web analytic professionals in news organizations (Tandoc and Thomas, 2015).

Intralopers are "non-traditional journalism actors working from within news organizations without journalism-oriented titles, they may be trained in journalism or be well versed in the craft of the profession." (Belair-Gagnon and Holton, 2018, p. 75) They work inside-out, for instance, by deploying new tools to assist journalists inside media organizations. With the dawn of the big data age, these new actors helped to bring the required data and technology knowledge into the newsrooms to empower them to perform data journalistic tasks.

This leads to observations about data journalism in what Borges-Rey (2017) describes as two directions: the newshound and the techie approach, indicating a tendency to describe and negotiate data journalism in an interplay between traditional journalistic values and computational processes, leading to an individual professional culture that aims to mitigate individual lack of skills (De-Lima-Santos, 2022).

The skill set required for data journalism has also been studied: Hermida and M. L. Young (2019, p. 71) propose a three-layer model to describe data journalistic work derived from the Science and Technology Studies literature. The first layer is objects of technology, which are data sets, visuals, software, and platforms for data journalists. A second layer describes the processes used, like web-scraping, data analysis, and visualizations. The third layer is the know-how to wrangle and clean data, interpret and combine it into a story.

Others described the skill set as having "high levels of data collection, analysis, and multimedia presentation skills" (Tandoc and Oh, 2017, p. 1003).

Bradshaw (2011) described the workflow as an inverted pyramid, where a lot of data comes in at the beginning and gets cleaned, contextualized, and combined until it ends in a communication to the readers.

Kandel et al. (2012) segmented data workers, as we could also regard data journalism and coherent with the variety of non-journalistic actors in data journalism, into three distinct categories — not based on the work steps taken, but on the tools they used: *application users*, who use Microsoft's Excel, spreadsheets or other click-based applications, *scripters*, who use software packages for data analysis (R or Matlab), and *hackers* who are fluent in the same analysis packages as scripters but also proficient in scripting languages (like Python or Perl) and data processing languages (like SQL).

Wrapping up this view into the actors and workflows in data journalism, we see that the required mix of skills — from crunching data to crafting stories visually appealing — has reshaped journalism. In the upcoming section, we will dive into how these practices are shaped by the past, exploring the evolution of data journalism from its early days to the present and highlighting how these changes have been absorbed in the industry.

## 2.2 Evolution of Data Journalism

To embed the COVID-19-related evolution of data journalism into perspective, it is worth giving a short rundown on the branch's historical root and early-stage developments.

### 2.2.1 Historical Perspective and Key Milestones

The idea of conveying quantitative information in non-textual form is not new. One of the first examples is a table in the Manchester Guardian (today known as The Guardian) of May 5th, 1821 (S. Rogers, 2011). It showed a list of schools in Manchester and Salford, how many pupils attended each, and their average annual spending. This list shows that public numbers about how many pupils received free schooling were much lower than the actual figures. This is a classic example of how journalists use numbers to compare and enable the audience to place themselves within the data if their local school was shown.

Another early example is John Snow's cholera map. In 1854, he used this data visualization technique to demonstrate that cholera is not spread over the air by mapping an outbreak in London's Soho in a chart. A bar chart representing a death was mapped on the deceased's house. This visualization showed a clear pattern: a cluster of deaths around a polluted pump in Broad Street (S. Rogers, 2013b). Four years later, the nurse Florence Nightingale created a rose chart (today, sometimes even called the Nightingale chart) to communicate the avoidable deaths of British soldiers during the Crimean War in 1853-1856 (S. Rogers, 2010).

However, today's data journalism was driven mostly by the advent of the computer. Its roots lie in the emergence of Computer-assisted reporting (CAR), which can be traced back to 1952 when CBS News used an early computer to predict the results of the presidential elections on election night (Cox, 2000). Philip Meyer, a pioneer in the field, authored a book on precision journalism, advocating for using computer-assisted social science methods like databases or surveys in the early 1970s (Meyer, 1973). In a later edition, he even called for journalists to be “database managers” (Meyer, 2002). The main idea of CAR was to use computational ways of information gathering and sense-making for investigative efforts. Data, however, was not central to the story, which was still human-centered; it was just necessary to bring up a story in the first place.

In data journalism and its closely related computational journalism, data becomes a source itself that could lead to a whole story. That opens the way for non-journalists to be included, whether opening up the whole dataset for analysis or collecting crowd-sourced information, leading to an active participation of the public, in comparison to a passive audience for CAR (Coddington, 2015, p. 338–343).

To distinguish between the three, Coddington (2015, p. 337) has developed distinctive features: Computer-assisted reporting is rooted in social science methods and traditional, human-story-centered investigative journalism. Computational journalism is focused on applying the abstract, computational processes of abstraction and automation to data. Data journalism is depicted by its openness to participation for non-journalists and cross-field hybridity of skills. While these distinctions are not shared entirely throughout the literature (Bounegru, 2012), a consensus emerged that data journalism is at least derived from computer-assisted reporting and might represent a form of update to the area with more data available that allows deeper focus on data as a source than had the journalists in the past.

In contrast to existing forms of journalism, data journalism blends data science, computer science, and journalism. This leads to data journalists having diverse backgrounds. Earlier work has shown that these programmer-journalists refocus from classical investigative reporting to a more inclusive approach that tries to advocate for changes through influence on public opinion using open source and open data (Parasie and Dagiral, 2012).

The beginning of data journalism as a distinct discipline can be placed around 2009, escorted by the establishment of governmental open data portals like data.gov, the popularity of open source, and the standardization of HTML5, which made additional plug-ins for videos and embedded charts unnecessary (Bravo and Tellería, 2020; S. Rogers, 2021). Initially, the players in data journalism were mostly “large, well-resourced news organizations that have made an institutional investment in this area” (Hermida and M. L. Young, 2019, p. 23), or driven forward by passionate individuals (De Maeyer et al., 2015). Media companies saw data journalism as a means to create more in-depth projects based on data analysis and ultimately to strengthen journalists’ role as gatekeepers (Appelgren and Nygren, 2014) and increase trust (Lorenz, 2012). Journalists saw the value in data journalism by being able to support claims with evidence and in its ability to present facts in a visual, non-textual form (Green-Barber, 2021).

According to various studies, the implementation of data journalism has been hindered by major barriers such as the lack of time and resources. These problems were identified by Fink and C. W. Anderson (2014) and are still prevalent today as stated in recent studies (B. R. Heravi and Lorenz, 2020; Bisiani et al., 2023). Furthermore, data availability remains an issue in sub-Saharan countries, which has slowed down the development of data journalism, as highlighted by Chiumbu and Munoriyarwa (2023).

Primarily, scholarly interest focused on Western democracies (Karlsen and Stavelin, 2013; Appelgren and Nygren, 2014; De Maeyer et al., 2015), often English-speaking (Fink and C. W. Anderson, 2014; Carl W. Anderson, 2018; Hermida and M. L. Young, 2019) where data journalism has its origins.

A new wave of research (Appelgren, Lindén, and Dalen, 2019) also included countries in Africa (Akinfemisoye-Adejare, 2019; Cheruiyot, Baack, and Ferrer-Conill, 2019; Chiumbu and Munoriyarwa, 2023), the Arab world (Mutsvairo and Bebawi, 2022; N. P. Lewis and Nashmi, 2019), Italy (Porlezza and Splendore, 2019), Hong Kong (Zhang and Chen, 2020), and Southern America (Palomo, Teruel, and Blanco-Castilla, 2019).

This observation holds true for Germany as well, which has been selected as the primary country of focus in this dissertation due to its role as a Western media system with a grown data journalistic community that has not been studied as extensively as the English-speaking world, especially not using computational methods.

### 2.2.2 Data Journalism in Germany

Data journalism is established in Germany as in many European countries (Meier et al., 2022).

In an early attempt to quantify the profession in the spring of 2013, Weinacht and Spiller (2014) identified 35 individuals working as data journalists in Germany and were able to interview them.

By 2020, Beiler, Irmer, and Breda (2020) estimated that data journalism is well-established in three-quarters of media outlets. A finding that Weinacht and Spiller (2022) could confirm in 2022 in their repeated study. They found data journalists working in all media sectors, not just national media, with the share of permanent employees rising from one to two-thirds. While the total number of data journalists remained low, the share of females increased from 3 out of 44 to 13 out of 57 respondents in the survey.

Weinacht and Spiller (2022) were able to identify three groups of German data journalists: analytical controllers — relatively many, slower researchers who aim to control the powerful, quick communicators with a focus on speedy outputs, and neutral informants who laid focus on data-driven explanations.

In another study, Haim (2022) increased the understanding of tools and role perceptions of data journalists in Germany. He extracted data journalists' names from publications, used Twitter and a mailing group of data journalists, and contacted media to identify 187 data journalists, of which 102 participated. A similar approach was used in this dissertation, which arrived at a comparable number. Haim found that data journalists' main tools were Excel or similar, the visualization tools Datawrapper, HTML and CSS, and the programming language R.

When comparing the role perceptions of data and non-data journalists, he found that the former has a much higher vision of being an adversary to the government and business, providing information to let people build their own views and to analyze and set the political agenda. They do not see themselves as entertainment providers, attract large readerships, promote cultural diversity, or let people express their views. The latter aligns with findings from Weinacht and Spiller (2014). Interestingly, data journalists consider it much more justified to pay people for data/information, use confidential business, government, or personal documents without authorization, exert pressure on unwilling informants, or publish stories with unverified content than their non-data peers.

Haim (2022) also found that data journalists mostly regard their profession as more fact-based and transparent than other forms, with a close majority agreeing that data journalism can help re-establish trust in journalism. Within their newsrooms, data journalists have reached a high level of institutionalization. Nineteen percent of respondents were part of a dedicated data team, 13 percent of an investigative team. However, only a third of the respondents spend more than 80 percent of his or her time in data journalism. Another third can only do data journalism in less than 20 percent of their time. It is, therefore, not surprising that 30 percent of the respondents considered their newsroom to be understaffed with data journalists.

Within their work, data journalists feel very autonomous and do not experience a lot of influence by editorial policy, advertising, or audience-research considerations, except a stated influence on the availability of resources and less influence from peers in the staff (Haim, 2022).

### 2.2.3 Perspectives on the Future

Zooming out of Germany, the future of data journalism seemed brighter than for many other areas in the newsroom — even before COVID-19.

Dissemination of the profession was described by de-Lima-Santos and Mesquita (2021) using two strategies: collaborations between media companies to promote the availability of open-data and data-driven storytelling and making data stories appealing to broad audiences, which increased visibility and created economic arguments in favor of data journalism.

Hermida and M. L. Young (2019, p. 63) argued that “data journalists take advantage of instability, contradictions and crises in the field with respect to credibility in journalism, newer competitors and a shifting relationship with the audience to advance their own professional interests, identity and community of practice.” Through this process, they observed data journalists gaining power and strategically reframing their roles and identities to respond to power and value changes in the media landscape (Hermida and M. L. Young, 2019, p. 53–54).

## 2.3 COVID-19 and Journalism

An example of a huge change in the media landscape was the appearance of the Coronavirus (COVID-19) in early 2020, which brought changes to journalism. A general summary of the influence of the virus on the media industry is given below, before focusing on changes in data journalism that have been found so far.

### 2.3.1 Impact of the Pandemic on the Media Industry

COVID-19 served as an accelerator for ongoing changes in journalism: the decline of print and other forms of traditional media, the rise of alternative news channels, restructured processes and altered skill requirements for journalists, changes in audience and their expectations and new approaches to journalism (Quandt and Wahl-Jorgensen, 2022).

Journalists attempted to be proactive and innovative during the uncertainty caused by COVID-19, although they were under massive economic constraints. M. F. Perreault and G. P. Perreault (2021) found in a discourse analysis that journalists placed themselves and their profession as a public service in danger due to tension to access sources while possibly contracting the virus, as well as through economic pressure and increased audience criticism on the selection of sources.

Local print news offered the possibility to break down international or national stories for less-connected audiences and provide an essential public service for the communities by appreciating the social powers of communities, coordinating political responses, and offering a feeling of cohesiveness (Hess and Waller, 2020).

José A. García-Avilés et al. (2022) observed most innovations during COVID-19 in the areas of product (like data visualizations or fact-checking), distribution (newsletters or podcasts), and commercialization (subscriptions and membership models). COVID-19, in particular, increased the numbers of visual journalism due to the availability of huge amounts of data about the virus and its consequences and an uncertain situation, which led the audience to focus on hard numbers. Infographics became an often-used, favorably-perceived way of conveying information about the pandemic (S. H. Lee et al., 2022). This has also led to criticism due to the bombardment with visualizations of infections and death tolls (José A. García-Avilés et al., 2022). A text analysis by Krawczyk et al. (2021) showed that around a quarter of all front-page news between January and October 2020 contained COVID-19 reporting, which is interpreted as indicative of an information overload.

### 2.3.2 Data Journalism Aids Changing Information Needs by COVID-19

Reporting on health and science issues was common in data journalism well before COVID-19, as these were data- and science-related topics (Loosen, Reimer, and De Silva-Schmidt, 2017; Cushion, J. Lewis, Sambrook, et al., 2016). However, the numbers in scientific literature exploded in 2020: researchers are estimated to have published up to 200,000 COVID-19-related papers (Else, 2020). In newsrooms, departments concentrating on science reporting became more relevant, cooperating with data journalism teams (José A. García-Avilés et al., 2022). This qualitative observation led to investigating authorship patterns of data and science journalists in several German newsrooms, reported in Paper 3.

In the early phases of the pandemic in the spring of 2020, data journalism tried to explain the pandemic as it unfolded, parallel to epidemiological models and policy responses. This also raised awareness of journalistic forecasting by employing and explaining different predictive models, which did not explain the past but made claims about a possible future. Pentzold, D. J. Fechner, and Zuber (2021) found three modes of data journalistic visualization and explanation in the early stages of the pandemic:

- comparison of different predictive models,
- comparison between a model and the way the pandemic unfolded,
- comparison of trajectories in relation to the measures taken.

They described the role of data journalists in this regard as those of knowledge brokers, fulfilling functions of awareness, accessibility, and fostering engagement for the general audience and policymakers. In single cases, they also linked actors and mobilized people to act.

The rise of computational models to explain possible future trajectories is a recent addition to the data journalistic toolbox (Pentzold and D. Fechner, 2019; Pentzold and D. Fechner, 2021). The audience's perceptions of this way of journalistic reporting of possibilities were poorly understood. Paper 1 explores the perceptions of news users towards journalistic predictions with the use case of elections that took place in 2021, after the pandemic had shaped the use of predictive reporting.

In an analysis of the Sigma Awards for data journalism after COVID-19, Auväärt (2023) found an immense increase in datafication of news journalism, driving reporters to analyze and visualize the statistical effects of COVID-19 on various societal sectors in quick statistical overview articles to in-depth features. Most projects centered on national aspects of the pandemic, highlighting the importance of geographical proximity in reporting. Interestingly, the COVID-19-related data sources were mostly limited to governmental data on around half of the projects, which is explained by the novelty of the virus and the official data count taking place, and data collection by the newsrooms on the other half, where official data was not regarded as being suitable or available for publication. Problems that came up were the time-consuming nature of data projects, difficulties in obtaining and handling data, the sensitivity of the topic, and technical challenges related to visualizing and updating data.

While there was a strong focus on the enumeration of the impact of the virus through global data collection and visualization of Coronavirus numbers, this also led to increased criticism of the sources of the pandemic reporting. On the one hand, a multi-national analysis of social media posts showed that the State was a main source for global pandemic reporting — even more intense in Latin American countries compared to the US, the UK, and Germany (Mellado et al., 2021), which points to a difference in the state's role in COVID-19 news. This small set of sources during COVID-19 led to a cyclical pattern of journalism-sources relationship "in which the source gains authoritative status through being used as a source while the news gains authority through having used this authoritative source" (Carlson, 2009, p. 530).

On the other hand, some areas were poorly prepared to handle COVID-19 data reporting: "Many communities at the margins, including many areas of the Global South, are virtually absent from this number-based narration of the pandemic", argued Milan and Treré (2020). This hinders the detection and prevention of COVID-19 cases and slows efforts to offer relief.

The observation that "letting the data speak for itself" (Tandoc and Oh, 2017, p. 1003) is not a new one for data journalism. Data is described as being subjective and created in the interest of its sponsor, which leads to an underrepresentation of the marginalized (Jeppesen, 2023).

To stand out, some media outlets performed their own data collection efforts, like the *Financial Times* for global Covid data, *The Econonomist* for excess deaths, or Germany's *Zeit* for more recent data on a more granular level than the one published by the main government body (Desai et al., 2021).

Not just regarding their output, data journalism was impacted by the COVID-19 pandemic. Also, professional decisions were made due to the coronavirus. A quarter of respondents to the Data Journalism Survey in 2021 stated that they had entered data journalism because of the pandemic. This effect was especially strong in the Global South (Bisiani et al., 2023). Concerning the field, the most commonly held view in the survey is that COVID-19 reporting has strengthened data journalism as a field (46%), it has increased audience data literacy (43%), and at least 28% of respondents also believe that access to data has improved.

However, the pandemic was also attributed to tightening existing media struggles: one in three respondents mentioned a decrease in resources due to the pandemic, 36% have felt an increase in time pressure, and 44% an increase in workload.

## 2.4 Computational Communication Science

The chapters above have detailed data journalism as the investigation object of this dissertation. The next chapter describes the foundation and expansion of computational methods for analyzing the impact of the COVID-19 pandemic on data-driven news reporting, described in the terms Computational Social, or Communication, Science (CSS).

Similarly to the advent of computer-assisted reporting and later data journalism described above, computational methods became increasingly common in social sciences (Lazer, Pentland, et al., 2009; Boyd and Crawford, 2012; Alvarez, 2016). Influenced by at least three drivers (Atteveldt and T.-Q. Peng, 2018): the vast amount of digital data that has become available — from newspaper archives to social media posts or digitalized archives, the availability of tools for collecting, handling, and analyzing this information, and the decrease in computation costs for IT infrastructure.

While definitions vary, a comprehensive one for Computational Social Science is given by Shah, Cappella, and Neuman (2015): "(1) the use of large, complex datasets, often—though not always—measured in terabytes or petabytes; (2) the frequent involvement of "naturally occurring" social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability to social theory in a variety of domains from the study of mass opinion to public health, from examinations of political events to social movements."

Hofman et al. (2021, p. 182–184) offer four distinctions on the research of digital data and computational methods:

- "Descriptive modeling" that aims to measure and describe relationships between categories of interest using methods like surveys, statistical analysis, topic modeling, and community detection.
- "explanatory modeling" describes methods to identify and estimate causal effects on an outcome used in sociology, political science, or psychology.
- "predictive modeling" tries to predict an outcome variable while not necessarily aiming to estimate all causal effects. This focuses on 'out-of-sample' predictions split into train and test data, like time series modeling and extensively supervised machine learning.
- "integrative modeling" attempts to predict 'out-of-distribution' data, which might change naturally or because of some intervention that causal connections can describe.

The main difference between the classical and Computational Social Science research processes lies in the methodological selection, data gathering, and data analysis, which is linear in the classical process but iterative and bi-directional in CSS. This is driven by the novelty of datasets, which require adaptions to the method and indicators in a continuous cycle of scientific analysis (Haim, 2023).

Data for Computational Social Science can be collected using various approaches or created by the researchers themselves. Furthermore, they can be retrieved using official interfaces. We will look at these three, starting with the last one: data collection via official interfaces, particularly from Social Media, as this is most central for this thesis.

### 2.4.1 Social Media Data

Since the advent of social media networks, social data, collected from users' interactions in social networks, has been a central foundation of Computational Social Science research.

Social data collected in a globalized society can offer novel insights into the social world by providing individual-level details of interactions available for a large population. But this also requires measurement changes to account for the problems that arise (Lazer, Hargittai, et al., 2021): the data might not be an optimal operationalization for the theory at hand, access to the data might be legally restricted, or the data might be limited in its temporal, spatial, or ethnical, structural integrity.

One shining example of this — and a showcase of its issues — is dealing with data retrieved from social media platforms like Twitter (now X) or Facebook as measurements for public sentiment or communication.

Data has grown from a by-product as users' payment for services of social networks to an essential good of social media platforms (Gillespie, 2010; Puschmann and Burgess, 2013; Helmond, 2015; Iyer and Getchell, 2018; Dijck, 2020), with offer access to their collected data via application programming interfaces (APIs), which third-party companies could use to increase their understanding of users and potential customers, but also for scientists in understanding social groups.

Over a period of several years, there was an increased research interest in topics like abuse, hate speech, trolling, and disinformation campaigns based on the data that was available from APIs (for a non-comprehensive but large list of those publications, see Bechmann (2018)). The data was also combined with other kinds of data collection to hold the platforms accountable for their offerings and actions — for instance, filter options on race categories for marketers offered by Facebook (Angwin and Parris, 2016).

In 2018, the Cambridge Analytics data scandal highlighted the possibility of generating many users' personal data without their consent for marketing and political purposes using Facebook's Open Graph API platform (Cadwalladr, 2018; Albright, 2018). This led to increased limitations in API access for several social media platforms. Leading to researchers claiming an 'APIcalypse', which hindered scientific scrutiny into the platforms (A. Bruns, 2019). Others described their hopes to end the 'Wild West of social media research', leading to negotiations for how social media data might be used in the future (Puschmann, 2019).

After Facebook shut down its APIs in April 2018, access and replication to Facebook data was largely impossible. Later, Facebook shared a vast dataset with a selection of independent scientists who had to apply for access — Social Science One. However, in 2021, it was found to contain flawed data, which raised doubts about its data quality (Timberg, 2021).

Insights into the platforms have worsened since the scaleback of the APIs (Rieder and Hofmann, 2020). While some data protection arguments brought up by the platforms might be justified, they might also be exaggerated to prevent public scrutiny of social media platforms' internal policies (Ausloos and Veale, 2021).

The remaining central network for Computational Social Science was Twitter, which still offered access to its tweets, mentions and hashtags, and metrics like likes or retweets.

It also allowed the observation of communication behavior between groups of users. Its data has been used vastly in so many different settings that it has been criticized for being leveraged by the "the principle of the drunkard's search" (Kaplan, 1964), sometimes also called the "streetlight effect". Twitter data was so easily available that researchers tried to use it for all kinds of studies, although it may not have been the ideal dataset.

For instance, Twitter data was often used to generate insights about populations of voters — although only a small and skewed subset of this population uses the platform (Haller, 2019). For Germany, only around a tenth of German internet users used Twitter at all (Newman, Fletcher, Schulz, et al., 2021; Newman, Fletcher, Robertson, et al., 2022).

Certain platforms also attract different social groups of users. Hargittai (2018) showed that users with a higher sociodemographic status and higher technical skills tended to be members of multiple social networking platforms, which might leverage their perspective when these data were used as the foundation for population-wide decisions.

With X, the former Twitter, announcing to shut down all free tiers of its APIs in early 2023 (@XDevelopers, 2023), the sunset of social network-driven social research, especially powered by tweets, seemed to have been reached.

Marres and Gerlitz (2016) have tried to solve this methodological problem by describing an approach called 'interface methods': "We explicitly recognize that social media data come in specific forms and formats and are informed by distinct use practices – which may steer social inquiry into specific directions, here that of proportional forms of analysis. On the other hand, adopting an interface methods approach means that we do not necessarily need to go along with these media effects: we can deploy our methodology to work against this type of bias, for example, by privileging the formation of new relations in our analysis" (Marres and Gerlitz, 2016, p. 40). Instead of adopting a fully new methodology to deal with the constraints embedded in the social media data, they argue to combine the restrictions from the data and the rationals of existing scientific methods into a new, grounded approach. It acknowledges that the data itself is affected by the environment in which it was collected, and it includes this in the analysis — a thought that is not new, thinking about previous data collection methods with their own limitations.

### 2.4.2 Non-Social Media Data

Since the descent of API-based methods has been ongoing for some time, escorted by increased restrictions of the "Post-API Age" (Freelon, 2018) increased focus was laid on the importance of already existing methods like web-scraping, the automated extraction of data from websites using text detection or parsing the document structure of the webpage (Document Object Model, DOM). It is even applicable when an API is unavailable, but it often violates the platform's terms of service. This is a second way of collecting data for computational methods in social science, especially useful for digital data not offered via an official channel.

Others called for a return to “digital fieldwork” (Venturini and R. Rogers, 2019) to search for and employ other, more direct data collection methods, possibly even with users’ direct consent or interaction, which might not be influenced by social media networks’ algorithms or business interests. Methods like this have been employed already, for instance, by asking users to donate their usage data (Araujo et al., 2022) or using data that tracked user behavior on websites (web tracking data) (Christner et al., 2021). Others have used agent-based tests to show personalization of prices or search engine results by modifying the characteristics of the perceived users of the online sites (Hannak et al., 2013; Hupperich et al., 2018; Haim, 2020).

Computational Social Science is not limited to online conduct. The methods can deal with all kinds of digital data and allow large-scale insights into real behavior within their actual social environments. Mas and Moretti (2009) have leveraged supermarket brand’s data to find productivity spillovers from very productive employees within a shift using scanner-level information. Others analyzed the behavior of New York City’s taxi drivers in data coming from electronic meters — and could find evidence for neoclassical economic theory in showing that taxi drivers tend to work longer hours on days when they made more money (Farber, 2015). Transactional-level financial data was used to explain household spending behavior at the onset of the COVID-19 pandemic, leading to an initially increased spending to stockpile goods at home before heavily decreasing once stay-at-home orders became in place — with money still being spent on groceries and food deliveries (Baker et al., 2020). These examples, however, show a huge difference from previous methods, which is very common in Computational Social Science. The data being used was not created for the research interest but for other purposes — and is being reused in research. This might lead to constraints like those already discussed in the context of social data from social media networks. Therefore, Computational Social Science must examine the possible implications and restrictions of datasets in detail or search for alternative ways of data provenance.

### 2.4.3 Self-Generation of Data

A third way to retrieve data for Computational Social Science methods is one’s own data generation. While not at the center of the field, methods like bootstrapping, Monte Carlo simulations, or agent-based modeling can help develop a deeper understanding of processes where data is missing and inaccessible or agents’ behavior needs to be modeled to be observable. A necessary prerequisite is a deep knowledge of the processes of the problem at hand to define the behavior and probabilities in the simulations.

These methods gained increased attention at the beginning of the COVID-19 pandemic when they were used to predict the diffusion of the virus (Xie, 2020) or the potential effects of policies on its spread (Kerr et al., 2021).

Monte Carlo simulations are founded on the law of large numbers and the central limit theorem. Based on historical averages and deviations, the computer script simulates a large amount of data, which can then be analyzed using statistical methods to gain insights. Communication Science has, for instance, used this method to gain insights into Intercoder reliability (Geiß, 2021).

Bootstrapping is a method to deal with data where essential knowledge about the population is missing or the distribution of an attribute is not normally distributed, which prevents common statistical methods from describing uncertainties of samples. Bootstrapping is a resampling technique used for estimating the sampling distribution of a statistic by repeatedly sampling, with replacement, from the observed data. It allows for robust statistical inference without making strong parametric assumptions that may not be satisfiable in the computational data (Mooney, 1996; Scharkow, 2017).

Agent-based modeling uses computer programs to simulate the individual or collective behavior of artificial agents with distinct attributes interacting with other agents in a controlled environment based on certain behavioral rules (Macy and Willer, 2002). Agent-based modeling is commonly used as a replacement for randomized experimental designs, where those may not be feasible (Gilbert et al., 2018). In Communication Science, the method was used to identify drivers of news waves (Waldherr, 2014), model commenters' behavior on media forums (Chmiel et al., 2011), or asses short-term media effects (Wettstein, 2020).

### 2.4.4 Summary

The methods described above share three overarching themes influencing and limiting their research: control, resources, and ethical considerations (Possler, S. Bruns, and Niemann-Lenz, 2019):

Researchers of Computational Social Science have to use self-collection methods like web scraping that leave them in control of data access. Using data from third parties, legal or technical barriers might restrict data access, making it harder or impossible to address data quality issues or be transparent in sharing the data, which is required for the reproducibility of their results.

The technical skills required for web scraping, the financial resources to buy access to data, or the personal network required to access secondary data analysis depend on the scholar's resources, which further limits Computational Social Science.

A third constraint concerns the ethical considerations that researchers aim to adhere to. On the one hand, datasets obtained from third parties, especially companies, are often limited by non-disclosure agreements or other contractual restrictions. Those may hinder scientific best practices, such as data sharing for reproducing results. On the other hand, self-employed methods like web scraping are often prohibited by websites. Therefore, collecting data using these methods comes with a certain legal risk and does not seek users' consent, as would be the case when using official data from social networks.

To summarize, the limitations faced by Computational Social Science researchers in data collection can be broadly categorized into issues related to control, resources, and ethical considerations. These constraints affect not only the quality and accessibility of data but also pose challenges to research's reproducibility and ethical integrity, setting the stage for discussions on methodological improvements and ethical frameworks.

Given the aforementioned constraints on data collection in Computational Social Science, this thesis's methodological approach is particularly designed to navigate these challenges. While a significant portion of the data is sourced from Twitter, the study acknowledges the platform's limitations and incorporates measures to address them. Furthermore, the research does not rely solely on Twitter data but integrates multiple data sources and methods, aiming to enhance the robustness of the findings and ethical considerations.

# Chapter 3

# Methods & Data

This thesis presents research results from several papers based on different datasets and methods, which will be described and explained in the following.

## 3.1 Methods

While some of the methods of Computational Communication and Social Science have been briefly described above, the sections below will focus specifically on the scientific methods used in the papers constituting this thesis.

The initial method is not grounded in Computer Science but has a longstanding tradition in Social Sciences — it is, however, used here to investigate the influence of the results of computational methods.

### 3.1.1 Survey And Qualitative Thematic Analysis

A survey was used to understand the perceptions of the work of predictive journalism on the German federal elections in 2021. As forward-looking journalism based on predictive models is still a newer development, the literature lacked research on the possible effects this kind of journalism may have on the users. Using this survey containing tests of understandability and open-ended questions on cognition might be an initial attempt to extract reactions and investigate the alignment between the designers' intent and the audience's perception.

The survey was designed to cover three aspects of the underlying inquiry into users' perceptions. The first set of questions was factually driven to test whether respondents could arrive at a reading consistent with the uncertainty conveyed. They were shown visualizations from the predictions also used in the reporting and had to give answers based on these charts. A second set of questions dealt with the perceived utility of the visualizations for which answers were provided using a Likert scale and given the opportunity to enter additional explanations in an open text field. For the third set of questions, a similar setup was offered: with Likert scale-based answers to questions about whether the visualizations influenced how the users thought about the campaign and the option to enter additional explanations in a text box. The use of Likert scale-type answer options offered a quick and simple understanding of the users' perceptions while keeping the questionnaire concise and within an established form, while the open text fields were an admittance to the lack of knowledge around the topic, which would benefit from additional collection of users' opinions and assessments.

The questionnaire was accessible via the top banner of an explanatory article about the predictive model. It was also shared on Twitter by the authors between September 2nd and September 28th, 2021, with the charts contained in the questionnaire remaining fixed on the values of September 1st. Three hundred ninety-nine users opened the survey, 192 started it, and 134 finished (33.6%). Users were aware of the questionnaire's scientific nature and consented to data collection. Personally identifiable data was not collected.

While this method collects actual user responses within a non-lab-setting, the non-random selection process is probably driven by the individual interest in the topic, which might have skewed the answers. This was considered sufficient as the study's primary aim was to collect users' opinions.

To understand users' comprehension of the actual visualizations, the responses to closed-ended questions were analyzed on their ability to arrive at a reading in line with the uncertainty displayed. This was done by a simple count of answers that aligned with that reading.

To derive themes in the data, Qualitative Thematic Analysis was used (Braun and Clarke, 2006). We practiced iterative constant comparison of themes and analyst triangulation between co-authors to develop the themes that we reported. We first applied qualitative thematic analysis to the textual data and afterward decided to count the prevalence of specific themes to show better how widespread certain themes were observed.

In terms of validity, we argue that employing a variety of question types—factual questions, Likert scales, and open-ended questions to assess different facets of user perception and cognition enhances construct validity as it allows for a comprehensive measurement of the complex construct of user perception towards predictive journalism. The factual questions aim to ascertain whether respondents accurately interpret the visualizations, thereby serving as a direct measure of the construct under investigation. For content validity, the open-ended questions allow for a broader capture of subjective user opinions and reactions, which more structured question types may not fully encapsulate.

The use of Likert scales contributes to the survey's reliability. These scales are widely recognized for producing reliable results when measuring attitudes or perceptions. Additionally, the factual questions with specific, verifiable answers can be easily scored for consistency, further enhancing the survey's reliability. To confirm this, you could run a test-retest reliability assessment, where the same survey is administered to the same audience at two different times, and the responses are then compared.

Collecting data through unstructured text fields presents specific challenges, including the likelihood of respondents opting not to provide answers, the acquisition of brief and ambiguous textual snippets, and the absence of an avenue for iterative or clarifying follow-up queries. While this approach to data collection offers merits in terms of sample size, diversity, and susceptibility to unforeseen interpretations, subsequent research endeavors should contemplate incorporating supplementary techniques, such as semi-structured interviews, to reduce these limitations. Using both Likert scales and open-ended questions attempts to mitigate this by allowing for a range of responses, from quick, instinctual reactions to more considered, detailed answers. However, the risk of bias remains an important consideration, and future research should aim to address this limitation, possibly through methodological triangulation or by employing mechanisms to ensure a more randomized participant selection.

The survey methodology employed in the study is intricately linked to the dissertation's broader aim of investigating the influence of predictive journalism in the context of the COVID-19 pandemic, particularly within German media landscapes. The survey's mixed-method approach, featuring both quantitative and qualitative elements, complements the dissertation's overarching computational methodology. While the dissertation primarily employs computational techniques, including open-ended questions in the survey adds a qualitative layer, allowing for a more nuanced understanding of user perceptions and cognition. This methodological pluralism enriches the dissertation's computational focus by integrating qualitative insights, offering a comprehensive view of the complex interplay between predictive journalism and user perception. Given the nascent state of research on predictive journalism, the survey's open-ended questions also serve as an exploratory tool to unearth unanticipated user reactions and opinions, which could be valuable for future computational analyses.

### 3.1.2 Semi-automated Infographic Detection

Computer vision has become essential to Computational Social Science's toolbox over the last decade (Williams, Casas, and Wilkerson, 2020). It has been used to discover gender and age discrimination in German TV channels (Jürgens, Meltzer, and Scharkow, 2022), to analyze the content of politicians' Instagram posts (Y. Peng, 2020), or to predict politicians' election chances given their rate of smiling in pictures (Horiuchi, Komatsu, and Nakaya, 2012). Commonly used approaches are based on deep learning methods, especially Convolutional Neural Networks, that aim to summarize the raw input level features — pixels in an image — to a label of the whole image by concatenating and aggregating several input features to a document level (Lecun et al., 1998). However, training those networks still requires solid training data to detect differences. When trying to develop a classifier for journalistic infographics on images posted on Twitter, we realized that coming up with a clearly defined training set might be challenging, as infographics are created in various visual ways.

In our case, further described in Paper 2, the analysis's originator was a dataset of images that media companies from six countries posted on Twitter. News media predominantly employ Twitter as a channel for disseminating their journalism (Malik and Jürgen Pfeffer, 2016). This trend guided our decision to collect a comparable sample from Twitter instead of web scraping or other API-based approaches, which may suffer from limited comparability.

USA, UK, Germany, France, Italy, and India's only three available English-speaking newspapers. They were selected to be part of their country's largest national, general-audience media based on circulation.

The first five are examples of Western-democratic media systems. However, non-Western media have also adopted infographics as a journalistic form. English-speaking newspapers from India were also included to account for this but still prevent language barriers. As the pandemic affected all these countries, we expected similar patterns. The number of selected countries is restricted to limit the manual effort required for the semi-supervised approach outlined below and to countries whose languages the authors master to such an extent that they can evaluate the results.

Tweets were retrieved for a list of usernames covering the media outlined above using Twitter's API v2 (Juergen Pfeffer et al., 2023) with the command "`from:USERNAME has:images`". All available 2,205,025 tweets for this query were collected for the time period between January 1st, 2018, and July 31st, 2022, between August 15th and September 3rd, 2022. However, not all these tweets contained images, contrary to the expected return from the API call. In total, we could download 1,911,496 images for analysis, all in either JPEG or PNG format.

We employed a series of filters based on hand-engineered features like image type, colors, and edges to narrow down the pool of images likely to be infographics. These features were selected based on a labeled test set of 600 infographics and 1000 non-infographics, which was used to optimize the image characteristic parameters to maximize the amount of true positives.

The workflow consisted of the following steps:

- Test Set Creation: Initially, a labeled test set of 600 infographics and 1000 non-infographics was created to identify characteristic differences.
- Feature Identification and Optimization: Using the test set, we identified common characteristics like image type, colors, and edges and optimized the parameters accordingly.
- Automated Extraction: All images exhibiting at least one of the identified characteristics were automatically extracted.
- Text Recognition: The pytesseract[1] optical character recognition package for Python was used to filter out images lacking text.
- Manual Inspection: Finally, trained coders manually examined a subset of 2,500 images, focusing on cartographic or statistical charts based on numeric data.

[1] https://pypi.org/project/pytesseract/

Our approach demonstrated impressive metrics — an accuracy of 0.99, an F1-score of 0.65, a sensitivity of 0.578, and a specificity of 0.997 — indicating high reliability over a subset of 2,500 labeled and manually-checked images. Out of the 1,911,496 images we analyzed, we found 25,813 infographics using the semi-automatic approach.

This semi-automated approach was preferred over machine learning techniques due to the scarcity of publicly available labeled infographics, their diversity in visual terms, and the surprisingly robust performance of our hand-engineered features. However, the method faces limitations in text detection, which is crucial for differentiating between infographics and non-infographics. This limitation particularly showed up in approximately 10% of the infographics, where text detection failed. Future research could use more sophisticated text detection techniques to account for small, hardly readable, and non-standard texts, such as text in word art. Due to these text detection limitations, the dataset may not include infographics that lack text. However, we anticipate that such instances would be extremely rare, as most data visualizations generally incorporate some textual elements for context and interpretation.

In the context of a dissertation examining the role and prevalence of data journalism in the era of COVID-19, this dataset serves as a resource for dissecting how mainstream media organizations across different cultural and national backgrounds leverage infographics on Twitter. This dataset allows for a comparative analysis that can unveil patterns, similarities, and differences in the prevalence of data-driven visual storytelling techniques, such as infographics, by top media companies in six distinct countries. The dataset's time frame, from January 2018 to July 2022, enables a longitudinal view that could reveal how the pandemic has possibly catalyzed or influenced changes in these trends. Given that Twitter is a significant platform for news dissemination, focusing on tweets containing infographics from major media organizations ensures that the dataset captures a relevant and influential aspect of data journalism.

### 3.1.3 Text Analysis

Text is an important source for social science: Whether it's political speeches (Quinn et al., 2009), news articles (L. Young and Soroka, 2012), or posts on Social Media Networks (A. Kim et al., 2013; Cheng, Niculescu-Mizil, and Leskovec, 2021). Understanding language is central to understanding processes and developments in these areas. To enable computers to deal with textual data, it has to be presented in some numerical form.

Over the previous decades, several approaches to text analysis — also labeled text mining or Natural Language Processing (NLP) — have been used in Computational Social Science (Grimmer and Stewart, 2013; Boumans and Trilling, 2015). At a very basic level, counting words or word shares has become a tool to study the prevalence of certain terms across texts. This approach yields high reproducibility, swift processing capabilities, and transparency of their outcomes. However, they are limited by their context-agnostic nature, reliance on fixed vocabularies (dictionaries), and a general neglect of word semantics. These straightforward methods might overlook the nuances of language evolution and polysemy — and are therefore often called bag of words approaches. However, in certain contexts, they provide quite reliable results when carried out with sufficient background knowledge on the matter. Bag-of-word approaches have been applied to various research, often in conjunction with interpreting sentiments based on pre-defined dictionaries. Examples stretch from the analysis of political reporting in the media (Fortuny et al., 2012), investigations of hostility in user comments (Ksiazek, Peer, and Zivic, 2014), movie reviews (Taboada, Brooke, and Stede, 2009), or suicide notes (Pestian et al., 2012).

Moving towards more context-sensitive techniques, concordances and collocations provide a window into the contextual use of words and common linguistic pairings, offering valuable qualitative insights. Despite their ability to uncover the immediate linguistic environment of words, they demand substantial manual effort to set up, scale poorly to larger and differing data sets, and lack predictive capabilities. For instance, this approach has been used to uncover the media narratives towards banks during the financial crisis 2007-2009 (Kleinnijenhuis et al., 2013).

While these approaches are still very focused on the occurrences of words or groups of words, machine learning models pose new opportunities for text analysis in the field, especially regarding methods like automated coding of topics in texts (Nelson et al., 2018; Osnabrügge, Ash, and Morelli, 2021).

Unsupervised learning aims to detect patterns and themes without needing labeled data, thus facilitating the discovery of underlying structures. The interpretation of their outputs can be complex, requiring expert judgment, and the models can be sensitive to the tuning of their parameters. Word2Vec (Mikolov et al., 2013) is a classical example, using a neural network model to learn word associations from a large text corpus. It creates a vector space, with each unique word in the corpus being assigned a corresponding vector in the space. It is learning to represent words by the context in which they appear. Potential downsides are represented in the static nature of the word meanings they encode, potentially overlooking the fluidity of language in different contexts (polysemy) or being unable to handle rare, unknown words. Like its predecessors, Word2Vec loses higher-level structures above words. Another often-used example for topic modeling of texts is Latent Dirichlet allocation (Blei, Ng, and Jordan, 2003), which aims to cluster topics by distinct terms.

The most recent advances in text analysis are transformer models like BERT (Devlin et al., 2018) and GPT (Radford, Narasimhan, et al., 2018; Radford, Wu, et al., 2019; Brown et al., 2020), which have set new benchmarks in performance thanks to their context-aware word representations and the advantages of transfer learning (Wankmüller, 2022). They are a mix of supervised and unsupervised learning. Initially, they are trained by learning from extensive amounts of text data without any human-provided labels — this is the unsupervised learning stage, where they determine the underlying language patterns on their own. Then, to specialize in particular tasks, they are fine-tuned with labeled data sets in a supervised fashion, where they learn from examples with known outcomes to make accurate predictions or classifications. These models require extensive computational resources, are complex in their fine-tuning, and require huge amounts of data for pre-training, which poses significant barriers.

This dissertation has deployed text analysis methods in multiple occurrences: In Paper 2, the methodology involved a manual categorization approach after employing frequency analysis to extract the 50 most used hashtags from tweets. These hashtags were manually coded into thematic categories such as COVID-19, election, Ukraine, sports, and politics. This manual coding, while labor-intensive, allowed for a nuanced understanding and classification of the hashtags based on expert judgment. Subsequently, the infographics associated with each category were quantified, connecting the quantitative extraction process to a qualitative content analysis.

We adopted a two-sided approach in Paper 4. Initially, using a regular expression `@[a-zA-Z0-9_]+` to identify and extract all retweeted usernames from the data set of tweets, distinguishing between political and data journalists and further differentiating by sex. After filtering out self-retweets, each group's 30 most retweeted accounts were identified and categorized into predefined groups such as German media, foreign media, politics, NGOs, and various journalistic domains. This classification provided a foundation for analyzing the retweet patterns and the sources that different journalist groups and sexes are more inclined to engage with.

Secondly, hashtags were extracted from tweet metadata and clustered into categories tailored to political and data journalists. Notably, an 'others' category was not required for the political journalist subset, which suggests a concentrated thematic focus within their tweets. A broader range of categories was employed for data journalists, reflecting the more diverse topics they tweet about.

The manual categorization is a methodological choice that leverages human interpretive skills over purely algorithmic sorting. This allows for a nuanced understanding of the context and the subtle differences between hashtags that may be lost in automated methods. By merging the precision of manual categorization with the breadth of computational techniques, research benefits from both the accuracy and depth of human analysis and the efficiency of algorithmic processing.

However, it introduces limitations related to the potential for subjective bias, the method's scalability, and the coding process's replicability. Furthermore, regular expressions and manual coding do not account for the semantic relationships between hashtags or the context within which usernames are mentioned, which could lead to an oversimplified understanding of the complex social interactions on Twitter. In addition, while frequency analysis effectively gauges the popularity of certain hashtags, it does not consider the context of their use, which can be vital for a true understanding of their role and significance in public discourse.

For the sake of the analysis, which aimed to generate insights into the topics tweeted about, hashtags were arguably valuable and easy-to-access data. Hashtags are designed to categorize content, making them a practical tool for identifying primary topics within tweets. They reflect community engagement, with users leveraging them to participate in larger conversations, thereby acting as indicators of collective attention.

### 3.1.4 Network Models

In Computational Social Science, social network analysis (SNA) has evolved as a methodology for understanding the complex relationships and patterns within various types of social systems. While the theoretical underpinnings of network analysis were already in place for some time — going back to early 20th-century sociologists and anthropologists, such as Georg Simmel, who declared that the essence of society exists in the pattern of relationships between actors, not merely in the actors themselves (Hollstein, 2021) — the advent of the internet and Social Media Networks and computational availability and methods drove the advances in the field by providing a vast amount of digital social interactions to study (Boase et al., 2006).

The method became particularly suited to studying the media's role in society, as communication networks could be directly observed and measured. Research has shown the potential influences our ties to friends have on our behavior (Christakis and Fowler, 2011) and how closely people might be connected (Travers and Milgram, 1969; Watts, 2004), even if they are not aware (Granovetter, 1973). Or how innovations — or novel information — are diffused through a network. A theoretical model by E. M. Rogers (2003) which gained renewed attention in the age of social media as researchers use SNA to track the viral spread of information and identify key influencers within networks. This might help explain the sharing of false news in Social media Networks (Vosoughi, Roy, and Aral, 2018). Others have used these methods to explain the spread of COVID-19 (Jo et al., 2021) or the structures of far-right networks on Telegram (Urman and Katz, 2020).

SNA provides a lens to examine the complex communications web underpinning social relations in each of these applications. Particularly in the age of big data, with the proliferation of digital communication channels, SNA's importance and utility continue to grow in discerning patterns that are otherwise not observable through traditional analytical methods.

A graph is the basis of a social network. Graphy theory is sometimes attributed back to Leonard Euler's solution of "The Seven Bridges of Königsberg" in the year 1736 (Euler, 1736): Königsberg, a city divided by the Pregel River connected by seven bridges, posed a challenge that intrigued the public: to find a walk through the city crossing each bridge exactly once. Euler demonstrated the impossibility of such a walk.

He abstracted the problem from a physical city layout into a graph, a collection of points (vertices) representing land and lines (edges) representing bridges. He proposed that for a walk to cross each bridge once, each land must have an even number of bridges. Königsberg's arrangement did not meet this criterion. Hence, no such walk existed. This insight led to the Eulerian path concept and laid the groundwork for the field of graph theory, which defined relationships as the connection between nodes or vertices via edges.

Around those foundational blocks, several metrics can be retrieved:

**Measures of Centrality**, like a node's number of direct connections (degrees). It indicates the activity level of a node in the network; How close a node is to all other nodes in the network, based on the average length of the shortest paths from the node to all others; The extent to which a node lies on the shortest path between other nodes, indicating its role as a 'bridge' within the network. **Connectivity measures**, like density, are the proportion of potential connections in a network that are actual connections, reflecting the overall 'tightness' of the network. Sometimes, it is also described by the path lengths. **Measures of Segmentation**, which describe the components — subsets of nodes within which each pair of nodes is connected —or the modularity — the strength of the division of a network.

The overall network structure is often described by reciprocity, the tendency for node pairs to form mutual connections in a directed network; transitivity, the probability that the adjacent nodes of a node are connected; or centralization, the degree to which one or a few nodes dominate a network.

These basic metrics can be supplemented with dynamic network metrics that describe changes in the network over time, like stability or evolution patterns.

Paper 4 uses social network analysis to describe differences in retweeting and mentioning behavior between male and female users of political and data journalists on Twitter to investigate journalistic amplification and legitimation via the platform. Retweeting and mentioning are considered edges, while users are modeled as nodes.

As network theory and computational capabilities evolve, they open new possibilities for dissecting the intricate tapestry of communication across various social arenas. Over time, the focus shifted towards understanding the dynamics of these networks, leading to the development of relational event models (REM). As defined by Butts (2008), a relational event is a discrete interaction directed from a social actor toward one or more targets, with the aim of REM being to utilize historical data of such interactions to predict future events. These models relied heavily on longitudinal data, typically time-stamped records of interactions, providing a richer and more dynamic understanding of social ties compared to traditional survey data (Corman and Scott, 1994). REMs have seen diverse applications, including analyzing virtual friendship formations, state-level interactions, and collaborative behaviors in various contexts (Juergen Lerner et al., 2013; Welles et al., 2014).

However, the REM framework is generally confined to dyadic interactions, suitable for one-to-one relational events but inadequate for capturing complex interactions involving multiple actors simultaneously (B. Kim et al., 2018). This limitation gave rise to the advancement of RHEM. This model extends the REM approach by including 'hyper-edges,' representing interactions between multiple actors (Jürgen Lerner, Tranmer, et al., 2019). Their work demonstrates the applicability of RHEMs in analyzing historical contact diaries and scientific coauthorship networks, providing a framework that captures the multi-dimensionality of interactions (Jürgen Lerner, Lomi, et al., 2021; Jürgen Lerner and Hâncean, 2023).

In this thesis, RHEMs model co-authorship networks among data journalists in Paper 3, where an article serves as a 'receiver' and the co-authors as 'senders.' The model is structured to account for hyperedges, where each co-authored article is a complex event involving multiple senders. This data is represented in two-mode networks with comprehensive metadata, including publication times, previous collaborations, and departmental affiliations. Through this approach, the study investigates the patterns and dynamics of co-authorship within the journalistic field. It offers insights into the collaborative process and its evolution in response to external events such as the COVID-19 pandemic.

### 3.1.5 Summary

The methodological section of the dissertation delineates a multifaceted approach to understanding the influence of COVID-19 on data journalism, incorporating both qualitative and quantitative research methods.

**Survey and Qualitative Thematic Analysis**: This subsection detailed the deployment of a survey to capture the nuances of user perception regarding predictive journalism in the context of the German federal elections of 2021. Employing a combination of visual-based questions, Likert scales, and open-text responses, the survey gauged users' understanding of uncertainty in predictive models, the utility of these visualizations, and their influence on users' thoughts about the election.

**Semi-automated Infographic Detection**: Here, the methodology advances into the computational domain, introducing a semi-automated process for detecting infographics in digital news content. This process combines computer algorithms and human oversight to identify infographic elements, establishing a dataset representative of visual data journalism from media companies' Twitter posts from the USA, UK, Germany, France, Italy, and English-speaking newspapers in India to analyze the prevalence of infographics.

**Text Analysis**: Two distinct papers employ text analysis through manual categorization methods to interpret Twitter data. The first paper utilizes frequency analysis to extract prominent hashtags from tweets, which are then manually sorted into thematic categories, blending quantitative data extraction with qualitative analysis. The second paper applies a two-sided approach, initially employing regular expressions to isolate usernames and cluster hashtags into predefined groups. This methodology allows for a detailed exploration of the content themes and retweet patterns within the journalistic discourse on Twitter.

**Network Analysis**: Using the structure of networks to explain social structures is used methodologically in two papers. The first one uses mentions and retweets between sexes and journalistic groups to study amplification and legitimation on Twitter. The second leverages the network structure to investigate the influence of COVID-19 on co-authorship patterns of data journalists using relational hyperevent models.

Looking ahead, the subsequent subsection will elaborate on the datasets compiled to operationalize the aforementioned methods. This will include a description of dataset sources, collection processes, and the criteria used for inclusion and exclusion. It will also discuss the data cleaning and preparation methods, ensuring that the datasets are robust and suitable for analysis. Furthermore, ethical considerations concerning data privacy and the use of proprietary or sensitive information will be addressed to maintain the integrity of the research process. This foundational step is critical as it contextualizes the methodological rigor and sets the stage for applying analytical techniques that follow in the papers.

| | Paper 1 | Paper 2 | Paper 3 | Paper 4 |
|---|---|---|---|---|
| Survey & QTA | X | | | |
| Semi-automated Infographic Detection | | X | | |
| Text Analysis | | X | | X |
| Network Models | | | X | |

Table 3.1: Overview on methods used in papers across this thesis.

## 3.2 Data

This dissertation uses an array of datasets, each with unique attributes and affordances, to pursue a comprehensive understanding of the influence of COVID-19 on data journalism.

These datasets serve as the empirical foundation for the study's computational analyses and interpretive insights. From survey responses that capture nuanced user perceptions to large-scale media content archives, they offer a multidimensional view of the phenomenon under scrutiny.

The following section provides a detailed exposition of each dataset utilized, elaborating on its origins, characteristics, and specific roles in addressing this dissertation's research questions.

### 3.2.1 Survey data

The dataset used in Paper 1 was collected using a questionnaire distributed to users of the German news website sueddeutsche.de. It contains the answers to 16 questions, some focusing on the perceptions towards works of predictive journalism, some showing users' understandability of the intention of the visualizations.

The dataset is primary data collected by the authors, using a web link that was posted in context to a news article that featured an explanatory article to election predictions, which the survey is investigating.

The data was collected between September 2nd and September 28th, 2021, through the final weeks of the German federal election race. The survey focused on a German audience but was technically accessible from everywhere through the web link.

The data is structured in a tabular form with 49 columns. The dataset contains a mix of categorical and numerical variables and free-text fields. Some of the key columns include:

- **Teilnehmer ID**: Participant ID, a unique identifier for each respondent.
- **Abgeschlossen**: Indicates whether the survey is complete ('ja' for yes, 'nein' for no).
- **Teilnahmebeginn and Teilnahmeende**: Start and end times for participation, formatted as timestamps.
- **Teilnahmedauer**: Duration of participation, potentially in the format of hh:mm:ss.
- **A series of questions (e.g., 1. Frage:, 2. Frage:, etc.)** that capture the core responses of the survey. These fields contain varied data types, including numerical ratings, categorical selections, and text-based answers.

As the dataset was populated while users filled out the questionnaire, terminators may have missing values. That also leads to some inconsistencies in the responses to questions, given that the survey allows for both multiple-choice and optional free-text answers, especially if the latter are not filled in for each participant. These have been accounted for in the analysis.

Regarding ethical considerations, the dataset does not contain overtly sensitive information. Furthermore, as there there is no way to attribute the answers to specific individuals. However, the presence of questions related to political attitudes and demographics necessitates careful handling to ensure anonymity and ethical compliance.

In light of the research question, the data is limited to the participants' entries in the free-text answers, their responses on Likert-scale answers, and factual accuracy on questions regarding the interpretation of the visualizations. Known biases mostly concern the non-random selection process that may favor affirmative or refusing voices.

This data is used to understand better the changes from an audience perspective that COVID-19 has brought to predictive journalism, which was used for both virus-prevalence- and electoral-focused predictions. Using a more qualitative approach helps collect various opinions compared to a more quantitative method.

### 3.2.2 Twitter data

In social media analytics, especially on Twitter (now known as X), this thesis will distinguish between two different types of data. They give us unique insights into user behavior and content dissemination.

The first category, termed "tweets-based Data," focuses on the intrinsic properties of tweets themselves, encompassing textual content, tweet metadata, and accompanying images. This facet illuminates the thematic and visual aspects of the discourses.

The second category, "User-based Data," is anchored in user interactions, specifically mentions and retweets. By studying this layer, we capture the relational dynamics and network structures that influence the spread and reception of information. This bifurcation provides a nuanced framework for analyzing different dimensions of social engagement and informational flows on Twitter.

#### Images

To analyze the changes in the prevalence of journalistic infographics across several geographies in Paper 2, we classified images retrieved from tweets by media organizations. This is particularly relevant in the context of the COVID-19 pandemic, which has impacted data journalism across different media landscapes. Infographical output is used to measure the prevalence of data-driven journalism within newsrooms.

We collected Twitter accounts for the five largest national, general-audience news media across six different countries by circulation[2]: USA, UK, Germany, France, Italy,

[2] Sources for circulation numbers: Alliance for Audited Media (USA) via pressgazette.co.uk/news/us-newspaper-circulations-2022, ABC (UK): www.abc.org.uk, IVW e. V. (Germany): www.ivw.de, ACPM (France): www.acpm.fr, FIEG (Italy): www.fieg.it, ABC (India): www.auditbureau.org, RNI: rni.nic.in.

and the only three available English-speaking newspapers in India. A search on Twitter manually retrieved the account names.

The final dataset comprises tweets collected via the Twitter API requesting all tweets containing images. It contains ten variables, each capturing different attributes of the tweets and the engagement they received.

- **author**: The Twitter account handle that authored the tweet.
- **tweet_id**: A unique identifier for each tweet.
- **created_at**: Timestamp indicating when the tweet was created, in the format 'YYYY-MM-DDTHH:MM:SS.ZZZZ'.
- **media_key**: A unique identifier for media attached to the tweet.
- **images/url**: URL to the image attached to the tweet.
- **source**: The platform or application used to publish the tweet (e.g., Twitter Web App, Twitter Media Studio).
- **text**: The actual text of the tweet.
- **likes_count**: The tweet's number of likes received.
- **rt_count**: The number of times the tweet has been retweeted.
- **quote_count**: The number of times the tweet has been quoted.
- **reply_count**: The number of replies the tweet received.

We further used image URLs included in the API response to download the 1,911,496 images connected to the tweets. Images we deduplicated and exposed to the semi-automated approach described in section 3.1.2. The results were returned in a CSV file containing the image file name and a categorical variable *infographic* that indicated if an image was classified as an infographic (yes/no). The images and resulting classification could be re-connected to their tweets via the file name. This allowed further analysis regarding the tweets' textual content (see section below) and interactions.

The data was taken from official Twitter accounts of media companies, so no personal user data that might have triggered ethical considerations for this dataset was involved. The results of the infographic detection can be shared. However, sharing the downloaded images is restricted by Twitter's terms.

Regarding the limitations of this dataset, we find potential issues, mostly regarding the selection of cases: The focus on the largest media companies by circulation may not provide a representative and comprehensive view of the media landscape in each country. Smaller, independent, or niche media outlets may use infographics differently but are excluded from this analysis. Larger media companies may have more resources to produce infographics, possibly skewing the prevalence of data journalism in the dataset — but individual editorial decisions may also influence the results. The countries selected for this study mainly represent Western democratic systems, which could introduce a cultural bias in understanding the global influence of COVID-19 on data journalism — which we, however, could not find in our data. Including only English-speaking newspapers in India may not capture the full scope of data journalism practices in a country with multiple languages and diverse media ecosystems.

In the broader landscape of this dissertation, this dataset offers opportunities for a multifaceted analysis. Specifically, it enhances our understanding of how data journalism, in the form of infographics, has been affected by or has evolved during the COVID-19 pandemic. The longitudinal nature of the data also provides a temporal dimension to the study, potentially allowing us to observe trend shifts in infographic use by major media outlets across several geographies. Consequently, the dataset substantiates the empirical facet of the dissertation.

### Text

To further understand the content of tweets, we have used text analysis in two papers for Hashtag extraction and clustering in tweets.

**Paper 2: Hashtag Clustering for Infographics Analysis:** The first use case for text analysis in this dissertation is a clustering of topics of the most used hashtags accompanying infographics on Social Media Network posts of media in Paper 2. As described above, we have collected tweets containing images of the largest media companies in six countries and aimed to identify the prevalence of infographics pre- and post-COVID-19. To compare and contrast the findings of infographics with the content of the selected tweets in general, we extracted all hashtags used in the text of the tweets using a regular expression `#\\w+`. We did not rely on the automatically detected hashtags that Twitter delivers with each API response. We then grouped and counted the hashtags for each country and filtered the 50 most used for each geography. Those were then manually clustered into several categories based on their overall topic:

- **COVID-19**: e.g. Coronavirus, vaccino, coronaviruslockdown
- **Politics**: e.g. Merkel, GiletsJaunes, RepublicDay
- **Election**: e.g. vote, Presidentielle2022, btw21
- **Ukraine**: e.g. Ukraine, UkraineRussiaWar, Putin
- **Sports**: e.g. Tokyo2020, Mondiali2018, SuperBowl

We then enabled a string detection across the whole dataset of tweets that checked for clustered hashtags to appear and labeled the tweet accordingly. Tweets with multiply categories were not attributed, which we also did not expect in our data due to the broad categories chosen.

The data was then used to identify the topics of tweets overall, particularly for tweets containing infographics.

In the broader landscape of this dissertation, this dataset connects the basic counting of detected infographics and a thematic analysis that indicates the influence COVID-19 might have had on the prevalence of infographics across our sample. This acts as a strong empirical reference regarding our research question on the influence of the pandemic on the output of data-driven journalism.

**Paper 4: Analysis of Retweet Sources and Hashtag Use Among Journalists:** A second use case for social media data was the hashtag extraction of tweets sent by female and male political and data journalists in Germany throughout 2021. The primary focus of the research was the tweeting differences between sexes across the two groups of journalists. Therefore, the dataset's creation will be described in detail below in the section on user-based data collection. However, the clustering of hashtags from the tweets' text column will already be described here: the point of origin is a dataset containing all the tweets of the two groups, joined by username with attributions of the sex of the user that was researched separately.

In this instance, we used the automatically detected hashtags provided via the nested entities field of Twitter's APIv2 JSON return. This research was conducted before the one in the preceding chapter, where we decided to extract the hashtags from the text fields ourselves rather than rely on Twitter's automated detection.

The hashtags were then grouped by group of journalists and sex and counted. The 30 most prevalent hashtags per group and sex were then dispatched to manual coding. The number of 30 was arbitrary, allowing for some variance across topics.

The clustered hashtags were based on a per-journalistic group basis, which provided these categories for political journalists, with no Other category required:

- **COVID-19**: e.g. Coronavirus, Covid19, Lockdown
- **Politics**: e.g. Merkel, Bundestag, EU
- **Elections**: e.g. btw21, Triell
- **Climate**: kidsfirst

For data journalists, these were the results of the clustered hashtags:

- **COVID-19**: e.g. Coronavirus, Covid19, Omikron
- **Politics**: e.g. Merkel, Bundestag, Scholz
- **Elections**: e.g. btw21, Bundestagswahl, Triell
- **Climate**: Klimakrise
- **Sports**: EURO2020
- **Data-driven journalism (ddj)**: e.g. OpenData, dataviz, AI
- **Other**: e.g. Nannenpreis, SciCAR, OSINT

These categories were then used to calculate the share of occurrences within the top 30 hashtags per group.

As this method is based on the communications provided by users on Twitter, it raises ethical concerns that require some scrutiny: the content of the tweets was certainly not provided for obvious scientific purposes, nor have the users given explicit consent for our work. However, they've acknowledged the terms of Twitter that inform about the possibility of this data being used for third-party processes. The nature of Twitter as a platform used to be very open and allowed non-members to read most of the tweets without logging on, which further emphasizes the publicity of messages published there. To somewhat anonymize the data, we focus not on individual accounts but their grouped sex attribution and affiliation to a journalistic area.

The limitations of this dataset arise in several ways: they are inherently dependent on the manual pre-selection of users, which might introduce severe bias if not done properly. It further depends on the availability of Twitter messages, which the users influence in not deleting historical tweets, and on Twitter in providing correct responses in its API.

In the broader landscape of this dissertation, this dataset enhances the opportunities for comparison between sexes within groups of journalists. It aims to add a qualitative perspective on the quantitative research results regarding amplification through mentions and retweets to understand better the influence of sex on the behavior of tweeting political and data journalists.

### User-based Data

Research on Twitter can be based on several approaches. The content of tweets, text, videos, or images — as already described above- and the network structure on a higher level could potentially serve as a data source. This data set is the primary foundation for Paper 4, which aimed to analyze sex-related differences between German political and data journalists' amplification behavior on Twitter.

Manual definition and user collection were necessary to identify these groups, which are not legally defined and openly accessible. We updated and used an already available list of political journalists on Twitter based on a definition to include journalists deployed to political departments or primarily working on political topics. Many larger newspapers offer an imprint with an overview of their authors and their positions, which often contains Twitter accounts. Smaller newspapers sometimes lack that information, which must be retrieved from the articles.

The selection of data journalists was done differently. Many German data journalists congregated in a professional group of the non-governmental reporters' advocacy group "Netzwerk Recherche" in the fall of 2020. The messaging platform Slack was used to facilitate easy communication. This platform was open to anyone who regards herself or himself as a data journalist. We assume the majority of data journalists to be members of this group, as there are no fees or further barriers to entry, and participation in the group offers incentives, like conversations on eminent topics in the field, information on upcoming gatherings, or job openings. However, this selection method introduces possible self-selection bias to the list. Compared to other simultaneously created research, Beiler, Irmer, and Breda (2020) and Haim (2022) showed similar numbers of identified individuals.

The attributed sex was extracted via the users' first name, and further research was employed where it was not clear initially. While this may not cover differences between the sex and gender of individuals, we argued that the external impression of a name's sex might affect the amplification behavior of others on Twitter, which we are focusing on.

The final dataset for political and data journalists was set up essentially like this — additional data like follower counts were not provided for both groups, but they were also not used in the analysis based on this dataset:

- **username**: The Twitter handle associated with the user's account.
- **id**: A unique numerical identifier assigned to the Twitter user.
- **Newspaper**: The name of the newspaper or media outlet the user is associated with.
- **Sex**: The user's sex as recorded in the dataset.

The attribution of sex data to an individual raises ethical issues. However, we have aggregated our results on a journalistic-group-sex-level to avoid having individual information released. The content of tweets was included in our analysis only so far as to use hashtags to capture topicality.

The tweets for all identified usernames were then retrieved on January 7, 2022, using Twitter API v2 (Juergen Pfeffer et al., 2023) for the time period between January 1 and December 31, 2021. The first dataset contains 430,451 tweets from 730 Twitter accounts of political journalists in Germany, and the second dataset with an identical structure contains 47,812 tweets from 149 Twitter accounts of data journalists.

Both datasets are structured in this way:

- **source**: Indicates the platform or device used to post the tweet.
- **conversation_id**: A unique identifier for the conversation to which the tweet belongs.
- **type**: Indicates the type of the tweet.
- **text**: Contains the full textual content of the tweet.
- **created_at**: The date and time when the tweet was originally posted.
- **lang**: The language code representing the language used in the tweet's text.
- **author_id**: The unique numerical identifier for the creator of the tweet.
- **id_tweet**: The unique identifier for the tweet itself.
- **possibly_sensitive**: A flag indicating whether the tweet may contain sensitive material.
- **entities/mentions**: detected mentioned usernames in the tweet text. Not used for identifications.
- **entities/hashtags**: detected Hashtags in the tweet text
- **entities/urls**: detected URLs in the tweet text.
- **entities/annotations**: detected annotations for entities or contexts in the tweet text.
- **retweet_count**: The number of times others have retweeted the tweet.
- **reply_count**: The count of replies to the original tweet.
- **like_count**: How many likes the tweet has accumulated.

- **quote_count**: The number of times other users have quoted the tweet.
- **author_username**: The Twitter handle of the user who posted the tweet.
- **author_name**: The display name of the user who authored the tweet.

To be able to analyze the amplification behavior, we extracted the usernames of mentioned and retweeted users from the textual portion of the tweets, using the text column to identify all usernames in the text following Twitter's standard of preceding the username with an @, which enabled us to use a regular expression in the form `@[a-zA-Z0-9_]+`. To detect retweets, we could leverage predefined API returns, using the entities/mentions column, which returned all detected usernames, and setting a filter on the type of the tweet to only return "retweeted" tweets. These usernames could then be left joined to the dataset containing the usernames of identified political and data journalists with their sexes attributed to starting the analysis.

Within this dissertation, these datasets enable a multifaceted exploration of internal communication inside groups in journalism, offering insights into networking tendencies, mutual engagement and amplifications, and the potential biases in sexes' visibility and influence. It shows the potential of computational methods to explore sex-related differences in communication behavior.

### 3.2.3 Web-Scraped data

Computational Social or Communication Science is not limited to data provided by Social media Networks. As already mentioned in section 2.4.2, there is even a certain push to access data sources that are independent of platforms — although this often comes with the caveat of having to create new ways of accessing this digitalized data. One potential road is to use web scraping computer tools that identify elements on webpages by their markup structure in the Hypertext Markup Language (HTML) or their content. This was the approach to collect articles published by German data journalists across several media companies before and after COVID-19 to compare their publication rates, but foremost, their cooperation within data journalism and with colleagues from different departments.

While some databases collect data on German news media, they are not necessarily comprehensive, restrictive in access, and potentially even harder to extract digitalized data from. Therefore, a web scraping approach was used to collect news media websites' authorships and article metadata information. The starting point was the author pages of data journalists, identified using the Slack group already described in section 3.2.2, and verifying the selection using imprints, where available.

Initially, a simple Javascript script was created that extracted elements (authors, date, title, URL) from author pages and concatenated and saved them into a JSON file. As each media has a different markup, the definitions in the script had to be adapted accordingly. Data for Bayerischer Rundfunk was obtained as an Excel file, as they did not provide author pages with required temporal dimensions. Data for Tagesspiegel was web scraped using the rvest-package (Wickham, 2022) for the statistical computing language R (R Core Team, 2022).

The JSON files were then individually cleaned to obtain unionizable tables, which were then brought together for analysis. For each author, a single line was created to allow them to be treated as individual observations — although technically, the granularity was on an article level. The following R-packages were used in the analysis: tidyverse (Wickham et al., 2019), lubridate (Grolemund and Wickham, 2011), and jsonlite (Ooms, 2014).

While some articles went back to 2011, a common starting point was chosen to be January 1st, 2019. Some data preprocessing had to take place, primarily on the byline: one media company used to combine multiple-authored articles into a single phrase, which did not give us any information on the actual authors and was therefore ignored. Some media companies used to abbreviate the first names of multiple authors, which had to be recreated to allow precise matching. In contrast, others placed locations into the author's byline, which had to be removed. The departments under which a certain article was published could be extracted from the articles' URLs and grouped when departments were identical. To identify whether an article was published before or after COVID-19 hit, March 16th, 2020, was set as the breakpoint.

The data structure of the final dataset is this:

- **title**: The title of the article.
- **date**: The publication date of the article.
- **authors**: The authors' names who contributed to the article.
- **url**: The URL where the article can be found.
- **author_individual**: The identified data journalist author of the article.
- **media**: The media outlet where the article was published.
- **department**: The department within the media outlet that published the article.
- **before_covid**: A Boolean value indicating whether the article was published before the COVID-19 pandemic.

This data focuses on the published work of identified data journalists, which has some limitations. The dataset may not represent all data journalism work, as it is collected from specific authors and media outlets. Articles by data journalists of other media, not included in the author pages, or those without an online presence, are missing, thus limiting the generalizability of the findings.

This is especially clear for German public media, which could only be included thanks to personal contact between the author and the data department of BR Data. However, while this is a quantitative method, the enormous qualitative pre-selection process limits the quantitative insights to certain carefully selected media companies.

Another limitation is to focus only on metadata. For articles with multiple authors, the dataset may not accurately reflect each author's contribution level, especially if the data journalist's role was not primary, which cannot be accounted for in other ways, as article contents were not analyzed due to accessibility issues like paywalls. This is acceptable as the research wants to highlight general changes to cooperation patterns, which should be observable in the metadata.

### 3.2.4 Summary

This dissertation's Method and Data section discussed various datasets that investigate the multifaceted nature of data journalism and its evolution, especially in the context of the COVID-19 pandemic. The datasets utilized can be summarized as follows:

**A survey dataset** was gathered through a questionnaire distributed to users of sueddeutsche.de, a leading German news platform. It encapsulates respondents' perceptions toward predictive journalism and their comprehension of the intentions behind data visualizations and is used in Paper 1.

To investigate potential shifts in the prevalence of journalistic infographics in Paper 2, an innovative **semi-automated infographic detection method** was applied to images sourced from tweets by media organizations. This approach provides a unique lens to assess the impact of the pandemic on the visual dimension of data journalism in various geographical locations.

The analysis of **Twitter texts**, through which hashtags were clustered, serves as a backbone for two distinct papers (3 and 4). This methodological choice allows for exploring thematic concentrations and discourse patterns within data journalism on social media.

A **user-based tweet dataset** forms the empirical basis for a study examining sex-related differences in the amplification behaviors of German political and data journalists on Twitter in Paper 4. By retrieving data on mentions and retweets, this dataset enables analysis of engagement and amplification dynamics.

Lastly, a self-curated **web-scraped dataset** provides insights into the authorship of data journalism articles within several German newsrooms, investigated in Paper 3. Given the limitations of existing databases, this proactive approach was essential for acquiring a comprehensive and accessible compilation of authorship and article metadata.

Each dataset has been carefully selected to ensure novel ways of measuring and gaining insights into data journalism practices. The methods and data collection strategies underscored in this section serve as the foundation for the analyses and discussions following through the upcoming papers of this thesis.

| | Paper 1 | Paper 2 | Paper 3 | Paper 4 |
|---|---|---|---|---|
| Survey | X | | | |
| Images | | X | | |
| Hashtags | | X | | X |
| Twitter Users | | | | X |
| Web-scraped Articles | | | X | |

Table 3.2: Overview of datasets used in papers across this thesis.

# Election Predictions in the News: How Users Perceive and Respond to Visual Election Forecasts

## Election Predictions in the News: How Users Perceive and Respond to Visual Election Forecasts

### Authors

Benedict Witzenberger, Nicholas Diakopoulos

### In

Information, Communication & Society, published as: Benedict Witzenberger and Nicholas Diakopoulos (July 2023). "Election predictions in the news: how users perceive and respond to visual election forecasts". In: *Information, Communication & Society*, pp. 1–22. DOI: 10.1080/1369118x.2023.2230267

### Abstract

Political journalism often tries to predict the future, especially the outcomes of elections. This has historically been accomplished through written articles or opinion pieces. A more recent development involves the publication of data-driven predictions in online news media. These news items not only contain an estimate for the election results but also often try to visualize potential uncertainties of the prediction. However, the ways in which users react to these forms of journalism have not yet been studied extensively. In this work, we survey users of a predictive journalism piece on the 2021 German federal elections published by the German newspaper Süddeutsche Zeitung to better understand their reactions. While we found an alignment between the designers' intention to show the inherent uncertainty of election predictions with the audience's reception, we encountered mixed results in users' ability to interpret the uncertainty visualizations presented. Most respondents indicated that the predictions did not influence their thinking about the race, and some remained skeptical toward such predictions published by journalists for various reasons. Based on these findings, we suggest the need for rigorous user testing of visualizations for election prediction and increased awareness and future research on ways to increase the transparency of methods and data to develop appropriate trust toward predictive journalism.

### Contribution of thesis author

Theoretical operationalization, survey design and realization, computational analysis, qualitative evaluation, contextualization, manuscript writing, revision, and editing.

### Publication Summary

Predictive journalism has gained popularity in the United States and, with COVID-19, has seen increased use globally in news reporting. Despite its growth, little research exists on how readers perceive these future-oriented articles, which is crucial for understanding their impact on election outcomes or health policies.

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2023.2230267

# Election predictions in the news: how users perceive and respond to visual election forecasts

Benedict Witzenberger [a] and Nicholas Diakopoulos [b]

[a]Technical University of Munich, Munich, Germany; [b]Northwestern University, Evanston, IL, USA

**ABSTRACT**

Political journalism often tries to predict the future, especially the outcomes of elections. This has historically been accomplished through written articles or opinion pieces. A more recent development involves the publication of data-driven predictions in online news media. These news items not only contain an estimate for the election results but also often try to visualize potential uncertainties of the prediction. However, the ways in which users react to these forms of journalism have not yet been studied extensively. In this work, we survey users of a predictive journalism piece on the 2021 German federal elections published by the German newspaper Süddeutsche Zeitung to better understand their reactions. While we found an alignment between the designers' intention to show the inherent uncertainty of election predictions with the audience's reception, we encountered mixed results in users' ability to interpret the uncertainty visualizations presented. Most respondents indicated that the predictions did not influence their thinking about the race, and some remained skeptical toward such predictions published by journalists for various reasons. Based on these findings, we suggest the need for rigorous user testing of visualizations for election prediction and increased awareness and future research on ways to increase the transparency of methods and data to develop appropriate trust toward predictive journalism.

**ARTICLE HISTORY**
Received 28 July 2022
Accepted 4 June 2023

**KEYWORDS**
Journalism; election predictions; visualization; uncertainty

## 1. Introduction

Time plays an important role for journalists, both as an organizing element like recency in news media (Bell, 1995; Schlesinger, 1978) and as a storytelling element in their reporting (Jaworski et al., 2004; Neiger, 2007; Neiger & Tenenboim-Weinblatt, 2016). In recent years another element of time has been displayed prominently by many outlets: predictions about the future in interactive dashboards, charts, or other forms of visualization (Diakopoulos, 2022). While election night forecasts have a long tradition in the US, with the first UNIVAC computer providing predictions on-air for CBS in 1952 (Shedden, 2014), today's election cycles are simultaneously accompanied by a multitude of election predictions in various forms by different news organizations. The COVID-19 pandemic added another aspect to this by presenting models that were expected to chart the possible course of the virus spread (Allaham & Diakopoulos, 2022; Pentzold et al., 2021).

**CONTACT** Benedict Witzenberger ✉ benedict.witzenberger@tum.de

These are instances of a shift in journalism that leverages expert-run (sometimes self-developed) models that use technical, visual, and statistical skills to create what has been referred to as 'predictive journalism'. We define this as the publication of data-driven projections of the future that were created by using or relying on computational modeling techniques (Diakopoulos, 2022; Pentzold & Fechner, 2021).

While some prior research has focused on the journalistic perspective of developing, creating and publishing pieces of predictive journalism (Pentzold & Fechner, 2019, 2021), there has been a relative lack of research on the reception and perceived impact of these pieces, particularly in the high-stakes domain of election predictions. As prediction of election results might lead to changes in voting or other politically-relevant behavior such as campaign donations or contributions, studying the effects of such visualizations on audiences is important, arguably even vital, to a democratic society. This work is therefore motivated by the following overarching question: *How do users perceive and respond to visual election predictions published by news media?*

To address this, we fielded a survey focused on user assessments of one such election prediction and its corresponding visualizations published by the German newspaper Süddeutsche Zeitung during the German federal election cycle in 2021. Such a self-selected yet ecologically valid survey sample offers a first glimpse into how at least some users of predictive journalism react and respond. Answers to factual questions in the survey offer an indication of understandability, while open-ended questions capture assessments and self-perceptions of the various constituent visualizations.

We found mixed results on the understandability of the visualizations. Many respondents suggested there was value in the publication of these charts, although some expressed a critical attitude toward any use of electoral models in journalism. Few respondents perceived any potential effect of the presentation on their voting intention or other political behaviors, although a sizeable minority indicated at least some influence on how they thought about the race. Beyond these insights about the response of some users of electoral predictive journalism, our findings offer suggestions for journalism practice in terms of the need for rigorous testing of possible politically relevant behaviors and the need for transparency of methods and data when preparing results of scientific methods for a critical audience that is skeptical of future predictions based on historical data.

## 2. Literature review

We position this work at the intersection between journalism more broadly and election prediction more specifically, including with respect to the display of uncertainty in the presentation of election predictions to broader audiences.

### *2.1. Predictive journalism*

Traditionally, news media is regarded as being past-focused (Lippmann, 1922; Zelizer & Tenenboim-Weinblatt, 2014). However, several authors have shown how journalists also tend to use the future for their reporting (Barnhurst & Mutz, 1997), going so far as to label journalists 'media oracles' (Neiger, 2007). Content analyzes have shown some increase in the levels of future speculation over time in Israeli newspaper headlines (Neiger, 2007), as well as differences in the type of communication medium (Neiger & Tenenboim-Weinblatt, 2016) with print news exhibiting a fuller narrative spectrum of temporal layers than online media. By changing the focus to the future, a news item might even gain news value and make it suitable for publication in the first place (Jaworski et al., 2004). Although by traditional standards, there may be no news to report since nothing has happened yet, the uncertainty of a future event could thus itself be seen as a selection criterion for which events become news items.

Predictive journalism is regarded as the data-driven form of future-oriented journalism in this work (Diakopoulos, 2022). It could be considered a genre of data journalism (Hermida & Young, 2019; Thurman, 2019), which more generally focuses on the visualization and interpretation of datasets, combining disciplines like statistical analysis, computer science, visualization, web design, and reporting (Coddington, 2015). A substantial portion of data journalism is focused on election reporting (Loosen et al., 2017; Solop & Wonders, 2016), and in recent years, it has begun to leverage its toolset to create related forecasts. Maycotte (2015) predicted that 'by using available data, journalists will be able to orchestrate predictions and write tomorrow's headlines and stories accordingly'. However, this is not the case in general: only a small share of data journalistic pieces (perhaps 5–6%) has some recognizable future outlook (Pentzold & Fechner, 2019).

Practitioners may downplay the journalistic relevance of such prognosis because it is limited by backward-facing data and creates time-consuming efforts to visualize data due to the complexity of showing multiple possible futures (Pentzold & Fechner, 2021). These are described as 'temporal exigencies', which form the historical patterns and plausibility of extrapolated future trends. As statistical predictions commonly result in a range of possible outcomes (which might be averaged to retrieve a single numerical point estimate), much effort is needed to display this variance in possible future outcomes, of which most will never come true. While data journalists describe the ambition to visualize uncertainty suitable for their audience, they often are limited by the data literacy of their readers and users, in what Pentzold and Fechner (2021) call 'probabilistic storytelling', which in most cases has been limited to reporting on a single thread out of the hairball of predictions.

The visualization of uncertainty itself is a frequent topic in data visualization literature (Spiegelhalter et al., 2011; van der Bles et al., 2019), although it is mostly focused on publications for expert audiences. Understanding uncertainty visualizations can be a challenging task. A lack of explanations can lead to misunderstandings of the uncertainty visualization (Broad et al., 2007), and there can be substantial variation in the perception of uncertainty intervals (Dieckmann et al., 2015). Some laypeople even have a critical attitude toward forecasts and expect biases when those are unjustified by data (Joslyn & Savelli, 2010).

Relatively little research has focused on the perspective of how a general audience of end-users interpret and make sense of predictive journalism, including the uncertainty, conveyed in predictions. One study compared hypothetical win probability to vote-share projections and found differences in user reception. However, the controlled experiments are somewhat lacking in ecological validity (Westwood et al., 2020). Another recent paper examined responses to COVID predictions in news media by qualitatively studying user comments, finding various affective and evaluative responses (Allaham &

Diakopoulos, 2022). In this work, we add to this nascent literature by pursuing a survey of users of a real, published piece of predictive journalism, specifically in the domain of elections.

### *2.2. Election prediction*

Election prediction has gained much interest during recent presidential elections in the US, fueled by the forecasts of FiveThirtyEight. Its founder Nate Silver described the advantages of his approach: 'Instead of spitting out just one number and claiming to know exactly what will happen, I instead articulate a range of possible outcomes' (Silver, 2012, p. 61). However, although FiveThirtyEight's election models use a lot of different input data, economic values, demographics, and COVID-19 measures -- fundamental are still poll averages that deliver a snapshot of the current state of the race (Silver, 2020). This focus on the results of polling has been criticized as the 'Nate Silver effect': 'overconfidence in election outcomes rooted in a reliance on quantitative measures of public opinion' (Toff, 2019, p. 874).

Journalists and pollsters can be described in a symbiotic relationship: Polling data contribute directly to campaign coverage patterns of journalists, while pollsters receive free advertising for their work (Strömbäck, 2012). This also applies to Germany, where the study presented in this paper takes place. The number of polls shown in a sample of Germany's leading dailies, including Frankfurter Allgemeine Zeitung, Frankfurter Rundschau, Süddeutsche Zeitung, and Die Welt increased from 65 in 1980 to 168 in 1994 (Brettschneider, 1997) and has only continued to increase in more recent elections (Holtz-Bacha, 2012). Whether publications of polls have an influence on voters' decisions is an ongoing debate. Brettschneider (1992) argued for a small effect, especially for supporters of smaller parties (which are always in danger of being elected out of parliament because of a 5 percent total-vote-share-threshold that parties have to cross). Faas and Schmitt-Beck (2007) showed at least a small influence on supporters of the FDP (liberals) in the 2005 election, while Schoen (2002) could not find evidence for tactical voting when controlling for party identification. Discovering the effect of election forecasts on voting intention has been tried before (Urminsky & Shen, 2020; Westwood et al., 2020), but is challenging because voters might be influenced by a variety of sources or socio-demographic factors.

While election predictions have been published by scientists and pollsters for quite some time, perhaps the largest and broadest audience is available through news media. Journalists amplify predictions from these sources, though they also increasingly use scientific methods to create and visualize their own such predictions. This leads to questions about the understandability of these predictions by a larger audience in light of uncertainty and the perceived impact such displays may have on politically relevant intentions or behaviors (Diakopoulos, 2022), which this work addresses by deploying a user survey.

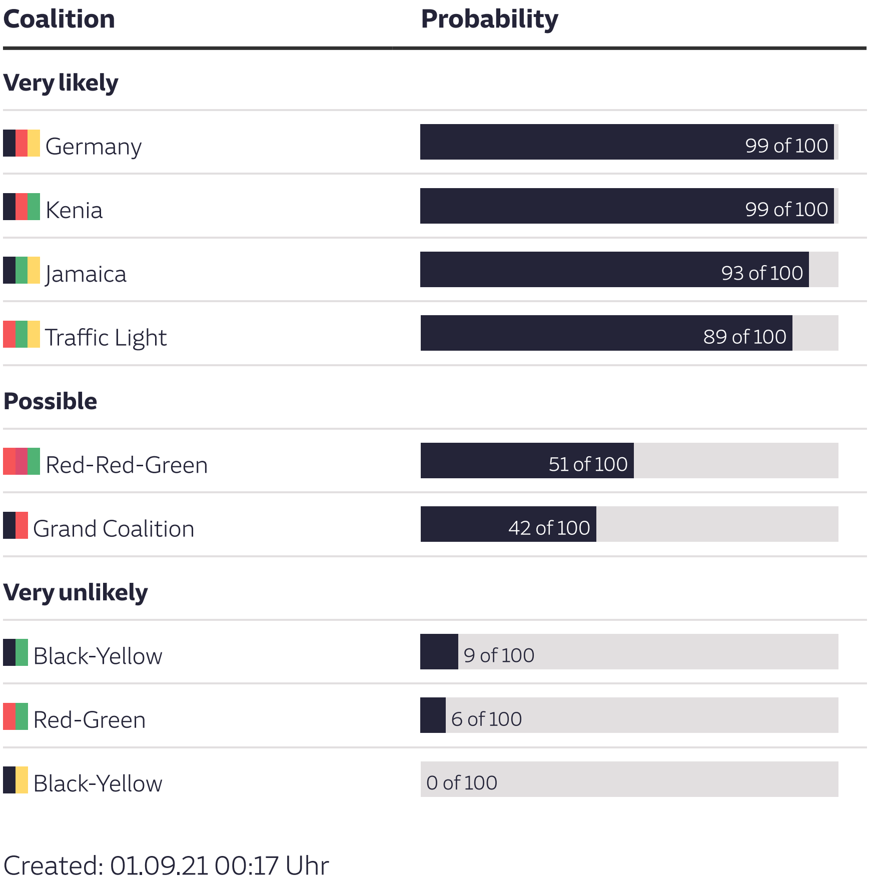
## 3. Study design and methodology

This study focuses on the perspective of users that were exposed to real-life predictive journalism in the context of the 2021 German federal parliamentary election, which took place on September 26th, 2021. In the following subsections, we describe the stimulus design of the predictions and visualizations, the survey design as it addresses our research questions, our recruited participants, and the data analysis approach we took.

### 3.1. Stimulus design

Predicting the multiparty election in Germany has evolved into an effort of multiple scholars to test their models and assumptions (Graefe & Jérôme, 2022). Süddeutsche Zeitung (SZ), one of Germany's largest privately-owned, nationwide newspapers and news websites, chose to cooperate with Zweitstimme.org which represents a team of political scientists from the University of Mannheim who created a predictive model. The model from Zweitstimme has two components, but we focus our evaluation here on the first component, which consists of a dynamic Bayesian forecasting model that predicts electoral results on a federal level. The model combines pre-election polls with fundamentals, which have been shown to be important predictors in the past, including historical results and polls as well as which party is incumbent. This information is already available far ahead of the election. The model is updated with each new public opinion poll during the race. The closer the election date, the more these polls are weighted. To derive probabilities, a Markov-Chain-Monte-Carlo-algorithm is used to simulate the election 9000 times (see Munzert et al., 2017; Stoetzer et al., 2019 for more details). In an ex-ante analysis of the 2017 elections, the model reached an RMSE of 1.88, an average difference between the real election result and model prediction of below 2 percentage points. Based on the outputs of this model, participants in the survey were shown two charts which we describe next.

The first chart (see Figure 1) presented probabilities for specific coalition options from simulations with the vote share model. Coalitions are a central element of a multiparty electoral system, like the German one, where power is mostly shared by several parties that must collaborate to form a government. As this leads to possible tradeoffs in possible post-election negotiations, it is an interesting element of predictive reporting. Most of the coalitions are named according to the colors of the parties involved, and thus, e.g., sometimes resemble countries (Kenia for the combination of black (CDU/CSU), red (SPD), and the Greens). The chart uses a combination of words, numeric indicators, and a visual bar chart representation to visualize the probabilities. In this case, the target audience is a broad public, which might not have a deep or formal education in lesser-known forms of visualization or in more sophisticated ways of conveying uncertainty. Bar charts thus offer a broadly understood way of showing magnitude and enabling comparisons between estimations of probability (Spiegelhalter et al., 2011, p. 1395). The data were transformed into a probability format by dividing the simulation results that allowed for a particular coalition by all simulations. This led from a distribution to a single percentage displayed in the bar chart and as a visual label. This label was presented in a frequency format (e.g., 89 out of 100), which has been shown to make these figures easier to understand (Gigerenzer & Hoffrage, 1995; Spiegelhalter et al., 2011). In addition, the data was also split into verbal categories to translate the numbers to a qualitative rating of uncertainty (see Table 1). Words alone can lead to many categories without clear differentiation, but combined with numbers, this should lead to higher interpretability (Spiegelhalter et al., 2011, p. 1394).



Created: 01.09.21 00:17 Uhr

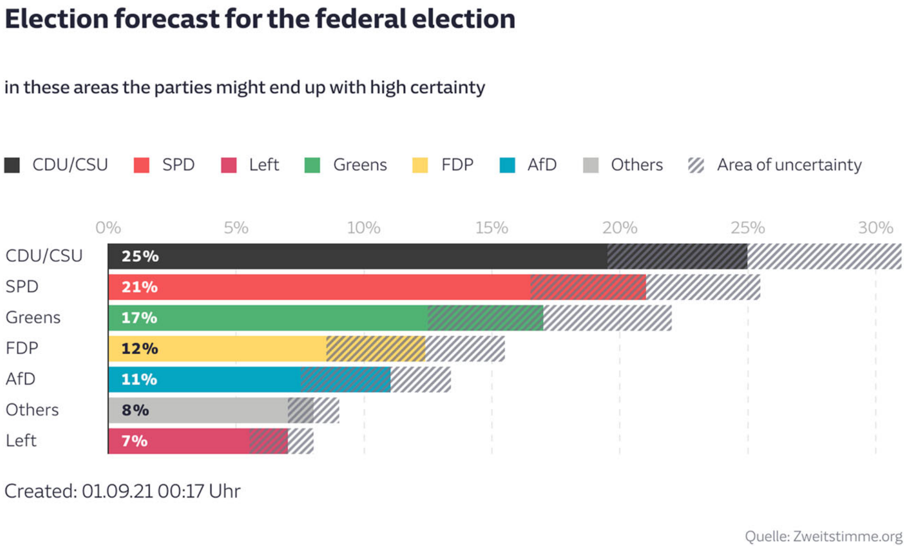
Source: Zweitstimme.org

**Figure 1.** Probabilities for certain coalition options, split into groups of verbal certainties.

The second chart (See Figure 2) reflects a traditional approach of directly visualizing vote share, as it is commonly presented in election race polls on election night. It uses point estimates derived from the model as the upper end of the bar chart, in combination with a shaded area of uncertainty around this value, which adds another uncertainty dimension to the data (Brodlie et al., 2012). This area uses a $\frac{5}{6}$ credibility interval instead of a more conventional 95 percent interval. This offers an easier way to communicate the probabilities of values falling inside, as it can be described as the result of ‘rolling anything other than a six on a fair die’ (Stoetzer et al., 2019), which seems to be more relatable to the general public. The interval is marked with gray hachures. Providing laypeople with a predictive interval forecast has been shown to be beneficial and

**Table 1.** Probabilities and their verbal expressions used to support the understandability of the visualizations.

| Probability | Verbal expression |
|---|---|
| $\geq$ 0.8 | ‘Very likely’ |
| 0.6–0.8 | ‘Likely’ |
| 0.4–0.6 | ‘Possible’ |
| 0.2–0.4 | ‘Unlikely’ |
| <0.2 | ‘Very unlikely’ |



Created: 01.09.21 00:17 Uhr

Quelle: Zweitstimme.org

**Figure 2.** Expected range and means of share per party.

understandable in other contexts (Savelli & Joslyn, 2013), but is also not undisputed (Gelman & Greenland, 2019).

### *3.2. Survey design*

This work is motivated by the overarching question of how users perceive and respond to visual election predictions published by news media. Here we detail three more specific research questions and describe how the survey was implemented to help answer them (See Appendix 1 for survey):

- **RQ 1**: Are users able to interpret the visualizations in ways consistent with the uncertainty portrayed?
- **RQ 2**: In what ways do users find value (or not) in the predictions presented?
- **RQ 3**: In what ways have the predictions presented impacted how the users think about the race?

The first question aimed to understand users' ability to interpret the presented visualizations-particularly in light of the uncertainty conveyed-by asking a set of multiple-choice questions. More specifically, we asked respondents to identify parties that might end up first (Figure 2), keeping in mind their uncertainty ranges. Another question asked users to select coalitions that are likely (Figure 1). The questions were oriented to assess users' ability to arrive at a range of interpretations consistent with the uncertainty portrayed. We acknowledge that there are several factors that might influence a respondent's interpretation of visualizations (Kennedy et al., 2016). However, we are using the questions to test if respondents are able to arrive at a reading that is consistent with the uncertainty conveyed. Although this is basic information about understandability and whether users were able to read the two charts in ways that are consistent with the uncertainty portrayed, it is also helpful to better understand the open-ended feedback provided in subsequent questions.

The second and third research questions are about the changes and thoughts such visualizations may spur in the thinking and reflection of the users. The second research question specifically addresses perceived utility of the charts, which might include aspects of usability, practicability, understandability, or other user-defined notions, and asks if the charts are meaningful, comprehensible, and reasonable. The final research question aims to examine perceived influences of these visualizations on respondents' awareness and consciousness about the electoral race. It tries to tackle the critical topic of whether presentations of election predictions create a perception of influence amongst voters. We asked users if (and if so, how) the visualizations influenced how they thought about the campaign. The survey combined multiple-choice selections with free-text inputs to allow us to collect respondents' opinions and assessments.

### 3.3. Participants

To recruit participants, a link to the survey was added to the top banner of an explanatory article about the predictive model (Witzenberger, 2021) and was also shared on Twitter by the authors. It was available between September 2nd and September 28th, 2021, with the charts in the survey remaining fixed on the values of September 1st. The tracking tool Linkpulse measured a total of 40227 page views on the explanatory article during this period. Three hundred ninety-nine users opened the survey, 192 started it, and 134 finished (33.6%).

Part of the survey asked for sociodemographic information to gain a deeper understanding of the sample (see Appendix 2). The sample was mainly male (76.6% male; 21.9% female; 1.6% other) and predominantly held a university degree (76.2%) or a high school diploma (18.5%) as the highest level of education. This is in line with the general audience of Süddeutsche Zeitung, which is regarded as left-liberal media in Germany (Hachmeister, 2012). Median age of respondents was 48 years. In terms of political favor, the sample was heavily skewed toward the Greens (49.6% in the sample, the party only got 14.8% of the vote in the most recent federal election), with disproportionately few participants who preferred the conservative parties (7.2% in the sample, 24.1% in the election) or right-wing AfD (0.8%; 10.3%). These figures caution an interpretation of our results scoped to a particular non-representative sample, a point which we return to when addressing limitations in the discussion section.

### 3.4. Analysis methods

To answer the three research questions, we used a mix-methods approach. RQ1 is addressed by analyzing the specific responses to survey responses about the visualizations. For RQs 2 and 3, we used qualitative thematic analysis to derive themes in the data, following the method outlined by Braun and Clarke (2006). It is well-suited to detecting key features of a textual dataset and to distinguishing similarities and differences in responses. We practiced iterative constant comparison of themes and analyst triangulation between co-authors to develop the themes that we report in this paper. We first applied qualitative thematic analysis to the textual data and afterward decided to count the prevalence of specific themes to better estimate how widespread certain themes were observed.

## 4. Findings

The following subsections detail our findings as they address the research questions stated above.

### 4.1. Visualization interpretation

To help answer RQ1 ‘Are users able to interpret the visualizations in ways consistent with the uncertainty portrayed?’ we analyzed quantitative answers to two basic questions about the charts. First, we asked respondents to identify all coalitions that were likely based on their reception of the chart. Secondly, we wanted respondents to identify parties that might end up in the first place of the election, keeping in mind possibly overlapping uncertainty ranges. As described below, we tried to assess respondents‘ ability to arrive at a reading that is consistent with the uncertainty displayed.

For the coalition options in chart one, we asked: ‘Based on this chart: Which are the three coalitions with the highest probability of winning a majority?’. This question was expected to be easy, as the answers were the top three coalitions in the chart. A large majority answered the questions as to be expected by the uncertainty portrayed (66.4%, $n = 89$), but 45 participants (33.6%) did not mark all three options.

A similar question was posed for the second chart, which showed the possible spread of party results. The question was: ‘Based on this chart: Which parties could win the most seats?’ Most users did not provide the expected answer (59.7%, $n = 80$), in most cases leaving out the Greens, which was a valid answer based on the uncertainty ranges depicted. Only 40.3% ($n = 54$) of users marked all three parties.

In sum, a majority of users were able to see coalition uncertainties in the first chart, whereas a majority of users had difficulty seeing and comparing the relative predicted votes shares of parties in the second chart.

### 4.2. Perceived value and perceived impact of the predictions

We next seek to develop a deeper sense of the perceived value (RQ2) and perceived impacts on how users think about the race (RQ3) of the election prediction visualizations. To do this, we analyzed three questions with free-text responses: ‘To what extent do you find the predictions presented here useful or not useful? Why?’, ‘Do you have any further feedback?’ and ‘Have the predictions presented here influenced how you think about the campaign?’. For the third question, we also present the results of a quantitative self-assessment. A broad set of answers describes the perceived value and perceived impact of the visualizations as experienced by the users. Overall, responses reflected how the displays enabled future-oriented cognition in light of uncertainty about the race and stimulated a critical, even skeptical response.

#### 4.2.1. Future-oriented cognition

A substantial minority of users ($n = 37$, 27.6%; see Figure 3) estimated that the visualizations influenced or tended to influence how they thought about the race. However, a majority ($n = 97$, 72.3%) fully or partially rejected this notion. Some ($n = 7$, 5.2%) suggested a possible influence of such displays on voting intents, referring to this as

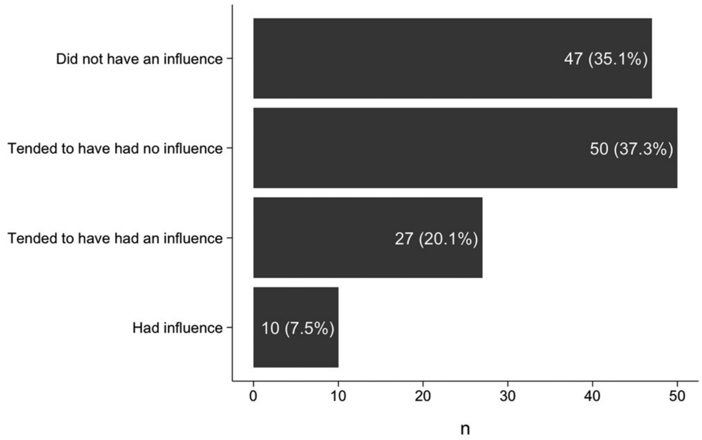


**Figure 3.** Self-attributed influence reported by the respondents on the question 'Have the predictions presented here influenced how you think about the campaign?'.

'tactical voting'. Respondents further describe those effects in ways that align with well-studied 'bandwagon' and 'reverse-bandwagon' effects (Mehrabian, 1998; Schmitt-Beck, 2015). Indeed a few respondents ($n = 3$, 2.2%) openly admitted to a change in their voting behavior by selecting another front-runner to vote for after seeing the visualizations, and another one described this as an option, though it remains unclear whether this was meant as an option for him- or herself. Another small group ($n = 3$, 2.2%) affirmed that they had already made their choice and that such predictions would not influence their vote. Some respondents also mentioned that such predictions could lead to impacts in the media, such as a reduction in issue-focused race reporting.

Some users ($n = 5$, 3.7%) pointed out that the charts 'form a picture of opinion and mood' of the current state of the race, providing a range of possible results, which makes the process of the race and its possibilities more concrete: 'it shows what is possible'. Two respondents appreciated the visualization of coalition options as possibilities because 'a government needs a majority in parliament and not necessarily a majority of the popular vote'. As the charts show a particular set of possible outcomes, some users ($n = 6$, 4.5%) mentioned that the visualizations provide a 'good source for speculation', which might be helpful for their own forecast, offer an anchoring point to compare their own opinion against, or provide a starting point for debates about possible election outcomes. More specifically, five respondents positively mentioned an increased insight into the electoral race. Getting a regularly updated chart shows 'what influences the parties' success in the election campaign, and to what extent.' Furthermore, it offers a feeling of the tension in which parties are campaigning. Two respondents described a change in perception of the front-runners, seeing shifts of public opinion that were suspected to be caused by the perception of the central party candidates.

A good number of responses ($n = 19$, 14.2%) suggested there was value in showing inherent uncertainty in election predictions, mentioning the realization that surveys and models are not an exact science. As one user put it: 'It is made clear that polls are not accurate forecasts.' Some respondents saw forecasts as more 'realistic' when presented with their uncertainty, allowing users to explore and cogitate on the range of political possibilities that may very well result from the election or understand 'the validity of the survey' ($n = 3$, 2.2%). In reflecting on uncertainty one respondent wrote, 'The election campaign [is] more open than suggested by other polls'. Two respondents argued in favor of the presentation of the probabilities of coalitions: One respondent explained that a 'panic campaign' of conservative CDU/CSU against a coalition led by a social democrat would make more sense to him or her after seeing the coalition comparison chart, which showed mainly options without CDU/CSU in the top positions.

### 4.2.2. Critical and skeptical responses

A number of users ($n = 24$, 17.9%) also expressed a more critical stance towards the presented predictions, including critiques of the polling methods ($n = 4$, 3.0%), forecast models ($n = 7$, 5.2%), the impossibility of prediction ($n = 17$, 12.6%), and calling for editorial responsibility ($n = 8$, 6.0%).

Some respondents argued that the surveys underlying the model are 'politically motivated' and could be modified to obtain the desired result. For example, two argued that polls favor the conservative CDU/CSU party. One respondent mentioned that polls are just a current snapshot of a selection of the electorate, while two other respondents doubted the methods of some polling firms, which still use landline phones and might underestimate younger parts of the voters and expect a shift in the results due to the high share of mail-in-voters.

Besides the negative judgment on the polling aspect of the model, some users targeted the model itself. Two users argued the uncertainty intervals were too broad to derive a helpful conclusion, while another respondent was surprised by the use of $\frac{5}{6}$ intervals, which are not as common as 95 percent. One respondent was missing a 'political probability' next to the coalitions, which might presuppose a comparison of matching and contradicting campaign promises between parties that could further illuminate whether those parties might actually form a government together. Another group of users ($n = 4$, 3.0%) wanted to see a more detailed methodological explanation of how the results are calculated, what the numbers are based on, or what the uncertainty intervals imply for the interpretation. One respondent questioned the model in its completeness, arguing it would not provide more clarity but show arbitrariness: 'You roll the dice for your desired results and retreat to the position with your uncertainty bars: 'I told you so' (if it goes wrong).'

A few users even questioned the premise of attempting to predict elections at all. Three pointed to the 2016 US-presidential race, which they perceived to be forecasted for a clear victory of the Democrats and Hillary Clinton, while in the end, Donald Trump became president. One respondent used COVID-19 forecasts as examples, arguing that their perceived inaccuracies led to disbelief toward forecasts more generally. Two others pointed to the past in arguing that all the data the model uses is based on historical data collection. This leads to 'overestimation of past results,' which is not justified since the election to be forecast is a future event that is influenced by current events not reflected in past data. A number of skeptics ($n = 11$, 8.2%) pointed to the impossibility of prediction caused by the expected openness of the race, with the incumbent Angela Merkel not standing for reelection, a diversification of the electorate, a lot of mail-in-voting, and a high share of undecided voters a few weeks ahead of election day. 'Predictions are shaky', wrote one respondent: 'I favor facts, clear election results.'

Finally, while some respondents ($n = 4$, 3.0%) praised the publication of methods and requested even further transparency in terms of raw data, others ($n = 5$, 3.7%) accused the visualizations of being manipulative, claiming 'the graphics are good, the intention is not', or they 'could lead voters to make wrong choices' and appealing for 'a more critical examination of election predictions and […] responsibility of the media'. One respondent put it more bluntly: 'Just stop this nonsense!'

## 5. Discussion

We have found varied results in our research on the perception of visualizations of election predictions, which lead us to three claims suggested by the data:

(1) The election predictions studied offered utility in terms of future-oriented cognition. However, we did not find a widespread perception of impact on how respondents were thinking about the campaign,
(2) There is a need for user testing of electoral predictions to ensure understandability,
(3) A not-insubstantial share of users is skeptical of data and methods, which need to be transparently explained and published.

### 5.1. No widespread perception of influence on how people thought about the campaign

An important theme that emerged from our analysis was how the election predictions supported future-oriented cognition (Szpunar et al., 2014). This includes using the visualizations as a ground for speculation and cogitation about the political situation as well as reflecting more worrisome expectations about the influence on voters' decisions. By reflecting on uncertainty and visualizing multiple options of how the election might end up, some users were stimulated to think about a multitude of outcomes, not just a single thread that is implied with the publication of singular poll results. This might help fuel more responsible speculations by offering a range of possible results grounded in data.

While a sizeable minority of respondents ($n = 37$, 27.6%) indicated that the presentations generally tended to or did influence how they thought about the campaign (including by informing opinion, increasing insight, and offering a basis for informed speculation), very few ($n = 3$, 2.2%) indicated that the predictions actually changed their voting intention or behavior. Slightly more respondents argued that these visualizations might impact others in their voting behavior ($n = 7$, 5.2%), leading to tactical voting like 'bandwagon' or 'reverse-bandwagon' effects. Interestingly, these respondents did not seem to see this problem for themselves but rather for a part of the electorate that was not described more closely. This points to an observation of a phenomenon called the 'third-person effect' (Davison, 1983), which describes a person exposed to mass media who attributes others with a greater effect from this exposure than themselves. In this case, the third-person effect may be helping to drive worry over the political implications of predictive journalism by inflating perceptions of real impacts.

Whether the presentation of polls prior to elections leads to tactical voting behavior and therefore changes the output of an electoral decision by a voter has been discussed in earlier work but has not been proven thoroughly (Schoen, 2002). Here, we also cannot offer conclusive experimental evidence of any sizeable impact of election predictions; however, our results do suggest that caution is warranted due to a small number of voters who perceived being influenced by these visualizations. Even such small numbers could have a political-relevant impact on certain elections if the race is close. As this research used self-report data, further research is needed to test the hypothesis in a more controlled way. In particular, such future studies might examine how emotional responses (Allaham & Diakopoulos, 2022) or affective forecasting (i.e., the prediction of one's future feelings in response to the predictions) might mediate outcomes. For instance, recent results from Tenenboim-Weinblatt et al. (2022) suggest that expecting a positive affective response to one's predictions could lead to increased political participation.

## 5.2. The need for evaluating electoral predictions

Our findings reflect a mixed result on the understandability of the visualizations. We did find an alignment between journalists' intentions and users' reception in the thematic analysis of the answers to the question of whether respondents found the visualizations useful. Based on our interactions with the design team, we knew that the intention was to display the current state of the race, show the inherent uncertainty of election models and their underlying data, and provide coalition probability estimates. These aspects have been favorably mentioned by respondents, indicating that a sizeable minority of respondents understood the intentions and found them useful.

At the same time, while most respondents were able to detect the most probable coalitions from a simple bar chart, a majority did not extract all the information on the possible spread of election results from the second chart. The design of the visualizations might have influenced this: the coalition chart was sorted in order, and therefore, the expected results were the top three options, while the chart on vote share required a closer look at the uncertainty intervals. A couple of respondents noted that these visualizations were not easy to see on a mobile device, which might increase the risk of overlooking the small portion of uncertainty interval that leaves the Greens a possibility to finish up first. We do not have information on the type of device used by respondents due to GDPR compliance and, therefore, cannot account for this possible explanation. Still, the results point to a widespread lack of understanding of a simple uncertainty interval for vote shares, which might benefit from further improvements drawing on recommendations from the literature (Correll & Gleicher, 2014; Hullman et al., 2015; Kale et al., 2019).

Designers of communicative visualizations often convey their main message to the audience without facilitating too many polysemic reading opportunities, clearly depicting the authors' main takeaways to convey (Segel & Heer, 2010). However, visualizations are perceived differently for a variety of reasons, particularly in broader audiences, making a single 'correct' reading of a visualization difficult to obtain (Cairo, 2012; Kennedy & Hill, 2017; Kennedy et al., 2016). Our findings above make clear that uncertainty reception can play a role in future-oriented cognition, underscoring the value of supporting a range of interpretations in light of that uncertainty. Thus we suggest here that some level of polysemy might be desirable in the domain of predictive journalism.

To show predictions and their uncertainty so that most users understand them as expected is a crucial element that can be improved through user testing, ideally before or at least iteratively with the design process after they are published. A focus should be on whether the uncertainty in the prediction was understood and whether the range of interpretations resulting from the conveyance of uncertainty was received (Adar & Lee, 2021). There might also be the need for additional training in numeracy and polling methods on the journalists' side, as proposed by Appelman and Schmierbach (2022) in regard to the presentation of opinion polls in news media.

### 5.3. Skepticism remains

Some user responses indicated that they did not find the displays useful. Various aspects of criticism and skepticism emerged, a response which has also been observed in other recent research on the perceptions of predictive journalism in the COVID-19 context (Allaham & Diakopoulos, 2022). Some respondents expressed hesitance for models based on election polling, others were skeptical about modeling an election, and still others questioned the predictability of the election at all because of the perceived openness of the race, pointing to historical mismatches between perceived assertions of the polls and election outcomes, or criticizing a model that is based on past data to predict the future. This points to the idea of 'temporal exigencies' that Pentzold and Fechner (2021) have described as constraints toward future predictions created based on historical data, which is attributed to inability to foresee upcoming developments. A small part of the audience also seems to be aware that they exist and might reduce the perceived validity of the visualization.

Some respondents even accused the journalists of attempted manipulation or called for editorial responsibility not to publish predictions to avoid influencing the audience, which has been described as an attempt at responsibility assignment in similar contexts (Allaham & Diakopoulos, 2022). These observations point to some high-level issues that have been increasingly discussed during the last couple of years: trust and credibility of journalism are declining in some audiences (Hanitzsch et al., 2017; Newman et al., 2021). Even though this project aimed at providing additional explanations, and even though the authors of the election model had published scientific papers to be transparent about the method, skepticism and disbelief remained. Additional research is needed on the general perception of journalistic forecasts as it relates to trust. It is not uncommon for journalists to look toward the future, not restricted to predictive journalism (Barnhurst & Mutz, 1997). How are users dealing with this information? Are they regarded as experts' analytic projections or as speculative opinions expressed for a broader audience?

A possible leverage point to tackle this issue might be increased transparency of methods and underlying data (Diakopoulos & Koliska, 2017). Some respondents mentioned that transparency positively affected their evaluation of the visualizations, and one asked for even more openness. As data and methods are publicly available, this is feasible -- although it would require additional design work to ensure the usability of the transparency information. This refers to a match between the aspiration for strong transparency of predictive journalism and reception by parts of the audience. Data journalists might share the code, data, and underlying assumptions more easily than classical journalists, who might not want to disclose their sources.

Others have proposed interactive visualizations that allow users to change assumptions or input data to see differences in outcomes (Pentzold & Fechner, 2019). This might help to understand the mechanics of the model, but it presumably needs an audience open for argument, not entirely rejecting electoral models. To find better ways to appeal to skeptics in the reception of visualizations of election dynamics is up to further research.

### 5.4. Limitations

Gathering answers via open text fields yields certain problems, namely giving respondents an easy option to refuse to answer, receiving only short, unclear fragments, and leaving out the possibility to ask again or more specifically. Although this form of data collection has advantages given the size of the sample, its variance, and its openness to unexpected interpretations, future work should consider complementary methods, such as guided interviews.

We must also acknowledge that our sample is not randomly selected. Users were shown a link to the survey if they visited an article that offered more information on the election model. This might appeal to and overrepresent users who found the visualization interesting but might also attract users who have a very negative sentiment toward the charts and are looking for a justification for this kind of reporting. Therefore, we expect to have included voices with either favorable or critical opinions of the figures but less moderate perspectives. Further research should construct a representative sample to draw more general conclusions on this topic. The sample was heavily skewed toward the readership of Süddeutsche Zeitung. The respondents had a higher level of education than the general population, which is often associated with more statistical training (Grotlüschen et al., 2016). As a result, we expect the understandability to be higher than for a general audience but might also have increased the sophistication of criticism, as respondents might have incorporated values taught in higher education. While this is a disadvantage on the generalizability of the results for a general population, these results still have validity for the readership of Süddeutsche Zeitung -- and lack other weaknesses that a controlled experiment might create with its artificial setting in terms of recruiting. We must also acknowledge the occasionally limited number of respondents who constituted a certain topic. We include counts and percentages to help faithfully portray the findings and acknowledge the need for future work to collect larger samples sizes for comparison and contrast.

## 6. Conclusion

There has been little ecologically valid prior research on how users respond to election predictions in news media. This work on a user survey of journalistic uncertainty displays of election models for the German parliamentary election in 2021 contributes a first analysis of users' perception of visualizations in predictive election journalism. It found a general alignment between a sizeable share of respondents and visualizations' intention to make uncertainties clearer. This indicates that a part of the audience values these charts in general, although we found a noticeable share of users who could not interpret the uncertainty conveyed in the charts effectively. We further found evidence that the predictions support future-oriented cognition, which describes the use of the visualizations by the users to make claims or fuel discussions about possible outcomes in the future. And while we did not see a widespread self-reported perception of influence on voters' decisions based on these charts, our findings suggest that more experimental research is warranted. A set of respondents expressed skepticism toward predictive data journalism, addressing criticism of polling and model methods and pointing to past perceived prediction failures, the openness of the race, or past-centric models. All of the above leads us to argue for rigorous user testing of such visualizations before publication. To address doubts or distrusts, we further suggest providing more resources for transparency and developing methods for effectively publishing and explaining models and data used to a more general audience.

## Acknowledgments

The authors would like to thank Süddeutsche Zeitung for supplying the survey data for scientific purposes.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

*Benedict Witzenberger* is a PhD student at the Professorship of Computational Social Science and Big Data at the Technical University of Munich, Germany. His research focuses on application of Computational Social Science methods on questions in Communicative Science, and the evolution of data journalism in particular.

*Nicholas Diakopoulos* is a professor in Communication Studies and Computer Science (by courtesy) at Northwestern University where he directs the Computational Journalism Lab and is director of Graduate Studies for the Technology and Social Behavior PhD program. His research focuses on computational journalism, including aspects of automation and algorithms in news production, algorithmic accountability and transparency, and social media in news contexts.

## ORCID

*Benedict Witzenberger* http://orcid.org/0000-0001-7476-3324
*Nicholas Diakopoulos* http://orcid.org/0000-0001-5005-6123

## References

Adar, E., & Lee, E. (2021). Communicative visualizations as a learning problem. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 946–956. https://doi.org/10.1109/TVCG.2020.3030375

Allaham, M., & Diakopoulos, N. (2022). Predicting COVID: Understanding audience responses to predictive journalism via online comments. *New Media & Society*, 14614448221135632. https://doi.org/10.1177/14614448221135632

Appelman, A., & Schmierbach, M. (2022). Coverage of public opinion polls: Journalists' perceptions and readers' responses. *Journalism Practice*, 1–20. https://doi.org/10.1080/17512786.2022.2058064

Barnhurst, K. G., & Mutz, D. (1997). American journalism and the decline in event-centered reporting. *Journal of Communication*, *47*(4), 27–53. https://doi.org/10.1111/j.1460-2466.1997.tb02724.x

Bell, A. (1995). News time. *Time & Society*, *4*(3), 305–328. https://doi.org/10.1177/0961463X95004003003

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brettschneider, F. (1992). Der taktische und rationale Wähler. Über den Einfluß von Wahlumfragen auf das Wählerverhalten bei den Bundestagswahlen 1983 bis 1990. *Politische Vierteljahresschrift*, *33*(1), 55–72. https://www.jstor.org/stable/24197902

Brettschneider, F. (1997). The press and the polls in Germany, 1980–1994 poll coverage as an essential part of election campaign reporting. *International Journal of Public Opinion Research*, *9*(3), 248–265. https://doi.org/10.1093/ijpor/9.3.248

Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007, May). Misinterpretations of the “cone of uncertainty” in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, *88*(5), 651–668. https://doi.org/10.1175/BAMS-88-5-651

Brodlie, K., Allendes Osorio, R., & Lopes, A. (2012). A review of uncertainty in data visualization. In J. Dill, R. Earnshaw, D. Kasik, J. Vince, & P.C. Wong (Eds.), *Expanding the frontiers of visual analytics and visualization* (pp. 81–109). Springer. https://doi.org/10.1007/978-1-4471-2804-5_6

Cairo, A. (2012). *The functional art*. New Riders Publishing.

Coddington, M. (2015). Clarifying journalism’s quantitative turn. *Digital Journalism*, *3*(3), 331–348. https://doi.org/10.1080/21670811.2014.976400

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, *47*(1), 1. https://doi.org/10.1086/268763

Diakopoulos, N. (2022). Predictive journalism: On the role of computational prospection in news media. *Tow Center for Digital Journalism*. https://doi.org/10.2139/ssrn.4092033

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dieckmann, N. F., Peters, E., & Gregory, R. (2015). At home on the range? Lay interpretations of numerical uncertainty ranges. *Risk Analysis*, *35*(7), 1281–1295. https://doi.org/10.1111/risa.2015.35.issue-7

Faas, T., & Schmitt-Beck, R. (2007). Wahrnehmung und Wirkungen politischer Meinungsumfragen. Eine Exploration zur Bundestagwahl 2005. In F. Brettschneider, O. Niedermayer, & B. Weßels (Eds.), *Die bundestagswahl 2005* (pp. 233–267). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90536-5_11

Gelman, A., & Greenland, S. (2019). Are confidence intervals better termed “uncertainty intervals”? *BMJ*, 366. https://doi.org/10.1136/bmj.l5381

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Graefe, A., & Jérôme, B.. (2022). Forecasting the 2021 German Federal Election. An Introduction. *PS: Political Science & Politics*, *55*(1), 61–63. https://doi.org/10.1017/S1049096521001001

Grotlüschen, A., Mallows, D., Reder, S., & Sabatini, J. (2016). Adults with low proficiency in literacy or numeracy. *OECD Education Working Papers* (131). Retrieved from https://www.oecd-ilibrary.org/content/paper/5jm0v44bnmnx-en

Hachmeister, L.. (2012). Süddeutsche Zeitung. Retrieved April 23, 2022, from https://www.mediadb.eu/forum/zeitungsportraets/sueddeutsche-zeitung.html.

Hanitzsch, T., Dalen, A. V., & Steindl, N. (2017). Caught in the nexus: A comparative and longitudinal analysis of public trust in the press. *The International Journal of Press/Politics*, *23*(1), 3–23. https://doi.org/10.1177/1940161217740695

Hermida, A., & Young, M. L. (2019). *Data journalism and the regeneration of news*. London, United Kingdom. Retrieved September 25, 2020, from https://www.routledge.com/Data-Journalism-and-the-Regeneration-of-News/Hermida-Young/p/book/9781138058934

Holtz-Bacha, C. (2012). Opinion polls and the media in Germany: A productive but critical relationship. In C. Holtz-Bacha & J. Strömbäck (Eds.), *Opinion polls and the media* (pp. 93–112). Palgrave Macmillan UK. https://doi.org/10.1057/9780230374959_5

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *P—LoS ONE*, *10*(11), e0142444. https://doi.org/10.1371/journal.pone.0142444

Jaworski, A., Fitzgerald, R., & Morris, D. (2004). Radio leaks: Presenting and contesting leaks in radio news broadcasts. *Journalism*, *5*(2), 183–202. https://doi.org/10.1177/146488490452003

Joslyn, S., & Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, *17*(2), 180–195. https://doi.org/10.1002/met.v17:2

Kale, A., Nguyen, F., Kay, M., & Hullman, J. (2019). Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 892–902. https://doi.org/10.1109/TVCG.2945

Kennedy, H., & Hill, R. L. (2017). The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology*, *52*(4), 830–848. https://doi.org/10.1177/0038038516674675

Kennedy, H., Hill, R. L., Allen, W., & Kirk, A. (2016). Engaging with (big) data visualizations: Factors that affect engagement and resulting new definitions of effectiveness. *First Monday*, *21*(11). https://doi.org/10.5210/fm.v21i11.6389

Lippmann, W. (1922). *Public opinion*. Free Press.

Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2017). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*, *21*(9), 1246–1263. https://doi.org/10.1177/1464884917735691

Maycotte, H. O. (2015). *Big data triggers predictive journalism*. Retrieved December 29, 2021, from https://www.niemanlab.org/2015/12/big-data-triggers-predictive-journalism/

Mehrabian, L. (1998). Effects of poll reports on voter preferences. *Journal of Applied Social Psychology*, *28*(23), 2119–2130. https://doi.org/10.1111/jasp.1998.28.issue-23

Munzert, S., Stötzer, L., Gschwend, T., Neunhoeffer, M., & Sternberg, S. (2017). Zweitstimme.org. Ein strukturell-dynamisches Vorhersagemodell für Bundestagswahlen. *Politische Vierteljahresschrift*, *58*(3), 418–441. https://doi.org/10.5771/0032-3470-2017-3

Neiger, M. (2007). Media oracles: The cultural significance and political import of news referring to future events. *Journalism*, *8*(3), 309–321. https://doi.org/10.1177/1464884907076464

Neiger, M., & Tenenboim-Weinblatt, K. (2016). Understanding journalism through a nuanced deconstruction of temporal layers in news narratives. *Journal of Communication*, *66*(1), 139–160. https://doi.org/10.1111/jcom.12202

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters institute digital news report 2021*. Reuters Institute for the Study of Journalism. Retrieved June 27, 2023 from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Pentzold, C., & Fechner, D. (2019, October). Data journalism's many futures: Diagrammatic displays and prospective probabilities in data-driven news predictions. *Convergence: The International Journal of Research into New Media Technologies*. https://doi.org/10.1177/1354856519880790

Pentzold, C., & Fechner, D. (2021). Probabilistic storytelling and temporal exigencies in predictive data journalism. *Digital Journalism*, *9*(6), 715–736. https://doi.org/10.1080/21670811.2021.1878920

Pentzold, C., Fechner, D. J., & Zuber, C. (2021). “Flatten the curve”: Data-driven projections and the journalistic brokering of knowledge during the COVID-19 crisis. *Digital Journalism*, *9*(9), 1367–1390. https://doi.org/10.1080/21670811.2021.1950018

Savelli, S., & Joslyn, S. (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, *27*(4), 527–541. https://doi.org/10.1002/acp.v27.4

Schlesinger, P. (1978). *Putting “reality” together: BBC news*. Constable.

Schmitt-Beck, R. (2015). Bandwagon effect. In *The international encyclopedia of political communication* (pp. 1–5). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118541555.wbiepc015

Schoen, H. (2002). Wirkungen von Wahlprognosen auf Wahlen. In T. Berg (Ed.), *Moderner wahlkampf* (pp. 171–191). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-95052-9_9

Segel, E., & Heer, J. (2010). November narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Shedden, D. (2014). *Today in media history: In 1952, a computer helped CBS predict the winner of the presidential election*. Retrieved February 20, 2022, from https://www.poynter.org/reporting-editing/2014/today-in-media-history-in-1952-a-univac-computer-helped-cbs-news-predict-the-winner-of-the-presidential-election/

Silver, N. (2012) *The signal and the noise: Why so many predictions fail–but some don't*. Penguin Press.

Silver, N.. (2020). How fivethirtyeight's 2020 presidential forecast works - and what's different because of COVID-19. Retrieved January 14, 2022, from https://fivethirtyeight.com/features/how-fivethirtyeights-2020-presidential-forecast-works-and-whats-different-because-of-covid-19/

Solop, F. I., & Wonders, N. A. (2016). Data journalism versus traditional journalism in election reporting: An analysis of competing narratives in the 2012 Presidential election. *Electronic News*, *10*(4), 203–223. https://doi.org/10.1177/1931243116656717

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400. https://doi.org/10.1126/science.1191181

Stoetzer, L. F., Neunhoeffer, M., Gschwend, T., Munzert, S., & Sternberg, S. (2019). Forecasting elections in multiparty systems: A Bayesian approach combining polls and fundamentals. *Political Analysis*, *27*(2), 255–262. https://doi.org/10.1017/pan.2018.49

Strömbäck, J. (2012). The media and their use of opinion polls: Reflecting and shaping public opinion. In C. Holtz-Bacha & J. Strömbäck (Eds.), *Opinion polls and the media* (pp. 1–22). Palgrave Macmillan UK. https://doi.org/10.1057/9780230374959_1

Szpunar, K. K., Spreng, R. N., & Schacter, D. L. (2014). A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition. *Proceedings of the National Academy of Sciences*, *111*(52), 18414–18421. https://doi.org/10.1073/pnas.1417144111

Tenenboim-Weinblatt, K., Baden, C., Aharoni, T., & Overbeck, M. (2022). Affective forecasting in elections: A socio-communicative perspective. *Human Communication Research*, *48*(4), 553–566. https://doi.org/10.1093/hcr/hqac007

Thurman, N. (2019). Computational journalism. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The handbook of journalism studies* (p. 16). Routledge/Taylor & Francis Group.

Toff, B. (2019). The 'Nate Silver effect' on political journalism: Gatecrashers, gatekeepers, and changing newsroom practices around coverage of public opinion polls. *Journalism*, *20*(7), 873–889. https://doi.org/10.1177/1464884917731655

Urminsky, O., & Shen, L. (2020). High chances and close margins: How equivalent forecasts yield different beliefs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3448172

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*(5), 181870. https://doi.org/10.1098/rsos.181870

Westwood, S. J., Messing, S., & Lelkes, Y. (2020). Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, *82*(4), 1530–1544. https://doi.org/10.1086/708682

Witzenberger, B. (2021). Wahlprognose: Wie die SZ die Bundestagswahl vorhersagt. *Süddeutsche Zeitung*. Retrieved December 31, 2021, from https://www.sueddeutsche.de/politik/bundestagswahl-prognose-vorhersage-1.5385101

Zelizer, B., & Tenenboim-Weinblatt, K. (2014). *Journalism and memory*. Palgrave Macmillan.

# Appendices

## Appendix 1. Questionnaire

### *A.1. English version*

The original German questionnaire is available from the authors.

(1) Based on this chart: Which are the three coalitions with the highest probability of winning a majority?
- Kenia (CDU/CSU, SPD and Greens)
- Deutschland (CDU/CSU, SPD and FDP)
- Jamaica (CDU/CSU, Greens and FDP)
- Ampel (*Traffic light*, SPD, FDP and Greens)
- Grand Coalition (*GroKo*, CDU/CSU and SPD)
- Red-Red-Green (SPD, Left and Green)
- Black-Green (CDU/CSU and Greens)
- Red-Green (SPD and Greens)
- Black-Yellow (CDU/CSU and FDP)

(2) Based on your opinion: Which coalition will form the next government? Why?
- Kenia (CDU/CSU, SPD and Greens)
- Deutschland (CDU/CSU, SPD and FDP)
- Jamaica (CDU/CSU, Greens and FDP)
- Ampel (*Traffic light*, SPD, FDP and Greens)
- Grand Coalition (*GroKo*, CDU/CSU and SPD)
- Red-Red-Green (SPD, Left and Green)
- Black-Green (CDU/CSU and Greens)
- Red-Green (SPD and Greens)
- Black-Yellow (CDU/CSU and FDP)

(3) Can you justify your opinion?

(4) Based on this chart: Which parties could win the most seats??
- CDU/CSU
- SPD
- Greens
- FDP
- AfD
- Left
- Others

(5) Based on your opinion: Which party will win the most seats?
- CDU/CSU
- SPD
- Greens
- FDP
- AfD
- Left
- Others

(6) Can you justify your opinion?

(7) With this presentation, we want to educate our readers as much as possible about the federal election, the current forecasts and the uncertainties associated with it. Does this make sense to you?
- Yes, I was aware of this.
- No, I was not aware of this.

  (a) You have indicated that the presentation has no added value for you. What could we do to improve it?

(b) You indicated that you did not perceive or understand the areas of uncertainty. What do we need to change to make this clearer?

(8) Please indicate whether you agree with the following statement: 'I like that transparency is created with the presentation of the uncertainty areas'.
- I agree fully.
- I agree partially.
- I disagree partially.
- I disagree fully.

(9) To what extent do you find the predictions presented here useful or not useful? Why?
- Userful.
- Rather useful.
- Rather not useful
- Not useful at all

(10) Can you justify your opinion?

(11) Do you have any further feedback?
- (a) Are there other aspects that you like about these graphics?
- (b) How could we improve these graphics for you?

(12) Have the predictions presented here influenced how you think about the campaign?
- Has influenced.
- Has rather influenced.
- Has rather not influenced.
- Has not influenced.

(a) How have the predictions affected you in your thinking about the campaign?

(13) Which of the following parties do you lean most toward?
- CDU/CSU
- SPD
- Greens
- FDP
- AfD
- Left
- Others

(14) What is your gender?
- male
- female
- other

(15) In what year have you been born?

(16) What is your highest educational diploma?
- I am still at school.
- School finished without graduation.
- Secondary school diploma.
- Realschule (middle school leaving certificate)
- Abitur or extended high school with graduation (12th grade) (Fach-/Hochschulreife)
- Study degree

## Appendix 2. Composition of sample

**Table B1.** Gender composition of the sampled users.

| Gender | Count | Share |
|---|---|---|
| Male | 98 | 76.56% |
| Female | 28 | 21.88% |
| Diverse | 2 | 1.56% |
| NA | 0 | 0% |
| **Sum** | **128** | **100.00%** |

**Table B2.** Age variation of the sample.

| Average | Median | 25-Percentile | 75-Percentile | *n* |
|---|---|---|---|---|
| 47.10 | 48 | 31.00 | 61.25 | 116 |

**Table B3.** Educational composition of the sample.

| Highest Education | Count | Share |
|---|---|---|
| University degree | 99 | 76.15% |
| High school diploma | 24 | 18.46% |
| Realschule (mittlere Reife) | 6 | 4.62% |
| Still in college | 1 | 0.77% |
| NA | 0 | 0% |
| **Sum** | **130** | **100.00%** |

**Table B4.** Party leaning of the sample.

| Party leaning | Count | Share | Result 2021 election |
|---|---|---|---|
| Grüne | 62 | 49.60% | 14.8% |
| SPD | 28 | 22.40% | 25.7% |
| FDP | 12 | 9.60% | 11.5% |
| Linke | 11 | 8.80% | 4.9% |
| CSU/CDU | 9 | 7.20% | 24.1% |
| Other | 2 | 1.60% | 8.7% |
| AfD | 1 | 0.80% | 10.3% |
| **Sum** | **125** | **100.00%** | |

# Popular and on the Rise - But Not Everywhere: COVID-19-Infographics on Twitter

# Popular and on the Rise - But Not Everywhere: COVID-19-Infographics on Twitter

### Authors

Benedict Witzenberger, Angelina Voggenreiter, Jürgen Pfeffer

### In

International Data Science Conference, Krems, Austria,
published as:
Benedict Witzenberger, Angelina Voggenreiter, and Jürgen Pfeffer (Jan. 4, 2024). "Popular and on the Rise—But Not Everywhere: COVID-19-Infographics on Twitter". In: *Data Science—Analytics and Applications. Proceedings of the 5th International Data Science Conference—iDSC2023*. Ed. by Peter Haber, Thomas J. Lampoltshammer, and Manfred Mayr. Cham: Springer Nature Switzerland, pp. 53–60. ISBN: 9783031421709. DOI: 10.1007/978-3-031-42171-6_7

### Abstract

The coronavirus pandemic has altered many industries around the world. Journalism is one of them. Especially data journalists have gained attention within and outside of their newsrooms. We aim to study the prevalence of journalistic data visualizations before and after COVID-19 in 1.9 million image posts of news organizations on Twitter across six countries using a semi-manual detection approach. We find an increase in the shares of tweets containing infographics. Although this effect is not consistent across countries, we find increases in the prevalence of COVID-19-related content and interactions in infographics throughout all geographies. This study helps to generalize existing qualitative research on a larger, international scale.

### Contribution of thesis author

Theoretical operationalization, data collection, computational analysis, manual coding, contextualization, manuscript writing, revision, and editing.

### Publication Summary

COVID-19 is attributed to having increased data journalistic reporting. However, this is mostly anecdotal evidence based on subjective observations, which are mostly limited to media systems that the authors are familiar with. This paper aims to understand better the prevalence of infographic journalism in six different regions. Computational Social Science methods enable standardized data analysis across countries, allowing for comparisons and determining how much the pandemic has affected data-driven journalism.

# Popular and on the Rise — But Not Everywhere: COVID-19-Infographics on Twitter

Benedict Witzenberger Angelina Mooseder Jürgen Pfeffer

April 11, 2023

The coronavirus pandemic has altered many industries around the world. Journalism is one of them. Especially data journalists have gained attention within and outside of their newsrooms. We aim to study the prevalence of journalistic data visualizations before and after COVID-19 in 1.9 million image posts of news organizations on Twitter across six countries using a semi-manual detection approach. We find an increase in the shares of tweets containing infographics. Although this effect is not consistent across countries, we find increases in the prevalence of COVID-19-related content and interactions in infographics throughout all geographies. This study helps to generalize existing qualitative research on a larger, international scale.

***Keywords***— COVID-19 data visualization data journalism.

## 1. Introduction

COVID-19 served as an accelerator for ongoing changes in journalism: the decline of print and other forms of "traditional" media, the rise of "alternative" news channels, altered skill requirements for journalists, changes in audience and their expectations [19]. Data journalists were central to some of these innovations, as they had the experience and technical means to create data visualizations and exploratory pieces on possible scenarios, which has increased awareness and accessibility to the numbers and fostered engagement of the audience [17]. We define data journalism as the use of data, quantitative analysis, and visualization methods to create journalism [2].

To shed light on the changes that were going on in journalistic data visualization around the world, we analyzed infographics shared by news media before and after COVID-19 hit as a proxy for the prevalence of data journalism. We define information graphics (short: infographics) as a graphical composition of one or more visualizations based on numerical data, images, and text [4].

We aim to answer the following research questions:

**RQ1**: How has the use of journalistic infographics changed during COVID-19?

**RQ2**: Which change in the prevalence of journalistic infographics can be found across different countries?

**RQ3**: How large is the portion of COVID-19 related infographics?

**RQ4**: How do tweet interactions change in tweets containing infographics compared to other image tweets?

## 2. Literature Review

We place our research on the influence of infographics in the news and the impact of COVID-19 on data journalism.

#### 2.0.1. Data journalism during COVID-19

Data journalism is regarded as a form of content or genre innovation in journalism [7], which might provide news companies with an increased reputation or a competitive advantage. Demand for data journalistic training has increased, and data journalists have grown in power and reframed their roles and identity in the newsroom [10]. A positive attitude towards data journalism in the newsrooms correlates with enjoying working with numbers and the belief that competency for data work is satisfactory [3].

The number of published data visualizations during COVID-19 led to criticism about an "information overload" [11], a "bombardment" with visualizations [8] and a very small number of sources, as governmental actors were the main data providers during COVID-19 [14]. Journalists use a small set of authoritative sources, which in turn gain authority by being used in media. To stand out, some media outlets performed their own data collection [5].

The general meaningfulness of leveraging COVID-19-related infographics has been examined in the literature. Visual communication can increase the understandability of highly-scientific topics like the spread of a virus for less educated groups [9], and might lead to higher acceptance for and adherence to protection measures. Research on the effects of infographics during COVID-19 could find positive effects on users' knowledge of mask-wearing techniques, and increased trust compared to text-only guidance, but no substantial effect on COVID-19-related anxiety [6]. Others found a positive effect of infographics on the intention to get a COVID-19 vaccination [20].

### 2.0.2. Infographics, journalism and COVID-19

Infographics are sometimes described as a method to integrate big data into journalism [21], although this does not necessarily points to the volume of data, which might not really be "big", but aspects like variety (of sources and data types) or veracity of information (in comparison to more traditional ways of reporting). Infographics seem to lead to increased news elaboration and increase more favorable news evaluation [12].

There are three common types of infographics — principle representation (or explanatory visualizations, which explain how things work), cartographic infographics (which show where things are), and statistics charts (which show how many things there are) [22]. We only focus on cartographic and statistic charts here, which are based on some form of numeric data, while principle representations are not necessarily grounded on datasets.

Yet, data journalists have been criticized for COVID-19 charts that might "make the world look more 'fixed' than it really is" [16] with maps not accounting for population densities, models not reporting their underlying assumptions, or exclusion of communities at the margins or the Global South [15].

## 3. Method

To analyze the diffusion of infographics, we collected data from Twitter, which news media uses mainly as a one-way communication channel to promote reporting [13]. We then implemented a semi-manual approach for infographic detection.

### 3.0.1. Twitter Collection

We collected accounts for the five largest, national, general-audience news media across six different countries by circulation: USA, UK, Germany, France, Italy[1]. To allow some variety of cultural backgrounds while still allowing authors to manually code and

[1] Sources for circulation numbers: Alliance for Audited Media (USA) via pressgazette.co.uk/news/us-newspaper-circulations-2022, ABC (UK): www.abc.org.uk, IVW e. V. (Germany): www.ivw.de, ACPM (France): www.acpm.fr, FIEG (Italy): www.fieg.it, ABC (India): www.auditbureau.org, RNI: rni.nic.in.

understand the content, three English-speaking newspapers from India were included. See Appendix A for accounts.

Tweets were retrieved using `from:USERNAME has:images` on Twitter's API v2 [18]. 2,205,025 tweets for this query were collected for the time range between January 1st, 2018, and July 31st, 2022, in the time between August 15th and September 3rd, 2022. However, contrary to the expected returns, not all these contained images. In total, we could download 1,911,496 images for analysis, either in JPEG or PNG format. The time range was selected to allow comparable time periods before and after the COVID-19 pandemic started.

### 3.0.2. Identifying Characteristics of Infographics

In line with the definition above, we defined infographics as images containing infographic elements, such as diagrams, maps, or explanatory illustrations, as well as text. To be able to detect infographics within our dataset, we first had to define the typical characteristics of an infographic. Therefore, we created a labeled test set consisting of 600 infographics and 1000 non-infographics (as typically more non-infographics than infographics are published by media accounts). The size of the test set allowed us to include a wide variety of images with very different image characteristics, as well as to include non-infographics, which looked very similar to infographics, and vice versa. Using this test set, we identified typical characteristics of infographics and optimized the image characteristic parameters in a way so that **all** infographics of the test set would be extracted (to the price of non-infographics being detected as well):

- **Image type**: If the image was a .png-file, it was likely to be created or edited on a computer and, therefore, likely no common photograph but an image containing graphical elements. For example, 39% of infographics in the test set were PNG-images, while only 14% of non-infographic were PNG-images. We consequently extracted all .png-images [1].

- **Colours**: While photographs typically contain a wide color range, as shadows and light conditions create many different shades of colors in objects, infographics typically consist of a few different graphical elements in a few different colors. It should be noted that transitions between graphical elements can also result in many different color shades, but these colors typically span only a few pixels. We thus extracted all images consisting of few colors spanning a wide area of pixels. In particular, we calculated an RGB histogram of the grayscale image, selected the maximum amount of pixels *pix_max* one color would span, detected the number of colors $n$ spanning at least one-quarter of *pix_max* pixels and extracted the image if $n > 30$. While 98% of infographics in the test set fulfilled this attribute, only 55% of non-infographics did so.

- **Edges**: Infographics mostly contain graphical elements such as text boxes or diagram axes, which can be detected by determining the existence of longer lines in the image. Thus, we extracted all images which included a line spanning at least one-sixth of the minimum of the image length and width. While 94% of infographics in the test set contained such a line, only 64% of non-infographics did so as well.

By extracting all images, which would include at least one of these image characteristics, **all** infographics of the test set could be extracted (resulting in 80% of non-infographics being pulled as well).

From these images, we extracted all images which contained text. We used the *pytesseract* optical character recognition package[2] to detect the existence of text within the image, but as sometimes text in specific fonts was not recognized by the system, around 10% of infographics in the test set were excluded by this step. At the same time, this step was crucial to exclude more non-infographics, as 64% of non-infographics in the test set could be excluded after this step.

### 3.0.3. Labelling Images

After identifying the characteristics of an infographic, we applied these to the dataset of Twitter images by using a semi-automatic approach, as illustrated in Figure 1: First, we extracted all images, including at least one of the infographic characteristics described above (image type, colors, edges), which reduced the initial dataset of our Twitter images to 71%. Second, we extracted all images containing text, which decreased the complete dataset size to 16% of the initial dataset. Thirdly, as mentioned above, this process would allow us to detect most infographics at the price of extracting a large number of non-infographics as well. These non-infographics had to be excluded by human inspection. Consequently, we distributed the remaining images to four trained human annotators, who manually excluded all images not being an infographic. The annotators were instructed to focus on cartographic or statistical charts, which needed to be grounded on numeric data and were not solely a text containing a single number, but a visual representation of data. The step of manual inspection reduced the dataset size to 1% of the initial size.

[2] https://pypi.org/project/pytesseract/

Figure 1: Infographic detection and labeling process.

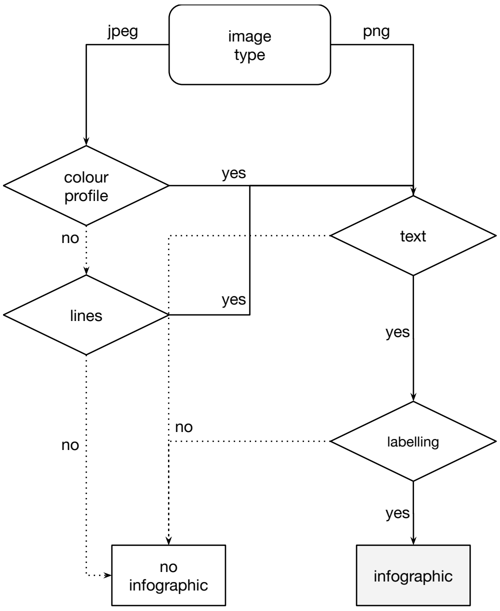


Table 1: Comparison of predicted labels (with semi-automatic approach) vs. human-coded, actual labels in a testset with n = 2500.

| | **Actual Condition** | |
|---|---|---|
| **Predicted Condition** | Infographics | Non-Infographics |
| Infographics | 26 | 4 |
| Non-Infographics | 8 | 2462 |

Finally, we evaluated this semi-automatic approach by creating a random subset of 2,500 images, labeling these images with the semi-automatic approach (to infer the predicted labels), manually inspecting these images (to infer the true labels), and comparing the labels. Our approach showed a sensitivity of 0.765 (26 out of 34), a specificity of 0.998 (2462 out of 2466), an accuracy of 0.995, and an F1-score of 0.813 (see Table 1).

## 4. Results

Out of the 1,911,496 images we analyzed, we found 25,813 infographics using the semi-automatic approach described above.

### 4.0.1. An increase in infographics — but not everywhere

The share of infographics within all shared media images increased significantly after the COVID-19 pandemic hit. We defined 'after COVID-19‘ by tweets published after March 1st, 2020 — when the pandemic as a journalistic topic had spread worldwide. We found an increase of 42 percent between pre- and post-pandemic infographic proportions from 1.2% (n = 10,652) to 1.6% (n = 15,161) ($\chi^2$ = 742.98, p <0.01, df = 1). Still, their share within all images was just around 1.6 percent, with most images (98.4%) remaining non-infographics. While this seemed to be a clear direction, we found differences when splitting the data into the observed countries.

Not all countries had similar increases in infographics. We found that media in the US had the largest absolute increase of infographics with 2.1 percentage points from 5.7% before the pandemic to 7.8% after. The highest proportional increase was found in India, which increased the share by 130 percent from 1% to 2.3%. In the UK, we found similar relative growths of around 125 percent.

In contrast, the share of infographics remained constant in German media at 0.7 percent. In Italy and France, we found fewer infographics after COVID-19: Italy decreased by around ten percent to 1.9%, and in France, the share dropped from 2.8% before COVID-19 to 1.9% after, reducing by around 32 percent.

### 4.0.2. COVID-19 is a prominent topic in tweets

To further understand the content of the infographics, we analyzed the 50 most-used hashtags for each country. These were manually coded into six categories: COVID-19, politics, ukraine, elections, sports, and others. Ambiguous terms were added to the most distinguishing category (“Biden” to ‘elections’, “putin” to ‘ukraine’). As COVID-19 brought up a set of new, distinctive words, its category seemed very unequivocal.

COVID-19 was found in between 8.1 and 23 percent of all tweets in our sample after the start of the pandemic. This also holds if only regarding infographics tweets, where COVID-19-related infographics made up 27.8 percent (n = 4,220) of all 15,161 detected infographics. While everywhere the share of tweets about COVID compared to other topics was higher for non-infographics than for infographics, this difference was very small in France (11.1% infographics, 13% non-infographics COVID-19-related). However, 63 percent of tweets could not be attributed.

When only regarding infographics after COVID-19 hit, we found that nearly half of infographics in German media were COVID-related, 33.2% in Italy, 30.8% in India, and 26.2% in France (see Figure 2). The UK and especially the US had smaller shares for

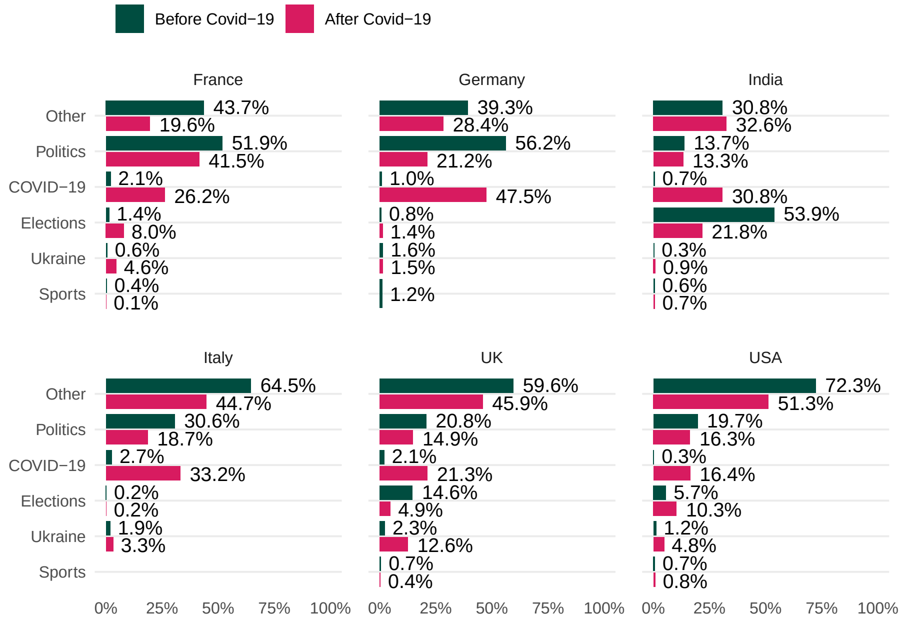


Figure 2: Share of infographic tweets before and after COVID by clustered topics.

COVID-19.

### 4.0.3. Infographic tweets receive higher interactions

From a social media perspective, infographics are a valuable tool for media organizations that seek to promote their content via Twitter. We found that mean counts for likes (37.12 versus 66.14), retweets (10.05 vs. 28.89), quotes (3.01 vs. 6.65), and replies (6.05 vs. 7.11) were significantly higher for tweets containing an infographic (p <0.001).

The audience's interest was also visible when comparing infographics before and after COVID-19. We found that after COVID, mean likes (31 versus 90), quotes (5.2 versus 7.6), and replies (4.8 versus 8.7) counts are significantly greater than before (p <0.001). Retweets showed a non-significant increase from 21.2 to 34.3. While we cannot account for changes in followers, as data collection took place at a fixed point in time, we could see that the increases in likes, retweets, and quotes were much higher for infographics than for non-infographics.

# 5. Discussion

COVID-19 had a solid effect on the prevalence of journalistic infographics on Twitter. We have found an increase in the use of infographics on media's Twitter pages after the start of the COVID-19 pandemic overall, addressing **RQ1**. This increased output of infographics is in line with qualitative literature [11, 8] that labeled these visualizations an "infodemic". Compared to the total number of tweets, however, infographics still only make up a small portion.

For **RQ2**, we found differences between countries. While there were substantial increases for US, Indian, and UK media, stagnation occurred in German media, and we surprisingly found slight declines in France and Italy. We can rule out some possible explanations for this: First, there might be a general difference in the reception of COVID-19 across the studied countries. However, we found much COVID-19-related content when analyzing the infographic tweets' texts across all countries — which also addresses **RQ3**. This leads us to believe that there was a consensus among journalistic infographic designers to produce COVID-19-related content in all observed countries. Second, some media might have different strategies for promoting their content on Twitter. While we cannot control for this from the outside perspective we have taken, we can show for **RQ4** that image tweets containing an infographic receive higher interactions. From a media distribution standpoint, it is rational to use these graphics on Twitter.

Nonetheless third, some media might have different approaches to posting data on social media. While manually labeling the automatically detected infographics, we found "text boards" in many instances. A computational, infographic-like image that contains numbers in textual format, mostly combined with images. These can be regarded as having an infographical appearance and could serve as a substitute for creating charts, which leads to higher requirements for data collection and analysis. From a definitional point of view, we decided not to include these images, as they do not contain charts but only display single numbers. As we only attributed binary labels to the infographics, we cannot control whether this has had a huge influence on the analysis.

# 6. Limitations

To explain precisely what led to the stagnation in Germany and the decline in Italy and France would require more insights into the newsrooms to rule out editorial decisions which are not visible from the outside. Some qualitative work has already been accomplished around this [5], restricted in scope and generalizability, however, by the efforts that a qualitative study requires.

This study is limited by several factors: Our approach left us relying on Twitter to detect images correctly. As we collected all data at one point in time, tweets that had been deleted could not be used for this study. The focus on six countries might have included strong influences of western-democratic media business that might not be applicable elsewhere. Although Indian media also followed the trend, restricting it to English-speaking media might have influenced the outcome, as others have discovered differences between western democracies and the Global South [14, 15].

Further research might focus on a larger variety of non-western countries to enhance understanding of possible differences in media cultures around the publication of infographics. It might also be beneficial to develop quantitative methods to detect publication differences within certain media markets, which is a field that is mostly covered by qualitative work and is hard to generalize and transfer to other populations.

In addition, our semi-automatic approach was restricted by the quality of text detection. As described in section 3.0.2, the existence of text within an image was a critical, required factor in differentiating between infographics and non-infographics, but at the same time, text detection failed in around 10% of infographics. Future research could use more elaborated text detection techniques to also take small, hardly readable, and non-standard texts (e.g. text in the form of word art) into account. This also limits our results which did not include infographics without text, which, however, we only expect to appear in very few cases, as data visualizations usually need some form of textual integration.

## 7. Conclusion

COVID-19 has influenced innovation in journalism in a lot of ways. We presented a quantitative study on the use of infographics on Twitter before and after COVID-19, which confirms earlier qualitative research. We saw an increase in infographics in our sample of image posts by the largest newsrooms in three of the six researched countries. However, in some countries, they declined. Nonetheless, we found an increase in COVID-19-related content, which is high across all geographies studied for image tweets in general, and infographic tweets in particular. Interactions for infographic tweets are higher than for images. This remains a topic for further research with deeper insights into newsroom practices. News organizations could adapt to COVID-19 in various ways. Increased use of infographics is just one of these developments — and might further influence reporting with potential consequences on the perception of journalism.

### 7.0.1. Acknowledgements

The authors would like to thank Sophie Brandt, Linus Mosch, and Sarah Talbi for their help with labeling the visualizations.

## References

[1] Adler, M., Boutell, T., Brunschen, C., Costello, A.M., Crocker, L.D., Dilger, A., Fromme, O., Gailly, J.l., Herborth, C., Jakulin, A., Kettler, N., Lane, T., Lehmann, A., Lilley, C., Martindale, D., Mortensen, O., Pickens, K.S., Poole, R.P., Randers-Pehrson, G., Roelofs, G., van Schaik, W., Schalnat, G., Schmidt, P., Wegner, T., Wohl, J.: PNG (Portable Network Graphics) Specification. Version 1.0. W3C (1996), https://www.w3.org/TR/REC-png-961001

[2] Anderton-Yang, D., Kayser-Bril, N., Howard, A., Teixeira, C.V., Slobin, S., Vermanen, J.: Why is data journalism important? In: Gray, J., Bounegru, L., Chambers, L., European Journalism Centre, Open Knowledge Foundation (eds.) The data journalism handbook 1. European Journalism Centre (2012), https://datajournalism.com/read/handbook/one/introduction/why-is-data-journalism-important

[3] Appelgren, E., Nygren, G.: Data Journalism in Sweden. Digital Journalism **2**(3), 394–405 (Jul 2014). https://doi.org/10.1080/21670811.2014.884344

[4] Cairo, A.: The Functional Art. New Riders Publishing (2012)

[5] Desai, A., Nouvellet, P., Bhatia, S., Cori, A., Lassmann, B.: Data journalism and the COVID-19 pandemic: opportunities and challenges. The Lancet Digital Health **3**(10), 619–621 (2021). https://doi.org/10.1016/S2589-7500(21)00178-3

[6] Egan, M., Acharya, A., Sounderajah, V., Xu, Y., Mottershaw, A., Phillips, R., Ashrafian, H., Darzi, A.: Evaluating the effect of infographics on public recall, sentiment and willingness to use face masks during the COVID-19 pandemic: a randomised internet-based questionnaire study. BMC Public Health **21**(1), 367 (2021). https://doi.org/10.1186/s12889-021-10356-0

[7] García-Avilés, J.A.: Reinventing television news: Innovative formats in a social media environment. In: Studies in Big Data, pp. 143–155. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-36315-4_11

[8] García-Avilés, J.A., Arias-Robles, F., de Lara-González, A., Carvajal, M., Valero-Pastor, J.M., Mondéjar, D.: How COVID-19 is Revamping Journalism: Newsroom Practices and Innovations in a Crisis Context. Journalism Practice pp. 1–19 (2022). https://doi.org/10.1080/17512786.2022.2139744

[9] Hamaguchi, R., Nematollahi, S., Minter, D.J.: Picture of a pandemic: visual aids in the COVID-19 crisis. Journal of Public Health **42**(3), 483–485 (2020). https://doi.org/10.1093/pubmed/fdaa080

[10] Hermida, A., Young, M.L.: Data Journalism and the Regeneration of News (2019), `https://www.routledge.com/Data-Journalism-and-the-Regeneration-of-News/Hermida-Young/p/book/9781138058934`

[11] Krawczyk, K., Chelkowski, T., Laydon, D.J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., Hadsund, J.T., Bhatt, S.: Quantifying online news media coverage of the COVID-19 pandemic: Text mining study and resource. Journal of Medical Internet Research **23**(6), e28253 (2021). https://doi.org/10.2196/28253

[12] Lee, E.J., Kim, Y.W.: Effects of infographics on news elaboration, acquisition, and evaluation: Prior knowledge and issue involvement as moderators. New Media & Society **18**(8), 1579–1598 (2016). https://doi.org/10.1177/1461444814567982

[13] Malik, M.M., Pfeffer, J.: A Macroscopic Analysis of News Content in Twitter. Digital Journalism **4**(8), 955–979 (Nov 2016). https://doi.org/10.1080/21670811.2015.1133249

[14] Mellado, C., Georgiou, M., Nah, S.: Advancing Journalism and Communication Research: New Concepts, Theories, and Pathways. Journalism & Mass Communication Quarterly **97** (Jun 2020). https://doi.org/10.1177/1077699020917204

[15] Milan, S., Treré, E.: The rise of the data poor: The COVID-19 pandemic seen from the margins. Social Media + Society **6**(3) (2020). https://doi.org/10.1177/2056305120948233

[16] Northwestern Buffett Institute for Global Affairs: Visualizing a World of COVID-19 Uncertainty (2020), `https://buffett.northwestern.edu/news/2020/visualizing-a-world-of-covid-19-uncertainty.html`

[17] Pentzold, C., Fechner, D.J., Zuber, C.: “flatten the curve”: Data-driven projections and the journalistic brokering of knowledge during the COVID-19 crisis. Digital Journalism **9**(9), 1367–1390 (2021). https://doi.org/10.1080/21670811.2021.1950018

[18] Pfeffer, J., Mooseder, A., Lasser, J., Hammer, L., Stritzel, O., Garcia, D.: This sample seems to be good enough! assessing coverage and temporal reliability of Twitter’s Academic API. Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM-2023) Forthcoming (2023)

[19] Quandt, T., Wahl-Jorgensen, K.: The Coronavirus pandemic and the transformation of (digital) journalism. Digital Journalism **10**(6), 923–929 (2022). https://doi.org/10.1080/21670811.2022.2090018

[20] Riggs, E.E., Shulman, H.C., Lopez, R.: Using infographics to reduce the negative effects of jargon on intentions to vaccinate against COVID-19. Public Understanding of Science **31**(6), 751–765 (2022). https://doi.org/10.1177/09636625221077385

[21] Smit, G., Haan, Y.D., Buijs, L.: Working with or next to each other? boundary crossing in the field of information visualisation. The Journal of Media Innovations **1**(2), 36–51 (2014). https://doi.org/10.5617/jmi.v1i2.875

[22] Zwinger, S., Langer, J., Zeiller, M.: Acceptance and usability of interactive infographics in online newspapers. In: 2017 21st International Conference Information Visualisation (IV). IEEE (2017). https://doi.org/10.1109/iv.2017.65

# A. Media Accounts

Table 2: Selected accounts for analysis.

| Country | Accounts | Followers |
|---|---|---|
| France | L'Humanite (@humanite_fr) | 396k |
| | Le Figaro (@le_figaro) | 3.6m |
| | Le Monde (@lemondefr) | 10.5m |
| | Le Parisien (@le_Parisien) | 3.2m |
| | Liberation (@libe) | 3.4m |
| Germany | Bild (@BILD) | 1.9m |
| | Die Welt (@welt) | 1.8m |
| | Frankfurter Allgemeine (@faznet) | 803k |
| | Handelsblatt (@handelsblatt) | 379k |
| | Sueddeutsche Zeitung (@SZ) | 1.8m |
| India | Hindustan Times (@httweets) | 8.6m |
| | The Hindu (@the_hindu) | 7.9m |
| | Times of India (@timesofindia) | 14.6m |
| Italy | Corriere della Sera (@Corriere) | 2.7m |
| | Il Resto del Carlino (@qn_carlino) | 55k |
| | Il Sore 24 Ore (@sole24ore) | 1.8m |
| | La Reppublica (@repubblica) | 3.5m |
| | La Stampa (@lastampa) | 1.3m |
| UK | Daily Mail (@mailonline) | 2.8m |
| | Daily Mirror (@dailymirror) | 1.3m |
| | The Daily Telegraph (@telegraph) | 3.3m |
| | The Sun (@thesun) | 2m |
| | The Times (@thetimes) | 1.7m |
| USA | Los Angeles Times (@latimes) | 4m |
| | New York Times (@nytimes) | 54.9m |
| | USA Today (@USATODAY) | 4.9m |
| | Wall Street Journal (@wsj) | 20.4m |
| | Washington Post (@washingtonpost) | 20m |

# Unleashing Data Journalism's Potential: COVID-19 as Catalyst for Newsroom Transformation

## Unleashing Data Journalism’s Potential: COVID-19 as Catalyst for Newsroom Transformation

### Authors

Benedict Witzenberger, Jürgen Pfeffer

### In

Pre-Print: Benedict Witzenberger and Jürgen Pfeffer (Jan. 29, 2024b). *Unleashing Data Journalism’s Potential: COVID-19 as Catalyst for Newsroom Transformation.* DOI: https://doi.org/10.48550/arXiv.2401.14816. eprint: arXiv:2401.14816, in Review for: Journalism Practice

### Abstract

In the context of journalism, the COVID-19 pandemic brought unprecedented challenges, necessitating rapid adaptations in newsrooms. Data journalism emerged as a pivotal approach for effectively conveying complex information to the public. Here, we show the profound impact of COVID-19 on data journalism, revealing a surge in data-driven publications and heightened collaboration between data and science journalists. Employing a quantitative methodology, including negative binomial regression and Relational hyperevent models (RHEM), on byline data of articles co-authored by data journalists, we comprehensively analyze data journalism outputs, authorship trends, and collaboration networks to address five key research questions. The findings reveal a significant increase in data journalistic pieces during and after the pandemic, in particular with a rise in publications within scientific departments. Collaborative efforts among data and science journalists intensified, evident through increased authorship and co-authorship trends. Prior common authorship experiences somewhat influenced the likelihood of future co-authorships, underscoring the importance of building collaborative communities of practice. These quantitative insights provide an understanding of the transformational role of data journalism during COVID-19, contributing to the growing body of literature in Computational Communication Science and journalism practice.

### Contribution of thesis author

Theoretical operationalization, data collection, computational analysis, manual coding, contextualization, manuscript writing, revision, and editing.

### Publication Summary

This study aims to increase quantifiable insights into newsroom practices by analyzing article bylines through network models to understand the publication networks of data journalists to look for shifts during the pandemic.

# Unleashing Data Journalism's Potential: COVID-19 as Catalyst for Newsroom Transformation

Benedict Witzenberger* Jürgen Pfeffer

Technical University of Munich, Munich, Germany

In the context of journalism, the COVID-19 pandemic brought unprecedented challenges, necessitating rapid adaptations in newsrooms. Data journalism emerged as a pivotal approach for effectively conveying complex information to the public. Here, we show the profound impact of COVID-19 on data journalism, revealing a surge in data-driven publications and heightened collaboration between data and science journalists. Employing a quantitative methodology, including negative binomial regression and Relational hyperevent models (RHEM), on byline data of articles co-authored by data journalists, we comprehensively analyze data journalism outputs, authorship trends, and collaboration networks to address five key research questions.

The findings reveal a significant increase in data journalistic pieces during and after the pandemic, in particular with a rise in publications within scientific departments. Collaborative efforts among data and science journalists intensified, evident through increased authorship and co-authorship trends. Prior common authorship experiences somewhat influenced the likelihood of future co-authorships, underscoring the importance of building collaborative communities of practice.

These quantitative insights provide an understanding of the transformational role of data journalism during COVID-19, contributing to the growing body of literature in computational communication science and journalism practice.

*CONTACT Benedict Witzenberger. Email: benedict.witzenberger@tum.de

***Keywords***— D ata journalism; COVID-19; Science Journalism; Collaborative Practices; Computational Communication Science

## 1 Introduction

The COVID-19 pandemic has brought unprecedented challenges to the field of journalism, compelling newsrooms to adapt swiftly to the rapidly evolving information landscape (Hanusch, 2022; Mellado et al., 2021; Quandt & Wahl-Jorgensen, 2021). In the wake of this crisis, data journalism has emerged as a powerful and vital approach to communicating complex information to the public (Danzon-Chambaud, 2021; García-Avilés et al., 2022; Pentzold & Fechner, 2019). This article explores the profound impact of COVID-19 on data journalism, focusing on the observable surge in data-driven publications and the heightened collaboration between data and science journalists.

Prior qualitative research has provided indications that data journalism gained deeper inclusion in newsrooms during the pandemic (Bisiani, Abellan, Robles, & García-Avilés, 2023; Wu, 2021). These preliminary findings suggest that the COVID-19 crisis may have served as a catalyst for news organizations to recognize the value and importance of data-driven reporting in effectively communicating critical information to the public. The qualitative insights have highlighted how newsrooms embraced data journalism as a means to make sense of complex data related to the pandemic, enabling them to provide audiences with accurate, visually engaging, and accessible information (Pentzold, Fechner, & Zuber, 2021).

To strengthen these findings, we will adopt a quantitative approach. This study seeks to complement prior qualitative research by systematically analyzing data journalism outputs, authorship patterns, and collaboration networks on a larger scale, using byline data — the short text snippet identifying the author of a text to allow attribution to the individual responsible for the piece — from articles that data journalists co-authored. We will regard two time periods: pre-COVID-19 — the time before the pandemic appeared, and post-COVID-19, which does not imply the eradication of the virus but the time period in the aftermath of the global occurrence of COVID-19. Results will be validated using common statistical inferences and modeling tools, namely $\chi^2$ tests and negative binomial regression models. To be able to also model author cooperations based on historical data, we will call on Relational hyperevent models (RHEM) that help explain network evolution in relational event history data — where an article is regarded as an event in this study.

We will ground this work on prior qualitative research on the influence of COVID-

19 on data journalism and aim to embed our findings within Communities of Practice (CoP) that are often used to explain knowledge sharing in group contexts. Communities of Practice can be described as groups of people who share a common interest and collaborate to learn from one another, develop their skills, and share knowledge and expertise within that specific domain. We argue that the collaboration between data journalists and their colleagues in the newsroom might be a form of common learning and sharing of knowledge during the pandemic.

## 1.1 COVID-19 and Data Journalism

Firstly, we want to provide some background on data journalism and its development during COVID-19 to show the gap in research that this paper aims to fill.

Data journalism is a young profession. It can be dated back to the first decade of the 21st century (Bravo & Tellería, 2020). However, its roots can be traced back to social-science methods proposed for precision journalism (Meyer, 1973) and computer-assisted reporting (CAR), with which it shares some connection to investigative reporting (Coddington, 2015). Since the early 2010s, data-driven storytelling has been on the rise around the world (Hermida & Young, 2019; Rogers, 2011; Segel & Heer, 2010), mostly in large, well-staffed news organizations (Beiler, Irmer, & Breda, 2020; Haim, 2022), but also occurs in local settings, with lower staffing (Stalph, Hahn, & Liewehr, 2022). Three factors have been found to be central in shaping data journalism (Appelgren, Lindén, & van Dalen, 2019): journalistic cultures define to what extend watchdog-transparency is regarded as a central value for strengthening the governing political system — democracies regard transparency to its processes and actors as central, while other autocratic systems may not (Hanitzsch, Hanusch, Ramaprasad, & de Beer, 2019; Lewis & Nashmi, 2019). This is a further factor regarding the political systems that data journalists operate in and whether the freedom of information leads to broad access to information (Appelgren & Salaverría, 2018; Porlezza & Splendore, 2019). A third factor is the media market structure, the availability or lack of resources that allows experimenting with innovative formats of journalism that have not yet proven to be successful (De Maeyer, Libert, Domingo, Heinderyckx, & Le Cam, 2015).

Data journalism is regarded as a form of content or genre innovation in journalism (García-Avilés, 2020; García-Avilés, Carvajal-Prieto, De Lara-González, & Arias-Robles, 2018), which might provide news companies with an increased reputation or another competitive advantage. Data journalists use data sets and their own data analysis as sources for their stories, often bundled with infographics or interactive elements.

Media innovation nearly always contains some societal effect, as media reflects society in its content and organizational and technological structures (Bruns, 2014; Pavlik, 2000). When and how this innovation takes place is shaped by internal factors like staff incentives or leaders' behavior (Ekdale, Singer, Tully, & Harmsen, 2015; García-Avilés, Carvajal-Prieto, Arias, & Lara-González, 2019; Paulussen, Geens, & Vandenbrande, 2011), but also by external influences, like technology changes, market opportunities or evolving industry norms, and audience behavior (Anderson, 2013; Bleyen, Lindmark, Ranaivoson, & Ballon, 2014; Ess, 2014; Storsul & Krumsvik, 2013).

COVID-19 served as an accelerator for ongoing changes and innovation in journalism: the decline of print and other forms of "traditional" media, the rise of "alternative" news channels, restructured processes and altered skill requirements for journalists, changes in audience and their expectations and new approaches to journalism (Quandt & Wahl-Jorgensen, 2022). There is, however, some debate on the extent of these changes (Hanusch, 2022).

Most innovations during COVID-19 were developed in the product (like data visualizations or fact-checking), distribution (newsletters or podcasts), and commercialization (subscriptions and membership models).

We will focus here on one innovation in particular: Science departments were more relevant and worked with data teams to create visualizations (García-Avilés et al., 2022). Data journalists were central to these innovations, as they had the experience and technical means to create data visualizations (Desai, Nouvellet, Bhatia, Cori, & Lassmann, 2021) and exploratory pieces on possible scenarios, which has increased awareness and accessibility to the numbers and fostered engagement of the audience (Pentzold et al., 2021). However, this also led to criticism about an "information overload" (Krawczyk et al., 2021), a "bombardment" with visualizations (García-Avilés et al., 2022) and a very small number of — often governmental — sources (Aula (2020); Mellado, Georgiou, and Nah (2020), see also Cawley (2016); Tandoc and Oh (2017)).

We will base this research on these empirical, qualitative observations and try to operationalize the relationship between data journalistic pieces — articles that were (co-)authored by a data journalist — and intra-newsroom cooperation between data and science journalists to be able to make quantitative conclusions about the extent that data journalism innovated during COVID-19. As a theoretical foundation, we will leverage the Communities of Practice model to help explain our findings.

## 1.2 Communities of Practice

At the beginning of the 1990ies, the focus in studies of social interaction started to move from individuals to groups. Zelizer (1993) described journalism as interpretive communities "united through shared discourse and collective interpretations of key public events" (Zelizer, 1993, p. 219). More prominently, Lave and Wenger (1991) developed the idea of Communities of Practice (CoP) as a social learning theory in organizations, "a set of relations among persons, activity, and world, over time and in relation with other tangential and overlapping communities of practice" (Lave & Wenger, 1991, p. 98). These communities exist in parallel to formal hierarchies of organizations as an informal social system.

Central elements of these communities are a common domain of knowledge, a community caring about, and a shared practice to be effective in this domain (Wenger, McDermott, & Snyder, 2002). The interactions within these communities are structured in three dimensions (Wenger, 1998): mutual engagement between community members fuelled by complementary or overlapping skills, mutual relationships that allow engagement. A joint enterprise is driven by negotiated responses to internal or external conditions, resources, or demands towards the community. And a shared repertoire in tools like language, routines, or tools that shape the work in the practice.

In later years, these basic sets have been amended by stages of community-building, which can be found in the field, but do not have to be passed through in this order (Wenger et al., 2002). This emphasizes the idea that CoP are dynamic and continually evolve through interactions and collaborations.

- Potential: when a first group of people starts to take on a certain topic.
- Coalescing: when the community starts to set up a basic structure.
- Maturing: when the community grows and continues to increase and share knowledge.
- Active: when there is an acceptable amount of members and the amount of added knowledge declines.
- Dispersing: when the community is no longer relevant due to other sources or a loss of relevance for the domain.

The concept of Communities of practice has been applied throughout organizational research, also, on journalism. Journalists often form informal networks within and across newsrooms centered around shared beats, interests, or expertise. These communities influence professional identity, newsroom culture, and journalistic norms. The formation of CoP in journalism demonstrates the significance of collective learning, fostering a culture of continuous improvement and adaptation in the face of evolving media trends.

Meltzer and Martik (2017) characterized newsroom journalism as a distributed community of practice or many subcommunities that constantly cycles between phases as new technologies appear. We will use this concept to embed our findings into a theoretical framework that helps describe the collaboration of journalists from different editorial departments.

Data journalism, characterized by data-driven storytelling and visualization, has rapidly emerged as an essential practice in modern journalism. Within data journalism, CoP could potentially play a pivotal role in sharing technical expertise, discussing data analysis techniques, and disseminating innovative storytelling approaches. These communities empower data journalists to explore new storytelling methods, employ cutting-edge tools, and interpret complex datasets, ultimately elevating the quality and impact of data-driven reporting.

In response to the COVID-19 crisis, data journalism may have gained deeper inclusion in newsrooms, reflecting the increasing relevance of data-driven reporting in crisis communication. During this time, CoP could conceivably play a critical role in facilitating collaborations between data journalists and science journalists. By leveraging their shared expertise and resources, these communities may have contributed to producing accurate, visually engaging, and accessible information on the pandemic.

## 1.3 Using Bylines to Study Journalism

To operationalize these cooperations, we will use byline data. Investigating bylines, the author attribution snippet above or below a journalistic article, to measure the implications of authorship have some tradition in journalism studies. They can mostly be distinguished by two aims: to show gender-related differences that the audience may perceive by looking at the author's name or to show the impact of computer-generated articles on the confidence and perception of the readers.

A famous initial study to measure gender-related attitudes with bylines is the work of Philip Goldberg, who showed in 1967 that female readers were likely to rate male authors more favorably than female authors (Goldberg, 1967). This work is often cited as a reference for bias against women authors. However, meta-research showed that the effect is negligible (Burkhart & Sigelman, 1990; Swim, Borgida, Maruyama, & Myers, 1989). Gender bias seems to be context dependent (Dogruel, Joeckel, & Wilhelm, 2021): Audiences still have a predefined vision of which areas females have to report on and rate them less credible if they leave those, for instance, sports journalism (Klaas & Boukes, 2022), which at least indicates persistent marginalization of female bylines over 15 years (Boczek, Dogruel, & Schallhorn, 2022). However, Boczek et al. (2022) could not confirm readers' biases against female reporters.

Another use case for byline methods are studies about the perception of computer-generated text in journalistic articles. These are already used in various applications, like trading or sports reporting. Legacy newsrooms may use computer-generated articles as baseline reporting. They may enrich this by reporters' inputs and increase the value for users, which some journalists regard as a complement, not a replacement of their work (Kunert, 2020). Attributing authorship for automated content is difficult and raises ethical questions about the responsibility for news content and the requirements for transparency towards the readers (Graefe, Haim, Haarmann, & Brosius, 2016; Henrickson, 2018; Montal & Reich, 2017; Van der Kaa & Krahmer, 2014; Waddell, 2018).

We will use bylines to measure the number of authors that contributed to an article and to allow for identification to which editorial department, mostly organized by subject area, a person belongs. Co-authorship analysis serves as a proxy for collaboration between journalists. However, it may not fully capture the intricacies of collaborative dynamics within newsrooms, as it does not account for informal exchanges and knowledge sharing that may occur without formal co-authorship.

## 1.4 Research Questions

The research questions outlined below will allow us to quantitatively explore the extent to which data journalism has proliferated in newsrooms during and after COVID-19, the changes in collaborative practices, and the role of science journalists in this context:

**Q1: Does the number of data journalistic pieces change after COVID-19?** This research question examines whether there has been a measurable increase in the frequency of data journalistic pieces published in news outlets during and after the COVID-19 pandemic. By analyzing data journalism outputs over time, we seek to identify potential shifts in journalistic practices in response to the pandemic.

**Q2: Does the number of authors on data journalistic pieces change after COVID-19?** This question explores whether there has been a change in the collaboration patterns among journalists working on data-driven pieces after the onset of the pandemic. Understanding how the number of authors involved in data journalism has evolved can provide insights into the intensification of collaborative efforts during times of crisis.

**Q3: Does the number of data journalistic pieces change across departments**

**after COVID-19?** In examining data journalism outputs across different departments within news organizations, we aim to assess whether the COVID-19 pandemic has influenced the distribution and shifted the focus of data-driven reporting.

**Q4: Does the authorship of science journalists change after COVID-19?** This research question delves into the involvement of science journalists in data-driven reporting during and after the COVID-19 pandemic. By analyzing the authorship patterns of science journalists in data journalism, we aim to understand the role of scientific expertise in the intensified collaboration between data and science journalists in the context of the pandemic.

**Q5: Does prior common authorship change the probability of future co-authorships for data journalistic articles?** Examining the co-authorship networks within data journalism, this question investigates whether prior common authorship experiences influence the likelihood of future collaborations. Understanding the dynamics of co-authorship relationships can provide valuable insights into collaboration and knowledge exchange patterns among data and science journalists during COVID-19.

By addressing these research questions, this study contributes to our understanding of the influence of COVID-19 on data journalism, shedding light on the changing landscape of journalism practice during times of crisis. The findings aim to advance scholarly knowledge on possibilities to generate quantitative insights in the field of computational communication science and journalism practice.

This article offers an initial quantitative analysis of the developments of data journalism in Germany during COVID-19 based on non-questionnaire data. We first present an overview of the literature on data journalism and the impact COVID-19 might have had on the practice. We then embed our findings within the existing research on communities of practice before presenting our results and discussing them in the light of the theory.

## 2 Materials and Methods

This research was conducted using computational methods to collect and analyze data. We will describe below how metadata on articles was collected on author pages and which statistical methods and network science models were used to derive the results.

## 2.1 Data

To identify data journalists, we started with a Slack group that was formed as an advocacy group for German data journalists by the non-governmental reporters' representation "Netzwerk Recherche" in the fall of 2020 (Netzwerk Recherche, 2020). While this may lead to potential self-selection bias, we assume the majority of data journalists to be members of this group, as there are no fees or further barriers, and participation in the group offers incentives, like discussions on current topics in the field, information on upcoming conferences or meet-ups, or a job market (Witzenberger & Pfeffer, 2022). We further acknowledge that we are extracting data from a somewhat closed group to which one of the authors had access, which raises potential privacy concerns. However, as the main purpose of this study is to analyze the authorships of public media articles, this information is already publicly accessible elsewhere, but detecting data journalists would be a lot more challenging.

We limited data collection on the largest four journalism teams in Germany, Süddeutsche Zeitung (SZ), Spiegel, Zeit, and public broadcaster Bayerischer Rundfunk, and included the quite small data team of the Berlin-based newspaper Tagesspiegel to allow variance in team sizes, resulting in 688 articles from 363 distinct authors (7.99 percent data journalists, n = 29). This sample was limited due to the availability of historical article and author data and the effort of manual processes that needed to be performed to code and validate the data. We created a list of data journalists and aimed to collect all articles they were (co-)authoring by web-scraping the article metadata of the authors' pages provided by news media to showcase works by individual journalists. A data-journalistic article or piece is, therefore, an article that was (co-)authored by a data journalist. Data collection took place in November 2022 for articles between January 2019 and December 2021 to allow for somewhat similar periods pre- and post-COVID-19. The titles, authors' names, dates, and URLs for each article were collected.

Data preparation tool place in R Core Team (2022) using the packages Wickham (2022); Wickham, François, Henry, Müller, and Vaughan (2023); Wickham, Vaughan, and Girlich (2023). While title, dates, and URL information were easily parseable, the way of specifying authors' names differed greatly between organizations. In a few instances, Süddeutsche Zeitung only describes authorship by "SZ-Autoren" ("SZ-authors"), which did not yield any relevant information for our analysis on authors, but the article was used in the analysis of the prevalence of data journalistic articles. Newsroom departments were retrieved using the URL file path and clustered across media organizations.

The public broadcaster Bayerischer Rundfunk (BR) posed a further challenge to data access. As the German media landscape is split into private and public media, the organization of the latter is guarded by a legal contract between German federal states ("Interstate Treaty on Broadcasting and Telemedia", *Staatsvertrag für Rundfunk und Telemedien (Rundfunkstaatsvertrag - RStV) vom 31. August 1991 in der Fassung des Zweiundzwanzigsten Staatsvertrages zur Änderung rundfunkrechtlicher Staatsverträge (Zweiundzwanzigster Rundfunkänderungsstaatsvertrag) in Kraft seit 1. Mai 2019* (2019)). It restricts the time periods that editorial publications of public service broadcasting companies are available online. The authors retained a dataset of data journalistic articles created by the data team of Bavarian public broadcaster Bayerischer Rundfunk due to a personal request, which will be used in the analysis. However, we found no authorship-cooperation with science journalists for BR, limiting the data's meaningfulness for part of the research questions, looking for evidence to find increased cooperation, but still allowing for investigation of the prevalence of data journalistic articles.

As we only focus on data and science departments in this research, only authors from those two have been manually coded in the data, using self-descriptions on author pages, imprints, or descriptions of Twitter. We found that relations to those departments were very stable and distinct and did not change over time, which might be caused by the high specialization in science reporting or the work with data that is required.

## 2.2 Methods

To derive conclusions, this paper will be two-fold: We will use common statistical inferences and modeling tools, namely $\chi^2$ tests and negative binomial regression models, to validate changes in the number of articles and authors and deferrals between different departments. But to show the usefulness of analyzing journalistic cooperation with network analysis methods, we will then apply Relational hyperevent models (RHEM) to investigate which changes of authorship can be found between departments (to answer Q4 and Q5).

Relational hyperevent models are an advancement of relational event models (REM). A relational event is defined as a "discrete event generated by a social actor and directed towards one or more targets" (Butts, 2008, p. 159). The central idea is to model the history of events to describe the probability rate for the next relational event, or to put it more bluntly: "how and why do relational events happen?" (Pilny, Schecter, Poole, & Contractor, 2016, p. 183). The data is collected in a longitudinal fashion and by design using time-stamped interactions, like e-mails or the publication data of articles in our case, which have been shown to give a more accurate reflection of interactions, compared with surveyal studies (Corman & Scott, 1994). This data can be enriched by adding individual attributes that may increase or hinder chances for interactions.

REMs have been used for a variety of use cases, from the formation of friendships in a virtual social network (Welles, Vashevko, Bennett, & Contractor, 2014), to study interactions between states (Lerner, Bussmann, Snijders, & Brandes, 2013), or the network structure of successful Wikipedia article editing collaboration (Lerner & Lomi, 2019).

However, REMs require the network data to be in a dyadic format, taking the form of a source/sender and target/receiver relationship. Networks, like cooperation between multiple journalists to collectively write an article, require a different set of models, which can model interactions between one sender and multiple receivers or, as in our case, between multiple senders (the authors) and a single receiver (the article) — called 'hyperedges' (Kim, Schein, Desmarais, & Wallach, 2018).

Lerner, Tranmer, Mowbray, and Hancean (2019) have proposed Relational hyperevent models, which can include the hyperedge structure in their output, and have shown its utility in studies on the network dynamics of contact diaries of former British Prime Minister Margarethe Thatcher (Lerner, Lomi, Mowbray, Rollings, & Tranmer, 2021), and analysis of scientific coauthor networks (Lerner & Hâncean, 2023).

We will model our data as two-mode networks between a set of one or multiple authors and a set of single articles — shown in Figure 1 — with encoded information on the time of publication, former cooperations between authors, and their respective departments (data journalism, science, investigative or other). One author can only be connected once to a particular article, but multiple authors could be connected to an article, which would correspond to a co-authorship, resembling sources in the network. In our model, articles are modeled as targets of interactions, therefore, not connected to authors or other articles. A connection is, therefore, authorship between an author and an article. This allows us to investigate interactions between different authors over time and compare expected with actual article publications.

## 3 Results

We will now move on to lay out the results of the analysis. We retrieved 688 articles from 363 distinct authors between January 4th, 2019, to May 15th, 2021, for the five media companies. The time period was adapted to include periods before and past COVID-19 occurred, which we set to be March 16th, 2020, when the first lockdown was decided in Germany. Figure 2a shows the monthly number of data journalistic articles with lockdowns highlighted, visualizing the increased publication rate in or close to lockdown

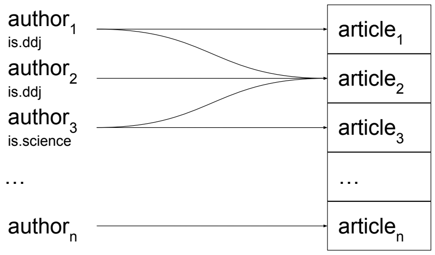
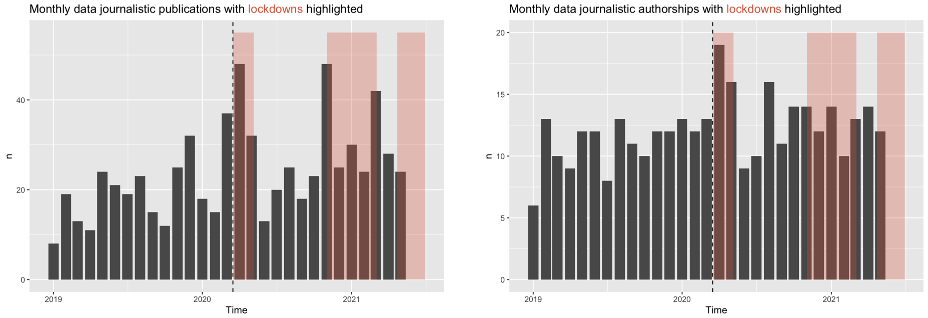


Figure 1: Data model of network analysis showing an exemplary two-mode network between one or more authors connected to one article.



(a) Showing the published data journalistic articles per month, with periods of nationwide lockdowns highlighted.

(b) Showing the authorships per month, with periods of nationwide lockdowns highlighted.

Figure 2: Comparing monthly number of publications (Fig. 2a) and authorships (Fig. 2b).

periods.

### 3.1 Some increase in articles, but not in authors

Between the time before COVID-19 hit and after, we found a 40 percent increase of data journalistic articles across all observed media from 287 to 401 articles ($\chi^2 < 0.01$), which seems to affirm Q1. We observed these changes for three newsrooms: Spiegel, SZ, and Bayerischer Rundfunk. The counts for Tagesspiegel (20 vs. 21) remained nearly the same. For Zeit (134 vs. 133), they decreased slightly. Figure 3 displays these counts.

During the observation period, we found a large increase in publications during the initial lockdown in March 2020 across all media, and after a short decline, an evenly increasing number of publications, as shown in Fig. 2a.

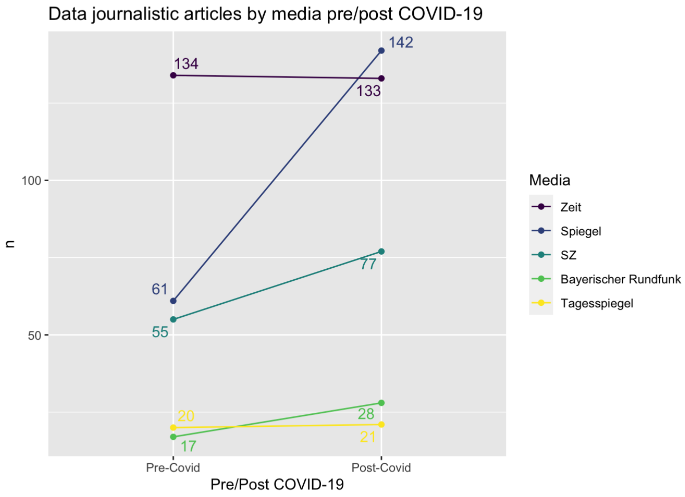


Figure 3: Absolute counts of data-journalistic articles per media pre- and post-COVID-19.

We then aimed to answer Q2, which shifted focus from publications to individual authors, to account for changes that might indicate an increased interest in the topic. Similar to the number of publications, we found an increase during the initial lockdown period but, in contrast, saw a decrease in the author numbers afterward when looking at Fig. 2b. This is backed up by results of a Wilcoxon signed rank test with continuity correction, which is used due to the sample's non-normal distribution, resulting in $p > 0.1$. A negative binomial regression, however, was used because the data indicated overdispersion, including media as a controlling factor, that yielded a 71 percent increase in author counts when holding all other variables constant (see Table 1). While this indicates some increase in author numbers, it also points to editorial or team-specific differences between media.

## 3.2 Science department becomes data journalistic

We then set to compare the prevalence of data journalistic articles across different departments in the newsroom. Departments were retrieved from the URL subdirectories of the articles and manually bucketed (i.e., politics, business, science, arts/culture) to align different naming conventions where a general link could be made. We investigated this question in two ways: First, we ran a negative binomial regression to investigate which department data-journalistic articles were presented pre- and post-COVID-19.

Table 1: Results of Negative Binomial regression model to predict the number of authors pre- and post-COVID, controlled by media.

| | *Dependent variable:* |
|---|---|
| | n |
| before_covidPost-Covid | $0.535^{***}$ |
| | (0.159) |
| mediaSpiegel | $1.627^{***}$ |
| | (0.256) |
| mediaSZ | $1.242^{***}$ |
| | (0.258) |
| mediaTagesspiegel | $-0.049$ |
| | (0.272) |
| mediaZeit | $2.474^{***}$ |
| | (0.253) |
| Constant | $3.635^{***}$ |
| | (0.210) |
| Observations | 10 |
| Log Likelihood | $-50.471$ |
| $\theta$ | $18.848^{**}$ (9.180) |
| Akaike Inf. Crit. | 112.942 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

The model (Table 2) found negative effects on data journalistic publications, digital and opinion departments, but a very strong positive effect on science with a 175 percent increase, answering Q3 and Q4.

To further investigate and examine the predictive capabilities of Relational hyperevent models, we included data, science, and investigative departments as attributes into an RHEM to investigate the probability of cooperation happening between journalists from these departments.

The three departments, data journalism, science journalism, and investigative journalism, were specifically modeled, as these were the ones that authors had been coded to before. All journalists in the data were regarded as available co-authors for the model. We used a so-called conditional size-directed hyperedge observation for the model, which samples events with non-events on a given time frame. While this does not allow us to model pre- and post-COVID-19 prevalences, it allows us to observe the authors' departments. These models were created for each media company and analyzed using the Cox proportional hazards model (Cox, 1972) that originates in survival modeling and is used to relate time, occurrences of events, and further co-variables. As Table 3 shows, we find some evidence for increased participation in data journalistic articles by science journalists. This is measured by the interaction $sender.avg.science : post\_cov$, which measures the occurrence of science journalists publishing with data journalists after COVID-19. This effect is not visible for data journalists, who seem to co-publish together very often as variable $sender.avg.ddj$ indicates.

### 3.3 Authorship collaborations change

An additional advantage of modeling the authorship events as a probabilistic network is the ability to gain deeper insights into the cooperation between authors. Using the RHEM model (see Table 3), we could analyze subset repetitions. Subset-repetition (modeled as sender.sub.rep) describes the probability of exact or partly identical authors across several articles. In our case, we limited the measure to subsets of size one — meaning two authors — and two — a subset of three authors, as we did not see any larger subsets.

The RHEM model showed that there was generally a significant subset repetition of size one for previous co-authorships between authors in SZ, Spiegel, and Zeit, indicating that authors tend to work together in smaller, prior constellations, affirming Q5. For Zeit, this also included subsets of three authors, indicating repeated cooperation between multiple journalists. For SZ, we only found an increase for those sets of three authors in the interaction with the post-COVID-19 variable.

Table 2: Results of Negative Binomial regression model to predict the number of publications pre- and post-COVID, controlled by department.

| | *Dependent variable:* |
|---|---|
| | n |
| before_covidPost-Covid | $0.793^{***}$ |
| | (0.074) |
| departmentbusiness | $-0.761^{***}$ |
| | (0.185) |
| departmentdigital | $-1.001^{***}$ |
| | (0.216) |
| departmentlocal | $-0.569^{***}$ |
| | (0.199) |
| departmentmobility | $-0.757^{***}$ |
| | (0.213) |
| departmentopinion | $-1.580^{***}$ |
| | (0.369) |
| departmentother | $-0.427^{**}$ |
| | (0.210) |
| departmentpolitics | $0.702^{***}$ |
| | (0.123) |
| departmentscience | $1.169^{***}$ |
| | (0.117) |
| departmentsports | $-0.651^{***}$ |
| | (0.197) |
| departmentwork | $-0.699^{**}$ |
| | (0.279) |
| Constant | $1.969^{***}$ |
| | (0.116) |
| Observations | 68 |
| Log Likelihood | −421.564 |
| Akaike Inf. Crit. | 867.127 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3: Results of Relational hyperevent model.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Article Publication | | | | |
| | SZ | SPIEGEL | BR | ZEIT | TAGESSPIEGEL |
| | (1) | (2) | (3) | (4) | (5) |
| sender.avg.ddj | 4.970$^{***}$ | 9.250$^{***}$ | 4.035 | 3.308$^{***}$ | 4.917$^{***}$ |
| | (0.960) | (1.185) | (3.021) | (0.459) | (1.093) |
| sender.avg.science | -0.877 | -0.590 | | 0.171 | -2.315 |
| | (1.707) | (1.867) | (0.000) | (0.874) | (2.525) |
| sender.avg.investigative | 1.448 | -3.764 | 4.870 | 1.427$^{**}$ | |
| | (2.548) | (3.210) | (3.022) | (0.682) | (0.000) |
| sender.sub.rep.1 | 6.333$^{***}$ | 2.474$^{***}$ | 1.790 | 2.974$^{***}$ | 2.274 |
| | (1.615) | (0.924) | (4.171) | (0.305) | (2.719) |
| sender.sub.rep.2 | -2.826 | 1.864 | 2.888 | 3.161$^{***}$ | -3.161 |
| | (1.754) | (3.012) | (4.065) | (0.964) | (18.278) |
| closure | 0.042 | -1.082$^{**}$ | 0.280 | 0.016$^{***}$ | -0.520 |
| | (0.071) | (0.492) | (0.219) | (0.006) | (0.733) |
| closure:post_cov | -0.044 | 1.069$^{**}$ | -0.263 | -0.009 | 0.474 |
| | (0.071) | (0.492) | (0.223) | (0.006) | (0.744) |
| sender.sub.rep.1:post_cov | -3.787$^{**}$ | -0.591 | 0.622 | -0.713$^{*}$ | 4.634 |
| | (1.718) | (0.999) | (4.886) | (0.418) | (4.942) |
| sender.sub.rep.2:post_cov | 3.628$^{*}$ | 1.490 | 0.812 | 0.110 | 1.278 |
| | (1.945) | (3.166) | (4.711) | (1.371) | (18.829) |
| sender.avg.ddj:post_cov | 1.112 | -2.521$^{*}$ | -0.248 | -0.038 | 0.357 |
| | (1.272) | (1.357) | (3.720) | (0.642) | (1.930) |
| sender.avg.science:post_cov | 4.309$^{**}$ | 3.994$^{**}$ | | 2.195$^{**}$ | 2.504 |
| | (1.930) | (1.992) | (0.000) | (1.048) | (3.204) |
| sender.avg.investigative:post_cov | -0.931 | 0.197 | -2.184 | -0.260 | |
| | (3.987) | (4.041) | (3.749) | (1.011) | (0.000) |
| Observations | 9,393 | 12,928 | 2,929 | 22,725 | 2,121 |
| $R^2$ | 0.050 | 0.051 | 0.010 | 0.040 | 0.026 |
| Max. Possible $R^2$ | 0.088 | 0.090 | 0.087 | 0.090 | 0.087 |
| Log Likelihood | -191.765 | -266.247 | -119.031 | -613.324 | -69.244 |
| Wald Test | 221.830$^{***}$ (df = 12) | 289.290$^{***}$ (df = 12) | 28.590$^{***}$ (df = 10) | 609.420$^{***}$ (df = 12) | 44.190$^{***}$ (df = 10) |
| LR Test | 480.407$^{***}$ (df = 12) | 682.140$^{***}$ (df = 12) | 29.615$^{***}$ (df = 10) | 925.125$^{***}$ (df = 12) | 55.347$^{***}$ (df = 10) |
| Score (Logrank) Test | 1,772.893$^{***}$ (df = 12) | 2,144.024$^{***}$ (df = 12) | 88.678$^{***}$ (df = 10) | 4,624.465$^{***}$ (df = 12) | 81.232$^{***}$ (df = 10) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | | | |

However, for some outlets, we also found negative effects on the *sender.sub.rep.*1 : *post_cov* interaction. This could indicate changes in author bylines after COVID-19 as the probabilities for partial subset repetitions of prior cooperation decreased. In short, existing cooperations might have been discontinued due to the changes in interests that COVID-19 brought.

A second variable to investigate authors' cooperations, called *closure*, was only small and significant for ZEIT and negative for Spiegel. It indicated the probability of co-authoring with another journalist if that journalist has already published together with a common third author. This triadic closure is common in social networks (Granovetter, 1973). Interestingly, while it was negative for Spiegel overall, the interaction with the post-COVID-19 time period turned this into an increase, opening an interpretation that this effect was initiated during the pandemic situation.

## 4 Discussion

The reporting on COVID-19 was data-driven. News media published visualizations on the prevalence and effects of the COVID-19 pandemic worldwide. Therefore, it is only a small step to argue that data journalism played a crucial role in enabling newsrooms to prepare and provide these charts and dashboards (Pentzold et al., 2021; Quandt & Wahl-Jorgensen, 2022). However, the extent to which this has led to new or increased cooperation between data journalism teams and other areas of the newsroom has not been studied thoroughly (García-Avilés et al., 2022).

As we have laid out in the Literature Review section, it might be worthwhile to ground this research into the existing Communities of Practice (CoP) literature to be able to regard the results in light of the theory of learning in organizations. The cooperation between data journalists and other departments can be viewed in this perspective as a Community of Practice between data specialists and science experts. As we have seen, there was a rapidly increasing number of data journalistic publications right at the start of the pandemic (Q1). This can be described as some maturing phase of communities (Wenger et al., 2002), which led into an active phase thereafter. The surge in data journalism during this crisis reflects the dynamic engagement within CoP, as data journalists leveraged their shared expertise to respond to the demand for data-driven reporting. As we have not seen a decrease in our data — there is some indication that the level of investment into the cooperations has not declined. The persistently high number of publications indicates that there is a general audience interest in data journalistic work or a generally higher acceptance of the work within the newsroom.

Both are classical examples of exogenous factors that drive and shape a joint enterprise in a Community of Practice.

Admittedly, we see a decline in authorships right after the first peak in the first lockdown (Q2), which seems to contradict this argument. However, there has to be a distinction between the authors and the publications. COVID-19 has drawn great interest from news audiences, which might have caused journalists from different departments to cooperate with data journalists initially. After the dust began to settle, those might have returned to their respective departments and abandoned the data journalistic partnerships. In Community of Practice phases, this can be described as a piece of evidence for the dispersing phase of communities, when members leave the community due to loss of relevance. However, as Meltzer and Martik (2017) pointed out, journalists tend to be members of different Communities of Practice. In our case, cooperation between data and science journalists has increased greatly.

Further research may investigate longitudinal effects on the prevalence of data-driven journalism and whether this explicitly takes place in cooperation with data journalism departments or has become detached and is now implicitly included in subject-specific editorial areas.

With the help of Relational hyperevent models (RHEM), we increased our understanding of journalistic author networks. In general, we found a strong sender subset repetition for data journalists, indicating that data journalists tend to cooperate with other data journalists. This aligns with the Communities of Practice understanding of creating a joint enterprise for a common domain of knowledge, a manifestation of continuous knowledge exchange and collaborative practices.

Interestingly, probabilities for sender subset repetition were negatively influenced by a post-COVID-19 dummy variable, which points to the possibility that COVID-19 has reshaped co-authorships in data journalism for some media companies. This observation is reinforced by a highly increased probability of publications featuring a science journalist after COVID-19, which points to a change through the pandemic. A new community was formed caring about the explainability of the pandemic activity through the use of scientific data and visualizations. This adaptation demonstrates the flexibility and adaptive capacity of CoP in journalism.

Significant effects on network closures were observed for ZEIT and Spiegel — and for those in two different directions. Network closures describe the probability that prior co-authorship of A and B with a third author, C, might increase chances for A and B to also cooperate in the future. One could argue that prior common authorship between a data journalist and a science journalist might also lead to a common publication with another data journalist and the same science journalist. While this might be true for ZEIT, the case for Spiegel is twofold. The number is below zero, indicating a lack of closure tendencies. However, after COVID-19, this probability is positive, which points to an increased chance for new cooperation. In combination with another observation from Spiegel on the decrease of data journalistic co-authorships after COVID-19, this could indicate less intra-data journalistic cooperation but increased co-authorships with other departments, particularly science (Q5).

The cooperation between data and science journalists can be described theoretically as combining different 'transactional expertise' from data to science. Journalists have been shown to be reluctant to communicate and convey uncertainty when reporting on scientific results (Witsen, 2019). When observations of official measurements contradict public experience, this may lead to an alienation between people and statistics, whereas some journalists may rely heavily on quantification and may blindly trust the numbers (Lugo-Ocando & Lawson, 2017). This led to calls for more direct reporting on the uncertainty in reporting (Anderson, 2018) or to enable journalists with 'transactional expertise,' which defines knowledge that enables an individual to converse about a certain topic without being able to practically work within the field (Collins, 2004; Witsen & Takahashi, 2018). The complementary combination of data and visualization skills and the ability to understand and report on scientific research can be combined together in a Community of Practice to fuel the journalistic output during the pandemic reporting, leading to increased publications of science and data journalists (Q3 and Q4).

To summarize, we found the number of data-journalistic articles changed between -0.7 and 133 percent (overall: percent) for similar time periods before and after COVID-19 hit, with four of five researched media increasing their numbers, and all media having high values for the time of the initial two lockdowns in March 2020, and Winter 2021. We also saw a huge decrease in publications across nearly all newsroom departments, with science being the sole outlier that increased its share clearly. This effect was also observable using RHEMs, which, in addition, indicated changes in the co-authorship structure pre- and post-COVID-19. At the same time, we found, generally, indications of subset repetition, which suggests recurring publication with previously co-authoring journalists. The appearance of COVID-19 reversed this effect, which implies the creation of new authorships.

## 4.1 Limitations

A number of factors limit this research. It is focused on a subset of the German data journalistic media landscape, covering the largest players in data journalism but leaving out most public broadcasters that do not provide author pages that could be scraped in a similar manner as private-owned media companies do. The study's analysis is based on data from five German newsrooms, which may limit the generalizability of the findings to a broader journalistic context. While the sample provides valuable insights into data journalism practices during COVID-19, caution should be exercised when extrapolating the results to other regions or news organizations with different characteristics and practices. The focus on Germany is also narrowing the view on a Western democracy, where data journalism has already well over a decade of history, and access to data is easier to achieve than it might in authoritarian political regimes, where COVID-19 might have also played a different role in public discourse.

A further limitation of the data is the perspective from which it was taken. During our observation period, we used the author pages of data journalists to build a ground dataset of data journalistic articles. This, however, defines data journalism as the work of data journalists, which were taken from a Slack channel, as described in the methods section. This implies that data journalistic work done completely by non-data journalists would not be included in our dataset. We expect this to be a very small number due to the focus on specific skills that might be bundled together in specific teams, which in turn form networks with other data journalists and should be visible in channels like the Slack group. In order to streamline our research efforts and allocate resources efficiently, we designated the coding of data to specifically target departments focused on journalism, science, and investigative reporting. We also observed far fewer department changes, as we suspected would happen in other areas of the newsroom and would affect the modeling effort.

Another effect on the prevalence of data journalism in Germany might have been the federal elections in 2021, as elections have traditionally been an important season for data journalists, which might have kept the number of publications higher than they might have been towards the end of the COVID-19 pandemic. However, we did not find a large decline in science data journalism towards the end of our observation period, just a small increase in political data journalism, which might indicate the expressed influence but also shows the limited extent it had. The research focuses on data journalism during the initial phase of the COVID-19 pandemic. It is essential to recognize that the pandemic's impact on data journalism and collaboration dynamics may continue to evolve over time. Future research could consider conducting longitudinal studies to examine how these trends develop over extended periods.

# 5 Conclusion

COVID-19 has influenced industries around the world, such as journalism. Data journalism especially came to increased attention, as many parts of pandemic reporting were based on data and visualizations for which data journalists had tools and knowledge. However, they also gained standing inside the newsroom and increased cooperation with science departments. We have analyzed co-authorships of German data journalists across five newsrooms.

We found that there was a significant increase of data journalistic pieces for most researched media during COVID-19, leading to more articles published, especially in scientific departments; the average number of authors per article also slightly increased during the initial phase of the pandemic, but since then decreased slightly. We found evidence of general recurring cooperation between previous (data journalistic) co-authors, which the occurrence of COVID-19 negatively influenced, which led to new, increased cooperation between data and science journalists and an increased number of publications in science departments during the pandemic.

The findings suggest that Communities of Practice play a vital role in facilitating collaborations, knowledge exchange, and innovation, enabling newsrooms to adapt to rapidly changing circumstances and produce credible data-driven reporting during challenging times. As journalism continues to evolve, the dynamics of Communities of Practice offer valuable insights for news organizations seeking to enhance journalistic cooperation.

## Acknowledgements

The authors would like to thank Termeh Shafie and Juergen Lerner for their valuable input and patient feedback on the network models and Florian Stalph for his feedback on the theoretical embedding of the work.

## Declaration Of Interest Statement

There are no potential conflicts of interest to disclose.

## Data availability statement

The data that support the findings of this study are openly available in ”Authorships German Data Journalists 2019-2021” at https://doi.org/10.7910/DVN/AGTEVS.

# References

Anderson, C. W. (2013). *Rebuilding the news: Metropolitan journalism in the digital age.* Temple University Press. Retrieved 2022-11-14, from `http://www.jstor.org/stable/j.ctt14bstt7`

Anderson, C. W. (2018). *Apostles of certainty: Data Journalism and the politics of doubt get access arrow.* Oxford University Press.

Appelgren, E., Lindén, C.-G., & van Dalen, A. (2019). Data journalism research: Studying a maturing field across journalistic cultures, media markets and political environments. *Digital Journalism*, *7*(9), 1191–1199.

Appelgren, E., & Salaverría, R. (2018). The Promise of the Transparency Culture. *Journalism Practice*, *12*(8), 986–996. Retrieved 2023-07-08, from `https://doi.org/10.1080/17512786.2018.1511823`

Aula, V. (2020, May). *The public debate around COVID-19 demonstrates our ongoing and misplaced trust in numbers.* Retrieved 2020-06-19, from `https://blogs.lse.ac.uk/impactofsocialsciences/2020/05/15/the-public-debate-around-covid-19-demonstrates-our-ongoing-and-misplaced-trust-in-numbers/`

Beiler, M., Irmer, F., & Breda, A. (2020). Data journalism at german newspapers and public broadcasters: A quantitative survey of structures, contents and perceptions. *Journalism Studies*, *21*(11), 1571-1589. Retrieved from `https://doi.org/10.1080/1461670X.2020.1772855`

Bisiani, S., Abellan, A., Robles, F. A., & García-Avilés, J. A. (2023). The data journalism workforce: Demographics, skills, work practices, and challenges in the aftermath of the COVID-19 pandemic. *Journalism Practice*, 1–21.

Bleyen, V.-A., Lindmark, S., Ranaivoson, H., & Ballon, P. (2014). A typology of media innovations: Insights from an exploratory study. *The Journal of Media Innovations*, *1*(1), 28–51. Retrieved from `https://doi.org/10.5617/jmi.v1i1.800`

Boczek, K., Dogruel, L., & Schallhorn, C. (2022). Gender byline bias in sports reporting: Examining the visibility and audience perception of male and female journalists in sports coverage. *Journalism*, 146488492110633.

Bravo, A. A., & Tellería, A. S. (2020). Data Journalism: From Social Science Techniques to Data Science Skills. *Hipertext.net*, *20*, 41–54. Retrieved 2020-06-12, from `https://www.raco.cat/index.php/Hipertext/article/view/361650`

Bruns, A. (2014). Media innovations, user innovations, societal innovations. *The Journal of Media Innovations*, *1*(1), 13–27.

Burkhart, F. N., & Sigelman, C. K. (1990). Byline bias? Effects of gender on news article evaluations. *Journalism Quarterly*, *67*(3), 492–500.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, *38*(1), 155–200.

Cawley, A. (2016). Is there a press release on that? the challenges and opportunities of big data for news media. In *Big data challenges* (pp. 49–58). Palgrave Macmillan UK.

Coddington, M. (2015, November). Clarifying journalism's quantitative turn. *Digital Journalism*, *3*(3), 331-348. Retrieved from https://doi.org/10.1080/21670811.2014.976400

Collins, H. (2004). Interactional expertise as a third kind of knowledge. *Phenomenology and the Cognitive Sciences*, *3*(2), 125–143.

Corman, S. R., & Scott, C. R. (1994). Perceived networks, activity foci, and observable communication in social collectives. *Communication Theory*, *4*(3), 171–190.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. Retrieved 2023-07-09, from http://www.jstor.org/stable/2985181

Danzon-Chambaud, S. (2021). Covering covid-19 with automated news. *Columbia Journalism Review. Available online: https://www. cjr. org/tow_center_reports/covering-covid-automated-news. php (accessed on 6 August 2021).* Retrieved 2023-07-29, from https://www.cjr.org/tow_center_reports/covering-covid-automated-news

De Maeyer, J., Libert, M., Domingo, D., Heinderyckx, F., & Le Cam, F. (2015). Waiting for Data Journalism. *Digital Journalism*, *3*(3), 432–446. Retrieved 2023-07-08, from https://doi.org/10.1080/21670811.2014.976415

Desai, A., Nouvellet, P., Bhatia, S., Cori, A., & Lassmann, B. (2021). Data journalism and the covid-19 pandemic: opportunities and challenges. *The Lancet Digital Health*, *3*(10), 619–621. Retrieved from https://doi.org/10.1016/S2589-7500(21)00178-3

Dogruel, L., Joeckel, S., & Wilhelm, C. (2021). Are byline biases an issue of the past? The effect of author's gender and emotion norm prescriptions on the evaluation of news articles on gender equality. *Journalism*, *24*(3), 560–579.

Ekdale, B., Singer, J. B., Tully, M., & Harmsen, S. (2015). Making change. *Journalism & Mass Communication Quarterly*, *92*(4), 938–958. Retrieved from https://doi.org/10.1177/1077699015596337

Ess, C. M. (2014). Editor's introduction: Innovations in the newsroom – and beyond. *The Journal of Media Innovations*, *1*(2), 1–9. Retrieved from https://doi.org/10.5617/jmi.v1i2.923

García-Avilés, J. A. (2020). Reinventing television news: Innovative formats in a social media environment. In *Studies in big data* (pp. 143–155). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-36315-4_11

García-Avilés, J. A., Arias-Robles, F., de Lara-González, A., Carvajal, M., Valero-Pastor, J. M., & Mondéjar, D. (2022). How COVID-19 is Revamping Journalism: Newsroom Practices and Innovations in a Crisis Context. *Journalism Practice*, 1–19. Retrieved 2022-11-14, from https://doi.org/10.1080/17512786.2022.2139744

García-Avilés, J. A., Carvajal-Prieto, M., Arias, F., & Lara-González, A. D. (2019). Journalists' views on innovating in the newsroom. Proposing a model of the diffusion of innovations in media outlets. *The Journal of Media Innovations*, *5*(1), 1–16.

García-Avilés, J. A., Carvajal-Prieto, M., De Lara-González, A., & Arias-Robles, F. (2018). Developing an Index of Media Innovation in a National Market. *Journalism Studies*, *19*(1), 25–42. Retrieved 2022-11-06, from https://doi.org/10.1080/1461670X.2016.1161496

Goldberg, P. A. (1967). Misogyny and the college girl. In *Meeting of the eastern psychological association.*

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2016, April). Readers' perception of computer-generated news: Credibility, expertise, and readability:. *Journalism.* Retrieved 2020-08-02, from `https://journals.sagepub.com/doi/10.1177/1464884916641269`

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. Retrieved 2023-09-10, from `http://www.jstor.org/stable/2776392`

Haim, M. (2022). The german data journalist in 2021. *Journalism Practice*, 1–20. Retrieved from `https://doi.org/10.1080/17512786.2022.2098523`

Hanitzsch, T., Hanusch, F., Ramaprasad, J., & de Beer, A. (Eds.). (2019). *Worlds of journalism.* New York Chichester, West Sussex: Columbia University Press. Retrieved 2023-07-08, from `https://doi.org/10.7312/hani18642`

Hanusch, F. (2022). Change and Continuity in Digital Journalism: The Covid-19 Pandemic as Situational Context for Broader Arguments about the Field. *Digital Journalism*, *10*(6), 1135–1140. Retrieved 2022-11-15, from `https://doi.org/10.1080/21670811.2022.2092020`

Henrickson, L. (2018). Tool vs. agent: attributing agency to natural language generation systems. *Digital Creativity*, *29*(2-3), 182–190. Retrieved from `https://doi.org/10.1080/14626268.2018.1482924`

Hermida, A., & Young, M. L. (2019). *Data Journalism and the Regeneration of News.* London, United Kingdom. Retrieved 2020-09-25, from `https://www.routledge.com/Data-Journalism-and-the-Regeneration-of-News/Hermida-Young/p/book/9781138058934`

Kim, B., Schein, A., Desmarais, B. A., & Wallach, H. (2018). The Hyperedge Event Model. *arXiv.* Retrieved from `https://arxiv.org/abs/1807.08225`

Klaas, E., & Boukes, M. (2022). A woman's got to write what a woman's got to write: the effect of journalist's gender on the perceived credibility of news articles. *Feminist Media Studies*, *22*(3), 571–587. Retrieved 2023-06-25, from `https://doi.org/10.1080/14680777.2020.1838596`

Krawczyk, K., Chelkowski, T., Laydon, D. J., Mishra, S., Xifara, D., Gibert, B., … Bhatt, S. (2021). Quantifying online news media coverage of the COVID-19 pandemic: Text mining study and resource. *Journal of Medical Internet Research*, *23*(6), e28253. Retrieved from `https://doi.org/10.2196/28253`

Kunert, J. (2020, July). Automation in Sports Reporting: Strategies of Data Providers, Software Providers, and Media Outlets. *Media and Communication*, *8*(3), 5–15. Retrieved 2020-08-02, from `https://www.cogitatiopress.com/mediaandcommunication/article/view/2996`

Lave, J., & Wenger, E. (1991). *Situated learning.* Cambridge University Press.

Lerner, J., Bussmann, M., Snijders, T. A. B., & Brandes, U. (2013). Modeling frequency and type of interaction in event networks. *Corvinus Journal of Sociology and Social Policy*, *4*(1), 3–32.

Lerner, J., & Hâncean, M.-G. (2023). Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks. *Network Science*, *11*(1), 5–35. Retrieved 2023-04-15, from `https://www.cambridge.org/core/`

journals/network-science/article/microlevel-network-dynamics-of-scientific-collaboration-and-impact-relational-hyperevent-models-for-the-analysis-of-coauthor-networks/375932B5B86D2033A0A290DE8198BB32

Lerner, J., & Lomi, A. (2019). The network structure of successful collaboration in wikipedia. In *Proceedings of the annual hawaii international conference on system sciences.* Hawaii International Conference on System Sciences.

Lerner, J., Lomi, A., Mowbray, J., Rollings, N., & Tranmer, M. (2021). Dynamic network analysis of contact diaries. *Social Networks*, *66*, 224–236. Retrieved 2023-04-15, from https://www.sciencedirect.com/science/article/pii/S0378873321000277

Lerner, J., Tranmer, M., Mowbray, J., & Hancean, M.-G. (2019). *Rem beyond dyads: relational hyperevent models for multi-actor interaction networks.* arXiv.

Lewis, N. P., & Nashmi, E. A. (2019). Data Journalism in the Arab Region: Role Conflict Exposed. *Digital Journalism*, *7*(9), 1200–1214. Retrieved 2023-07-08, from https://doi.org/10.1080/21670811.2019.1617041

Lugo-Ocando, J., & Lawson, B. (2017). Poor numbers, poor news: the ideology of poverty statistics in the media. In A. Nguyen (Ed.), *News, numbers and public opinion in a data-driven world* (pp. 62–77). London, UK: Bloomsbury. Retrieved from https://eprints.whiterose.ac.uk/109706/

Mellado, C., Georgiou, M., & Nah, S. (2020, June). Advancing Journalism and Communication Research: New Concepts, Theories, and Pathways. *Journalism & Mass Communication Quarterly*, *97*. Retrieved 2020-06-05, from https://doi.org/10.1177/1077699020917204

Mellado, C., Hallin, D., Cárcamo, L., Alfaro, R., Jackson, D., Humanes, M. L., ... Ramos, A. (2021). Sourcing pandemic news: A cross-national computational analysis of mainstream media coverage of covid-19 on facebook, twitter, and instagram. *Digital Journalism*, *9*(9), 1261–1285. Retrieved from https://doi.org/10.1080/21670811.2021.1942114

Meltzer, K., & Martik, E. (2017). Journalists as Communities of Practice: Advancing a Theoretical Framework for Understanding Journalism. *Journal of Communication Inquiry*, *41*(3), 207–226. Retrieved 2023-07-16, from https://doi.org/10.1177/0196859917706158

Meyer, P. (1973). *Precision journalism: A reporter's introduction to social science methods.* Bloomington, Indiana, USA: Indiana University Press.

Montal, T., & Reich, Z. (2017). I, robot. you, journalist. who is the author? *Digital Journalism*, *5*(7), 829–849. Retrieved from https://doi.org/10.1080/21670811.2016.1209083

Netzwerk Recherche. (2020, December). *Fachgruppe Datenjournalismus.* Retrieved 2021-10-29, from https://netzwerkrecherche.org/ueber-uns/vorstand/fachgruppe-datenjournalismus/

Paulussen, S., Geens, D., & Vandenbrande, K. (2011). Fostering a culture of collaboration: Organizational challenges of newsroom innovation. In *Making online news. volume 2* (pp. 3–14). Peter Lang Publishing Inc. Retrieved from http://hdl.handle.net/1854/LU-1152182

Pavlik, J. (2000). The impact of technology on journalism. *Journalism Studies*, *1*(2), 229–237. Retrieved from https://doi.org/10.1080/14616700050028226

Pentzold, C., & Fechner, D. (2019). Data journalism's many futures: Diagrammatic displays and prospective probabilities in data-driven news predictions. *Convergence: The International Journal of Research into New Media Technologies*. Retrieved 2020-05-15, from https://doi.org/10.1177/13548565198807

Pentzold, C., Fechner, D. J., & Zuber, C. (2021). "Flatten the Curve": Data-driven projections and the journalistic brokering of knowledge during the covid-19 crisis. *Digital Journalism*, *9*(9), 1367-1390. Retrieved from https://doi.org/10.1080/21670811.2021.1950018

Pilny, A., Schecter, A., Poole, M. S., & Contractor, N. (2016). An illustration of the relational event model to analyze group interaction processes. *Group Dynamics: Theory, Research, and Practice*, *20*(3), 181–195.

Porlezza, C., & Splendore, S. (2019). From Open Journalism to Closed Data: Data Journalism in Italy. *Digital Journalism*, *7*(9), 1230–1252. Retrieved 2022-11-06, from https://doi.org/10.1080/21670811.2019.1657778

Quandt, T., & Wahl-Jorgensen, K. (2021). The coronavirus pandemic as a critical moment for digital journalism. *Digital Journalism*, *9*(9), 1199–1207.

Quandt, T., & Wahl-Jorgensen, K. (2022). The coronavirus pandemic and the transformation of (digital) journalism. *Digital Journalism*, *10*(6), 923–929. Retrieved from https://doi.org/10.1080/21670811.2022.2090018

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rogers, S. (2011, July). Data journalism at the Guardian: what is it and how do we do it? *The Guardian*. Retrieved 2020-06-12, from https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism

Segel, E., & Heer, J. (2010, November). Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148. Retrieved from https://doi.org/10.1109/TVCG.2010.179

*Staatsvertrag für Rundfunk und Telemedien (Rundfunkstaatsvertrag - RStV) vom 31. August 1991 in der Fassung des Zweiundzwanzigsten Staatsvertrages zur Änderung rundfunkrechtlicher Staatsverträge (Zweiundzwanzigster Rundfunkänderungsstaatsvertrag) in Kraft seit 1. Mai 2019.* (2019). Retrieved from https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/RStV_22_nichtamtliche_Fassung_medienanstalten_final_web.pdf

Stalph, F., Hahn, O., & Liewehr, D. (2022). Local data journalism in germany: Data-driven reporting amidst local communities and authorities. *Journalism Practise*, 1–20. Retrieved from https://doi.org/10.1080/17512786.2021.2019089

Storsul, T., & Krumsvik, A. H. (2013). What is media innovation? In *Media innovations* (pp. 13–26). Nordicom. Retrieved from https://www.nordicom.gu.se/en/publications/media-innovations

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, *105*(3), 409–429.

Tandoc, E. C., & Oh, S.-K. (2017, August). Small Departures, Big Continuities? *Journalism Studies*, *18*(8), 997–1015. Retrieved 2021-03-13, from https://doi.org/10.1080/1461670X

.2015.1104260

Van der Kaa, H., & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. In *Proceedings of the Computation+Journalism conference.* (Computation + Journalism Symposium 2014 ; Conference date: 24-10-2014 Through 25-10-2014)

Waddell, T. F. (2018, February). A Robot Wrote This? *Digital Journalism*, *6*(2), 236–255. Retrieved 2020-08-14, from https://doi.org/10.1080/21670811.2017.1384319

Welles, B. F., Vashevko, A., Bennett, N., & Contractor, N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures*, *8*(4), 223–243.

Wenger, E. (1998). *Communities of Practice.* Cambridge University Press.

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating Communities of Practice.* Harvard Business School Press.

Wickham, H. (2022). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=stringr (R package version 1.5.0)

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 1.1.2)

Wickham, H., Vaughan, D., & Girlich, M. (2023). tidyr: Tidy messy data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tidyr (R package version 1.3.0)

Witsen, A. V. (2019). How daily journalists use numbers and statistics: The case of global average temperature. *Journalism Practice*, *14*(9), 1047–1065.

Witsen, A. V., & Takahashi, B. (2018). Knowledge-based journalism in science and environmental reporting: Opportunities and obstacles. *Environmental Communication*, *12*(6), 717–730.

Witzenberger, B., & Pfeffer, J. (2022). Gender dynamics of german journalists on twitter. In *2022 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM).* IEEE.

Wu, S. (2021). Data "objectivity" in a time of coronavirus: Uncovering the potential impact of state influence on the production of data-driven news. *Digital Journalism*, *9*(9), 1303–1320. Retrieved from https://doi.org/10.1080/21670811.2021.1942111

Zelizer, B. (1993, September). Journalists as interpretive communities. *Critical Studies in Mass Communication*, *10*(3), 219–237. Retrieved 2022-01-25, from https://doi.org/10.1080/15295039309366865

# More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany

# More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany

### Authors

Benedict Witzenberger, Jürgen Pfeffer

### In

Journalism and Media, published as: Benedict Witzenberger and Jürgen Pfeffer (Mar. 21, 2024a). "More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany". In: *Journalism and Media* 5.1, pp. 412–431. DOI: 10.3390/journalmedia5010027

### Abstract

Women are underrepresented in many areas of journalistic newsrooms. In this paper, we examine if this established effect persists in the new forms of journalistic communication, namely social media networks. We use mentions, retweets, and hashtags as measures of journalistic amplification and legitimation. Furthermore, we compare two groups of journalists in different stages of development: political and data journalists in Germany in 2021. Our results show that journalists identified as women tend to favor other women journalists in mentions and retweets on Twitter (now called X), compared to men. While both professions are dominated by men, with a high share of tweets authored by men, women mention and retweet other women more than their male colleagues. Female data journalists also leverage different sources than men. In addition, we found data journalists to be more inclusive of non-member sources in their networks compared to political journalists.

### Contribution of thesis author

Theoretical operationalization, data collection, computational analysis, manual coding, contextualization, manuscript writing, revision, and editing.

### Publication Summary

Social media networks provide a modern platform for exchanging information, much like old town squares, especially for professionals who gather there. Previous research has shown that US political journalists' use of Twitter has decreased the amplification of female journalists' voices. However, no comparison has been made between other geographies or other groups of journalists. A comparative study of German political and data journalists of both sexes can yield valuable insights into communication structures on social media platforms. It can also help increase the possibility of insights being created using the Computational Social Science toolbox.

*journalism and media*

MDPI

*Article*

# More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany

**Benedict Witzenberger * and Jürgen Pfeffer**

School of Social Sciences and Technology, Technical University of Munich, 80333 Munich, Germany
* Correspondence: benedict.witzenberger@tum.de

**Abstract:** Women are underrepresented in many areas of journalistic newsrooms. In this paper, we examine if this established effect persists in the new forms of journalistic communication, namely social media networks. We use mentions, retweets, and hashtags as measures of journalistic amplification and legitimation. Furthermore, we compare two groups of journalists in different stages of development: political and data journalists in Germany in 2021. Our results show that journalists identified as women tend to favor other women journalists in mentions and retweets on Twitter (now called X), compared to men. While both professions are dominated by men, with a high share of tweets authored by men, women mention and retweet other women more than their male colleagues. Female data journalists also leverage different sources than men. In addition, we found data journalists to be more inclusive of non-member sources in their networks compared to political journalists.

**Keywords:** journalism; social networking (online); gender issues; information retrieval

check for updates

**Citation:** Witzenberger, Benedict, and Jürgen Pfeffer. 2024. More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany. *Journalism and Media* 5: 412–431. https://doi.org/10.3390/journalmedia5010027

Academic Editors: Andreas Veglis and Charalampos Bratsas

Received: 16 December 2023
Revised: 4 March 2024
Accepted: 11 March 2024
Published: 21 March 2024

## 1. Introduction

Social media networks (SMNs) such as Twitter have had a significant impact on journalism. Researchers have focused on how Twitter (now called X) has challenged key values of journalism, such as objectivity, gatekeeping, and transparency (Hermida 2010; Lasorsa et al. 2012; Lawrence et al. 2014). Twitter and other microblogging platforms have also changed the news cycle by creating a hybrid system of new actors and news-sourcing habits (Chadwick 2011). Journalists commonly use Twitter as a source of information (Paulussen and Harder 2014). Some have noted changes in the way private and professional personae are presented on Twitter, which may collide with corporate brands (Hanusch 2018; Ottovordemgentschenfelde 2017).

A rigid selection of information shapes the world of SMNs. This is not a new development. Lippmann (1922) described the bias between reality and perception—or mental image—around 100 years ago, referring to it as a pseudo-environment. This explains the selective way of processing information shaped by social constructs surrounding the individual, which has been researched since then (Lazarsfeld 1944).

A sender-based selection form was described by Lewin (1947), showing that disseminators tend to spread information that aligns with their values. The foundation of the gatekeeping theory (White 1950) has shaped journalism over the decades but has become a more general phenomenon since the global spread of information is no longer restricted to journalists but open to everyone on social media platforms. This has led to an increase in data, which might help to shed light on processes that have so far taken place behind closed doors. In this article, we attempt to enhance our understanding of journalistic discourses on social media, focusing mostly on gender and journalistic areas as differentiators.

The history of women's journalism is much older than social media. Female journalists were first hired during the second half of the 19th century out of financial interests. They were needed to help create so-called "women's pages" (Chambers et al. 2004; Hunter 2019; Kay 2012; Steiner 2008) with topics like fashion, art, or societal gossip. These "women's pages" targeted a female audience, which the newspapers wanted to attract because of the increasing revenues from advertising in newspapers (Lang 1999). Currently, women journalists are underrepresented in many newsrooms. Therefore, they are less visible in the media (Hannis and Strong 2007; Kian and Hardin 2009; North 2016; Smith 1981), which might lead to distorted "news-is-for-men" perceptions in the audience (Sui et al. 2022). Other channels of public appearance could provide new platforms for women journalists to promote their work or build reputations in their beats. Twitter, as a platform with few barriers to entry, would naturally be expected to serve as an enhancement to building a platform. However, previous work has shown that this is not necessarily the case (Lasorsa 2012; Usher et al. 2018); political journalists, in particular, have been shown to form male-dominated, elitist networks (Lawrence et al. 2014; Matusitz and Breen 2012).

We build on an emerging body of literature that uses Twitter data to analyze networks of journalists to find out if there are sex-related differences between journalists on Twitter in general and groups of journalists in particular. We focus on two groups: political and data journalists in Germany. Journalists, as a profession, play an important role in disseminating information and shaping public opinion. However, this is most visible for political journalists, who often cover issues of profound societal and political significance. Their presence on social media platforms like Twitter can have policy implications and influence public discourse on critical topics. Investigating the behaviors of political journalists on Twitter contributes to a broader understanding of the interplay between journalism, politics, and society, which has become increasingly important in the digital age.

Data journalism, in particular, has witnessed significant growth and innovation in recent years due to the big data revolution, driven by the availability of large, behavior-based datasets, improved computational resources, and new and accessible analytic techniques (Mayer-Schönberger and Lenneth 2013), which also took place in media and journalism (Howard 2014; Lewis and Westlund 2015). Data visualization, interactive storytelling, and data-driven investigations have become increasingly prevalent in journalism. By focusing on data journalists, we aim to capture emerging trends in journalism practices and explore how these innovations manifest on Twitter.

We analyzed 478,263 tweets from political and data journalists in Germany in 2021 to compare the communication styles within these communities and between sexes. Men dominate the number of tweets, whereas women tend to favor other women journalists in mentions in general and retweets of political journalists. We also find men to be self-retweeting themselves a lot more than women, and there are different sourcing behaviors in both groups of journalists.

The contributions of this article are as follows:

- We provide a comparative quantitative analysis of communicative differences between and within German political and data journalists, offering a non-US-centric perspective.
- We found manifestations of existing sex-related norms in the Twitter behaviors of both journalist groups, confirming prior studies on U.S. political journalists.
- Our analysis of sources and hashtags reveals a broader spectrum of sources for data journalists and different sharing behaviors of men and women.

As part of our research into the dynamics of German journalists on Twitter, preliminary findings were presented at the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). This earlier work laid the groundwork for the extended analyses presented in this manuscript (Witzenberger and Pfeffer 2022).

We will start our argument by laying out the related literature on gender issues in journalism, followed by an overview of political and data journalists in Germany; we will then present our methods and results.

## 2. Literature Review

Prior studies have looked at women in journalism and on SMNs, as well as the behaviors of political journalists on Twitter.

### *2.1. Journalism and Gender*

Journalism and gender have been studied from various perspectives over the last two decades. Most research has focused on comparative perspectives and the possible influence on journalistic style between men and women. (Craft and Wanta 2004; Hannis and Strong 2007; Kian and Hardin 2009; North 2016).

While parity in lower-level editorial positions has nearly been achieved, there is still a discrepancy within higher-level jobs in newsrooms (Andi et al. 2020; Byerly 2011; Chambers et al. 2004; Ziamou 2000). This "glass ceiling" impacts the editorial policy, as higher-ranking positions dictate the editorial ethos. This leads to a limited perspective on issues (Fleras 2003; Smith 1981) and differences in beat assignments (Craft and Wanta 2004).

Several studies exist on the size of the gender gap in Germany, all with slightly different methods, yet none provide current figures. A tally by the initiative "ProQuote", which lobbies for a women's quota of 30%, found that the share of women in power (ranging from editors-in-chief to deputy section leaders) varied between 16.1 and 50.8% for national newspapers in June 2019 (von Garmissen and Biresch 2019), totaling 25.1% for women compared to 74.9% for men. Across all positions, the "Worlds of Journalism Study" in 2016 found a proportion of 40.1% for women (Hanitzsch et al. 2016; Steindl et al. 2017), while a 2013 study by the European Institute for Gender Equality, which only included a few media corporations, estimated at around 44% (European Institute for Gender Equality (EIGE) 2013).

### *2.2. Twitter for Journalists*

Since its creation in 2006, Twitter and its implications on journalism have been studied in multiple dimensions, as previously mentioned. It has primarily been described as a platform for breaking news (Kwak et al. 2010), with its users mainly talking about headlines and current affairs (Asur et al. 2011; Kwak et al. 2010). Twitter is a medium for professional communicators—like politicians and celebrities. This aspect of Twitter seems to make it more appealing to journalists in comparison to other professions (von Nordheim et al. 2018).

This is especially true of political journalists: The platform is, if not a central source of news in Washington D.C. (Hamby 2013; Kreiss 2016), a central source of news in several Westminster democracies (Hanusch 2018) and Germany (Degen and Olgemöller 2021; Nuernbergk 2016; Nuernbergk and Schmidt 2020).

Over time, the use of Twitter has increasingly normalized for journalists (Lasorsa et al. 2012). However, this process takes time and requires corporate policies that prevent the fast, widespread adoption of newer features (Molyneux and Mourão 2019). Others have described this process less as normalization and more as a negotiation between traditional gatekeeping roles and editorial decision-making and the new influences injected by users on SMNs (Tandoc and Vos 2016).

But why do journalists participate on Twitter at all? Viewed from Bourdieu's field theory (Bourdieu 1993), journalists compete over attention in their spaces, leveraging their networks of connections—memberships in one or many groups—which they may potentially mobilize through their social capital (Bourdieu 1986). While this was already the case before SMNs existed, they offer a new space for validation (Carlson 2017) or to validate their "gut feelings" (Schultz 2007), or"interpretive communities", as Zelizer (1993) called them.

These connections have often been described through the lens of homophily— an old concept that suggests that human ties are formed if they share attributes. An early scientific example is Lazarsfeld et al. (1954), which investigated the formation of friendships within two communities, finding evidence that social status and shared values are drivers for forming or dissolving friendships.

Homophily has been identified across various social areas, such as race, sex, gender, age, religion, education, occupation, network position, behavior, attitudes, abilities, and beliefs (McPherson et al. 2001). While some of these aspects may not play an essential role in journalism, for instance, education, where levels are potentially higher for journalists than in the general population (Josephi et al. 2019), some aspects of homophilous influence have already been found in journalists' behavior on SMNs, such as sharing a common beat, shared values, and common geographies.

Homophily has also been used to show that journalists on SMNs prefer to form connections with other journalists who share similar journalistic values but not necessarily ideological intersections (Li et al. 2023).

Vergeer (2014) found that journalists covering similar geographical areas were more likely to connect, even if they were not working for the same outlets.

Multiple studies have shown that political journalists form elitist circles on social media. Research on the tweeting behaviors of reporters covering the 2012 Republican and Democratic conventions showed that journalists tended to express more opinions in their writing on Twitter than in journalistic media. A study involving a list of 430 reporters and commentators was conducted and manually coded. Reporters consistently maintained a closed gate-keeping level by mainly linking and retweeting themselves and their fellow reporters and rarely reacting to their followers (Lawrence et al. 2014).

Further research using a similar dataset from the 2012 presidential race showed that reporters focused their tweets on the main topics and rarely questioned their peers' views but used Twitter as a "space for collective interpretation of political events" (Mourão 2015). This view describes the journalists as creating a virtual "bubble" (Zelizer 1993).

Evidence from the 2016 presidential race in the U.S. suggested similar results, although the study was limited to retweets, quoted tweets, and replies (Molyneux and Mourão 2019). This observation was made even after Twitter's user base had stabilized. Several other scholars have shown that journalists mainly discuss issues with other journalists or politicians (Maares et al. 2021; Mourão et al. 2016).

Further research dealt with the impact of additional characteristics on tweeting behavior.

### *2.3. Twitter and Gender Dynamics*

The behavior of journalists in SMNs has been the subject of several studies. On the one hand, women journalists on Twitter tend to share more about their personal lives and link to external websites more often, indicating more transparency than their male peers (Lasorsa 2012). On the other hand, women journalists frequently encounter sexual harassment in online environments (Stahel and Schoen 2020), especially when covering topics that are somewhat regarded as male territory (Sarikakis et al. 2021). This has been shown to limit their ability to communicate with their audience (Chen et al. 2020), lead to avoidance (Adams 2018; Stahel and Schoen 2020), and is described as being aimed at disciplining journalists (Waisbord 2020).

Regarding amplification through retweets and mentions on Twitter, an analysis of political reporters in Washington, D.C., showed that male journalists tend to amplify and engage with their male peers almost exclusively. Women engage with each other but retweet men more often in absolute terms than they retweet women (Usher et al. 2018).

Similarly, in 2019, Fincham (2019) found strong homophily when comparing U.S. and U.K. political journalists' Twitter behaviors. However, he also found gender-related discrepancies, i.e., strong homophily in male interactions, women journalists retweeting more men than other women, and a higher likelihood of using replies when interacting with one's own gender.

These findings are mirrored in Hanusch and Nölleke (2019), who investigated Australian journalists. They have been found to share a significant degree of homophily in characteristics like organization, geographic proximity, and gender. The largest amount of homophily, however, is attributed to their beat. This leads to a tightly knit, homogenous, elitist community, mainly interacting with itself.

### *2.4. Twitter Use of Political Journalists in Germany*

For reporters in the German parliament, the Bundestag, Twitter is the most used social media network for journalists covering federal politics in Berlin used to observe sources and topics and gather information (Nuernbergk 2016; Nuernbergk and Schmidt 2020).

Research has suggested that the interpretive standpoints chosen in their reporting can already be concluded by looking at the tweets of political journalists (Degen and Olgemöller 2021). Furthermore, Twitter interactions between politicians and journalists can lead to different assessments of Twitter, compared to journalists with no interactions, indicating that the network also plays a role in relationship management (Nuernbergk and Schmidt 2020).

Research from 2014 has shown that correspondents incorporate politicians into their communicative circles but stick together when debating, not reacting with other users attempting to contribute to the discussion (Nuernbergk 2016). This is consistent with other authors, as previously mentioned above.

### *2.5. Data Journalists in Germany*

Data journalism is a new playing field in journalism. While its roots are mostly dated back to the 1970s idea of "precision journalism" (Bravo and Tellería 2020; Coddington 2015; Meyer 1973, 2002), some sources even go as far as defining its provenance to the use of tables in The Guardian in 1821 or visualizations by Florence Nightingale and Jon Snow in the 1850s (Rogers 2010); however, it is mainly regarded as having been started around 2009 (Bravo and Tellería 2020). Its primary focus involves combining data analytical approaches to find and extract information from data and tools to visualize the results and tell stories with it, enhancing traditional reporting (Anderton-Yang et al. 2012; Antonopoulos and Karyotakis 2020; Berret and Phillips 2016; Coddington 2015).

Data journalism in Germany has been enumerated twice. In the spring of 2013, Weinacht and Spiller (Weinacht and Spiller 2014) identified 35 individuals working as data journalists in Germany and were able to interview them, and in 2020, Beiler et al. (2020) estimated that data journalism is well-established in three-fourths of media outlets.

While there is no published data on the gender distribution of data journalists in Germany, an analysis of the 2013 study by Weinacht and Spiller, which aimed to cover all data journalists in the country at that time, shows that 3 out of 35 interviewees had women's first names (Weinacht and Spiller 2014). Likewise, a study on data journalists in Sweden in 2014 found that 46% of the respondents were women, 53% were men, and 2% declined to answer (Appelgren and Nygren 2014).

Compared to other areas, data journalism is regarded as a new field not guarded by "old boys" networks, thus being more open to all genders (De Vuyst 2018). This allows data journalism access to journalistic areas that were formerly more exclusive, like investigative reporting. On the downside, there is a lack of women in technical positions, which spills over into a lack of women in data journalism because they lack the skills to apply. This is seen as a lack of women in computer sciences (De Vuyst 2018). In a self-assessment study, male data journalists rated themselves as more experienced than their female counterparts (Appelgren and Nygren 2014). However, it is unclear if this is due to men's overconfidence or the women's understatement.

### *2.6. Hypotheses*

To structure this research, we present three questions, split into five hypotheses we aim to answer.

#### 2.6.1. The "Boys on the Bus" Are Now on Twitter

The first hypothesis is centered around the idea of an elitist community of political journalists, which has been identified multiple times in the past (Lippmann 1922). Twitter could, by default, have an opening effect on those groups. Political journalists have been shown to form elitist circles on social media (Lawrence et al. 2014; Molyneux and Mourão 2019; Mourão 2015; Nuernbergk 2016). We want to compare them to data journalism as a newer form of journalism. Because the latter is derived from a more technical, computer science-driven background—referred to as "programmer-journalists" (Parasie and Dagiral 2012)—they may have different approaches to communication. Data journalism is often regarded as more transparent in its underlying data and methods (Diakopoulos 2016), which might be conveyed differently in social media discourses. Furthermore, data journalists have been recognized for incorporating several versatile discourses around technology, transparency, and democratic values, which may further increase the diversity of topics and users they interact with (Hannaford 2022; Tong and Zuo 2019). Our first research question is as follows:

**RQ1**: Are data journalists engaging differently with non-peers on Twitter compared to political journalists?

The primary hypothesis is as follows:

**H1:** *Data journalists have a more open discourse than political journalists.*

### 2.6.2. Journalistic Gender Dynamics on Twitter

Another set of hypotheses is aligned with the question of gender dynamics in the Twitter behavior of journalists.

Twitter plays a vital role in publicly providing journalistic legitimation (Carlson 2017) or dominance in a specific field (Barnard 2014). This has also been argued above when excluding outsiders from discourses but is also true within the field when establishing a hierarchy (Mourão 2015).

As shown by Usher et al. (2018), men have dominated the use of Twitter within Washington D.C.'s political journalism scene. Not only do male journalists amplify their gender, but women also tend to mention and retweet male correspondents more than their peers in absolute terms. Relatively, women retweet other women much more than expected based on the raw share of genders. This selective behavior, as an inherent trait in SMNs, has already been described earlier, with researchers showing that men primarily retweet men and women mostly retweet women (Xiao et al. 2012). In journalism, the extent of the observed gender gap is striking, being described as a "gendered echo chamber" (Usher et al. 2018, p. 338).

These results were retrieved by calculating so-called power users based on typical Twitter activities attributed to specific categories, namely replying or following as measures of engagement, mentioning as a form of legitimation, and retweeting and quoting for amplification. Our analysis uses mentions and retweets as indicators for legitimizing or amplifying behavior.

**RQ2**: Are there differences in gender bias in the mentions and retweets of German politics and data journalists on Twitter?

We created two hypotheses for our tweet analysis:

**H2**: *Women journalists are mentioned less than men.*

**H3**: *Women journalists are retweeted less than men.*

### 2.6.3. Differences in Sources

A third perspective is based on the content of the tweets that are shared by both sexes in the studied journalistic disciplines. Earlier research has suggested that women journalists tend to be assigned to types of stories regarded as being 'soft,' like arts, education, or health (North 2016), and use different sources in their reporting (Armstrong 2004). As this research already focuses on a narrow subset of journalism, we want to understand if these observations hold on Twitter, making it easier for journalists to elevate sources and focus on topics important to them without having to clear editorial processes.

As data journalism is derived from a very broad set of backgrounds, we would expect data journalists to leverage a more diverse set of sources than political journalists.

Therefore, we ask the following:

**RQ3**: Can we identify differences in retweeted sources or hashtags between sexes and German political and data journalists on Twitter?

To answer this question, we raise two hypotheses:

**H4**: *Women journalists amplify different sources and hashtags than their male counterparts.*

**H5**: *Data journalists have a more diverse set of topics than political journalists.*

## 3. Materials and Methods

To provide an accurate and detailed snapshot of **German political journalists** on Twitter, we based the selection on the circulation and sizes of German newspapers. We attempted to identify journalists who were clearly deployed to political sections or mainly worked on political topics. This approach limited the proportion of regional newspapers, which use news agencies more extensively in their political reporting than larger newspapers and have no apparent political reporters. Many larger newspapers offer imprints with an overview of their authors and their vitae, which often contain Twitter accounts. Smaller newspapers sometimes lack this information, which must be retrieved from the articles. From these 730 accounts, all tweets between 1 January and 31 December 2021 were retrieved on 7 January 2022, using Twitter API v2 (Pfeffer et al. 2023) (in total, 430,451 tweets).

Journalists working for T.V. or radio stations—largely public corporations in Germany—have been omitted. The importance of newspapers has been assessed by two publications: the quarterly circulation data provided by the so-called "Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V." (abbreviated IVW), which corresponds to the "Audit Bureau of Circulation", and the recurring "Media-Analyse", a research survey that attempts to evaluate the media consumption habits of the German population. We utilized data from 2019 for regional newspapers and from 2020 for national newspapers to identify which publications to further investigate for journalists who used Twitter.

This approach differs from Nuernbergk's (Nuernbergk 2016), which used a predefined set of political journalists who are members of the official German Federal Press Conference. As a result, we expected our sample to include more journalists in areas other than the German capital of Berlin.

Accounts that were obviously only private—meaning they showed no connection to the newspaper or regularly mentioned its stories—were discarded. This list was compiled in July 2020 and updated on 7 January 2022.

To identify **data journalists**, we used an advocacy group as a starting point. A significant number of German data journalists have decided to congregate as a so-called "Fachgruppe" (professional group) within the non-governmental reporters' representation, "Netzwerk Recherche", in the fall of 2020. "Netzwerk Recherche" sees itself "as general representatives of the interests of the entire field of data journalism and all its manifestations" (Netzwerk Recherche 2020). To simplify communications, a group on the messaging platform Slack was created, open to anyone identifying as a data journalist. The restriction on this platform introduces a form of self-selection, which may lead to bias in this research. However, we assume the majority of data journalists to be members of this group, as there are no fees or further barriers. Participating in the group offers incentives, like discussions on current topics in the field, information on upcoming conferences or meet-ups, and a job market. We acknowledge potential privacy concerns introduced by using a somewhat non-public data source. However, we did not analyze data on an individual level. We identified 167 members at the time of our data collection, similar to what has been collected in previous studies (Beiler et al. 2020; Haim 2022). Twitter usernames and affiliations were manually added whenever mentioned in the profile's description text; 148 data journalists could be connected with a Twitter account, and 47,812 tweets were downloaded on 11 March 2022 for 2021. See Table 1 for a comparison of the extracted numbers.

**Table 1.** Comparison of the shares of sexes in our sample by a group of journalists with a total from the "Worlds of Journalism Study" (Steindl et al. 2017).

| | Women | Men |
|---|---|---|
| Total | 40.1% | 59.9% |
| Political | 28.6% | 71.4% |
| Data | 32.2% | 67.8% |

### 3.1. *Adding a Gender Attribution*

We assigned a binary gender category to all users on our lists by manually coding the authors' first names into traditional male or female first names. This approach may result in misspecifications if someone identifies as a different gender, as expected by the name. It has to be noted that this reliance on a binary gender framework may also not adequately capture the complexities of gender identity. It may exclude non-binary and transgender journalists, whose interactions on social media could offer valuable insights into the broader conversation about gender dynamics in journalism.

However, as this work attempts to identify a potential divergence between users who appear as women and men for outsiders and aims to be comparable to prior studies, we consider this issue approach sufficient. In unclear cases, we attempted to deduce the gender using profile pictures.

No names were found that were not explicit enough to be assigned to a gender. In our data, 28.6% of political and 32.2% of data journalist users were regarded as women, while 71.4% of political and 67.8% of data journalists identified as men. Both groups had fewer shares of women than the "Worlds of Journalism Study" found in Germany in 2016, with 40.1%.

### 3.2. *Clustering Sources and Hashtags*

Incorporating the clustering of retweet sources and hashtags into our study constitutes an approach that enriches the depth of our analysis by providing a more comprehensive understanding of the content of tweets within the context of journalistic communication.

We first extracted all retweeted usernames throughout our dataset, totaling 20,937 accounts for political journalists and 6519 for data journalists. After removing self-retweets, we extracted the 30 most retweeted accounts for both groups and genders of journalists and labeled them into categories (see Tables S1 and S2 for the cluster results). Political journalists were categorized into German media, foreign media, politics, NGOs, and political journalism. For data journalists, the following categories were used: media, NGOs, data visualization advocates, politics, foreign media, data journalists, non-date journalists, and others.

In the second step, we extracted the hashtags used in tweets across the data. These resulted in 32,200 hashtags for politics and 4918 for data journalists. These were clustered into categories (see Tables S3 and S4 for cluster results). We used COVID-19, politics, elections, climate, and journalism to cover internal discourses and others for political journalists. For data journalists, these were COVID-19, politics, elections, DDJ (data-driven journalism), journalism (for non-data journalism-related internal discourses), climate, sports, journalism, and others.

## 4. Results

Men are not only over-represented in our sample, but they also tweet significantly more (724.47 tweets per man/289.23 tweets per woman across both groups on average). Consequently, men created a large majority of tweets. Women wrote less than 19% of data journalist tweets and only 13% of political journalists' tweets (Table 2). This is also consistent with data journalists, although not in a similar dimension. Male political journalists also use more mentions on average, measured by extracting all strings prefixed by an 'at' sign, which is Twitter's specification for tagging usernames. Women political and data journalists receive slightly more retweets on average.

**Table 2.** Summary statistics of gender, retweets, and mentions of political (P) and data (D) journalists' tweets.

| | | n | Share | $\overline{Tweets}$ | $\overline{Retweets}$ | $\overline{Mentions}$ |
|---|---|---|---|---|---|---|
| P | m | 375,582 | 0.87 | 4803.0 | 248.6 | 1.14 |
| | f | 55,211 | 0.13 | 1415.5 | 293.6 | 1.25 |
| D | m | 38,815 | 0.81 | 1425.2 | 484.1 | 1.34 |
| | f | 8997 | 0.19 | 1358.0 | 506.0 | 1.41 |

### *4.1. Data Journalists Have a More Open Discourse*

Part of our research focused on a general question about the arena of debate that takes part on Twitter. By extracting all mentions and comparing these users to our pre-compiled lists by cross-tabulation, we can show the share of references that stay within the political and data journalistic network.

Of all mentions by political journalists, 10.7% are referenced within our sample, and 89.3% are outside of our sample. This number is even lower for data journalists. Only 8.2% of mentions are within the data journalistic community, and nearly 92% are elsewhere. We found a statistically significant difference ($\chi^2$ = 429.06, $p$ < 0.001, df = 1) between the two groups, with data journalists incorporating more outsiders into their discourses, therefore representing a less closed network compared to political journalists, which confirms H1.

### *4.2. Women Favor Their Peers in Mentions*

We already show in Table 2 that there is a gap between the gender share of tweets and the gender share of users. This divergence can also be observed in the cross-tabulated share of mentions. This analysis only applies to tweets among our observed journalist users because we cannot derive the gender of others.

Women users tend to favor their peers when mentioning others within the journalistic bubble. Political journalists mention their peers in 27.4% of mentions, which is close to their share in the sample but more than their share on all tweets in the sample. This effect is even more pronounced for mentioning female data journalists; they mentioned other female data journalists in 35.9% of intra-data journalistic discourses, which is even higher than the share of women in the sample.

> "RT @mjKolly: Open question: How could and should people in the media industry credit each other's work?"—@datentaeterin (1 February 2021 03:41:12 p.m.)

> "RT @datentaeterin: "Anyone who wants to work in journalism should be able to handle data," says @ChElm in an interview with @journocode. That's why she wants to anchor data skills more firmly in education, for example, at the @IJ_Online #ddj"—@daten_drang (6 October 2021 07:37:58 p.m.)

Men, in comparison, only mentioned women in 17.0% of cases for political journalists and 20.7% for data journalists; see Table 3. A chi-squared analysis showed statistically significant results for both groups ($p$ < 0.001). The effect size $\phi$ is 0.10 for political journalists ($\chi^2$ = 549.78, df = 1) and 0.13 for data journalists ($\chi^2$ = 112.03, df = 1), demonstrating a small effect. A contribution analysis shows that the mentioning of women by women composes 66.48% of the measured effect for political journalists and 63.76% for data journalists. We are, therefore, able to confirm H2.

**Table 3.** Gender of mentioned users by author's gender.

| | | Mentioned | | |
|---|---|---|---|---|
| | | **m** | **f** | |
| Mentioning Political | m | 83.0% | 17.0% | $\chi^2$ = 549.78, $p < 0.01$ |
| | f | 72.6% | 27.4% | |
| Mentioning Data | m | 79.3% | 20.7% | $\chi^2$ = 112.03, $p < 0.01$ |
| | f | 64.1 % | 35.9% | |

### *4.3. Retweets Are More Evenly Distributed for Data Journalists*

While mentions are unevenly shared between genders in both groups, this is not identical concerning retweets. Women political journalists are only retweeted by men in 13.3% of intra-journalistic retweets; male data journalists only share tweets of women in 18.4% of cases, as shown in Table 4. Again, the share of women retweeting other women is higher in both groups but lower than their share of users in both cases. Pearson's chi-squared test shows a statistically significant result for political journalists. For data journalists, the results are not significant. The effect size of 0.10 is small for political journalists and even smaller for data journalists. While residues and contributions favor an effect among women for political journalists, this is not the case for data journalists. The effect on them seems to be much smaller. H3 can certainly be confirmed for political journalists but not for data journalists. See Figure 1 for a full-size network illustration of retweets among political and data journalists.

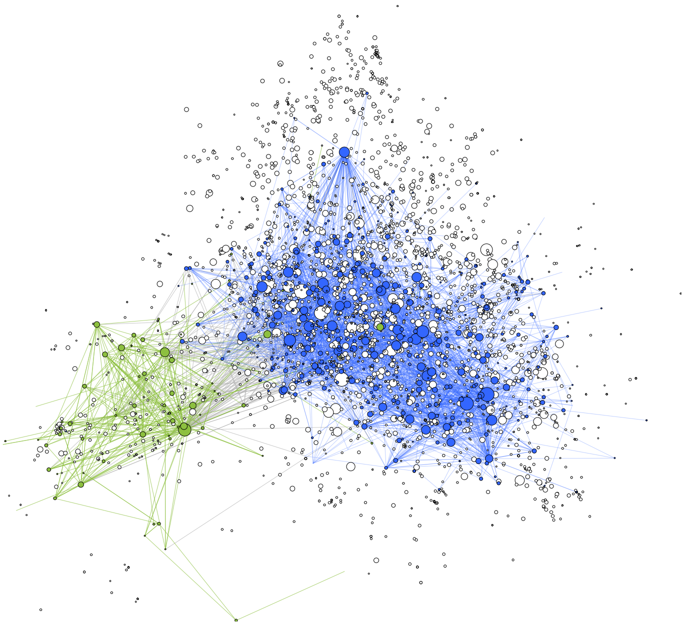


**Figure 1.** Retweet network of political and data journalists created using a force-directed layout (Kamada and Kawai 1989): showing edges when at least two retweets were sent, node sizes by degrees. Green nodes and edges: data journalists ($n$ = 118), blue nodes and edges: political journalists ($n$ = 546), white nodes: unknown (users were not part of pre-defined lists, $n$ = 6765), gray edges: connections between political and data journalists.

**Table 4.** Gender of retweeted user by author's gender.

| | | Retweeted | | |
|---|---|---|---|---|
| | | m | f | |
| Retweeting Political | m | 86.7% | 13.3% | $\chi^2 = 326.42$,<br>$p < 0.01$ |
| | f | 76.5% | 23.5% | |
| Retweeting Data | m | 81.7 % | 18.3% | $\chi^2 = 1.56$,<br>$p > 0.05$ |
| | f | 78.5% | 21.5% | |

### *4.4. Sources Differ Between Genders and Journalistic Disciplines*

To analyze the source references, we compare all usernames retweeted by our population.

When examining retweeted accounts, we find that only a portion of the sources is shared between genders for both groups of journalists. Political journalists used 20,937 users as sources, with 19% common across both genders for political journalists and around 14% for data journalists, out of a total of 6519, confirming H4.

We also find differences in self-sourcing shares, where users retweet their own tweets for their audience. The share of self-retweets is 2.57 times significantly higher for political journalists than for data journalists. In particular, male political journalists significantly self-retweet themselves a lot more than their female colleagues. In contrast, this finding is exactly the opposite for data journalists—although in a much smaller size. A logistic regression was used to analyze the relationship between gender, the area of journalism, and self-retweeting behavior. It was found that—holding all other predictor variables constant—the odds ratio of a self-retweet occurring increased on average by 6.56 (95% CI 5.19, 8.3) for the occurrence of the male sex. It was also found that, under the same conditions, the odds ratio of a self-retweet occurring increased on average by 2.43 (95% CI 2.02, 2.93) for the occurrence of political journalism.

To further understand possible clusters of sources, we added a content analysis at this stage:

First, we manually coded the top 100 retweeted sources for each sex and field. While political journalists used German media and other political journalists as their primary sources across the sexes, with a few men referring to foreign media, data journalists also leveraged non-peer journalists in their retweets. Female data journalists relied strongly on political, science, or other sources, while men seemed to strongly emphasize their peers (see Table 5).

**Table 5.** Shares of clusters of the top 100 sources by the sex (female/male) of political and data journalists.

| | Political | | Data | |
|---|---|---|---|---|
| Cluster | F | M | F | M |
| Intra-group journalists | 49.00 | 48.49 | 22.75 | 29.93 |
| Extra-group journalists | 5.66 | 1.17 | 25.14 | 27.84 |
| Media | 37.31 | 33.86 | 19.27 | 19.36 |
| Foreign media | 1.24 | 10.55 | 0.5 | 1.98 |
| Politic | 1.21 | 2.46 | 5.87 | 1.58 |
| Science | 0.55 | 0.97 | 5.13 | 3.94 |
| NGO | 3.44 | 2.08 | 10.09 | 11.12 |
| Visual | - | - | 3.14 | 2.13 |
| Others | 1.59 | 0.43 | 8.11 | 2.11 |

Second, an analysis of the 100 most used hashtags revealed little differences for political journalists but a much more diverse set of topics for data journalists, confirming H5, with female data journalists seemingly communicating different topics than their male colleagues, like politics, intra-journalistic discourses, and an increased share of other topics, but less COVID-19 and no sports coverage (see Tables 6 and 7).

**Table 6.** Shares of clusters of the top 100 hashtags by the sex (female/male) of political journalists.

| Cluster | F | M |
|---|---|---|
| Politics | 41.20 | 40.06 |
| COVID-19 | 32.53 | 39.65 |
| Elections | 19.04 | 10.42 |
| Climate | 3.78 | 0.41 |
| Others | 2.43 | 8.23 |
| Journalism | 1.01 | 1.23 |

**Table 7.** Shares of clusters of the top 100 hashtags by the sex (female/male) of data journalists.

| Cluster | F | M |
|---|---|---|
| COVID-19 | 24.41 | 37.19 |
| Data journalism | 31.07 | 30.70 |
| Politics | 17.89 | 11.18 |
| Elections | 11.58 | 11.59 |
| Journalism | 7.56 | 2.37 |
| Others | 4.99 | 2.49 |
| Climate | 2.50 | 2.13 |
| Sports | - | 2.35 |

## 5. Networks

To confirm our insights and show the utility of network analysis for this task, we modeled the data as four distinct networks for each profession: an internal profession retweet network, a retweet network that incorporates all internal and external retweets, an internal network of mentions, and a network of hashtags used in tweets.

### *5.1. Internal Retweets and Mentions*

To further understand the dynamics of retweets, we created a retweet network. Purple nodes represent women, and green nodes represent men. Gray edges represent at least two retweets between men in both directions. Purple edges represent at least two retweets between women, and green edges represent a retweet connection between a male and a female user. For data journalists, we show edges for at least one retweet in both directions, as the network is much smaller.

While women's networks are hard to spot on the network of political journalists (see Figure 2), we find clusters of affiliations between different publishers. While reporters and editors for the media company Axel Springer and its outlets are closely connected on the left, journalists of Der Spiegel or Süddeutsche Zeitung are found on the lower right.

The data journalists' network does not show similar patterns, which the smaller team sizes in the field could influence (see Figure 3).

We found the values of in-degree ($t(642) = 2.0341, p < 0.05$), out-degree ($t(440.09) = 2.9155, p < 0.01$), and Kleinberg's authority centrality score ($t(642) = 2.0108, p < 0.05$) to be statistically significant for political journalists, but not for data journalists. See Table 8 for network property metrics.

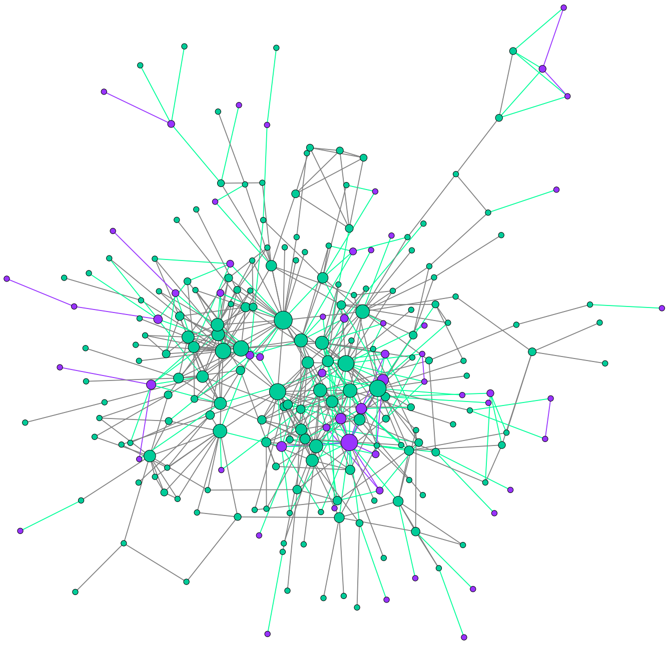


**Figure 2.** Internal retweet network of political journalists by gender, created using a force-directed layout (Kamada and Kawai 1989); graph network of retweets by German political journalists, showing edges when both nodes send at least 2 mutual retweets, node sizes by degrees. Purple nodes: women ($n$ = 58), green nodes: men ($n$ = 233), green edges: men–women, purple edges: both women, gray edges: both men.

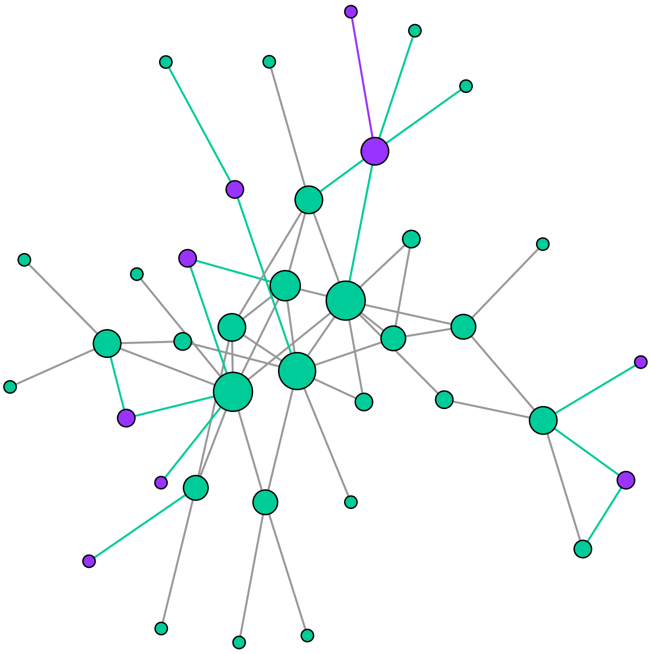


**Figure 3.** Internal retweet network of data journalists by gender created using a force-directed layout (Kamada and Kawai 1989); graph network of retweets by German data journalists, showing edges when both nodes send at least one mutual retweet, node sizes by degrees. Purple nodes: women ($n$ = 9), green nodes: men ($n$ = 29), green edges: men–women, purple edges: both women, gray edges: both men.

**Table 8.** Properties for internal networks of mentions and retweets for political (P) and data journalists (D).

| | Mentions | | Retweets | |
|---|---|---|---|---|
| **Property** | **P** | **D** | **P** | **D** |
| Mean Dist. | 2.81 | 2.34 | 3.05 | 2.76 |
| Edge Density | 0.03 | 0.088 | 0.02 | 0.05 |
| Reciprocity | 0.48 | 0.53 | 0.33 | 0.28 |
| No. of Nodes | 644 | 134 | 569 | 116 |
| No. of Edges | 11,515 | 1565 | 7255 | 731 |

### 5.2. External Retweets

Creating networks of sources for both sexes and professions leads to further conclusions. We compare both sexes and areas of journalism in all their retweeted messages. This also included outsiders of their journalistic circles, different from the analysis above. The intent was not only to understand the journalistic communities' internal structures but also their differences in leveraging different external actors. The network was constructed by defining all users as nodes and the retweets of each user as edges.

Women's source networks for political journalists have a higher mean distance between nodes than male political journalists (see Table 9). This indicates that these networks are further spread out, while female data journalists form a much more compact source network.

Reciprocity numbers indicate a more coherent sourcing behavior for political journalists, who seem to reference themselves more than data journalists (as described above). This can also be seen in transitivity metrics describing how likely adjacent nodes are connected, revealing tightly connected communities. Political journalists of both sexes have metrics that are an order of magnitude higher than data journalists.

To further enable the comparison, we calculated a weighted E-I index (Krackhardt and Stern 1988) to show the edge density between the internal and external connections for the source (retweet) network, combining both groups of journalists and all their retweeted accounts. The formula is as follows:

$$E - I\ Index = \frac{E - I}{E + I}$$

$E$ is the sum of external retweet ties (the number of times a username was retweeted from a journalist), and $I$ is the sum of internal retweet ties (the count of retweets internal to the network of journalists).

**Table 9.** Properties for networks of sources for political and data journalists of women (F) and men (M).

| | Political | | Data | |
|---|---|---|---|---|
| **Property** | **F** | **M** | **F** | **M** |
| Mean Dist. | 4.69 | 3.57 | 2.50 | 3.46 |
| Reciprocity | 0.016 | 0.036 | 0.0026 | 0.0217 |
| Transitivity | 0.009 | 0.017 | 0.0012 | 0.0079 |
| Mean Indegree | 556.31 | 2863.24 | 492.09 | 632.40 |
| Mean Outdegree | 149.73 | 239.31 | 19.54 | 24.03 |
| No. of Nodes | 6063 | 19,024 | 1985 | 5499 |
| No. of Edges | 23,084 | 136,086 | 3872 | 14,382 |
| E-I Index | 0.594 | 0.623 | 0.837 | 0.751 |

For the whole network of retweets, this returns 0.6235, indicating a strong connection to outside groups. As this is an overarching network of multiple subgroups, this is not surprising. For the individual retweet networks of political journalists, this number is

0.6184. For the retweet network of data journalists, it is 0.7688, indicating more external ties for data journalists. The numbers for the sex-separated networks are reported in Table 9.

### *5.3. Hashtags*

Lastly, a two-mode network of hashtags helps to deepen the understanding of the diversity of topics that the journalistic domains mention in their tweets. This network was created by defining hashtags and users as nodes and tweets containing certain hashtags as edges.

While we have found a consensus within the most prevalent hashtags for political journalists above, we can see that women political journalists form a hashtag network with nine components, which indicates that there are separate, unconnected parts of the network, which we can identify as users that do not use hashtags at all. This is also observable for male political journalists, but only with two components, not for data journalists, who seem to form a joint network connected via shared hashtags.

We use the average ratio $\frac{UsedHashtagsByUser}{HashtagUsedByUsers}$ to compare groups for an indication of the prevalence of hashtag use. It is important to note that indegrees and outdegrees are not typically reported for two-mode networks. However, we included this comparison in this particular case to provide a more comprehensive analysis. We find that women political journalists, on average, create 285.12 outdegrees, with hashtags averaging 34.75 indegrees, on average, a ratio of 108.46. On the contrary, male political journalists have an average of 679.95 outdegrees, their hashtags' degrees being on a similar level to women with 27.52 indegrees, at a ratio of 280.01. Data journalists have much lower values, with women creating, on average, 138.31 outdegrees and 6.46 indegrees (ratio: 88.46), and males ending up with 178.23 outdegrees and 6.43 indegrees, at a ratio of 108.35. This shows that political journalists use, on average, much more hashtags than data journalists, with men being ahead in each case.

## 6. Discussion

We have shown differences in mentioning and retweeting behavior between the sexes among political and data journalists in Germany, confirming H2 and H3 for political journalists.

Women tend to mention their peers more often in tweets than men. Since men comprise the larger share of the Twitter network, they tend to be more visible. The differences could make women and their work less apparent on Twitter, therefore receiving less amplification and legitimization. This work provides a non-US perspective on the differences in Twitter communication styles between sexes and different groups of journalists (Maares et al. 2021). The results indicate existing norms within newly created public communication spheres, pointing to a selective, gatekeeping process on the disseminator side of information (White 1950).

This effect can also be found in retweets of political journalists, although it is not similarly strong for data journalists. This might indicate that data journalists share the work of others with less regard for the sexes compared to their colleagues in political reporting. However, women journalists show higher rates of retweet behavior than their peers. Since the effect appears in both groups, this indicates that women pay greater attention to tweets by other women; however, this effect is much more solid for political journalists. These results confirm the earlier work by (Usher et al. 2018).

Although their share of identical sources across sexes is low, political journalists tend to focus their retweets of central issues mainly on direct peers or media sources. In contrast, data journalists seem to convey their most prevalent sources from a more diverse spectrum of backgrounds and less from other data journalists. This might indicate the broader background that data journalism has as a discipline (Hannaford 2022) and again points to the homophilous network structures of political journalists that have been described before (Hanusch and Nölleke 2019; Molyneux 2015; Molyneux and Mourão 2019; Mourão 2015; Mourão et al. 2016; Nuernbergk 2016). However, this finding needs to be regarded in combination with a general contrast between sexes, which points to the fact that female and male journalists retweet in parts different voices on Twitter.

Women political journalists' source networks have a higher mean distance but also seem to be more closed than those of men, given the lower E-I index. There seems to be a contrasting finding regarding the higher number of average hashtags for political journalists, which might indicate a higher diversity of topics for this area but could also point to the fact that hashtags emerge quickly for breaking-news political events rather than for data journalist topics, which rarely involve reporting news up to the minute (Lin et al. 2021; Vicari et al. 2018; Zhang 2017), as we do not find indications for this when analyzing the most prevalent hashtags that have a strong focus on politics, elections, and COVID-19. This, however, might be concealed for this method and is a starting point for further research.

While analyzing retweet networks, we found visual evidence of clusters of affiliations that might impact tweet behavior, which might also be a vantage point for further research.

We found relatively high percentages for external mentions and retweets for both journalistic groups but with data journalists having a greater share of external ties than political journalists. This is partly in line with Nuernbergk (2016), who identified a journalistic–political Twittersphere, mainly referencing each other, and is consistent with similar findings by Mourão (2015) and Molyneux and Mourão (2019). However, we need to point out that, as this work did not include a broad set of politicians in the sample or account for different groups of non-journalistic actors, apart from highlighting the most commonly retweeted users and focusing more on gender differences; therefore, the results are not fully comparable with those earlier results. We also did not include Twitter accounts of media companies in our analysis, which might represent a large share of mentioned or retweeted users, as seen by visual inspections and source clustering.

Acknowledging the methodological limitations of our study, we must highlight the challenges encountered in the data collection process. Due to the inherent differences between political journalism and data journalism—the former being a distinct beat and the latter encompassing both a beat and a method applicable across various beats—identifying and collecting samples for each group presented unique challenges. Political journalists, typically identifiable through imprints, allowed for a more straightforward manual collection. In contrast, data journalists, whose roles may not be explicitly specified within a media organization's data department, require a more inclusive approach to ensure representation.

To this end, we used a dual-strategy approach to data collection: manual collection for political journalists and utilization of a Slack working group for data journalists. While an identical collection mechanism for both groups would have been ideal, the nature of data journalism and the absence of a comparable list for political journalists dictated our methodology. We believe that leveraging the Slack working group not only facilitated a better inclusion of data journalists—who are often less visibly defined within organizational structures—but also addressed the challenge of adequately representing this diverse group in our study. However, this methodological divergence might introduce some limitations in the direct comparison between the two groups, especially as public broadcasting journalists are not included in the political journalists' sample, therefore, reflecting the differences in how these journalistic practices are embedded within media organizations.

We also need to clarify that the networks we investigated were formed by retweets and mentions, aiming to create the internal social media amplification of messages, ignoring the latent network of followers, and following relations that might lead to messages being transported outside the social media network. Our research is further limited by the influence of Twitter's algorithms, which might have shaped the tweets shown to users. However, at the time of analysis, Twitter still used a chronological order for displaying tweets. This study focuses exclusively on German journalists, which restricts the generalizability of the findings to other countries or cultures.

Due to the practical constraints of manual coding, hashtag clustering was confined to the top 100 by count, leaving out the long tail. Our objective was to capture the most significant drivers of communication in journalist networks on Twitter, focusing on the central themes and sources that are the most impactful and representative of the discourse. This could be an area for further research, exploring the long tail of hashtags and sources and potentially using automated categorization or machine learning techniques to manage more extensive datasets efficiently.

As often with quantitative methods, they lack depth in understanding the reasons behind the observed behaviors. Qualitative methods, such as structured interviews, could provide further research opportunities into the motivations, perceptions, and challenges that journalists face on Twitter, as could the additional collection of data. For instance, differences in age structures may influence communication behavior as well, which might have affected this analysis due to the age differences between the groups of journalists. This could enhance the understanding of how gender dynamics manifest in the digital interactions of journalists.

We could, however, find a significant difference in shares of those internal discourses between the two groups of journalists, with data journalists being less locked than political journalists. That might indicate a greater openness to the influence of others in the data journalistic community, which is a finding that could need closer examination.

## 7. Conclusions

We compared the Twitter (now X) networks of German political and data journalists to analyze the differences in communication between women and men. We found a difference in the proportions of internal discourses within the two groups of journalists. Data journalists tended to have fewer internal discussions on Twitter than political journalists. However, we could not reproduce earlier findings, which showed an elitist network of political journalists on Twitter.

This study showed that men dominated the number and share of tweets in networks of political and data journalists in Germany. Women were much less mentioned and retweeted by men, while other women tended to favor their peers. This effect was visible in both groups for mentions and was also observable for retweets by political journalists and, to a lesser degree, for data journalists. This indicates a different perception of the work and arguments made by colleagues on Twitter between genders, which might lead to less amplification and legitimization of women's voices on Twitter. Further research is required to extract the causes behind this effect and the possibilities of countering this behavior.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/journalmedia5010027/s1. Table S1: Clusters of Sources: Political Journalists, Table S2: Clusters of Sources: Data Journalists, Table S3: Cluster of Hashtags: Political Journalists, Table S4: Cluster of Hashtags: Data Journalists.

**Author Contributions:** Conceptualization, B.W. and J.P.; data curation, B.W.; formal analysis, B.W.; investigation, B.W.; methodology, B.W.; project administration, B.W.; software, B.W.; Supervision, J.P.; validation, J.P.; visualization, J.P.; writing—original draft, B.W. and J.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data are available upon request.

**Acknowledgments:** The authors would like to thank Rebecca Leipsic for helping to compile the list of political journalists.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest.

## References

Adams, Catherine. 2018. "They Go for Gender First". *Journalism Practice* 12: 850–69. [CrossRef]

Anderton-Yang, David, Nicolas Kayser-Bril, Alexander Howard, César Viana Teixeira, Sarah Slobin, and Jerry Vermanen. 2012. Why is data journalism important? In *The Data Journalism Handbook 1*. Edited by Jonathan Gray, Liliana Bounegru, Lucy Chambers, European Journalism Centre and Open Knowledge Foundation. Sebastopol: OŔeilly.

Andi, Simge, Meera Selva, and Rasmus Kleis Nielsen. 2020. *Women and Leadership in the News Media 2020: Evidence from Ten Markets*. Oxford: Reuters Institute for the Study of Journalism.

Antonopoulos, Nikos, and Minos-Athanasios Karyotakis. 2020. Data journalism. In *The SAGE International Encyclopedia of Mass Media and Society*. Edited by Debra L. Merskin. Thousand Oaks: SAGE Publications, Inc., pp. 440–41. [CrossRef]

Appelgren, Ester, and Gunnar Nygren. 2014. Data Journalism in Sweden. *Digital Journalism* 2: 394–405. [CrossRef]

Armstrong, Cory L. 2004. The Influence of Reporter Gender on Source Selection in Newspaper Stories. *Journalism & Mass Communication Quarterly* 81: 139–54. [CrossRef]

Asur, Sitaram, Bernardo A. Huberman, Gabor Szabo, and Chunyan Wang. 2011. Trends in Social Media: Persistence and Decay. Paper presented at the International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17–21. vol. 5, pp. 434–37.

Barnard, Stephen R. 2014. 'Tweet or be sacked': Twitter and the new elements of journalistic practice. *Journalism* 17: 190–207. [CrossRef]

Beiler, Markus, Felix Irmer, and Adrian Breda. 2020. Data Journalism at German Newspapers and Public Broadcasters: A Quantitative Survey of Structures, Contents and Perceptions. *Journalism Studies* 21: 1571–89. [CrossRef]

Berret, Charles, and Cheryl Phillips. 2016. *Teaching Data and Computational Journalism*. New York: Columbia School of Journalism.

Bourdieu, Pierre. 1986. The Forms of Capital. In *Handbook of Theory and Research for the Sociology of Education*. Edited by John G. Richardson. Westport: Greenwood Press, pp. 241–58.

Bourdieu, Pierre. 1993. *The Field of Cultural Production*. New York: Columbia University Press.

Bravo, Adolfo Antón, and Ana Serrano Tellería. 2020. Data Journalism: From Social Science Techniques to Data Science Skills. *Hipertext.net* 20: 41–54. [CrossRef]

Byerly, Carolyn. 2011. *Global Report on the Status of Women in the News Media*. Washington, DC: International Women's Media Foundation (IWMF).

Carlson, Matt. 2017. *Journalistic Authority: Legitimating News in the Digital Era*. New York: Columbia University Press. [CrossRef]

Chadwick, Andrew. 2011. The Political Information Cycle in a Hybrid News System: The British Prime Minister and the "Bullygate" Affair. *The International Journal of Press/Politics* 16: 3–29. [CrossRef]

Chambers, Deborah, Linda Steiner, and Carole Fleming. 2004. *Women and Journalism*. London: Routledge.

Chen, Gina Masullo, Paromita Pain, Victoria Y. Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger. 2020. 'You Really Have To Have a Thick Skin': A Cross-Cultural Perspective on How Online Harassment Influences Female Journalists. *Journalism* 21: 877–95. [CrossRef]

Coddington, Mark. 2015. Clarifying journalism's quantitative turn. *Digital Journalism* 3: 331–48. [CrossRef]

Craft, Stephanie, and Wayne Wanta. 2004. Women in the Newsroom: Influences of Female Editors and Reporters on the News Agenda. *Journalism & Mass Communication Quarterly* 81: 124–38. [CrossRef]

Degen, Matthias, and Max Olgemöller. 2021. German Political Journalists and the Normalization of Twitter. *Journalism Studies* 22: 1317–38. [CrossRef]

De Vuyst, Sara. 2018. Cracking the coding ceiling: Looking at gender construction in data journalism from a field theory perspective. *Journal of Applied Journalism & Media Studies* 7: 387–405. [CrossRef]

Diakopoulos, Nicholas. 2016. BuzzFeed's Pro Tennis Investigation Displays Ethical Dilemmas of Data Journalism. *Columbia Journalism Review*. Available online: https://www.cjr.org/tow_center/transparency_algorithms_buzzfeed.php (accessed on 10 March 2024).

European Institute for Gender Equality (EIGE). 2013. *Women and the Media. Advancing Gender Equality in Decision-Making in Media Organisations*. Luxembourg: Publications Office of the European Union. [CrossRef]

Fincham, Kelly. 2019. Exploring political journalism homophily on Twitter: A comparative analysis of U.S. and U.K. elections in 2016 and 2017. *Media and Communication* 7: 213–24. [CrossRef]

Fleras, Augie. 2003. *Mass Media Communication in Canada*. Toronto: Nelson.

Haim, Mario. 2022. The German Data Journalist in 2021. *Journalism Practice* 1–20. [CrossRef]

Hamby, Peter. 2013. *Did Twitter Kill the Boys on the Bus? Searching for a Better Way to Cover a Campaign*. Harvard: Joan Shorenstein Center on the Press, Politics, and Public Policy.

Hanitzsch, Thomas, Nina Steindl, and Corinna Lauerer. 2016. *Country Report: Journalists in Germany*. Munich: Universitätsbibliothek der Ludwig-Maximilians-Universität München. [CrossRef]

Hannaford, Liz. 2022. The discourses of data journalism. *Journalism* 24: 2397–417. [CrossRef]

Hannis, Grant, and Cathy Strong. 2007. The visibility of female journalists in Australian and New Zealand newspapers: The good news and the bad news. *Australian Journalism Review* 29: 115–25.

Hanusch, Folker. 2018. Political Journalists' Corporate and Personal Identities on Twitter Profile Pages: A Comparative Analysis in Four Westminster Democracies. *New Media & Society* 20: 1488–505. [CrossRef]

Hanusch, Folker, and Daniel Nölleke. 2019. Journalistic Homophily on Social Media. *Digital Journalism* 7: 22–44. [CrossRef]

Hermida, Alfred. 2010. Twittering the News. *Journalism Practice* 4: 297–308. [CrossRef]

Howard, Alexander. 2014. *The Art and Science of Data-Driven Journalism*. New York: Tow Center for Digital Journalism. [CrossRef]

Hunter, Andrea. 2019. From the Woman's Page to the Digital Age: Women in Journalism. In *Working Women in Canada: An Intersectional Approach*. Edited by Leslie Nichols. Toronto: Canadian Scholars, pp. 287–302.

Josephi, Beate, Folker Hanusch, Martin Oller Alonso, Ivor Shapiro, Kenneth Andresen, Arnold de Beer, Abit Hoxha, Sonia Virgínia Moreira, Kevin Rafter, Terje Skjerdal, and et al. 2019. *Profiles of Journalists: Demographic and Employment Patterns*. New York: Columbia University Press, vol. 4, pp. 67–102. [CrossRef]

Kamada, Tomihisa, and Satoru Kawai. 1989. An algorithm for drawing general undirected graphs. *Information Processing Letters* 31: 7–15. [CrossRef]

Kay, Linda. 2012. *The Sweet Sixteen: The Journey That Inspired the Canadian Women's Press Club*. Montreal: McGill-Queen's University Press.

Kian, Edward M., and Marie Hardin. 2009. Framing of Sport Coverage Based on the Sex of Sports Writers: Female Journalists Counter the Traditional Gendering of Media Coverage. *International Journal of Sport Communication* 2: 185–204. [CrossRef]

Krackhardt, David, and Robert N. Stern. 1988. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly* 51: 123. [CrossRef]

Kreiss, Daniel. 2016. Seizing the moment: The presidential campaigns' use of Twitter during the 2012 electoral cycle. *New Media & Society* 18: 1473–90. [CrossRef]

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York: Association for Computing Machinery, pp. 591–600. [CrossRef]

Lang, Marjory. 1999. *Women Who Made the News: Female Journalists in Canada, 1880–1945*. Montreal: McGill-Queen's University Press.

Lasorsa, Dominic. 2012. Transparency and Other Journalistic Norms on Twitter. *Journalism Studies* 13: 402–17. [CrossRef]

Lasorsa, Dominic, Seth C. Lewis, and Avery E. Holton. 2012. Normalizing Twitter. *Journalism Studies* 13: 19–36. [CrossRef]

Lawrence, Regina G., Logan Molyneux, Mark Coddington, and Avery Holton. 2014. Tweeting Conventions. *Journalism Studies* 15: 789–806. [CrossRef]

Lazarsfeld, Paul F. 1944. *The People's Choice*. New York: Columbia University Press.

Lazarsfeld, Paul F., and Robert K. Merton. 1954. Friendship as a Social Process: A Substantive and Methodological Analysis. *Freedom and Control in Modern Society* 18: 18–66.

Lewin, Kurt. 1947. Frontiers in group dynamics. *Human Relations* 1: 143–53. [CrossRef]

Lewis, Seth C., and Oscar Westlund. 2015. Big Data and Journalism: Epistemology, expertise, economics, and ethics. *Digital Journalism* 3: 447–66. [CrossRef]

Li, Qin, Hans J. G. Hassell, and Robert M. Bond. 2023. Journalists' networks: Homophily and peering over the shoulder of other journalists. *PLoS ONE* 18: e0291544. [CrossRef] [PubMed]

Lin, Yu-Ru, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. 2021. #Bigbirds never die: Understanding social dynamics of emergent hashtags. Paper presented at the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA, July 8–11. Palo Alto: Association for the Advancement of Artificial Intelligence, vol. 7, pp. 370–79. [CrossRef]

Lippmann, Walter. 1922. *Public Opinion*. New York: Free Press.

Maares, Phoebe, Fabienne Lind, and Esther Greussing. 2021. Showing off Your Social Capital: Homophily of Professional Reputation and Gender in Journalistic Networks on Twitter. *Digital Journalism* 9: 500–17. [CrossRef]

Matusitz, Jonathan, and Gerald-Mark Breen. 2012. An Examination of Pack Journalism as a Form of Groupthink: A Theoretical, and Qualitative Analysis. *Journal of Human Behavior in the Social Environment* 22: 896–915. [CrossRef]

Mayer-Schönberger, Viktor, and Cukier Lenneth. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27: 415–44. [CrossRef]

Meyer, Philip. 1973. *Precision Journalism: A Reporter's Introduction to Social Science Methods*, 1st ed. Bloomington: Indiana University Press.

Meyer, Philip. 2002. *Precision Journalism: A Reporter's Introduction to Social Science Methods*, 4th ed. Lanham: Rowman & Littlefield.

Molyneux, Logan. 2015. What journalists retweet: Opinion, humor, and brand development on Twitter. *Journalism* 16: 920–35. [CrossRef]

Molyneux, Logan, and Rachel R. Mourão. 2019. Political Journalists' Normalization of Twitter. *Journalism Studies* 20: 248–66. [CrossRef]

Mourão, Rachel. 2015. The Boys on the Timeline: Political Journalists' Use of Twitter for Building Interpretive Communities. *Journalism* 16: 1107–23. [CrossRef]

Mourão, Rachel, Trevor Diehl, and Krishnan Vasudevan. 2016. I Love Big Bird: How journalists tweeted humor during the 2012 presidential debates. *Digital Journalism* 4: 211–28. [CrossRef]

Netzwerk Recherche. 2020. Fachgruppe Datenjournalismus. Available online: https://netzwerkrecherche.org/ueber-uns/vorstand/fachgruppe-datenjournalismus/ (accessed on 29 October 2021).

North, Louise. 2016. The Gender of "soft" and "hard" news. *Journalism Studies* 17: 356–73. [CrossRef]

Nuernbergk, Christian. 2016. Political Journalists' Interaction Networks. *Journalism Practice* 10: 868–79. [CrossRef]

Nuernbergk, Christian, and Jan-Hinrik Schmidt. 2020. Twitter im Politikjournalismus. *Publizistik* 65: 41–61. [CrossRef]

Ottovordemgentschenfelde, Svenja. 2017. 'Organizational, professional, personal': An exploratory study of political journalists and their hybrid brand on Twitter. *Journalism* 18: 64–80. [CrossRef]

Parasie, Sylvain, and Eric Dagiral. 2012. Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New Media & Society* 15: 853–71. [CrossRef]

Paulussen, Steve, and Raymond A. Harder. 2014. Social Media References in Newspapers. *Journalism Practice* 8: 542–51. [CrossRef]

Pfeffer, Juergen, Angelina Mooseder, Jana Lasser, Luca Hammer, Oliver Stritzel, and David Garcia. 2023. This sample seems to be good enough! assessing coverage and temporal reliability of Twitter's Academic API. Paper presented at the Seventeenth International AAAI Conference on Web and Social Media (ICWSM-2023), Limassol, Cyprus, June 5–8.

Rogers, Simon. 2010. Florence Nightingale, datajournalist: Information has always been beautiful. *The Guardian*, August 13.

Sarikakis, Katharine, Bruktawit Ejigu Kassa, Natascha Fenz, Sarah Goldschmitt, Julia Kasser, and Laura Nowotarski. 2021. "My haters and I": Personal and political responses to hate speech against female journalists in Austria. *Feminist Media Studies* 23: 67–82. [CrossRef]

Schultz, Ida. 2007. The journalistic gut feeling: Journalistic doxa, news habitus and orthodox news values. *Journalism Practice* 1: 190–207. [CrossRef]

Smith, Roger. 1981. Women and Occupational Elites: The Case of Newspaper Journalism in England. In *Access to Power: Cross-National Studies of Women and Elites*. Edited by Rose Laub Coser and Cynthia Fuchs Epstein. London: Routledge.

Stahel, Lea, and Constantin Schoen. 2020. Female Journalists Under Attack? Explaining Gender Differences in Reactions to Audiences' Attacks. *New Media & Society* 22: 1849–67. [CrossRef]

Steindl, Nina, Corinna Lauerer, and Thomas Hanitzsch. 2017. Journalismus in Deutschland. *Publizistik* 62: 401–23. [CrossRef]

Steiner, Linda. 2008. Gender in the Newsroom. In *The Handbook of Journalism Studies*. Edited by Karin Wahl-Jorgensen and Thomas Hanitzsch. London: Routledge.

Sui, Mingxiao, Newly Paul, Caley Hewitt, Jessica Maki, and Kathleen Searles. 2022. Is news for men?: Effects of women's participation in news-making on audience perceptions and behaviors. *Journalism* 25: 41–60. [CrossRef]

Tandoc, Edson C., and Tim P. Vos. 2016. The Journalist Is Marketing the News. *Journalism Practice* 10: 950–66. [CrossRef]

Tong, Jingrong, and Landong Zuo. 2019. The Inapplicability of Objectivity: Understanding the Work of Data Journalism. *Journalism Practice* 15: 153–69. [CrossRef]

Usher, Nikki, Jesse Holcomb, and Justin Littman. 2018. Twitter Makes It Worse: Political Journalists, Gendered Echo Chambers, and the Amplification of Gender Bias. *The International Journal of Press/Politics* 23: 324–44. [CrossRef]

Vergeer, Maurice. 2014. Peers and Sources As Social Capital in the Production of News: Online Social Networks As Communities of Journalists. *Social Science Computer Review* 33: 277–97. [CrossRef]

Vicari, Stefania, Laura Iannelli, and Elisabetta Zurovac. 2018. Political hashtag publics and counter-visuality: A case study of #fertilityday in italy. *Information, Communication & Society* 23: 1235–54. [CrossRef]

von Garmissen, Anna, and Hanna Biresch. 2019. *Welchen Anteil haben Frauen an der publizistischen Macht in Deutschland?* Hamburg: ProQuote Medien e.V. Available online: https://www.pro-quote.de/wp-content/uploads/2019/11/ProQuote-Studie_print_online_digital-2019.pdf (accessed on 10 March 2024).

von Nordheim, Gerret, Karin Boczek, and Lars Koppers. 2018. Sourcing the sources. *Digital Journalism* 6: 807–28. [CrossRef]

Waisbord, Silvio. 2020. Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism. *Digital Journalism* 8: 1030–46. [CrossRef]

Weinacht, Stefan, and Ralf Spiller. 2014. Datenjournalismus in Deutschland. *Publizistik* 59: 411–33. [CrossRef]

White, David Manning. 1950. The "gate keeper": A case study in the selection of news. *Journalism Quarterly* 27: 383–90. [CrossRef]

Witzenberger, Benedict, and Jürgen Pfeffer. 2022. Gender dynamics of German journalists on Twitter. Paper presented at the ASONAM '22: Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, November 10–13. pp. 226–30. [CrossRef]

Xiao, Chunjing, Ling Su, Juan Bi, Yuxia Xue, and Aleksandar Kuzmanovic. 2012. Selective Behavior in Online Social Networks. Paper presented at the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, December 4–7. Washington, DC: IEEE Computer Society, vol. 1, pp. 206–13. [CrossRef]

Zelizer, Barbie. 1993. Journalists as interpretive communities. *Critical Studies in Mass Communication* 10: 219–37. [CrossRef]

Zhang, Xinzhi. 2017. Visualization, technologies, or the public? *Digital Journalism* 6: 737–58. [CrossRef]

Ziamou, Theodora. 2000. *Women Make the News: A Crack in the "Glass Ceiling"?: A Report on the 8 March 2000 Initiative*. Paris: UNESCO. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000123003 (accessed on 20 January 2022).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

# Chapter 4

# Discussion

As the world was hit with the unprecedented challenges of the COVID-19 pandemic, a transformative wave was sweeping across journalism. This dissertation aims to dive into this transformation through its hybrid studies, uncovering how the pandemic has altered the practice and perception of data journalism.

The papers in this dissertation offer a multifaceted study of this intersection, revealing the challenges and opportunities that have emerged. From the surge in data-driven reporting and the innovative use of computational tools to the evolving public understanding of complex information, the impact of the pandemic on journalism is profound and far-reaching.

This discussion synthesizes the findings of the individual studies, aligning them with the overarching research questions and objectives. It offers an analysis of the changes COVID-19 has made to the field of data journalism and a reflection on the effectiveness of Computational Social Science methods in capturing these shifts.

In doing so, we aim to contribute to a deeper understanding of the role and responsibility of data journalism in times of uncertainty and change. This discussion extends beyond the academic sphere, offering insights and recommendations pertinent to practitioners and educators in journalism.

## 4.1 Comparative Analysis of Findings

This thesis is framed around two research questions: RQ 1 asked about the influence of COVID-19 on data journalism, and RQ 2 aimed to study the effectiveness of Computational Social Science Methods in investigating this influence. Three research objectives further accompanied these questions.

### 4.1.1 Influence of COVID-19 on the Practice of Data Journalism (RQ 1)

In answering research question 1, these objectives could already be fulfilled. Paper 2 and Paper 3 contributed to increased quantitative insights into COVID-19's Influence on Data Journalism, with Paper 2 focusing on the prevalence of infographics in media and Paper 3 on the increased collaborative efforts and outputs in data journalism. Paper

1 studied the effects of reporting on predictive models on the audience, another aspect that rose to some prominence during the pandemic.

Paper 2 used images from media's Twitter posts to analyze the infographics shared using a semi-automated method based on the image attributes. It found that across six studied countries, three countries (USA, UK, India) increased their amount of shared infographics compared to pre-COVID-19 times, one country — Germany — remained on a similar level, and two countries (Italy and France) decreased their shares. While this was an ambiguous result initially, an additional analysis of the tweet's hashtags showed that the COVID-19 pandemic was hugely discussed, especially through infographics, across all geographies. The paper could also show that using infographics in a tweet led to increased interactions compared to other image tweets — an important driver for coverage that news media aims for.

These results indicate that COVID-19 led to an enormous increase in infographics in some countries, and the topic dominated parts of the reporting across all investigated countries, which is in line with other findings (Auväärt, 2023). The differences in the publication rates of those visualizations are open to discussion. It is striking to see a large increase in the US and the UK, which already have a prominent data journalistic history (S. Rogers, 2011; Carl W. Anderson, 2018; Hermida and M. L. Young, 2019), while France and Italy even had a decline in these visualizations — although the pandemic had similarly hit them. These might indicate differences in audience expectations, in which media in English-speaking countries placed a larger emphasis on visual forms of reporting. In contrast, media in other countries might have reduced the workforce or moved capacities from infographic generation to other areas (Bisiani et al., 2023). That COVID-19 played an important role in the reporting using infographics could be shown by the hashtag cluster analysis that pointed to enormous shares of COVID-19-related tweets containing infographics. In short, we argue for a consensus across infographic designers (Auväärt, 2023) on the importance of the topic while acknowledging the discrepancy in the prevalence of actual infographic outputs.

Paper 3 also fits into this discussion. It investigated the collaboration between data and science journalists in Germany and reveals how these interactions have evolved during the COVID-19 crisis. The research is based on bylines of data journalistic articles.

It found increased data journalistic output for most of the observed newsrooms after COVID-19, in combination with an initial increase in the number of authors that published during the first months of the pandemic, which later declined. Further analysis found a shift towards increased cooperation between science and data journalists while decreasing corporations with all other departments. Network model analysis helped to understand temporal interactions between authors, concluding that journalists in some newsrooms tended to repeat previous co-authorships.

These results give another empirical indication of the increase of data journalistic work, in this case geographically limited to data journalistic teams in some of Germany's largest newsrooms. While the initial increase in authors shows additional resources, the latter decline might indicate a refocus on expert personnel. This research further demonstrates the shifts in co-authorships across newsroom departments and the increased cooperation between data and science reporters. As discussed in Paper 3, this might be explained by the emergence of Communities of Practice between health reporters and data specialists, which were driven by the exogenous audience expectations during the Coronavirus pandemic (Wenger, McDermott, and Snyder, 2002).

Using the results of Relational hyperevent models, we were able to deepen our understanding of the co-authorship networks between data journalists and others. They indicated that data journalists used to author articles with the same other journalists repeatedly. This habit would be typical for a joint enterprise within a common domain in a Community of Practice. Interestingly, in some instances, the previous cooperation between journalists was negatively affected by the COVID-19 pandemic, indicating that new cooperations might have been formed during the pandemic. For one newsroom, we also found increased closure tendencies after the pandemic started, indicating the formation of new co-authorships through triadic closures (Granovetter, 1973).

COVID-19 led to an increase in data journalistic output around the world. This claim has been raised in several publications already (Desai et al., 2021; Pentzold, D. J. Fechner, and Zuber, 2021; José A. García-Avilés et al., 2022). However, the extent of this change had not been studied thoroughly on a generalizable scale.

We have collected quantifiable evidence on the prevalence of data journalistic output using the number of infographics in media Twitter posts. We have further amended this with a byline analysis of data from journalistic articles in German news media. They indicate an overall increase in data journalistic output, however, with some differences across countries and newsrooms.

We further found evidence for increased co-authorships between data and science departments. Cooperations between those two were nothing new, as science was generally a data-savvy topic (Loosen, Reimer, and De Silva-Schmidt, 2017; Cushion, J. Lewis, Sambrook, et al., 2016). However, the increase through the pandemic was striking and could be quantified using the approaches above. The formation of cooperation across different fields of the newsroom and the insights into and alterings of roles and tasks have been shown to drive journalistic developments during a crisis like COVID-19 (Konow-Lund, Mtchedlidze, and Barland, 2022).

Differently, Paper 1 analyzed an actual journalistic output that gained momentum during the pandemic: predictive journalism. While they had been used before primarily for election predictions, during COVID-19, predictive models were also leveraged in reporting. Paper 1 tried to understand the impact of the uncertainty inherent in these models on their audience for electoral predictive reporting after the pandemic hit — and increased the amount of data visualizations that users might have been exposed to and might have influenced their data literacy.

We found a general alignment between the visualization designers' intentions in displaying the models' uncertainty and their audiences. However, not all users could accurately interpret some of the visualizations they were shown, and we could not find a large perception of influence on user's thinking about the campaign. A larger share of users voiced disagreement towards these visualizations, assuming the unpredictability of elections based on historical data, the unreliability of polling, or political desirability manipulation on the models. Similar concerns have been found around COVID-19 predictive modeling (Allaham and Diakopoulos, 2022).

These results indicated a general interest in the results of the predictive models as a form of future-oriented cognition (Szpunar, Spreng, and Schacter, 2014) to aid users in embedding their speculations into some data. Only a small fraction of users saw the threat to potentially influence voters in their decisions. However, not all users could correctly identify all parties that could potentially end up first from a bar chart of vote share, which displayed areas of uncertainty. This indicated the need for rigorous user testing before publishing visualizations like these to enable a large share of the audience to interpret the visualizations of uncertainty in line with the designer's intentions.

The large share of criticism towards these predictive models led to arguments for increased transparency on the data and methods that were used to create the models (Diakopoulos and Koliska, 2017), or the use of interactive visualizations that enable users to affect the main drivers of the model to understand potential influence factors on the prediction (Pentzold and D. Fechner, 2019).

Paper 1 offered a more nuanced view of a single form of data journalistic reporting that gained further interest during the COVID-19 pandemic. In addition to the existing literature (Pentzold and D. Fechner, 2019; Pentzold and D. Fechner, 2021), it aimed to increase the understanding of its reception and perceived impact on the audience of predictive journalism.

### 4.1.2 Effectiveness of Computational Social Science Methods (RQ 2)

The second research question was centered around the possible insights Computational Social Science methods could offer in analyzing data journalistic practices and the impact of COVID-19 on them, which was also specified in research objective 2. Papers 2 and 3 contributed to this research question and objective, while Paper 4 offered additional insights into using computational methods to analyze data and political journalistic networks.

Paper 2, "Popular and on the Rise - But Not Everywhere: COVID-19-Infographics on Twitter," employed semi-automated infographic detection, providing a case study using computational methods to analyze data journalism trends.

The research revealed increased shares of tweets containing infographics about COVID-19, though this increase was inconsistent across different countries. The method effectively combined image attributes analysis and manual labeling to identify specific content within social media datasets. We further found evidence for increased reporting on COVID-19 in infographic tweets and a generally higher rate of interactions for infographic tweets.

This approach demonstrated how computational methods can be accurate and comprehensive in analyzing large-scale social media data. It revealed the effectiveness of combining automated processes based on image attributes with human oversight to ensure nuanced data interpretation and prevent algorithmic black boxes. However, the variation in results across different countries suggested a need for context-sensitive computational methods and highlighted the balance between automated efficiency and the necessity for manual intervention — indicating that while the comprehensiveness in terms of collected images might have been sufficient, additional insights were required actually to understand the differences between countries. Paper 2 can, therefore, motivate further qualitative research into the different markets (which has already been going on, e.g., in Mellado et al. (2021)) or could potentially fuel other quantitative methods that might capture the reasons behind those differences in greater detail.

Paper 3's use of negative binomial regression and Relational hyperevent models (RHEM) also exemplifies the application of advanced computational methods in journalism research.

Using these methods, we observed a significant increase in data journalism outputs and intensified collaboration among data and science journalists in Germany during and after the pandemic. This approach enabled a detailed analysis of authorship trends and collaborative networks of and with data journalists.

The findings underscored the depth of computational methods in uncovering complex relational patterns through journalistic bylines, which have been primarily used in more qualitative or attitudinal contexts until now (Burkhart and Sigelman (1990) and Dogruel, Joeckel, and Wilhelm (2021), see Boczek, Dogruel, and Schallhorn (2022) as an exception).

The study illustrated how Computational Social Science could provide insightful analysis into the evolving dynamics of newsrooms, especially under transformative conditions like the COVID-19 pandemic. It also raised discussions about the importance of context-specific studies and the potential need for interdisciplinary approaches to enrich computational analyses. The accuracy of these methods is highlighted in their ability to detect subtle shifts in collaboration patterns and authorship trends within data journalism, which more traditional analytical approaches might overlook. They allow for a detailed examination of a large dataset, enabling exploration and quantification of trends over time and across different newsrooms. This approach is crucial in understanding the full scope of the pandemic's influence on data journalism practices and potentially allows the comparison of magnitudes of change in similar future events.

Paper 4 used tweets to understand amplification behavior between different sexes and groups of journalists. It found sex-based differences in Twitter interactions among German political and data journalists, with women journalists showing a tendency to engage more with other women. It also revealed that data journalists were more inclusive towards non-member sources than their political journalism counterparts. This study highlighted computational analyses' societal and professional implications, particularly in understanding sex dynamics and inclusivity within groups of journalists. It showed the capability of computational methods to uncover potential biases and patterns in professional networks, contributing to discussions on diversity and representation in media (Usher, Holcomb, and Littman, 2018). The ability to identify specific patterns of interaction (mentions, retweets) among groups of journalists of different sexes reflects some high degree of accuracy, acknowledging that the interactions are still an operationalized form of the actual variable of interest, amplification, and cognition. The study also points to the need for broader analyses across various social media platforms to fully grasp the scope and impact of journalistic practices in the digital era.

### Methodological Reflections, Limitations, and Challenges

Each of the four papers in this dissertation has its own empirical approach, which comes with a distinct set of opportunities and limitations.

The survey-based approach to user perception of predictive journalism in Paper 1 aimed to capture user sentiment toward this newer form of journalism. Using a questionnaire shared directly with actual users of predictive journalism in a non-lab environment provided a realistic setting to gauge user perceptions and understanding. The study's focus on a specific predictive model in a particular newspaper for a single election limits its generalizability. This specificity can be valuable for detailed insights but restricts the broader applicability of the findings. Users were not drawn randomly but chose to self-select, limiting the results. However, we found a lack of understanding even within this group of potentially interested persons on the topic. While rich in detail, the qualitative approach to understanding perceptions using many open-text fields may not capture the breadth of user experiences and interpretations across different predictive journalism models.

The combination of automated image attribute analysis and manual labeling in Paper 2 allowed for a fast yet trustworthy analysis of many images downloaded from Social Media. This method enabled a balanced approach between computational efficiency and human insight, reducing insights drawn from algorithmic black boxes, but based on first, an understandable computational analysis, and secondly, human oversight, which is a valuable combination for the validity of the results. Technical hurdles arose, particularly with text detection in non-standard formats, such as word art, which could have impacted the accuracy of infographic detection. The study offered an external perspective based on data, lacking insights from within newsrooms, which could provide a more comprehensive understanding of the creation and dissemination of infographics, which could be facilitated using a more qualitative approach. The approach to combining multiple newsrooms across multiple countries enables a quantification of the influence of COVID-19 on the prevalence of infographics.

Measuring article output and byline attribution as primary and direct indicators of journalistic activity in Paper 3 provided a direct and quantifiable approach to understanding changes in data journalism practices through the Coronavirus pandemic. The models used were advanced but offered interpretability of the coefficients while allowing deep insights into cooperation structures. Similar to Paper 2, this paper also views the subject from an external data perspective, which lacks internal newsroom insights and does not account for byline rules within newsrooms. It also just regards articles, not infographics, which often are not directly attributable to an individual journalist. The complexity of the statistical methods used, while robust, may pose interpretative challenges, requiring a deep understanding of the models to draw conclusions from the data accurately.

Paper 4 provided an analysis of social media interactions among journalists on Twitter with insights into recognition and amplification structures within elite journalistic circles. The study reflects behavior on Twitter, which may not accurately represent real-life interactions and dynamics. Moreover, the platform's elitist nature may skew findings, particularly when examining broader societal trends. As the study focuses on exactly those circles, the use of Twitter data seems reasonable — but it must be acknowledged that the measurements are still only valid for behavior on the platform. Compiling comprehensive lists of journalists, particularly data journalists identified through a Slack group and political journalists manually collected from imprints, was a significant challenge due to the lack of official records.

While distinct in their approaches, each paper grapples with the balance between the depth of qualitative insights and the breadth of quantitative analysis. The external perspectives provided by the computational methods offer valuable but limited insights and underscore the need for an integrated approach combining data-driven analysis with internal newsroom perspectives to understand the evolving landscape of data journalism fully.

## 4.2 Broader Implications And Contributions to the Field of Data Journalism Research

To summarize this discussion, outlined below are contributions, implications for (data) journalism research and practitioners, and possible future research directions that might be derived from this thesis.

### 4.2.1 Advancement of Knowledge in the Influence of COVID-19 on Data Journalism

Collectively, the four papers contribute to an advanced understanding of how the COVID-19 pandemic has influenced data journalism. The key findings from these studies provide valuable insights into the changes and challenges faced by data journalism during this period.

The rise in the number of infographics is indicative of an increased public interest in COVID-19-related topics. This demand likely necessitated the creation of more infographics, posing challenges regarding resource allocation and capacity within newsrooms (Bisiani et al., 2023). This trend suggests a shift in journalistic priorities and resource management, focusing on delivering complex information in an accessible visual format.

The observed increase in infographics and data journalism articles was not a temporary phenomenon, fueling arguments for lasting changes in the industry (Jose A. García-Avilés, 2021; Quandt and Wahl-Jorgensen, 2021). It remained consistently high until at least 2021, marking a significant shift in journalistic practices. This sustained production level implies a long-term impact of the pandemic on data journalistic output and audience expectations, possibly leading to a permanent shift in the landscape of data news reporting.

This is in line with other observations on the innovations going on in journalism during COVID-19, blending existing knowledge and technological solutions to generate rapidly developed and iteratively adapted journalistic products (Konow-Lund, Mtchedlidze, and Barland, 2022), strategically using data journalism for satisfying audience demands (José A. García-Avilés et al., 2022). As discussed in Hermida and M. L. Young (2021), innovations like this can be driven by concerns about competition and survival during societal and technological trends in the news industry but, in many cases, not following a clearly defined strategy.

The pandemic spurred the development of new or increased collaborations, particularly between data journalists and science journalists, leading to the formation of Communities of Practice. These collaborations were essential for effectively reporting on the complex health science aspects of the pandemic. However, this focus might have reduced cooperation with other newsroom departments, suggesting a reallocation of resources and attention within newsrooms. Understanding this shift is crucial for comprehending the evolving nature of journalistic collaborations in response to large-scale global events.

Predictive reporting emerged as a notable new form of journalism during the pandemic. While innovative, this approach brought challenges related to interpretability and audience skepticism. Addressing these challenges necessitates increased transparency and rigorous user testing of visualizations to ensure audience comprehension and trust. This development highlights the need for journalistic practices to evolve in terms of content and how the audience presents and understands this content.

### 4.2.2 Practical Implications for Data Journalism Practitioners

When the results of the four papers in this thesis are combined, valuable insights emerge into the practical implications for data journalism practitioners. These implications can help journalists enhance the effectiveness and reach of their reporting, particularly in the context of evolving newsroom dynamics and audience engagement strategies.

Paper 1 showed clear implications for practitioners regarding the output of predictive reporting: The necessity for rigorous user testing of visualizations is critical. This ensures that the audience accurately understands the intended message of predictive models. The paper also showed the clear need for transparency in reporting on predictive models, including the underlying data and methodologies. This transparency is crucial for countering criticism and building trust with the audience.

Paper 2 analyzed the infographic output of media across several countries. The observation that some countries had less infographic output despite the dominance of COVID-19 as a topic and the increased interactions infographics garnered on Twitter is puzzling. This discrepancy suggests that media executives should assess the potential of infographics in enhancing content distribution on social media platforms.

Paper 3 concluded that science departments may have risen as the primary partner for the co-publication of data journalistic articles during the pandemic. While this form of cooperation did exist before COVID-19, it might be valuable to discover further ways of cooperation within science-data-journalistic teams — outside of COVID-19 reporting — in areas like climate reporting. There's a need to explore ways to rejuvenate or enhance cooperation with other newsroom departments, like politics or economics, which may have been sidelined during the pandemic. This reintegration could diversify and strengthen journalistic content beyond science-related reporting.

Paper 4 aimed to compare communication behavior on Social Media. The identified differences in how journalists of different sexes amplify each other's voices on Twitter highlight an area for reflection. Journalists should be conscious of these biases and actively work towards a more balanced amplification of voices across sexes. Understanding and addressing these differences can contribute to a more equitable and diverse journalistic landscape, particularly in digital and social media.

As the field continues to evolve, these insights can guide practitioners in navigating the challenges and opportunities that arise. Additional research demonstrating the practical application of these findings in various journalistic contexts would benefit a more comprehensive understanding of these implications.

### 4.2.3 Future Research Directions

The papers' findings suggest several avenues for future research in data journalism. These directions address the limitations of the studies presented here and expand our understanding of journalistic practices.

**Audience Responses to Election Predictions:** Future research based on the results of Paper 1 should focus on enhancing the generalizability and depth of understanding regarding audience responses to predictions in news media. To achieve this, employing more targeted and nuanced methods, such as guided interviews, could provide deeper insights into individual interpretations and responses. Additionally, collecting and analyzing larger, cross-media samples would allow for a comparative approach, enabling a broader evaluation of how the publications in different media companies affect audience perceptions of predictions. This expanded methodology could significantly enrich our understanding of audience engagement with predictive journalism across diverse media landscapes.

**Quantitative and Mixed Methods for Comparison of Media Systems:** For

Papers 2 and 3, future research should aim to refine quantitative methods for analyzing publication behaviors across different media systems or countries to address the current gap in generalizability and depth. The field, often dominated by qualitative approaches, could benefit significantly from developing quantitative techniques that provide a wider scope of analysis while maintaining depth. Also, integrating mixed methods approaches, combining quantitative breadth with qualitative depth, would offer a more comprehensive understanding of the reasons for differences. This would provide a balanced perspective, capturing both the macro-level trends and the nuanced, context-specific aspects of publication behaviors in various countries.

**Evolution of COVID-19's Influence on Data Journalism:** Investigating the enduring impact and potential evolution of COVID-19 on the publication behavior of data journalism forms a critical avenue for future research following Papers 2 and 3. As the immediate crisis of the pandemic recedes from public focus, it is essential to explore whether the changes observed during the pandemic have persisted, altered, or vanished. Longitudinal studies could be particularly beneficial in this context, tracking the long-term effects of the pandemic on journalistic practices and audience engagement trends toward infographics or data reporting. This research would provide valuable insights into the lasting influence of global crises on journalism, particularly in data journalism and visual communication.

# Chapter 5

# Conclusion

This dissertation has contributed to understanding the influence of COVID-19 on data journalism, offering a comprehensive analysis through a series of studies that merge the fields of data journalism and Computational Social Science. Key findings from these studies reveal a marked increase in the use of infographics and data journalism during the pandemic, indicating a heightened public interest in visual and data-driven news content. The research has shown that the pandemic catalyzed new collaborations, particularly between data and science journalists, while also prompting the rise of innovative journalistic forms like predictive reporting, which were not necessarily welcomed by parts of the audience.

This dissertation's contributions to the field of data journalism highlight the evolving dynamics within newsrooms, underscore the importance of interdisciplinary collaboration, and emphasize the need for rigorous user testing and transparency in reporting. In Computational Social Science, the dissertation showcases the utility and effectiveness of quantitative methods, such as semi-automated infographic detection and advanced statistical models, in analyzing larger-scale journalistic data.

Several recommendations emerge for future research. Studies that employ mixed-methods approaches should combine the depth of qualitative insights with the breadth of quantitative analysis. Investigating the long-term impacts of the pandemic on journalistic practices and audience engagement remains a crucial area of exploration. Additionally, further research should aim to understand the persisting changes in newsroom dynamics and the potential for applying the lessons learned during the pandemic to other critical reporting areas, such as climate change.

In conclusion, this dissertation not only illuminates the immediate impacts of COVID-19 on data journalism but also sets the stage for ongoing research and practice in this rapidly evolving field.

It also highlights the need to combine quantitative and qualitative approaches. Returning to the quote in the introduction: The plural of anecdotes might be data, but the collection, examination, and addition of anecdotes to data is critical for the greatest possible insight into opaque systems like newsrooms.

The findings and insights provide a foundation for future explorations and innovations in data journalism and Computational Social Science.

# Bibliography

@XDevelopers (2023). *Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead.* URL: https://twitter.com/XDevelopers/status/1621026986784337922 (visited on 02/02/2023).

Akinfemisoye-Adejare, Motilola (2019). "Data Journalism in Nigeria: Interrogating the Nigerian Mainstream Media's Data-Driven Reporting of the Elections of 2019". In: *Data Journalism in the Global South*. Ed. by Bruce Mutsvairo, Saba Bebawi, and Eddy Borges-Rey. Palgrave Studies in Journalism and the Global South. Cham: Springer International Publishing, pp. 73–84. ISBN: 9783030251772. DOI: 10.1007/978-3-030-25177-2_5. URL: https://doi.org/10.1007/978-3-030-25177-2_5.

Albright, Jonathan (2018). *The Graph API: Key Points in the Facebook and Cambridge Analytica Debacle*. The Tow Center for Digital Journalismat the Columbia Graduate School for Journalism. URL: https://medium.com/tow-center/the-graph-api-key-points-in-the-facebook-and-cambridge-analytica-debacle-b69fe692d747 (visited on 10/22/2023).

Allaham, Mowafak and Nicholas Diakopoulos (2022). "Predicting COVID: Understanding audience responses to predictive journalism via online comments". In: *New Media & Society*. ISSN: 1461-4448. DOI: 10.1177/14614448221135632. URL: https://doi.org/10.1177/14614448221135632.

Alvarez, R. Michael (2016). *Computational Social Science Discovery and Prediction. Discovery and Prediction*. Cambridge, UK: Cambridge University Press, p. 312. ISBN: 9781107107885.

Anderson, Carl W. (2018). *Apostles of Certainty: Data Journalism and the Politics of Doubt Get access Arrow*. Oxford, UK: Oxford University Press. ISBN: 9780190492335. DOI: 10.1093/oso/9780190492335.001.0001.

Anderton-Yang, David et al. (2012). "Why Is Data Journalism Important?" In: *The data journalism handbook 1*. Ed. by Jonathan Gray et al. European Journalism Centre, O'Reilly. URL: https://datajournalism.com/read/handbook/one/introduction/why-is-data-journalism-important (visited on 05/23/2020).

Angwin, Julia and Terry Jr. Parris (2016). *Facebook Lets Advertisers Exclude Users by Race. Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads*. ProPublica. URL: https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race (visited on 10/22/2023).

Antonopoulos, Nikos and Minos-Athanasios Karyotakis (2020). "Data Journalism". en. In: *The SAGE International Encyclopedia of Mass Media and Society*. Ed. by Debra

L. Merskin. Thousand Oaks,, California: SAGE Publications, Inc., pp. 440–441. DOI: 10.4135/9781483375519.n175. URL: https://sk.sagepub.com/Reference//the-sage-encyclopedia-of-mass-media-and-society/i5415.xml.

Appelgren, Ester (2016). "Data Journalists Using Facebook". In: *Nordicom Review* 37.1, pp. 156–169. DOI: 10.1515/nor-2016-0007. eprint: https://doi.org/10.1515/nor-2016-0007. URL: https://doi.org/10.1515/nor-2016-0007.

Appelgren, Ester, Carl-Gustav Lindén, and Arjen van Dalen (2019). "Data Journalism Research: Studying a Maturing Field across Journalistic Cultures, Media Markets and Political Environments". In: *Digital Journalism* 7.9, pp. 1191–1199. DOI: 10.1080/21670811.2019.1685899.

Appelgren, Ester and Gunnar Nygren (2014). "Data Journalism in Sweden". In: *Digital Journalism* 2.3, pp. 394–405. ISSN: 2167-0811. DOI: 10.1080/21670811.2014.884344. URL: https://doi.org/10.1080/21670811.2014.884344.

Araujo, Theo et al. (2022). "OSD2F: An Open-Source Data Donation Framework". In: *Computational Communication Research* 4.2, pp. 372–387. DOI: 10.5117/ccr2022.2.001.arau.

Atteveldt, Wouter van and Tai-Quan Peng (2018). "When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science". In: *Communication Methods and Measures* 12.2-3, pp. 81–92. ISSN: 1931-2458. DOI: 10.1080/19312458.2018.1458084. URL: https://doi.org/10.1080/19312458.2018.1458084.

Ausloos, Jef and Michael Veale (2021). "Researching with Data Rights". In: *Technology and Regulation*. DOI: 10.26116/TECHREG.2020.010.

Auväärt, Liis (2023). "Fighting COVID-19 with data: An analysis of data journalism projects submitted to Sigma Awards 2021". In: *Central European Journal of Communication* 15.3(32), pp. 379–395. ISSN: 1899-5101. DOI: 10.51480/1899-5101.15.3(32).3.

Baack, Stefan (2018). "Practically Engaged". In: *Digital Journalism* 6.6, pp. 673–692. ISSN: 2167-0811. DOI: 10.1080/21670811.2017.1375382. URL: https://doi.org/10.1080/21670811.2017.1375382.

Baker, Scott R. et al. (2020). "How Does Household Spending Respond to an Epidemic? Consumption during the 2020 COVID-19 Pandemic". In: *The Review of Asset Pricing Studies* 10.4. Ed. by Jeffrey Pontiff, pp. 834–862. DOI: 10.1093/rapstu/raaa009.

Barouki, Robert et al. (2021). "The COVID-19 pandemic and global environmental change: Emerging research needs". In: *Environment International* 146, p. 106272. ISSN: 0160-4120. DOI: 10.1016/j.envint.2020.106272. URL: https://www.sciencedirect.com/science/article/pii/S0160412020322273.

Bechmann, Anja (2018). *Publications that could not have existed without access to API data*. URL: https://docs.google.com/document/d/15YKeZFSUc1j03b4lW9YXxGmhYEnFx3TSy68qCrX9BEI.

Beiler, Markus, Felix Irmer, and Adrian Breda (2020). "Data Journalism at German Newspapers and Public Broadcasters: A Quantitative Survey of Structures, Contents and Perceptions". In: *Journalism Studies* 21.11, pp. 1571–1589. DOI: 10.1080/1461670X.2020.1772855. URL: https://doi.org/10.1080/1461670X.2020.1772855.

Belair-Gagnon, Valerie and Avery E. Holton (2018). "Strangers to the Game? Interlopers, Intralopers, and Shifting News Production". In: *Media and Communication* 6.4, pp. 70–78. ISSN: 2183-2439. DOI: https://doi.org/10.17645/mac.v6i4.1490. eprint: https://doi.org/10.17645/mac.v6i4.1490. URL: https://doi.org/10.17645/mac.v6i4.1490.

Berret, Charles and Cheryl Phillips (2016). *Teaching data and computational journalism.* New York City, NY, USA: Columbia School of Journalism. ISBN: 9780692637456.

Bisiani, Simona et al. (2023). "The Data Journalism Workforce: Demographics, Skills, Work Practices, and Challenges in the Aftermath of the COVID-19 Pandemic". In: *Journalism Practice*, pp. 1–21. DOI: 10.1080/17512786.2023.2191866.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022. ISSN: 1532-4435.

Boase, Jeffrey et al. (2006). *The Strength of Internet Ties. The internet and email aid users in maintaining their social networks and provide pathways to help when people face big decisions.* Pew Internet & American Life Project. URL: https://www.pewresearch.org/internet/2006/01/25/the-strength-of-internet-ties/ (visited on 11/04/2023).

Boczek, Karin, Leyla Dogruel, and Christiana Schallhorn (2022). "Gender byline bias in sports reporting: Examining the visibility and audience perception of male and female journalists in sports coverage". In: *Journalism.* DOI: 10.1177/14648849211063312.

Borges-Rey, Eddy (2017). "Towards an epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity". In: *Journalism* 21.7, pp. 915–932. DOI: 10.1177/1464884917693864.

Boumans, Jelle W. and Damian Trilling (2015). "Taking Stock of the Toolkit". In: *Digital Journalism* 4.1, pp. 8–23. DOI: 10.1080/21670811.2015.1096598.

Bounegru, Liliana (2012). "Data Journalism in Perspective". In: *The data journalism handbook 1.* Ed. by Jonathan Gray et al. European Journalism Centre. URL: https://datajournalism.com/read/handbook/one/introduction/data-journalism-in-perspective (visited on 12/09/2023).

Boyd, Danah and Kate Crawford (2012). "Critical Questions for Big Data". In: *Information, Communication & Society* 15.5, pp. 662–679. DOI: 10.1080/1369118x.2012.678878.

Bradshaw, Paul (2011). *The inverted pyramid of data journalism.* URL: https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/ (visited on 06/19/2020).

Braun, Virginia and Victoria Clarke (2006). "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2, pp. 77–101. ISSN: 1478-0887. DOI: 10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

Bravo, Adolfo Antón and Ana Serrano Tellería (2020). "Data Journalism: From Social Science Techniques to Data Science Skills". In: *Hipertext.net* 20 (20), pp. 41–54. ISSN: 1695-5498. DOI: 10.31009/hipertext.net.2020.i20.04. URL: https://www.raco.cat/index.php/Hipertext/article/view/361650 (visited on 06/12/2020).

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: DOI: 10.48550/ARXIV.2005.14165. arXiv: 2005.14165 [cs.CL].

Bruns, Axel (2019). "After the 'APIcalypse': social media platforms and their fight against critical scholarly research". In: *Information, Communication & Society* 22.11, pp. 1544–1566. DOI: 10.1080/1369118x.2019.1637447.

Burkhart, Ford N. and Carol K. Sigelman (1990). "Byline bias? Effects of gender on news article evaluations". In: *Journalism Quarterly* 67.3, pp. 492–500.

Burns, Lynette Sheridan and Benjamin J. Matthews (2018). "First Things First: Teaching Data Journalism as a Core Skill". In: *Asia Pacific Media Educator* 28.1 (1), pp. 91–105. ISSN: 1326-365X. DOI: 10.1177/1326365X18765530.

Butts, Carter T. (2008). "A Relational Event Framework for Social Action". In: *Sociological Methodology* 38.1, pp. 155–200. DOI: 10.1111/j.1467-9531.2008.00203.x.

Cadwalladr, Carole (2018). "'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower". In: *The Guardian*. URL: https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump (visited on 10/22/2023).

Carlson, Matt (2009). "Dueling, Dancing, or Dominating? Journalists and Their Sources". In: *Sociology Compass* 3.4, pp. 526–542. DOI: https://doi.org/10.1111/j.1751-9020.2009.00219.x. eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9020.2009.00219.x. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9020.2009.00219.x.

Cheng, Justin, Cristian Danescu Niculescu-Mizil, and Jure Leskovec (2021). "Antisocial Behavior in Online Discussion Communities". In: *Proceedings of the International AAAI Conference on Web and Social Media* 9.1, pp. 61–70. DOI: 10.1609/icwsm.v9i1.14583.

Cheruiyot, David, Stefan Baack, and Raul Ferrer-Conill (2019). "Data Journalism Beyond Legacy Media: The case of African and European Civic Technology Organizations". In: *Digital Journalism* 7.9, pp. 1215–1229. ISSN: 2167-0811. DOI: 10.1080/21670811.2019.1591166. URL: https://doi.org/10.1080/21670811.2019.1591166.

Chiumbu, Sarah and Allen Munoriyarwa (2023). "Exploring data journalism practices in Africa: data politics, media ecosystems and newsroom infrastructures". In: *Media, Culture & Society* 45.4, pp. 841–858. DOI: 10.1177/01634437231155341.

Chmiel, Anna et al. (2011). "Negative emotions boost user activity at BBC forum". In: *Physica A: Statistical Mechanics and its Applications* 390.16, pp. 2936–2944. DOI: 10.1016/j.physa.2011.03.040.

Christakis, Nicholas A. and James H. Fowler (2011). *Connected The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends' Friends'*

*Friends Affect Everything You Feel, Think, and Do. The Surprising Power of Our Social Networks and How They Shape Our Lives – How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do.* Little Brown & Company, p. 368. ISBN: 9780316036139.

Christner, Clara et al. (2021). "Automated Tracking Approaches for Studying Online Media Use: A Critical Review and Recommendations". In: *Communication Methods and Measures* 16.2, pp. 79–95. DOI: 10.1080/19312458.2021.1907841.

Coddington, Mark (2015). "Clarifying Journalism's Quantitative Turn". In: *Digital Journalism* 3.3 (3): *Journalism in an Era of Big Data: Cases, Concepts, and Critiques*, pp. 331–348. DOI: 10.1080/21670811.2014.976400. URL: https://doi.org/10.1080/21670811.2014.976400.

Corman, Steven R. and Craig R. Scott (1994). "Perceived Networks, Activity Foci, and Observable Communication in Social Collectives". In: *Communication Theory* 4.3, pp. 171–190. DOI: 10.1111/j.1468-2885.1994.tb00089.x.

Cox, Melisma (2000). "The development of computer-assisted reporting". In: Newspaper Division, Association for Education in Journalism and Mass Communication, Southeast Colloquium (University of North Carolina, Mar. 17, 2000). Chapel Hill, NC, USA. URL: https://web.archive.org/web/20110928144238/http://com.miami.edu/car/cox00.pdf (visited on 10/08/2023).

Cushion, Stephen, Justin Lewis, and Robert Callaghan (2016). "Data Journalism, Impartiality And Statistical Claims". In: *Journalism Practice* 11.10, pp. 1198–1215. ISSN: 1751-2786. DOI: 10.1080/17512786.2016.1256789. URL: https://doi.org/10.1080/17512786.2016.1256789.

Cushion, Stephen, Justin Lewis, Richard Sambrook, et al. (2016). *Impartiality Review of BBC Reporting of Statistics: A Content Analysis.* BBC. URL: http://downloads.bbc.co.uk/bbctrust/%20assets/files/pdf/our_work/stats_impartiality/content_analysis.pdf (visited on 07/21/2023).

Daly, Michael, Angelina R. Sutin, and Eric Robinson (2022). "Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study". In: *Psychological Medicine* 52.13, pp. 2549–2558. ISSN: 0033-2917, 1469-8978. DOI: 10.1017/S0033291720004432. URL: https://www.cambridge.org/core/journals/psychological-medicine/article/longitudinal-changes-in-mental-health-and-the-covid19-pandemic-evidence-from-the-uk-household-longitudinal-study/3076D6D9BA396B94D02E67FEBF7C66D8.

Danzon-Chambaud, Samuel (2021). "Covering COVID-19 with automated news". In: *Columbia Journalism Review.* URL: https://www.cjr.org/tow_center_reports/covering-covid-automated-news (visited on 07/29/2023).

Davies, Kayt (2018). "Getting Started with Data Journalism: A Baby Steps Approach:" in: *Asia Pacific Media Educator.* DOI: 10.1177/1326365X18767460.

De Maeyer, Juliette et al. (2015). "Waiting for Data Journalism". In: *Digital Journalism* 3.3, pp. 432–446. ISSN: 2167-0811. DOI: 10.1080/21670811.2014.976415. URL: https://doi.org/10.1080/21670811.2014.976415.

Desai, Angel et al. (2021). "Data journalism and the COVID-19 pandemic: opportunities and challenges". In: *The Lancet Digital Health* 3.10, pp. 619–621. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(21)00178-3. URL: https://doi.org/10.1016/S2589-7500(21)00178-3.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: DOI: 10.48550/ARXIV.1810.04805. arXiv: 1810.04805 `[cs.CL]`.

Diakopoulos, Nicholas and Michael Koliska (2017). "Algorithmic Transparency in the News Media". In: *Digital Journalism* 5.7, pp. 809–828. DOI: 10.1080/21670811.2016.1208053. eprint: https://doi.org/10.1080/21670811.2016.1208053. URL: https://doi.org/10.1080/21670811.2016.1208053.

Dijck, José Van (2020). "Seeing the forest for the trees: Visualizing platformization and its governance". In: *New Media & Society* 23.9, pp. 2801–2819. DOI: 10.1177/1461444820940293.

Dogruel, Leyla, Sven Joeckel, and Claudia Wilhelm (2021). "Are byline biases an issue of the past? The effect of author's gender and emotion norm prescriptions on the evaluation of news articles on gender equality". In: *Journalism* 24.3, pp. 560–579. DOI: 10.1177/14648849211012176.

Else, Holly (2020). "How a torrent of COVID science changed research publishing — in seven charts". In: *Nature* 588.7839, pp. 553–553. DOI: 10.1038/d41586-020-03564-y.

Engebretsen, Martin, Helen Kennedy, and Wibke Weber (2018). "Data Visualization in Scandinavian Newsrooms: Emerging Trends in Journalistic Visualization Practices". In: *Nordicom Review* 39.2, pp. 3–18. DOI: 10.2478/nor-2018-0007. eprint: https://doi.org/10.2478/nor-2018-0007. URL: https://doi.org/10.2478/nor-2018-0007.

Euler, Leonhard (1736). "Solutio problematis ad geometriam situs pertinentis". In: *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8, pp. 128–140.

Farber, Henry S. (2015). "Why you Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers". In: *The Quarterly Journal of Economics* 130.4, pp. 1975–2026. DOI: 10.1093/qje/qjv026.

Fernández-Pedemonte, Damián, Felicitas Casillo, and Ana Inés Jorge-Artigau (2020). "Communicating COVID-19: Metaphors We "Survive" By". In: *Trípodos* 2.47, pp. 145–160–160. ISSN: 2340-5007. URL: http://www.tripodos.com/index.php/Facultat_Comunicacio_Blanquerna/article/view/820 (visited on 08/14/2020).

Fink, Katherine and C. W. Anderson (2014). "Data Journalism in the United States". In: *Journalism Studies* 16.4, pp. 467–481. DOI: 10.1080/1461670x.2014.939852.

Fortuny, Enric Junqué de et al. (2012). "Media coverage in times of political crisis: A text mining approach". In: *Expert Systems with Applications* 39.14, pp. 11616–11622. DOI: 10.1016/j.eswa.2012.04.013.

Freelon, Deen (2018). “Computational Research in the Post-API Age”. In: *Political Communication* 35.4, pp. 665–668. DOI: 10.1080/10584609.2018.1477506.

García-Avilés, José A. (2021). “Review article: Journalism innovation research, a diverse and flourishing field (2000-2020)”. In: *El profesional de la información*. DOI: 10.3145/epi.2021.ene.10.

García-Avilés, Jose A. (2021). “Journalism as Usual? Managing Disruption in Virtual Newsrooms during the COVID-19 Crisis”. In: *Digital Journalism* 9.9, pp. 1239–1260. ISSN: 2167-082X. DOI: 10.1080/21670811.2021.1942112.

García-Avilés, José A. et al. (2022). “How COVID-19 is Revamping Journalism: Newsroom Practices and Innovations in a Crisis Context”. In: *Journalism Practice*, pp. 1–19. ISSN: 1751-2786. DOI: 10.1080/17512786.2022.2139744. URL: https://doi.org/10.1080/17512786.2022.2139744.

Geiß, Stefan (2021). “Statistical Power in Content Analysis Designs: How Effect Size, Sample Size and Coding Accuracy Jointly Affect Hypothesis Testing – A Monte Carlo Simulation Approach.” In: *Computational Communication Research* 3.1, pp. 61–89. DOI: 10.5117/ccr2021.1.003.geis.

Gilbert, Nigel et al. (2018). “Computational Modelling of Public Policy: Reflections on Practice”. In: *Journal of Artificial Societies and Social Simulation* 21.1. DOI: 10.18564/jasss.3669.

Gillespie, Tarleton (2010). “The politics of ‘platforms’”. In: *New Media & Society* 12.3, pp. 347–364. DOI: 10.1177/1461444809342738.

Granovetter, Mark S. (1973). “The Strength of Weak Ties”. In: *American Journal of Sociology* 78.6, pp. 1360–1380. ISSN: 00029602, 15375390. URL: http://www.jstor.org/stable/2776392.

Green-Barber, Lindsay (2021). “Beyond Clicks and Shares: How and Why to Measure the Impact of Data Journalism Projects”. In: *The Data Journalism Handbook*. Ed. by Liliana Bounegru and Jonathan Gray. Amsterdam, The Netherlands: Amsterdam University Press, pp. 370–378. DOI: 10.5117/9789462989511_ch50. URL: https://datajournalism.com/read/handbook/two/situating-data-journalism/beyond-clicks-and-shares-how-and-why-to-measure-the-impact-of-data-journalism-projects (visited on 12/10/2023).

Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3, pp. 267–297. DOI: 10.1093/pan/mps028.

Grolemund, Garrett and Hadley Wickham (2011). “Dates and Times Made Easy with lubridate”. In: *Journal of Statistical Software* 40.3, pp. 1–25. URL: https://www.jstatsoft.org/v40/i03/.

Haber, Peter, Thomas J. Lampoltshammer, and Manfred Mayr, eds. (2024). *Data Science—Analytics and Applications. Proceedings of the 5th International Data Science Conference—iDSC2023*. Cham: Springer. 104 pp. ISBN: 9783031421716.

Haim, Mario (2020). "Agent-based Testing: An Automated Approach toward Artificial Reactions to Human Behavior". In: *Journalism Studies* 21.7, pp. 895–911. DOI: 10.1080/1461670x.2019.1702892.

– (2022). "The German Data Journalist in 2021". In: *Journalism Practice*, pp. 1–20. DOI: 10.1080/17512786.2022.2098523. eprint: https://doi.org/10.1080/17512786.2022.2098523. URL: https://doi.org/10.1080/17512786.2022.2098523.

– (2023). *Computational Communication Science.* Studienbücher zur Kommunikations- und Medienwissenschaft (STBKUM). Wiesbaden, Germany: Springer Fachmedien. 358 pp. DOI: 10.1007/978-3-658-40171-9.

Haller, André (2019). "Die Online-Kampagnen im Bundestagswahlkampf 2017". In: *Die (Massen-)Medien im Wahlkampf.* Springer Fachmedien Wiesbaden, pp. 49–72. DOI: 10.1007/978-3-658-24824-6_3.

Hannak, Aniko et al. (2013). "Measuring personalization of web search". In: *Proceedings of the 22nd international conference on World Wide Web.* ACM. DOI: 10.1145/2488388.2488435.

Hargittai, Eszter (2018). "Potential Biases in Big Data: Omitted Voices on Social Media". In: *Social Science Computer Review* 38.1, pp. 10–24. DOI: 10.1177/0894439318788322.

Helmond, Anne (2015). "The Platformization of the Web: Making Web Data Platform Ready". In: *Social Media + Society* 1.2, p. 205630511560308. DOI: 10.1177/2056305115603080.

Heravi, Bahareh R. and Mirko Lorenz (2020). "Data Journalism Practices Globally: Skills, Education, Opportunities, and Values". In: *Journalism and Media* 1.1, pp. 26–40. ISSN: 2673-5172. DOI: 10.3390/journalmedia1010003.

Hermida, Alfred and Mary Lynn Young (2019). *Data Journalism and the Regeneration of News.* London, New York. URL: https://www.routledge.com/Data-Journalism-and-the-Regeneration-of-News/Hermida-Young/p/book/9781138058934.

– (2021). "Journalism Innovation in a Time of Survival". In: *News Media Innovation Reconsidered: Ethics and Values in a Creative Reconstruction of Journalism.* Ed. by María Luengo and Susana Herrera-Damas. John Wiley & Sons, Ltd, pp. 40–52. ISBN: 9781119706519. DOI: 10.1002/9781119706519.ch3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119706519.ch3.

Hess, Kristy and Lisa Jane Waller (2020). "Local newspapers and coronavirus: conceptualising connections, comparisons and cures". In: *Media International Australia* 178.1, pp. 21–35. DOI: 10.1177/1329878x20956455. eprint: https://doi.org/10.1177%2F1329878X20956455. URL: https://doi.org/10.1177/1329878X20956455.

Hofman, Jake M. et al. (2021). "Integrating explanation and prediction in computational social science". In: *Nature* 595.7866, pp. 181–188. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03659-0. URL: https://www.nature.com/articles/s41586-021-03659-0.

Hollstein, Betina (2021). "Georg Simmel's Contribution to Social Network Research". In: *Personal Networks.* Cambridge University Press, pp. 44–59. DOI: 10.1017/9781108878296.003.

Holovaty, Adrian (2005). *Announcing chicagocrime.org.* URL: https://www.holovaty.com/writing/chicagocrime.org-launch/.

– (2006). *A fundamental way newspaper sites need to change.* URL: http://www.holovaty.com/writing/fundamental-change/.

– (2008). *In memory of chicagocrime.org.* URL: https://www.holovaty.com/writing/chicagocrime.org-tribute/.

– (2009). *The definitive, two-part answer to "is data journalism?"* URL: http://www.holovaty.com/writing/data-is-journalism/ (visited on 06/12/2020).

Horiuchi, Yusaku, Tadashi Komatsu, and Fumio Nakaya (2012). “Should Candidates Smile to Win Elections? An Application of Automated Face Recognition Technology”. In: *Political Psychology* 33.6, pp. 925–933. DOI: 10.1111/j.1467-9221.2012.00917.x.

Howard, Alexander (2014). “The Art and Science of Data-Driven Journalism”. In: *Tow Center for Digital Journalism.* DOI: 10.7916/D8Q531V1.

Hupperich, Thomas et al. (2018). “An Empirical Study on Online Price Differentiation”. In: *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy.* ACM. DOI: 10.1145/3176258.3176338.

Iyer, Bala and Kristen Getchell (2018). “Why APIs Should Be Regulated”. In: *MIT Sloan Management Review.* URL: https://sloanreview.mit.edu/article/why-regulate-digital-organizations-apis/ (visited on 10/22/2023).

Jeppesen, Sandra (2023). “Radical Data Journalism”. In: *Radical Journalism. Resurgence, Reform, Reaction.* Ed. by Seamus Farrell, Eugenia Siapera, and George Souvlis. London, UK: Routledge. ISBN: 9781003221784.

Jo, Wonkwang et al. (2021). “A social network analysis of the spread of COVID-19 in South Korea and policy implications”. In: *Scientific Reports* 11.1. DOI: 10.1038/s41598-021-87837-0.

Jürgens, Pascal, Christine E. Meltzer, and Michael Scharkow (2022). “Age and Gender Representation on German TV”. In: *Computational Communication Research* 4.1. DOI: 10.5117/ccr2022.1.005.jurg.

Kalatzi, Olga, Charalampos Bratsas, and Andreas Veglis (2018). “The Principles, Features and Techniques of Data Journalism”. In: *Studies in Media and Communication* 6, p. 36. DOI: 10.11114/smc.v6i2.3208. eprint: https://doi.org/10.11114/smc.v6i2.3208. URL: https://doi.org/10.11114/smc.v6i2.3208.

Kandel, Sean et al. (2012). “Enterprise Data Analysis and Visualization: An Interview Study”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12, pp. 2917–2926. DOI: 10.1109/tvcg.2012.219.

Kaplan, Abraham (1964). *The Conduct of Inquiry. Methodology for Behavioural Science.* San Francisco, CA, USA: Chandler Pub. Co. 428 pp.

Karlsen, Joakim and Eirik Stavelin (2013). “Computational Journalism in Norwegian Newsrooms”. In: *Journalism Practice* 8.1, pp. 34–48. DOI: 10.1080/17512786.2013.813190.

Kerr, Cliff C. et al. (2021). “Covasim: An agent-based model of COVID-19 dynamics and interventions”. In: *PLOS Computational Biology* 17.7. Ed. by Manja Marz, e1009149. DOI: 10.1371/journal.pcbi.1009149.

Kim, Annice et al. (2013). “Can Tweets Replace Polls? A U.S. Health-Care Reform Case Study”. In: *Social Media, Sociality, and Survey Research.* Ed. by Craig A. Hill, Elizabeth Dean, and Joe Murphy. John Wiley & Sons, Inc., pp. 61–86. DOI: 10.1002/9781118751534.ch3.

Kim, Bomin et al. (2018). “The Hyperedge Event Model”. In: *arXiv.* DOI: 10.48550/ARXIV.1807.08225. URL: https://arxiv.org/abs/1807.08225.

Kleinnijenhuis, Jan et al. (2013). “Financial news and market panics in the age of high-frequency sentiment trading algorithms”. In: *Journalism* 14.2, pp. 271–291. DOI: 10.1177/1464884912468375.

Kniffin, Kevin M. et al. (2021). “COVID-19 and the workplace: Implications, issues, and insights for future research and action”. In: *American Psychologist* 76.1, pp. 63–77. ISSN: 1935-990X. DOI: 10.1037/amp0000716.

Konow-Lund, Maria, Junai Mtchedlidze, and Jens Barland (2022). “Organizational and Occupational Innovation when Implementing a Covid-19 Live Tracker in VG Newsroom”. In: *Journalism Practice*, pp. 1–17. ISSN: 1751-2794. DOI: 10.1080/17512786.2022.2116592.

Krawczyk, Konrad et al. (2021). “Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource”. In: *Journal of Medical Internet Research* 23.6, e28253. DOI: 10.2196/28253. eprint: https://doi.org/10.2196/28253. URL: https://doi.org/10.2196/28253.

Ksiazek, Thomas B., Limor Peer, and Andrew Zivic (2014). “Discussing the News”. In: *Digital Journalism* 3.6, pp. 850–870. DOI: 10.1080/21670811.2014.972079.

Lazer, David, Eszter Hargittai, et al. (2021). “Meaningful measures of human society in the twenty-first century”. In: *Nature* 595.7866, pp. 189–196. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03660-7. URL: https://www.nature.com/articles/s41586-021-03660-7.

Lazer, David, Alex (Sandy) Pentland, et al. (2009). “Life in the network: the coming age of computational social science”. In: *Science* 323.5915, pp. 721–723. ISSN: 0036-8075. DOI: 10.1126/science.1167742. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/.

Lecun, Yann et al. (1998). “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Lee, Eun-Ju (2021). “Making Sense of Pandemic-Induced Changes in Journalism and Beyond”. In: *Digital Journalism* 9.9, pp. 1431–1437. ISSN: 2167-0811. DOI: 10.1080/21670811.2021.1997149. URL: https://doi.org/10.1080/21670811.2021.1997149.

Lee, Seung Heyck et al. (2022). “Perceptions of using infographics for scientific communication on social media for COVID-19 topics: a survey study”. In: *Journal of*

*Visual Communication in Medicine* 45.2, pp. 105–113. ISSN: 1745-3054. DOI: 10.1080/17453054.2021.2020625. URL: https://doi.org/10.1080/17453054.2021.2020625.

Lerner, Juergen et al. (2013). "Modeling frequency and type of interaction in event networks". In: *Corvinus Journal of Sociology and Social Policy* 4.1, pp. 3–32. DOI: 10.14267/cjssp.2013.01.01.

Lerner, Jürgen and Marian-Gabriel Hâncean (2023). "Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks". In: *Network Science* 11.1, pp. 5–35. ISSN: 2050-1242, 2050-1250. DOI: 10.1017/nws.2022.29. URL: https://www.cambridge.org/core/journals/network-science/article/microlevel-network-dynamics-of-scientific-collaboration-and-impact-relational-hyperevent-models-for-the-analysis-of-coauthor-networks/375932B5B86D2033A0A290DE8198BB32.

Lerner, Jürgen, Alessandro Lomi, et al. (2021). "Dynamic network analysis of contact diaries". In: *Social Networks* 66, pp. 224–236. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2021.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0378873321000277.

Lerner, Jürgen, Mark Tranmer, et al. (2019). *REM beyond dyads: relational hyperevent models for multi-actor interaction networks.* DOI: 10.48550/ARXIV.1912.07403.

Lewis, Norman P. and Eisa Al Nashmi (2019). "Data Journalism in the Arab Region: Role Conflict Exposed". In: *Digital Journalism* 7.9, pp. 1200–1214. ISSN: 2167-0811. DOI: 10.1080/21670811.2019.1617041. URL: https://doi.org/10.1080/21670811.2019.1617041.

De-Lima-Santos, Mathias-Felipe (2022). "ProPublica's Data Journalism: How Multidisciplinary Teams and Hybrid Profiles Create Impactful Data Stories". In: *Media and Communication* 10.1, pp. 5–15. ISSN: 2183-2439. DOI: 10.17645/mac.v10i1.4433.

de-Lima-Santos, Mathias-Felipe and Lucia Mesquita (2021). "Data Journalism Beyond Technological Determinism". In: *Journalism Studies* 22.11, pp. 1416–1435. ISSN: 1461-670X. DOI: 10.1080/1461670X.2021.1944279. URL: https://doi.org/10.1080/1461670X.2021.1944279.

Loosen, Wiebke, Julius Reimer, and Fenja De Silva-Schmidt (2017). "Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013-2016". In: *Journalism* 21.9 (9), pp. 1246–1263. ISSN: 1464-8849. DOI: 10.1177/1464884917735691. URL: https://doi.org/10.1177/1464884917735691 (visited on 03/13/2021).

Lorenz, Mirko (2012). "Business Models for Data Journalism". In: *The data journalism handbook 1.* Ed. by Jonathan Gray et al. European Journalism Centre. URL: https://datajournalism.com/read/handbook/one/introduction/why-is-data-journalism-important (visited on 05/23/2020).

Macy, Michael W. and Robert Willer (2002). "From Factors to Actors: Computational Sociology and Agent-Based Modeling". In: *Annual Review of Sociology* 28.1, pp. 143–166. DOI: 10.1146/annurev.soc.28.110601.141117.

Malik, Momin M. and Jürgen Pfeffer (2016). "A Macroscopic Analysis of News Content in Twitter". In: *Digital Journalism* 4.8, pp. 955–979. ISSN: 2167-0811. DOI: 10.1080/21670811.2015.1133249. URL: https://doi.org/10.1080/21670811.2015.1133249.

Marres, Noortje and Carolin Gerlitz (2016). "Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology". In: *The Sociological Review* 64.1, pp. 21–46. DOI: 10.1111/1467-954x.12314.

Mas, Alexandre and Enrico Moretti (2009). "Peers at Work". In: *American Economic Review* 99.1, pp. 112–145. DOI: 10.1257/aer.99.1.112.

Meier, Klaus et al. (2022). "Examining the Most Relevant Journalism Innovations: A Comparative Analysis of Five European Countries from 2010 to 2020". In: *Journalism and Media* 3.4, pp. 698–714. ISSN: 2673-5172. DOI: 10.3390/journalmedia3040046. URL: https://www.mdpi.com/2673-5172/3/4/46.

Mellado, Claudia et al. (2021). "Sourcing Pandemic News: A Cross-National Computational Analysis of Mainstream Media Coverage of COVID-19 on Facebook, Twitter, and Instagram". In: *Digital Journalism* 9.9, pp. 1261–1285. ISSN: 2167-0811. DOI: 10.1080/21670811.2021.1942114. URL: https://doi.org/10.1080/21670811.2021.1942114.

Meyer, Philip (1973). *Precision Journalism: A Reporter's Introduction to Social Science Methods.* 1st ed. Bloomington, Indiana, USA: Indiana University Press.

– (2002). *Precision journalism: a reporter's introduction to social science methods.* English. 4th ed. Lanham, Maryland: Rowman & Littlefield. ISBN: 9780742510876.

Mierzejewska, Bozena (2011). "Media Management Theory and Practice". In: Deuze, Mark. *Managing media work.* SAGE, pp. 13–30. ISBN: 9781412971249.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: DOI: 10.48550/ARXIV.1301.3781. arXiv: 1301.3781 [cs.CL].

Milan, Stefania and Emiliano Treré (2020). "The Rise of the Data Poor: The COVID-19 Pandemic Seen From the Margins". In: *Social Media + Society* 6.3. DOI: 10.1177/2056305120948233. eprint: https://doi.org/10.1177/2056305120948233. URL: https://doi.org/10.1177/2056305120948233.

Mooney, Christopher Z. (1996). "Bootstrap Statistical Inference: Examples and Evaluations for Political Science". In: *American Journal of Political Science* 40.2, p. 570. DOI: 10.2307/2111639.

Morini, Francesca, Marian Dörk, and Ester Appelgren (2022). "Sensing What's New: Considering Ethics When Using Sensor Data in Journalistic Practices". In: *Digital Journalism*, pp. 1–19. DOI: 10.1080/21670811.2022.2134161.

Mutsvairo, Bruce and Saba Bebawi (2022). "Journalism and the Global South: Shaping Journalistic Practices and Identity Post "Arab Spring": Special Issue: Remembering the Arab Spring: Pursuing Possibilities and Impediments in Journalistic Professional Practice across the Global South". In: *Digital Journalism* 10.7, pp. 1141–1155. ISSN: 2167-082X. DOI: 10.1080/21670811.2022.2107551.

Nelson, Laura K. et al. (2018). “The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods”. In: *Sociological Methods & Research* 50.1, pp. 202–237. DOI: 10.1177/0049124118769114.

Newman, Nic, Richard Fletcher, Craig T. Robertson, et al. (2022). *Reuters Institute Digital News Report 2022.* Oxford, United Kingdom: Reuters Institute for the Study of Journalism.

Newman, Nic, Richard Fletcher, Anne Schulz, et al. (2021). *Reuters Institute Digital News Report 2021.* Oxford, United Kingdom: Reuters Institute for the Study of Journalism. URL: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.

Ojo, Adegboyega and Bahareh Heravi (2017). “Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences”. In: *Digital Journalism* 6.6, pp. 693–718. ISSN: 2167-082X. DOI: 10.1080/21670811.2017.1403291.

Ooms, Jeroen (2014). “The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects”. In: *arXiv:1403.2805 [stat.CO].* URL: https://arxiv.org/abs/1403.2805.

Osnabrügge, Moritz, Elliott Ash, and Massimo Morelli (2021). “Cross-Domain Topic Classification for Political Texts”. In: *Political Analysis* 31.1, pp. 59–80. DOI: 10.1017/pan.2021.37.

Palomo, Bella, Laura Teruel, and Elena Blanco-Castilla (2019). “Data Journalism Projects Based on User-Generated Content. How La Nacion Data Transforms Active Audience into Staff”. In: *Digital Journalism* 7.9, pp. 1270–1288. ISSN: 2167-082X. DOI: 10.1080/21670811.2019.1626257.

Parasie, Sylvain and Eric Dagiral (2012). “Data-driven journalism and the public good: ”Computer-assisted-reporters” and ”programmer-journalists” in Chicago:” in: *New Media & Society* 15 (6), pp. 853–871. DOI: 10.1177/1461444812463345.

Peng, Yilang (2020). “What Makes Politicians’ Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision”. In: *The International Journal of Press/Politics* 26.1, pp. 143–166. DOI: 10.1177/1940161220964769.

Pentzold, Christian and Denise Fechner (2019). “Data journalism’s many futures: Diagrammatic displays and prospective probabilities in data-driven news predictions”. In: *Convergence: The International Journal of Research into New Media Technologies.* DOI: 10.1177/1354856519880790. URL: https://doi.org/10.1177/13548565198807.

– (2021). “Probabilistic Storytelling and Temporal Exigencies in Predictive Data Journalism”. In: *Digital Journalism* 9.6, pp. 715–736. DOI: 10.1080/21670811.2021.1878920. eprint: https://doi.org/10.1080/21670811.2021.1878920. URL: https://doi.org/10.1080/21670811.2021.1878920.

Pentzold, Christian, Denise J. Fechner, and Conrad Zuber (2021). ““Flatten the Curve”: Data-Driven Projections and the Journalistic Brokering of Knowledge during the COVID-19 Crisis”. In: *Digital Journalism* 9.9, pp. 1367–1390. DOI: 10.1080/21670811.

2021.1950018. eprint: https://doi.org/10.1080/21670811.2021.1950018. URL: https://doi.org/10.1080/21670811.2021.1950018.

Perreault, Mildred F. and Gregory P. Perreault (2021). "Journalists on COVID-19 Journalism: Communication Ecology of Pandemic Reporting". In: *American Behavioral Scientist* 65.7, pp. 976–991. DOI: 10.1177/0002764221992813.

Pestian, John P. et al. (2012). "Sentiment Analysis of Suicide Notes: A Shared Task". In: *Biomedical Informatics Insights* 5s1, BII.S9042. DOI: 10.4137/bii.s9042.

Pfeffer, Juergen et al. (2023). "This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API". In: *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM-2023).* Forthcoming.

Porlezza, Colin and Sergio Splendore (2019). "From Open Journalism to Closed Data: Data Journalism in Italy". In: *Digital Journalism* 7.9, pp. 1230–1252. ISSN: 2167-0811. DOI: 10.1080/21670811.2019.1657778. URL: https://doi.org/10.1080/21670811.2019.1657778.

Possler, Daniel, Sophie Bruns, and Julia Niemann-Lenz (2019). "Computational Methods for Communication Science— Data Is the New Oil—But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science". In: *International Journal of Communication* 13.0. ISSN: 1932-8036. URL: https://ijoc.org/index.php/ijoc/article/view/10737.

Puschmann, Cornelius (2019). "An end to the wild west of social media research: a response to Axel Bruns". In: *Information, Communication & Society* 22.11, pp. 1582–1589. DOI: 10.1080/1369118x.2019.1646300.

Puschmann, Cornelius and Jean Burgess (2013). "The Politics of Twitter Data". In: *SSRN Electronic Journal.* DOI: 10.2139/ssrn.2206225.

Quandt, Thorsten and Karin Wahl-Jorgensen (2021). "The Coronavirus Pandemic as a Critical Moment for Digital Journalism". In: *Digital Journalism* 9.9, pp. 1199–1207. DOI: 10.1080/21670811.2021.1996253.

– (2022). "The Coronavirus Pandemic and the Transformation of (Digital) Journalism". In: *Digital Journalism* 10.6, pp. 923–929. DOI: 10.1080/21670811.2022.2090018. eprint: https://doi.org/10.1080/21670811.2022.2090018. URL: https://doi.org/10.1080/21670811.2022.2090018.

Quinn, Kevin M. et al. (2009). "How to Analyze Political Attention with Minimal Assumptions and Costs". In: *American Journal of Political Science* 54.1, pp. 209–228. DOI: 10.1111/j.1540-5907.2009.00427.x.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Radford, Alec, Karthik Narasimhan, et al. (2018). "Improving language understanding by generative pre-training". In: URL: https://openai.com/research/language-unsupervised (visited on 11/04/2023).

Radford, Alec, Jeffrey Wu, et al. (2019). "Language Models are Unsupervised Multitask Learners". In: URL: https://openai.com/research/better-language-models (visited on 11/04/2023).

Rieder, Bernhard and Jeanette Hofmann (2020). "Towards platform observability". In: *Internet Policy Review* 9.4. DOI: 10.14763/2020.4.1535.

Rogers, Everett M. (2003). *Diffusion of Innovations, 5th Edition*. Fifth edition. New York: Simon & Schuster. 576 pp. ISBN: 0743222091.

Rogers, Simon (2010). "Florence Nightingale, datajournalist: information has always been beautiful". In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/news/datablog/2010/aug/13/florence-nightingale-graphics (visited on 06/12/2020).

– (2011). "Data journalism at the Guardian: what is it and how do we do it?" In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism (visited on 06/12/2020).

– (2013a). "Data journalism in action: what is Facts are Sacred about?" In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/news/datablog/2013/apr/04/data-journalism-facts-are-sacred (visited on 05/23/2020).

– (2013b). "John Snow's data journalism: the cholera map that changed the world". In: *The Guardian*. ISSN: 0261-3077. URL: https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map (visited on 06/12/2020).

– (2021). "From The Guardian to Google News Lab: A Decade of Working in Data Journalism". In: *The Data Journalism Handbook*. Ed. by Liliana Bounegru and Jonathan Gray. Amsterdam, The Netherlands: Amsterdam University Press, pp. 279–285. DOI: 10.5117/9789462989511_ch38. URL: https://datajournalism.com/read/handbook/two/organising-data-journalism/a-decade-of-data-journalism-2009-2019 (visited on 12/09/2023).

Scharkow, Michael (2017). "Bootstrapping". In: *The International Encyclopedia of Communication Research Methods*. Ed. by Jörg Matthes, Christine S. Davis, and Robert F. Potter. John Wiley & Sons, Ltd. ISBN: 9781118901731. DOI: https://doi.org/10.1002/9781118901731.iecrm0017.

Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman (2015). "Big Data, Digital Media, and Computational Social Science". In: *The ANNALS of the American Academy of Political and Social Science* 659.1, pp. 6–13. DOI: 10.1177/0002716215572084.

Spyridou, Lia-Paschalia et al. (2013). "Journalism in a state of flux". In: *International Communication Gazette* 75.1, pp. 76–98. DOI: 10.1177/1748048512461763.

Szpunar, Karl K., R. Nathan Spreng, and Daniel L. Schacter (2014). "A taxonomy of prospection: Introducing an organizational framework for future-oriented cognition". In: *Proceedings of the National Academy of Sciences* 111.52, pp. 18414–18421. DOI: 10.1073/pnas.1417144111. URL: https://www.pnas.org/doi/10.1073/pnas.1417144111.

Taboada, Maite, Julian Brooke, and Manfred Stede (2009). “Genre-Based Paragraph Classification for Sentiment Analysis”. In: *Proceedings of the SIGDIAL 2009 Conference.* Ed. by Patrick Healey et al. London, UK: Association for Computational Linguistics, pp. 62–70. URL: https://aclanthology.org/W09-3909.

Tandoc, Edson C. and Soo-Kwang Oh (2017). “Small Departures, Big Continuities?” In: *Journalism Studies* 18.8 (8), pp. 997–1015. ISSN: 1461-670X. DOI: 10.1080/1461670X.2015.1104260. URL: https://doi.org/10.1080/1461670X.2015.1104260.

Tandoc, Edson C. and Ryan J. Thomas (2015). “The Ethics of Web Analytics”. In: *Digital Journalism* 3.2, pp. 243–258. ISSN: 2167-0811. DOI: 10.1080/21670811.2014.909122. URL: https://doi.org/10.1080/21670811.2014.909122.

Timberg, Craig (2021). “Facebook made big mistake in data it provided to researchers, undermining academic work”. In: *Washington Post.* ISSN: 0190-8286. URL: https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/ (visited on 10/15/2023).

Toff, Benjamin et al. (2020). *What we think we know and what we want to know: perspectives on trust in news in a changing world — Reuters Institute for the Study of Journalism.* URL: https://reutersinstitute.politics.ox.ac.uk/what-we-think-we-know-and-what-we-want-know-perspectives-trust-news-changing-world (visited on 10/14/2023).

Travers, Jeffrey and Stanley Milgram (1969). “An Experimental Study of the Small World Problem”. In: *Sociometry* 32.4, p. 425. DOI: 10.2307/2786545.

Urman, Aleksandra and Stefan Katz (2020). “What they do in the shadows: examining the far-right networks on Telegram”. In: *Information, Communication & Society* 25.7, pp. 904–923. DOI: 10.1080/1369118x.2020.1803946.

Usher, Nikki, Jesse Holcomb, and Justin Littman (2018). “Twitter Makes It Worse: Political Journalists, Gendered Echo Chambers, and the Amplification of Gender Bias”. In: *The International Journal of Press/Politics* 23.3, pp. 324–344. DOI: 10.1177/1940161218781254. URL: https://doi.org/10.1177/1940161218781254.

Venturini, Tommaso and Richard Rogers (2019). ““API-Based Research” or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach”. In: *Digital Journalism* 7.4, pp. 532–540. DOI: 10.1080/21670811.2019.1591927.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). “The spread of true and false news online”. In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.

Waite, Matt (2007). *Announcing PolitiFact.* URL: http://www.mattwaite.com/posts/2007/aug/22/announcing-politifact/ (visited on 10/08/2023).

Waldherr, Annie (2014). “Emergence of News Waves: A Social Simulation Approach”. In: *Journal of Communication* 64.5, pp. 852–873. DOI: 10.1111/jcom.12117.

Wankmüller, Sandra (2022). “Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis”. In: *Sociological Methods & Research.* DOI: 10.1177/00491241221134527.

Watts, Duncan J. (2004). *Six degrees. The science of a connected age.* [Repr.] First published as a Norton paperback 2004. - Includes bibliographical references and index. New York , NY [u.a.]: Norton. 374 pp. ISBN: 0393041425.

Weinacht, Stefan and Ralf Spiller (2014). "Datenjournalismus in Deutschland". de. In: *Publizistik* 59.4, pp. 411–433. ISSN: 1862-2569. DOI: 10.1007/s11616-014-0213-5. URL: https://doi.org/10.1007/s11616-014-0213-5.

– (2022). "Datenjournalismus in Deutschland revisited". In: *Publizistik* 67.2-3, pp. 243–274. DOI: 10.1007/s11616-022-00747-7.

Weiss, Amy Schmitz and Jéssica Retis (2018). "'I Don't Like Maths, That's Why I am in Journalism': Journalism Student Perceptions and Myths about Data Journalism". In: *Asia Pacific Media Educator* 28.1, pp. 1–15. DOI: 10.1177/1326365X18780418.

Welles, Brooke Foucault et al. (2014). "Dynamic Models of Communication in an Online Friendship Network". In: *Communication Methods and Measures* 8.4, pp. 223–243. DOI: 10.1080/19312458.2014.967843.

Wenger, Etienne, Richard McDermott, and William M. Snyder (2002). *Cultivating Communities of Practice. A guide to managing knowledge.* Brighton, Massachusetts, USA: Harvard Business School Press. ISBN: 9781578513307.

Wettstein, Martin (2020). "Simulating hidden dynamics". In: *Computational Communication Research* 2.1, pp. 1–33. DOI: 10.5117/ccr2020.1.001.wett.

Wickham, Hadley (2022). *rvest: Easily Harvest (Scrape) Web Pages.* R package version 1.0.3. URL: https://CRAN.R-project.org/package=rvest.

Wickham, Hadley et al. (2019). "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. DOI: 10.21105/joss.01686.

Williams, Nora Webb, Andreu Casas, and John D. Wilkerson (2020). *Images as Data for Social Science Research.* Cambridge University Press. DOI: 10.1017/9781108860741.

Witzenberger, Benedict and Nicholas Diakopoulos (2023). "Election predictions in the news: how users perceive and respond to visual election forecasts". In: *Information, Communication & Society*, pp. 1–22. DOI: 10.1080/1369118x.2023.2230267.

Witzenberger, Benedict and Jürgen Pfeffer (2024a). "More Inclusive and Wider Sources: A Comparative Analysis of Data and Political Journalists on Twitter (Now X) in Germany". In: *Journalism and Media* 5.1, pp. 412–431. DOI: 10.3390/journalmedia5010027.

– (2024b). *Unleashing Data Journalism's Potential: COVID-19 as Catalyst for Newsroom Transformation.* DOI: https://doi.org/10.48550/arXiv.2401.14816. eprint: arXiv:2401.14816.

Witzenberger, Benedict, Angelina Voggenreiter, and Jürgen Pfeffer (2024). "Popular and on the Rise—But Not Everywhere: COVID-19-Infographics on Twitter". In: *Data Science—Analytics and Applications. Proceedings of the 5th International Data Science Conference—iDSC2023.* Ed. by Peter Haber, Thomas J. Lampoltshammer, and Manfred Mayr. Cham: Springer Nature Switzerland, pp. 53–60. ISBN: 9783031421709. DOI: 10.1007/978-3-031-42171-6_7.

Xie, Gang (2020). “A novel Monte Carlo simulation procedure for modelling COVID-19 spread over time”. In: *Scientific Reports* 10.1. DOI: 10.1038/s41598-020-70091-1.

Young, Lori and Stuart Soroka (2012). “Affective News: The Automated Coding of Sentiment in Political Texts”. In: *Political Communication* 29.2, pp. 205–231. DOI: 10.1080/10584609.2012.671234.

Zhang, Xinzhi and Minyi Chen (2020). “Journalists’ Adoption and Media’s Coverage of Data-driven Journalism: a Case of Hong Kong”. In: *Journalism Practice* 16.5, pp. 901–919. DOI: 10.1080/17512786.2020.1824126.