Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections

BRIAN MCINNIS, University of California San Diego, USA
LU SUN, University of California San Diego, USA
JUNGWON SHIN, Cornell University, USA
STEVEN P. DOW, University of California San Diego, USA

News websites can facilitate global discussions about civic issues, but the financial cost and burden of moderating these forums has forced many to disable their commenting systems. In this paper, we consider the role that data visualizations play in online discussion around a civic issue, through an analysis of how people talk about climate change data in the comment threads at three news websites (i.e., Breitbart news, the Guardian, the New York Times). We find that out of 6,525 comments, only 2.4% reference data visualizations in the articles. While rare, the paper presents illustrative examples of how people refer to data—their collection, analysis, and visual representation—to engage with an article's narrative. Using text classification techniques we identify several features related to the content of comments that contain data-centered talk, such as article cosine similarity, hyperlinks, and comparison terms. Finally, we discuss potential ways that newsrooms might apply this analysis to promote data literacy, data science, and to foster community around shared experiences.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Online discussion, civic issues, data visualization, data literacy

ACM Reference Format:
Brian McInnis, Lu Sun, Jungwon Shin, and Steven P. Dow. 2020. Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 174 (October 2020), 27 pages. https://doi.org/10.1145/3415245

1 INTRODUCTION

Online systems have the potential to facilitate informed discussion about civic issues with large and globally distributed audiences [16, 69], but the commenting sections at news websites rarely meet this expectation [7, 15, 85, 99]. Popular news sites regularly receive a high volume of comments, which can be challenging to moderate [25]. Commenting sections also tend to elicit a substantial degree of incivility [40] whether due to the relative anonymity of users [44, 99], regular bad actors [7, 37], or to controversial topics that stir up readers' emotions [14, 15]. However, even when people are civil they are also not often talking with each other [20, 48]. In light of these factors, many news sites have opted to close their commenting sections altogether [31, 56].

In this paper, we explore the value of online conversations about data that emerge on news articles that present evidence about civic issues through data visualizations. A well-crafted visualization can convey meaningful insights about data that are hard to express in text [51]. Data visualizations

Authors' addresses: Brian McInnis, bmcinnis@ucsd.edu, University of California San Diego, USA; Lu Sun, l5sun@ucsd.edu, University of California San Diego, USA; Jungwon Shin, js2888@cornell.edu, Cornell University, USA; Steven P. Dow, spdow@ucsd.edu, University of California San Diego, USA.

2573-0142/2020/10-ART174 $15.00
https://doi.org/10.1145/3415245

offer an entry point for people to use their automatic perceptual skills to notice, wonder, and talk about observations [43, 95]. In data journalism, visualizations are not intended to end conversation, but rather to aid informed discussion about issues raised by the article [52].

However, we know little about the extent and the ways that people weigh observations about data in online discussions. To deepen understanding of data-centered talk in the wild, we examine the following research questions through a mixed-methods analysis of online comments to a collection of news articles:

- **RQ1**: How prevalent is online discussion about data at news websites?
- **RQ2**: What communication themes emerge in online discussions about data journalism?
- **RQ3**: What predictive factors help to automatically recognize comments about data?

The paper presents a deep dive into the commenting behaviors around visualizations of climate change data posted on articles by the Guardian and the New York Times, as well as articles by Breitbart news that largely contest climate change. The selected articles reflect three common topics across the news sites: (1) dietary choices and carbon dioxide emissions, (2) state-level transition towards renewable energy, and (3) living with severe weather. While similar in topic, the news articles make use of climate change data to emphasize different narratives.

Our analysis found that just 10.9% of 6,525 comments mention data related to an article's topic and a mere 2.4% referenced the data visualizations in an article (157 comments). We analyzed these *data-centered* comments and discovered eleven illustrative themes that depict the range of ways that people refer to and engage with a civic issue through data journalism. Given the potential for data to foster discussion about an article, we explored to what extent we can identify data-centered talk through text classification techniques. Our text analysis of the comments identified several predictive factors that can be used to automatically classify comments that contain data-centered talk with 90.36% accuracy. These factors include comments that are linguistically similar with the article text, and contain hyperlinks, numbers, and comparison terms.

This article presents evidence that data-centered talk is rare, yet valuable in online discussion about civic issues. The emergent themes signify instances where people question the data provenance, make sense of visualizations, and share personal stories about data. We show how automated text classification techniques could help moderators (and the readers) recognize and build on top of data-centered reasoning. Finally, we discuss several possible ways that newsrooms might apply this analysis to promote objectives related to data literacy, data science, and toward fostering community around shared experiences through online discussion about news articles.

2 RELATED WORK

There are many reasons why people choose to participate in online discussion about news articles [25, 91]. As a result the patterns of discourse in an online discussion can vary widely from monologues issued by one-time participants to banter among people who comment together frequently. This paper focuses on discourse about data analysis as one of many potential focal points for the informed online civic discussion that can follow from a news article.

2.1 How rhetorical elements provide context in a data visualization

How people communicate about data is an important topic in research about the practice of data science. For data science teams, data management and analysis are social activities often involving large groups of people [103]. As data work progresses—from collection to stages of processing, exploring, and interpreting—data science teams deliberate at each step in order to maintain a shared understanding of the data as a basis for their trust in the analysis and to avoid downstream problems, since each stage of analysis builds on the prior [75, 77]. While deliberation is common in data science work, it is less apparent if and how readers exhibit these social processes around data visualizations on popular news websites.

Deliberation in data science often centers around visualizations of an analysis. In order for people to talk with each other about a data visualization, they need to have a shared understanding about how to read the visualization by interpreting its labels, captions, visual chunks, and other guides (called *rhetorical elements*) [51]. Stasko [95] argues that the value of a data visualization, "goes beyond the ability to support answering questions about data—it centers upon a visualization's ability to convey a true understanding of the data, a more holistic broad and deep innate sense of the context" [95, pg. 48]. To meet this objective, an effective data visualization *should* fulfill several informational capabilities [95]: (1) minimize the time to answer diverse questions, (2) spur the generation of insights and insightful questions, (3) convey the essence of the data, and (4) generate confidence and knowledge about the data's domain and context.

Some rhetorical elements require more cognitive effort to interpret than others. For example, proportions are commonly used to communicate the probability of a health risk, whether graphically (e.g., as an icon array, along a color-scale) or numerically (e.g., as a percentage, or ratio). In a systematic review of health-risk communication, Ancker et al. [1] offer several recommendations about how to improve quantitative judgement with proportion-based risk communication, such as making the numerators and denominator visually salient, sizing each graphical element in proportion to the numbers they portray. These designs make use of automatic perceptual skills and learned skills, like calculation, to help people interpret proportions on a graph [1].

Rhetorical elements can also be used to communicate specific narratives. Often health-risk communication promotes behavior change, but as discussed in Ancker et al. [1], this emphasizes different rhetorical elements than a visualization intended to promote quantitative judgment. For example, emphasizing the numerator of a risk ratio is more likely to promote a behavior change, but deemphasizing (or removing) the denominator can also inflate a viewers' perception of risk [1]. Rhetorical elements in a data visualization can be manipulated to different intents [51].

Unlike communicating a risk, where there is a known probability, statistical uncertainty refers to situations where the probability of an event is less clear, because information about the event is only partially observable [53]. Error-bars, confidence intervals, and forecast lines are several common ways to represent uncertainty on a data visualization [90]. However uncertainty is more often expressed in text than it is graphically [51].

Data science training programs are designed to prepare people to recognize these and other considerations related to the analysis and visual representation of data [76]. Our study investigates how online news readers make sense of, refer to, and are motivated by data presented in an article; however, online news readers may not have access to the time and training necessary to develop their data literacy (e.g., [4, 38]).

2.2 How data journalism is structured to introduce people to an analysis

News articles that present an issue through data analysis (called *data journalism*) are typically framed so that they are accessible to a general audience, who may have a low-level of data literacy (e.g., [4, 38]). In order to help readers into an analysis, data journalists will structure a news article to present data in specific ways. Based on interviews with 26 major news organizations in Europe, Weber et al. [107] describe three main patterns for presenting data with a news article: (1) *linear*, guided-tour through an analysis, (2) *nonlinear*, reader-driven exploration, and (3) *linear-nonlinear hybrid*, where a basic linear story is presented, so that the audience has a sense of how to explore the data on their own.

Several considerations guide newsroom decisions about how to present data journalism [32]. First, a lot of care can go into the production of a data visualization. Passi and Jackson [77] describe this work as a continuous negotiation about trust in a data analysis, e.g., What numbers matter? How might we rationalize a result? How credible is the data? What assumptions underlie our modeling approach? In the newsroom, journalists and data science teams sometimes deliberate about how much of this negotiation to present to readers [32].

Second, there is an open question about whether data journalism should offer readers a nonlinear (or unguided) exploration of a data set, or not [32]. An unguided tour opens the data narrative broadly to audience interpretation, whereas with a linear structure, each visualization of the data is presented as evidence to support an article's narrative. As many people lack access to data literacy skills (e.g., [4, 38]), it may be easier for people on a nonlinear tour to get sidetracked from the narrative or stumble into a false interpretation; however, a linear approach may also stifle discussion by limiting the focus to specific data observations.

Some newsrooms promote data literacy by hosting expert-facilitated online discussions about news articles [43]. As an example, in 2017 the New York Times Learning Network partnered with the American Statistical Association (ASA) to facilitate discussion about articles that contain data visualizations. Every week, students and teachers are encouraged to consider the data presented in an article, with the following prompts: (1) What do you notice? (2) What do you wonder? And (3) What do you think is going on in this graph? The prompts are intended to invite students to think critically about the presented data analysis and to share those thoughts in the online discussion.

Third, newsrooms also think about how readers may relate to the data analysis presented in an article, as some readers may have deeply personal experiences related to the issues. Through an interview-based study of how people in rural Pennsylvania, USA are motivated towards or away from different data visualizations, Peck et al. [78] find that people gravitate toward visualizations that depict issues related to their personal experience, beliefs, and values. A key implication of this observation is that first impressions matter, as people are less likely to look twice at a visualization that does not reflect them and their interests.

Data visualizations can also stir up an emotional response among readers [47, 59]. In a diary and focus-group based study of data visualizations that people find in their everyday, Kennedy and Hill [59] draw attention to the complex pattern of emotions a visualization can evoke—from pleasure in its smooth lines, to feeling lost in numbers, and empathy for the people reflected by the data. While a newsroom may structure data journalism to elicit discussion about specific topics, people may not have the skills to engage with the intended topics, but may also prefer to talk about other topics that stem from their personal experience of the data, or may just want to talk about the look and feel of a particular visualization.

2.3 How commenting systems represent and elicit discussion about data journalism

Many news websites have closed off their commenting forums due to concerns about the quality of discussion about news articles [31, 36, 56]. Proponents of online commenting at news websites have referred to these systems as enabling *participatory journalism*, where professional reporters work in tandem with their audience to gather, organize, edit, and communicate timely information about news events [24]. Critics of participatory journalism point to instances of incivility [15] and misinformation in online discussion [92], as well as to the real threats that commenting can have on a reporter's reputation [25, 46].

System design can help to address some of these concerns. For example, there are several exciting HCI/CSCW systems specifically designed to support online discussion about data analysis and visualizations in particular (e.g., [45, 104, 109]). Common features in these systems include linking and tagging to help ground the communication about data in specific observations [45, 109], micro-tasks to elicit personal insights (e.g., explanations for a trend or outliers) [108], and pathways to follow an observation from one visualization to the next in a narrative sequence [62]. These features are intended to help people communicate observations of data, so that they can collaborate with a mutual understanding of the data.

However, finding mutual understanding with others through discussion is not always what motivates people to post a comment. Survey based research has reported that people comment online about news articles because they are looking for social interaction, because they like asking and answering questions about a specific topic, or because they want to correspond with a particular reporter [91, 99]. Commenters also value commenting sections as a space to share personal stories and beliefs about a topic [25]. Above all, commenters (as well as lurkers) report that they enjoy reading comment threads at news websites, even low-quality discussions. However, non-users—people who don't comment and who don't read comments—tend to view online commenting as a waste of time [91].

The content of a comment can also influence whether and how people respond to it [2, 11]. Arguello et al. [2] present a computational analysis of factors related to an online discussion context and comment content that can effect the likelihood of receiving a response. Comments that are on-topic, include testimonials, make requests, and ask questions are more likely to receive a response than those that do not. Comments posted to active forums are also likely to receive a response, particularly if the commenter has participated in the discussion previously [2, 11]. While commenting system features and factors related to participant motivation affect online discussion, the focus of this paper is on understanding the content and context for data-centered talk.

Our research builds on an existing study into the ways that people have talked about data at data visualization blogs. Hullman et al. [52] found that 42% of comments sampled (N=1,100) from the Economist's data blog, *Graphic Detail*, "contribute information from external sources, data, and personal experiences that may be useful for interpreting the presentation" [52, pg. 1173]. This existing research demonstrates the potential that online discussion comments may add value for data journalists. Our work extends this earlier study by investigating data-centered talk in new settings and with greater depth.

3 METHOD

This paper investigates the prevalence and potential value of data-centered talk across three different news sites that support online discussions of civic topics. Our research seeks to understand how prevalent and what kinds of commenting behaviors emerge around data. We also explore the potential for text analysis technology to automatically recognize comments about data and to highlight these as building blocks within online discussion systems.

3.1 Context and data selection

3.1.1 Selection of news websites. For the analysis, we chose to focus on news articles that either use data visualizations to present evidence of climate change or to contest climate change science. Climate change refers to the Earth's atmospheric temperature due to the release of gasses that trap heat radiating from the planet, the consequences of which are difficult to predict, but increase the likelihood of extreme weather events [29]. As personal experience with the long-term and global consequences of climate change is still rare, the gap between experience and evidence can be cognitively jarring and can translate into a motivated rejection of evidence [9, 66]. Despite nearly universal agreement among the scientific community about the causes of climate change, many people remain skeptical [58].

Data journalism about climate change is a useful lens for research about discussion systems for several reasons. First, climate change has the potential to affect the lives of people around the world, and online systems offer an exciting potential to facilitate such large and global discussions [16, 54, 69]. Second, evidence about climate change is tricky to present visually, as the effects of climate change are time-delayed and uncertain [90, 106]; however, online discussion may offer data journalists an opportunity to refine their visualizations by corresponding with their audience [52]. Finally, creating opportunities for people to communicate different knowledge about divisive topics, like climate change, is a core tenet of CSCW research [28].

We initially considered 30+ news sites, but only included sites that met the following criteria: host an online discussion, regularly report about climate change, and use data visualizations to present climate change data. Additionally, it was important to select news sites that report from different perspectives on climate change, since personal beliefs about a civic issue can motivate people to engage with specific news sources [9, 66] and data visualizations [78]. Personal beliefs about an issue can influence whether people are willing to participate in an online discussion [25, 91] and can affect how they engage with topics already under discussion [70].

Based on these considerations we chose to include articles published by **The Guardian** and **The New York Times** (NYT) that report findings from climate change science, as well as articles published by **Breitbart News** that use data visualizations to contest climate change [3].

3.1.2 News website discussion context. Each of the news sites present a different commenting system and moderation practices; these choices play into norms that emerge in online discussion (e.g., [34, 64, 70, 83, 110]). Each of the news sites require participants to login before they are allowed to post a comment, a reply, or recommend comments; however, participants are not required to register if they want to share a discussion comment in social media (e.g., Twitter, Facebook) or flag inappropriate comments in the thread. All of the news sites encourage participants to use their real names, but people can join in discussion under a pseudonym.

The NYT and the Guardian systems are custom-built. Commenting is not enabled on all articles at these websites, because both the NYT and Guardian rely on human moderators to manage the discussion. At the NYT, every comment that is published is moderated.[1] At the Guardian, moderators pay attention to the comment stream and rely on users to report abusive, offensive, or otherwise inappropriate comments, so that they can be removed after appearing on the site.[2] In addition to removing abusive content, moderators at both news sites look for comments to highlight as "NYT Picks" and "Guardian Picks" that reflect high-quality contributions. The most active commenters at the NYT have referred to the coveted NYT Pick badge as akin to "the gold medal of the commenting olympics" [67]. Since discussion moderation can be tedious [34], the NYT and the Guardian only enable commenting on a selection of articles and only for a brief period (e.g., 24-hours, a few days).

Commenting is enabled in perpetuity on nearly all articles at the Breitbart news site and is managed through a third-party comment hosting service called *Disqus*.[3] Disqus has been criticized for comments posted to Breitbart news articles that violate the Disqus terms of service by posting prohibited content, including blackmail, extortion, intimidation of users, spam, and unlawful activities [12]. The Breitbart news terms of service also prohibit such commenting behaviors.[4]

3.1.3 Selection of climate change articles and discussions. Each of the news sites have designated sections (or "tags") to identify articles that are about climate change. Two members of our research team manually reviewed every article associated with climate change that was published during a one-year period between July 9, 2018-2019. The review involved identifying articles that include data visualization(s) related to climate change and allow online commenting; our analysis found 73 articles (2.3%) that met these criteria (see Table 1).

[1] The NYT - Comment FAQ: https://help.nytimes.com/hc/en-us/articles/115014792387-Comments

[2] The Guardian - Comment FAQ: https://www.theguardian.com/community-faqs

[3] https://disqus.com/

[4] Breitbart news - Comment Policy: https://www.breitbart.com/terms-of-use/

	Articles	Contain a data viz	Commenting enabled
Breitbart	408	9	9
Guardian	1,937	167	45
NYT	831	37	19
Total	3,176	213 (6.7%)	73 (2.3%)

Table 1. Number of articles published between July 9, 2018-2019 about climate change on three news sites. Only 6.7% of those article contain a data visualization and a subset of those enable online commenting as well (2.3% of all articles).

Two researchers read each of the news articles to categorize their data visualization(s), by noting the data sources (e.g., U.S. Energy Information Administration, National Oceanic and Atmospheric Administration), display types (e.g., bar chart, line graph, heat map), and visualized relationships (e.g., carbon-emissions by energy sector by year). The most frequent variables among the visualized relationships included carbon-emission, temperature change, and a categorical variable capturing various energy sectors (e.g., coal, natural gas, solar, wind).

We grouped the 73 articles that include a data visualization and enable online commenting into common discussion topics about climate change, which we identified by examining the most frequent variables among the visualized relationships. This analysis resulted in three discussion topics about climate change (i.e., CO_2, Energy, Weather), represented by twelve total articles that draw from each of the news sites. While each discussion topic includes just a few articles, they generated a combined 6,525 comments (Table 2).

	Climate change sub-topics			
	CO_2	Energy	Weather	Total
Breitbart	284	560	88	932
Guardian	2,217	1,075	98	3,390
NYT	127	253	1,823	2,203
Total	2,628	1,888	2,009	6,525

Table 2. Total comments posted between July 9, 2018-2019 to each news site on 12 articles and about the three climate change sub-topics. Number of comments on sub-topics vary from three news sites.

3.1.4 CO_2 Topic: Dietary choices and carbon-dioxide emissions. Articles published by the Guardian [13] and the NYT [73] present evidence that shifting away from meat and toward a plant-based diet could have a market effect on food production that reduces the carbon-footprint of this industry. By contrast, Delingpole [22] suggests that estimates of the effect of dietary choices on carbon emissions are overstated and do not account for a “rebound effect” associated with the increased purchasing power of people who do reduce their meat consumption.

3.1.5 Energy Topic: State-level transition towards renewable energy. Articles published by the NYT [81, 82] and the Guardian [30, 49] present data about how states have shifted towards energy sources that emit a lower level of greenhouse gasses than coal production (e.g., natural gas, biomass, solar). While states do generate energy, they also consume imported energy from other states and countries. An opinion piece published by Breitbart news [23] promotes investment in natural gas as a way to curb energy imports.

3.1.6 Weather Topic: Living with severe weather. Articles published by the NYT [79, 80] and the Guardian [18] present data visualizations that depict an increasing oceanic temperature and a higher frequency of high temperature days. Offering a counter-narrative, a Breitbart news opinion piece [21] presents evidence that ocean temperature and extreme weather events are not related to greenhouse gas emissions, but rather due to long-duration changes in the sea surface temperature.

3.2 Coding comments for the presence of data-centered talk

In order to investigate how people talk about specific sources of evidence, we hand-coded all comments from a collection of twelve articles drawn from the selected news sites and about specific topics related to climate change. The hand-coding was to identify whether comments speak to an article's topic (called "topic coherence" [70]), reference data, and reference specific data visualizations.

- *Topically coherent (Binary)*: The topic of the comment relates to the argument of the article.
- *References data (Binary)*: The comment includes data, such as numbers, relationships among variables, trends, or references to data sources and collection procedures.
- *References to data visualizations (Binary)*: The comment references the title, figure number, or data observations presented in a specific data visualization.

In our definition, the opposite of being topically coherent was not *incoherent*, but could include comments that introduce new ideas or information. However, comments that are not topically coherent inherently direct attention away from the evidence presented in a news article. We refer to *data-centered talk* as comments that are both topically coherent and that engage with the data, such as by referencing specific visualizations in an article or introducing new evidence to the discussion. For each article, the research team constructed a flow-chart of the main arguments in the article and used this artifact to determine the topic coherence of each comment to the topics structured in the article. Figure 1 presents the example of Holmes à Court [49] about how Australian states are shifting their energy consumption away from coal and towards a mix of renewable energy sources (e.g., wind power).

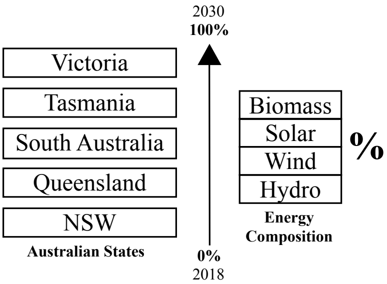

Fig. 1. Flowchart to summarize the main arguments in Holmes à Court [49], that the renewable energy composition of different Australian States is increasing.

Three researchers conducted the coding for each article. For articles with more than 200 comments, the team iteratively trained and tested on 10% samples until an acceptable inter-rater reliability was reached for each code (a Cohen's Kappa above 0.70). For articles with fewer than 200 comments, each researcher coded all of the comments independently and the team discussed each code to arrive at a consensus. In our report of RQ1, we present observations and inter-rater reliability for Topically Coherent and References Data (§4, Table 3).

3.3 Thematic analysis of data-centered talk

A range of analytical lenses have been applied in prior research about online discussion. Concepts from public deliberation are common in analyses of how people comment about civic issues at news websites (e.g., [7, 15, 40, 41, 68, 85, 96, 111]). During a deliberation people refer to their situated and domain knowledge as evidence to justify their opinions about a civic issue [83, 97]. Our research focuses on the specific case of how people comment about articles that present data analyses of a civic issue. When data scientists deliberate about a data analysis they justify their opinions by raising questions about the provenance [75, 77] and visual representation of data [51, 95]. As an initial scaffold for the thematic analysis, our research team found it useful to bridge these areas of literature by constructing the following types of evidence that people might reference when talking about data and civic issues.

- **Situated and Domain Knowledge**: During a discussion, people draw evidence from their own experiences living with an issue (called *situated knowledge*) as well as their own expertise (referred to as *domain knowledge*) [83]. Situated knowledge can elicit storytelling during a discussion [33], which can help participants to identify with an issue and with each other [5, 86]. Regardless, when new evidence is introduced to a discussion, other people need to familiarize themselves with the evidence, which may take additional time and training [88].
- **Data Provenance**: Forte et al. [39] characterize information about the provenance of data as "metadata that make it possible to trace the ownership or origin of an artifact or dataset" [39, pg. 2474]. Strategies to represent data provenance in a visualization include [51]: linking sources, citing methodological choices, as well as annotating exceptions and corrections. Presenting data provenance can promote the credibility of an analysis [39, 51]. As data scientists (and data science teams) continue to process a data set towards an analysis, each intermediary decision is also part of the provenance of the data [75, 77].
- **Visual Representation**: There are fundamental decisions about how to present data in digital media, such as the use of contrast, color, similarity, and proximity to emphasize data observations [84]. Such stylistic decisions can improve the accuracy of reading a data visualization [8]. Experiments that vary the presentation of data (e.g., tables, bar charts, scatter plots) demonstrate that visualization type significantly impacts their interpretation [87, 89]. These stylistic decisions can also influence reader behaviors [1]. For example, interactivity in a data visualization can nudge people to explore more of the visualizations interface [6].
- **Data Narrative**: Visualizations can help readers to discover insights about an issue that are hard to express in text; this process of discovery can also prompt readers to read further [27, 51]. Data journalists make choices about what data to present, how to visualize it, and in what order within an article [27, 62]. For example, Erete et al. [35] document how community-based non-profits make narrative choices about whether to communicate broad trends or a deep look into specific cases, through their selection of data and levels of analysis.

Three researchers independently read through the comments containing data-centered talk and constructed affinity diagrams to capture the ways that people refer to data in the discussion (N=714 comments). The team reviewed each others work to arrive at a consensus around the major and minor themes (§5, Table 4).

3.4 Automated classification of Data-Centered Talk

Computational methods are commonly applied in online discussion research to investigate the prevalence and influence that text can have on a discussion [2, 11]. To complement our thematic analysis, we fit a series of logistic regressions to examine the likelihood that comments contain data-centered talk based on factors related to the discussion context and comment content. While thematic analysis can offer a rich understanding of the communication behaviors that can emerge around data in a discussion, computational text analysis can also offer a glimpse into the potential for systems to classify these comments automatically. Such automated text classification systems are now necessary in the practice of moderating online discussions at news websites [36].

To control for factors related to the discussion context that may affect the presence of data-centered talk, we included categorical variables in the models to reflect differences among the news site platforms (i.e., NYT, Breitbart, Guardian) and topics (i.e., CO_2, Energy, Weather). There is little that we can say about how specific news site design and moderation strategies relate to the presence of data-centered talk, as each of the platforms are quite different and on many dimensions, and our sample of comments is drawn from just twelve news articles. Additionally, we carefully selected the news articles so as to consider topically similar discussions, but the articles differ in various ways: e.g., written by different authors, for different audiences, and their narratives depend on (some) different data sources.

We included several factors related to the comment content, which we constructed with standard information extraction methods, such as regular expression-based string matching, term frequency-inverse document frequency (TF-IDF), and text cosine similarity. These factors related to the comment content include hyperlink references as well as text similarity with the article, as adopting the language of an article can be an indication that comments are topically coherent [44, 98]. To account for longer comments being more likely to include data components than shorter comments, we included the log of the word count for each comment as a control variable.

Additionally, we incorporated the linguistic inquiry and word count dictionary (LIWC) [101] in our analysis as a way to consider the influence of keywords. As our qualitative analysis identified examples of how people make sense of data, by themselves and with others, we included indicators of *cognitive processes* from LIWC (e.g., causation, insight, tentative, certain, differentiation). Words associated with *numbers*, *quantifiers*, and *comparisons* were also included as a way to signal possible references to data. Public discussions about climate change also tend to break down over in-group/out-group tensions [58, 66], so we considered the extent to which *identity references* may factor into commenting about data. To this end, we measured expressions of identity as the proportion of personal and impersonal pronouns included in a comment.

The following control variables were included to account for comment length as well as external factors in the discussion context that may influence the presence of data-centered talk:

- **Platform**: i.e., NYT, Breitbart, Guardian.
- **Topic**: i.e., CO_2, Energy, Weather.
- **Word count**: (Logged) Number of words in each comment.

The following factors were included to evaluate the influence of specific textual characteristics on the likelihood that comments contain data-centered talk:

- **Hyperlinks**: (Binary) Regular expression-based string matching.
- **Article cosine similarity**: (Proportion) Calculated consine similarity between each comment and the article text weighted by TF-IDF.
- **Cognitive Processes**: (Proportion) e.g., cause, know, ought.
- **Numbers**: (Proportion) e.g., second, thousands.
- **Quantifiers**: (Proportion) e.g., few, many, much.
- **Comparisons**: (Proportion) e.g., greater, best, after.
- **Pronouns**: (Proportion) 1st, 2nd, & 3rd person pronouns, both singular and plural.

While using these variables to construct the logistic regressions, we paid attention to how the inclusion and exclusion of predictors affected each model's goodness-of-fit, using the Hosmer-Lemeshow test [50] and log likelihood ratio test as guides [50]. This procedure provides both a higher threshold than comparison with just the intercept and consistency in treatment of the control characteristics through the modeling process. To assess whether the model fit improved because of adding more parameters, we also calculated the Akaike Information Criterion (AIC) score which penalizes based on the number of estimated parameters. Additionally, we report the relative agreement between the predicted and observed data (Cohen's Kappa score) as well as McFadden's pseudo-R^2 to further evaluate the goodness-of-fit.

We present the regression coefficients in Table 5. The regression coefficient for a given variable is the estimate of the increase of an occurrence given a one unit increase in the variable holding the other predictor variables constant. For example, the coefficient of word count reported in the Final Model (Table 5) indicates that a one standard deviation change in word count is associated with a 1.25 times increase in the odds that a comment contains data-centered talk since $exp(.81) = 2.25$.

Evaluating the effects associated with each level of a categorical variable is not easy. We applied a Tukey-based pairwise comparison test of the estimated marginal means (EMMs) to evaluate each categorical variable post-hoc, e.g., platform, topic. While the coefficients reported for each model in Table 5 reflect the estimates associated with each predictor, where appropriate we also present the Tukey-based estimates in the text of our findings. Finally, ten-fold cross-validation was used to confirm the models were reasonably accurate at predicting the hand-coded data (above 90%).

4 RQ 1. HOW PREVALENT IS ONLINE DISCUSSION ABOUT DATA AT NEWS WEBSITES?

Out of 6,525 comments on articles with data visualizations about climate change, nearly a third of all posts were topically coherent with the articles (1,895 comments) and a third of those also reference data 10.9% (714 comments). A mere 2.4% of posts referenced at least one of the 26 data visualizations in the article collection (157 comments). The number of observations per article is reported in Table 3.

Our analysis identified differences in the prevalence of DCT and VCT among the articles, which may stem from differences in the presentation of data. Five of the articles include just one visualization (i.e., [13, 22, 79–81]), all of which elicited low levels of VCT (≤ 4%). Some articles include several variations on the same chart. For instance, Popovich [82] includes a percent stacked area chart for each U.S. State to depict how electricity generation has shifted over time, which observed the highest level of VCT (45%). Several articles mix multiple display types (e.g., bar chart, line graph), to present separate arguments associated with an article's narrative. Holmes à Court [49] presents trends in Australian energy generation through a third-party data visualization system called *Datawrapper*[5] which offers an interactive display and hyperlinks to data files; however, we note a low level of DCT (9%) and VCT (0.5%) in response to the article. These choices speak to the journalistic style of each author and each newspaper.

Each of the news articles reference data sources associated with the visualizations. We found that the references point towards academic journals, government and industry reports, public data portals, and popular blogs.[6] As we examined the data sources for each article, we noticed that many of the visualizations were not created by the newspaper. Five of the news articles include static visualizations that originally appear in other sources (i.e., [18, 21–23, 30]). As an example,

[5] https://www.datawrapper.de/

[6] Evidence reported throughout Delingpole [21] references the "Not a lot of people know that" blog, which has repeatedly been recognized by FactCheck.org as a source of misinformation about climate science [65]

		Topic coherent		Ref. data		DCT		VCT	
Citation	Comments	*Obs.*	*IRR*	*Obs.*	*IRR*	*Obs.*	*%*	*Obs.*	*%*
CO_2 Topic: *Dietary choices and carbon-dioxide emissions*									
Delingpole [22]	284	17	0.83	16	0.92	6	(2%)	0	(0.0%)
Carrington [13]	2,217	414	0.75	268	0.88	140	(6%)	32	(1%)
Moskin et al. [73]	127	84	-	49	-	41	(32%)	26	(20%)
Energy Topic: *State-level transition towards renewable energy*									
Delingpole [23]	560	48		44	0.86	15	(3%)	4	(0.7%)
Holmes à Court [49]	622	133	0.86	129	0.79	53	(9%)	3	(0.5%)
Edis [30]	453	110	0.94	79	0.89	35	(8%)	8	(2%)
Popovich [82]	92	85	-	57	-	56	(61%)	41	(45%)
Plumer [81]	161	128	-	88	-	81	(50%)	6	(4%)
Weather Topic: *Living with severe weather*									
Delingpole [21]	88	10	-	12	-	3	(3%)	0	(0.0%)
Cox [18]	98	23	-	15	-	14	(14%)	5	(5%)
Pierre-Louis [79]	1,189	666	0.74	255	0.80	187	(16%)	19	(2%)
Pierre-Louis [80]	634	177	0.79	143	0.82	83	(13%)	13	(2%)
Total	6,525	1,895		1,155		714	(11%)	157	(3%)

Table 3. Article and topic-level report of Observations and Cohen's Kappa score for inter-rater reliability (IRR). Articles with fewer than 200 comments do not have an IRR, as the codes were agreed to by consensus. We adopt the terms *data-centered talk* (DCT) to reflect the union of comments that are Topically Coherent and References Data as well as *visualization-centered talk* (VCT) to reflect the union of comments that are Topically Coherent and add References to specific data visualizations.

Cox [18] depict changes in surface temperature from 1910-2018 through a set of line and bar charts that appear to be copied from the biennial State of the Climate report (2018 release) produced by the Australian Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) [19]. All other articles include static graphics and interactive visualizations created by the newspaper (i.e., [49, 73, 82]). This distinction speaks to the different levels of editorial control that journalists may have over visualization design choices.

5 RQ 2. WHAT COMMUNICATION THEMES EMERGE IN ONLINE DISCUSSIONS ABOUT DATA JOURNALISM?

While data-centered talk is rare, in this section we examine what value these contributions add to online discussions about a civic issue. Our thematic analysis identified eleven ways that comments reference data to engage with the different types of evidence (i.e., situated knowledge, domain knowledge, data provenance, visual representation, data narrative). We reference each comment with unique identifiers that range from N1-N448 for comments to the NYT, G1-G242 for comments to the Guardian, and B1-B24 for Breitbart news comments.

5.1 How people use data to convey their situated and domain knowledge

Peck et al. [78] raise a general concern that there is limited research about how personal experience and interests motivate people towards or away from discussions about data. Such gaps between personal experience and evidence about a civic issue can widen existing political divides, as they have in discussions about climate change [58]. Our analysis found that when people are motivated by an issue enough to post a comment at a news site, they may share personal data and hypothetical

Category	Theme	Count
Situated Knowledge	1. Adds personal experience related to the narrative	119
Domain Knowledge	2. References external sources related to the narrative	423
Data Provenance	3. Identifies problems with the data collection	20
	4. Identifies missing elements in the analysis	7
Visual Representation	5. Offers feedback about the data presentation	13
	6. Difficulty reading the data (e.g., units, axes, titles, captions)	38
	7. Asks about specific observations (e.g., trends, outliers)	15
Data Narrative	8. Adds personal reflection about the data	17
	9. Identifies differences between the article and data	7
	10. Suggests data changes that might shift the narrative	18
	11. Interprets the data in support or opposition to the narrative	37
Total comments		**714**

Table 4. Number of comments clustered into eleven illustrative themes of data centered talk.

scenarios, and may pull in other sources (e.g., articles, reports, data) to substantiate their values and beliefs related to an issue.

Most of the commenters draw on situated (Theme 1) or domain knowledge as they engage with data (Theme 2). For example, people processed the visualized trends in temperature change by reflecting on events from their own lives "diving and snorkelling at various places around the world" (N242) or as an "older [person]," speaking from life experience (N252) or professional experience working in an affected industry, such as "automobile manufacturing" (N376) and "forestry" (N293). These comments demonstrate situated knowledge of a topic.

Comments also included hypothetical data scenarios constructed to support an argument. "Imagine sitting in a 100F apartment with disposable income and watching people like you in AC bliss. What would you do? Probably buy an AC [...] nearly a billion more people will have resources for these 1st world choices in 20-30 years" (N413). Such scenarios suggest causal relationships by drawing on personal and hypothetical data.

Comments that refer to domain knowledge often allude to external data, reports, and articles, but include few references. For example, comments cite nuclear power as contributing the largest share of France's energy grid—e.g., "75%" (N288), "80%" (N196), "81%" (N160), though we found only one comment that cites a data source. It may be feasible to fact-check the most frequently referenced external data in a discussion, but difficult to identify the long-tail of less common statistics that people reference in their comments.

The comments in Themes 1 and 2 engage with the narrative of an article, but do so by directing attention toward external data, whether personal, hypothetical, or reported elsewhere. Some comments include a hyperlink and a few words about what observations to look for in the reference. Several commenters also copied tab delimited data directly into the text of their comments. Many comments made claims about the data without citing a source.

5.2 How people fact-check sources and identify missing data

Fact-checking comments is tedious work for online discussion moderators [63], but at some social computing platforms, like Wikipedia, a community of people work collaboratively to fact-check content [102]. We did not fact-check each comment in the collection (e.g., hyperlink, statistic), but our analysis found that some commenters raised concerns about questionable claims and provide references in response.

Comments about data provenance add (new) external references to the discussion that may highlight variables, assumptions, and observations that might be missing (or intentionally excluded). In total, we found 27 comments that reference problems with the data collection (Theme 3) or missing elements in the analysis (Theme 4). Some responses offer a reference along with their correction.

> "A while back a fellow poster put me onto the Department of Environment and Energy and their yearly reports. Their information does not correspond to the graphs supplied here [link]" (G33).

Other responses note the error and offer data as evidence, but do not provide a hyperlink back to their reference for the correction.

> "One other note is a data error. Minnesota reached its renewable requirement of 25% last year which is 7 years before the deadline of 2025" (N69).

A few commenters also fact-check comments about the data provenance: "@[username] I looked at the paper cited by the Popular Mechanics article. They do NOT say to give up on renewables due to metals demand, instead they [...]" (N124). As another example, the following claims that a key variable is missing from a visualization in the article, "[w]hat the California graph doesn't show is an additional 8% of customer-based solar (and growing) that isn't counted in official records" (N37). In this instance, however, another commenter took a closer look at the caption and replied, "@[username], behind-the-meter (home) solar is included in the graph. See: [link]" (N38).

As we catalogued each data visualization, we did find a few that required additional searching to identify a (plausible) data source for the visualization. Commenters noted these reference errors as well, "I don't know where [author] got that figure - I can't see it in the source [they] cite, which doesn't deal in pie charts. We're closer to 30% not 3%, for renewables" (B2). Most comments in Themes 3 and 4 argue that: if there are flaws in the data collection or analysis procedures, then the narrative is flawed as well.

5.3 How people debate data visualization design choices

Choices about how to present a data analysis inherently suggest ways to interpret the analysis [1, 51, 95]. Our analysis found that commenters will critique visual design decisions, highlight misunderstandings, and will engage with each other to address these issues. This observation confirms prior findings that data-centered talk can yield insights for data journalists [52]. Additionally, social interactions among commenters can emerge in response to these comments.

Comments critique design choices about data visualizations (Theme 5). An important data visualization choice is what units to present. In response to Moskin et al. [73] about how dietary decisions affect carbon emissions in food production, some debate ensued around the units used to compare food groups: "OK, but who gets 50 grams of protein from cheese?" (N17) arguing that "comparing the protein/pound between plant and meat is a false equivalent" (N24), but others countered that "it is possible to get 50 grams of protein from plant foods" (N18). This example demonstrates how comments may reference specific elements of a data visualization as they engage with an article narrative.

Some comments also highlight points of misunderstanding in a visualization (Theme 6). Labels can help people to interpret units and other visual elements. Commenters to Carrington [13] grappled with how to interpret a chart about the food groups people need to eat less of to keep global temperatures under 2C by 2050. A strong contingent of commenters were bothered by an annotation on the chart (27 comments): "UK citizens will need to eat nine times less pork." Numerous commenters responded with criticism of the unfortunate phrasing, "WTF is nine times less?" (G1), "why not say one ninth" (G4) or more practically, "[i]f you eat beef every day, drop that down to every nine days" (G7). As evident in this case, stylistic decisions in a visualization can distract attention away from a narrative.

People also help each other to make sense of unexpected elements in a data visualization (Theme 7). In response to a visualization showing oceanic temperature change from 1940-2018, a commenter observed that "[t]he curve for the deeper ocean warming is steeper. Intuitively I would have expected the opposite. Are the descriptive captions for each curve reversed?" (N74) Another commenter replied with a hyperlink to the European Environment Agency's heat content report, "I expected what you did, and I'm not sure why it's different, but the trend is sharper when the depths to 2000 meters are included" (N75). This reply adds domain knowledge to the discussion that might help others to make sense of the data and help the author to improve the article.

Choices about what to measure may also appear to present a biased narrative. In response to Pierre-Louis [79], which is a NYT article about global greenhouse gas emissions, many commenters took issue with a decision to list countries in a static chart by total change in carbon-dioxide emissions instead of per-capita emissions. Commenters argued that, "properly measured, that is calibrated relative to population, China is NOT the world's worst polluter" (N96) and "in the United States we emit 16.5 metric tons per person per year compared to a worldwide average of about five metric tons" (N307). An enhanced discussion forum could add an interactive view of this data, enabling people to toggle between carbon-dioxide emissions measured in the aggregate or per-capita (e.g., *CommentSpace* [109]), but it is less clear whether this design change would elicit meaningful discussion or more entrenched polarization.

A few comments offer feedback about the data presentation at an article. "[F]ascinating charts [...] not sure they graphically need to swap position to maintain order of overall production by year, might be clearer and eas[ier] to assess if positions remained constant as they grew or dwindled" (N56). However, the feedback often amounted to simple praise or grief: "literally the worst graph ever made, but I wholeheartedly agree" (G14).

5.4 How people refer to visualizations to engage with the narrative

In the context of a newspaper article, each visualization contributes to a central argument about how readers should interpret the data presented in the article [32, 107]. We found instances where commenters recognize themselves in a data visualization and refer to specific observations to share their personal narrative of an issue. Our analysis also found comments that interrogate the data journalist's visualization choices, by raising questions about reporting bias and offering alternative narratives from the data.

Commenters use data visualizations as a point of personal reflection guided by observations and trends presented in the data (Theme 8).

> "Fascinating, but locally very disappointing. I was of course unhappy to see that my state—with 300 days/yr of often unmitigated sunshine obtains only 5% from solar, while neighboring AZ gets only 6, just matching that figure from MA, where I lived for 15+ years, and which sees lots of gray cloudy skies" (N55).

Unlike situated knowledge that is generally related to the article topics and data (Theme 1), these personal stories are prompted by specific data points presented in the reporting.

A few commenters challenge the narratives in articles by identifying discrepancies between an article's text and visualizations (Theme 9). Some call attention to possible *cherry-picking* from the observations, "[w]hen I look at the chart, the two biggest impacts are beef and farmed crustaceans; however, in the text the authors only mention beef and dairy. Why?" (N13) Other comments suggest that the text overstates or misleads.

> "the overlay says natural gas has edged out coal (in bold) which isn't what you see in the graph [...] a number of small states have dramatic shifts in their energy mix in short period of times. With small user bases, shifting just one big power plant changes the profile of the whole state" (N67).

Commenters also suggest data changes that might shift the narrative (Theme 10). For example, articles in the Energy topic present the electricity *generated* by energy source [23, 49, 81] and by state [82] or region [30], but comments raised that a better comparison might be *consumed* energy by source (e.g., coal, natural gas, solar). The examples suggest that the production, export, and import of energy by source may offer a different story about state-level commitment to renewable energy (N63; G33). While similar to comments that call attention to issues related to the data provenance (Themes 3-4), these examples show how a missing variable in the visualization may shift the narrative in new directions.

Lastly, some of the data-centered talk comments synthesize observations about data to seemingly motivate a call-to-action around a civic problem or possible solution (Theme 11).

> "What was most interesting was that many states use resources that are readily available, with one exception. States with coal use that. States benefiting from mountain snow melt into rushing rivers use hydroelectric power. States in that massive wind tunnel in the middle of the country are increasingly making use of wind power. The one exception is solar power in the sunniest states. We'll hope that is changing" (N64).

In this example, the comment calls attention to a perceived under-utilization of solar power depicted in a particular data visualization (i.e., [82]). The comment was *Recommended* by other readers nine times and was highlighted in the NYT interface as among the *Reader Picks*. It is less clear how many times this comment was shared via the social media tools embedded within the commenting system. Commenters also use data to call for inaction. "3% renewables, and all these idiot climate changers want everyone to get rid of hydrocarbons and get electric cars! What a sick joke" (B1).

A data visualization can bring to light factors that underlie a civic problem. Our analysis finds that the discussion surrounding a data visualization at news sites may also call attention to issues beyond what is visualized, to how choices about the data collection, analysis, and visual representation play into the narrative.

6 RQ 3. WHAT PREDICTIVE FACTORS HELP TO AUTOMATICALLY RECOGNIZE COMMENTS ABOUT DATA?

When people comment on the data in a news article they may share their experiences, ideas, and feelings about an issue, as evident in the eleven themes in data-centered talk (RQ2). Our analysis also identified instances of social interaction around the ways data is visually represented; as people help each other to fact-check and make sense of an analysis, or when they gleefully gang up on a poorly phrased data label (e.g., "WTF is nine times less?"). This observation suggests that commenting not only offers critique for a data journalist [52], but these comments also help to further the conversation about an article.

While potentially valuable opportunities for informed discussion about data presented in an article, instances of data-centered talk are particularly rare (RQ1). In face-to-face settings, a facilitator might help move a civic discussion about data forward using well-crafted conversation prompts [72, 74], but such facilitation is less feasible in online discussions [71]. To explore how discussion systems could be improved through text classification methods, we investigated automated methods to filter through a noisy online discussion for instances of data-centered talk (RQ3). Such techniques are commonly used to remove abusive content in social media [57] and at news sites [36]. Such intelligent filtering for data-centered talk may offer moderators (and readers) new ways to highlight, fact-check, and facilitate these comments.

	Baseline	Final
Variables	Coefficient (SE)	Coefficient (SE)
Control Variables		
(Intercept)	-5.96(0.25) ***	-6.05 (0.32) ***
Word Count	1.01(0.05) ***	0.81 (0.06) ***
Discussion Context		
Platform-Breitbart	-2.51(0.24) ***	-2.31 (0.25) ***
Platform-Guardian	-1.72(0.13) ***	-1.91 (0.14) ***
Topic-Energy	1.43(0.13) ***	1.37 (0.15) ***
Topic - CO2	1.40(0.15) ***	1.46 (0.16) ***
Sourcing		
Article Cosine Similarity		0.08 (0.01) ***
Hyperlink		1.02 (0.16) ***
Cognitive Processes		
Cognitive processes		-0.36 (0.85)
Quantifiers		3.07 (1.57) .
Comparisons		5.01 (1.45) ***
Number		16.92 (1.15) ***
Identity References		
1st Person Singular		-2.58 (2.15)
1st Person Plural		-4.39 (2.31) .
2nd Person Pronouns		-4.39 (2.00) *
3rd Person Singular		-17.05 (7.65) *
3rd Person Plural		-12.20 (3.42)
Impersonal Pronouns		-0.72 (1.40)
Model performance		
Cohen's Kappa	0.1819	0.3416
Log Likelihood	-1807	-1578
Pseudo-R^2	0.1981	0.2996
Accuracy	0.8947	0.9036
Model selection		
AIC	3625.4	3192.2
Goodness-of-fit	$\chi^2(df = 12, N = 6,525) = 457.22^{***}$	

Table 5. Logistic regression model on comments of data-centered talk. As *Platform* and *Topic* are categorical variables, the coefficients reflect comparisons with the NYT and Weather topic, which are included in the model intercept. To evaluate the performance of each model, the table presents the Cohen's Kappa score, McFadden's pseudo R^2, and model accuracy based on ten-fold cross-validation. To compare the goodness-of-fit of the final model, we ran the log likelihood ratio test and report the χ^2 statistic. The log likelihood, AIC score, and χ^2 statistic all indicate that incorporating features related to the comment text significantly improve the model fit. In the final model, the following textual features have a significant effect on the likelihood that a comment contains data-centered talk: Cosine similarity, Hyperlinks, Comparisons, Numbers. p-value significance codes: 0.0001 '***', 0.001 '**', 0.01 '*', 0.05 '.'

6.1 The model provides a reasonably accurate classification (with caveats)

As described in the Method section (Section 3.4), the analysis involved an initial Baseline model to predict the likelihood that comments contain data-centered talk, while controlling for several factors, such as the news site platform, discussion topic, and comment length (logged word count). The Baseline model was used as a point of comparison with a Final model (see Table 5), which we expanded from the Baseline to include a series of factors related to a comment's content (e.g., article text cosine similarity, inclusion of hyperlinks, comparison terms).

Ten-fold cross-validation was used to confirm the models were reasonably accurate at classifying comments that contain data-centered talk (above 90%). While the gain in accuracy between the Baseline and Final model is relatively small ∼1%, we find that the Final model has a greater agreement with the observed data (Cohen's Kappa) and offers a better fit to the data as indicated in the pseudo-R^2 and the AIC score, which does include a penalty term to account for complex models (those with more parameters) being more likely to over fit the training data by comparison to a simple model (with fewer parameters). The log likelihood ratio test χ^2 statistic further confirms that the Final model offers a significantly better fit to the data, by comparison to the Baseline.

However, a limitation of our binary classifier is that it tends to yield more false negatives, as our prediction of data-centered talk is less accurate (63.8%) than our prediction of comments that do not include data-centered talk (91.7%). This is due to the small proportion of comments coded as including data-centered talk in our training set (714 of 6,525 or 10.9%), relative to the volume of comments that are not topically coherent and reference data (5,811 of 6,525 or 89.1%).

These results suggest cautious optimism about the potential for systems that can automatically detect comments that contain data-centered talk. Our analysis found several text features that increase the likelihood of data-centered talk being present in a comment, but future analyses should consider larger data collections as well as more robust statistical approaches, such as bootstrapping, to improve the model prediction by sub-sampling. The following review the most promising predictive features, which add nuance to the thematic analysis of data-centered talk (RQ2).

6.2 Predictive factors that help to identify data-centered talk in comments

6.2.1 Platform and topic affect the likelihood of data-centered talk. The news sites selected for this study differ in terms of their system design, moderation practices, and reporting about climate change. Our analysis found that the news sites are not equal in their levels of data-centered talk. Using a Tukey-based pairwise comparison in the Final model we found that posts to Breitbart news and the Guardian were 0.10 and 0.15 times likely to contain data-centered talk than posts to the NYT. These differences are important to account for in analysis, but it is less clear what factors related to the platforms may have caused this difference.

6.2.2 The discussion topics also elicited unequal levels of data-centered talk. Articles about the topic "living with the new reality of global warming" (Weather) were less likely to elicit posts about data-centered talk than other topics. Using a Tukey-based pairwise comparison in the Final model we found that posts to the topic "state-level shifts toward green energy" (Energy) were 2.93 times more likely to include data-centered talk than posts to the Weather topic and posts to "dietary contributions to carbon emission" (CO2) were 3.30 times more likely. We caution against reading too deeply into this observation, as the topics not only differ in terms of their content, but also in data sources, visualization, and interactivity. The selected articles also differ in length, publish date, and author among other factors.

6.2.3 External links can indicate that the data is under discussion. Similar to our hand-coding of topically coherent comments, we calculated a comment's cosine similarity with terms in the article weighted by their TF-IDF to consider the extent to which a comment speaks to information within an article. Our analysis found that a one standard deviation increase in cosine similarity increases the odds that a comment is about relevant data by 8.3% in the Final model. As many of the communication behaviors presented in section 5 revolved around external sources of evidence, it is not surprising to see that the odds of comments which include a hyperlink were 1.77 times more likely than the comments without a hyperlink to contain data-centered talk since the exponential of the coefficient of Hyperlink is $exp(1.02) = 2.77$.

6.2.4 Specific language can be predictive of data-centered talk. The commenting behaviors captured by the themes in data-centered talk reflect a range of cognitive processes. Our quantitative analysis included several predictors based on cognitive indicators in LIWC, such as words commonly associated with causation, insight, and certainty. Our analysis also found that a one standard deviation increase in the volume of words commonly associated with the cognitive process of comparison, words like "greater and least", significantly increased the odds of a comment containing data-centered talk. Additionally, we examined the presence of numbers, quantifiers, and terms commonly related to quantitative comparisons. The volume of number words, like "twenty and thousand", increased the odds that a comment was about data-centered talk significantly.

Comments were also less likely to contain data-centered talk when they include identity references, which we operationalized as the volume of personal pronouns. Specifically, comments containing more 2nd person (e.g., you, your) and 3rd person plural pronouns (e.g., they, them) were marginally significant in our statistical models.

7 DISCUSSION AND LIMITATIONS

7.1 What value might designing for data-centered talk yield?

Our analysis indicates that data-centered talk is rare at news websites, yet offers a valuable window into how people engage with each other around data journalism. Comments that contain data-centered talk are somewhat similar linguistically to those that Arguello et al. [2] found generally likely to receive a response in online discussion. Our analysis of RQ3 finds that the discussion context matters (i.e., platform, topic) and that comments containing data-centered talk tend to be topically coherent with the article; these findings echo those of Arguello et al. [2]. However, data-centered talk is distinct in that hyperlinks, comparison terms, and numbers are more prevalent, while cognitive processes and identity references play a less pronounced role than they do in comments that are generally likely to generate a response.

The eleven themes (RQ2) offer illustrative evidence of the range of responses that a discussion about data can generate (Table 4). People use data as a prompt to share personal stories and reference data to argue a point. People also raise questions that help to clarify and generate insights about data: e.g., Where did this data come from and how did it get here? How do the rhetorical elements of a visualization guide its interpretation? What story does the visualization present and are there alternative narratives? By promoting such mutual understanding, data might become a valuable focus for informed discussion about civic issues. Due to our limited sample of comments, we were unable to explore these social behaviors computationally (e.g., [2, 10, 11]), but hope that future research will investigate what social roles people may adopt and develop over time as they continue to read data journalism and follow specific topics in the news [100].

Our qualitative analysis surfaced a rich discourse around data and we think that the themes might also serve as a guide for system designers to think about facilitating deliberation about data analysis in online settings. In this paper, each theme was presented independently, but our interpretation of the findings is that there are inter-dependencies among the themes. For example, a data-discussion facilitator might first direct participants to consider questions about the data collection process (Theme 3), and then gradually move the conversation towards clarifying how to read the visualizations (Theme 6), so that participants have a shared sense of how the data was collected and represented, in order to promote informed discussion about data analyses presented in a news article (Theme 10). As discussion participants explore and generate new questions, a skilled facilitator might recognize when to pause the discussion, so that participants can re-calibrate their shared understanding, perhaps by revisiting a theme in data-centered talk.

A natural next step might be to develop ways of eliciting data-discussion based on the themes and then practice facilitating them in face-to-face contexts. The contexts for such evaluation might build on existing research about how people experience data in daily life [47, 59, 78] and how data scientists deliberate about an analysis [77]. Through practice facilitating data-discussions in face-to-face settings, we might recognize new facilitation strategies, evaluation criteria, and system designs that we could then apply to promote informed online discussion about civic issues.

7.2 Design considerations for data-centered talk

We argue that data-centered talk offers a valuable opportunity for facilitating online civic discussion, but our analysis of RQ1 clearly indicates that the opportunities for data-centered talk are rare. Just 73 articles in 3,176 (2.3%) include both data visualizations and enable online commenting. System designers can propose many novel ways to promote data-centered talk at news sites, but whether to enable or disable commenting about an article is typically a newsroom decision.

Newsrooms weigh numerous considerations when determining how to present data [32, 107] and (whether to) support participatory journalism [24, 25, 46]. To illustrate this decision process, consider the following system design idea that extends from our finding that data-centered talk often includes questions about data provenance and visual choices:

- *Facilitate data-centered Q&A around an article*: Inspired by the *Storia* [60] workflow for synthesizing social media posts into a narrative summary of an event, we might design a system that simply extracts open data-centered questions from a discussion and lists each alongside an article, so that other readers can offer answers.

For discussion newcomers, responding to a data-centered question might offer an easy way into the conversation [71]; experts might regularly peruse these lists for opportunities to be useful, as in *Stack Overflow*, and regular commenters could get recommended questions based on their commenting history on specific topics or data [17]. While increasing the salience of data-centered questions might foster discussion, doing so may have adverse consequences for a journalist, by making it easier for commenters to cultivate uncertainty about the reporting [25, 46], and easier for data-centered disinformation campaigns to generate attention [92].

CSCW research is often tangled in what Jackson et al. [55] refer to as a *policy knot*: we see opportunities for system design to advance research and collaboration, but find that there are unintended consequences in practice, and policy considerations that affect and are affected by system design choices. As newsrooms determine their approach to participatory data journalism, existing CSCW research offers a number of system design ideas to weigh along with other considerations, such as journalist reputation and moderation practices. The following are a few examples of designing for data-centered talk to promote various outcomes, which we offer for consideration by newsrooms as well as the third-party hosting services that enable online discussion about news articles (e.g., Disqus, Datawrapper):

- *Promote statistical literacy among readers.* Much of data-centered talk is asking and answering questions about how to interpret a data visualization. Recognizing that many people do not have access to the skills necessary to make sense of data, some newsrooms have opted to promote statistical literacy by hosting expert facilitated discussions about data journalism

[43]; however, these efforts depend on human moderators, who have a limited bandwidth to respond to comments [33]. *CrowdSCIM* [105] demonstrates how novice crowds can learn expert skills, such as historical thinking, by performing scaffolded micro-tasks. To help readers to develop these skills without direct attention from a moderator, a news site might integrate the task scaffolding techniques applied by CrowdSCIM with existing statistical literacy training materials (e.g., [4, 38]).

- *Generate public insights for data science teams.* Recall that many of the visualizations of climate change data were not created in the newsroom, but reproduced from other sources (e.g., academic journals, government organizations). Rather than promote data-centered talk between readers and journalists, a newsroom might opt to facilitate data-centered talk between readers and the data science teams that produced the analysis. Newsrooms might offer programmatic access to data-centered talk around specific data sources, such as the Global Carbon Budget[7] [42], so that data scientists can integrate public discussion about data directly into their analysis pipeline, as discussed in Willett et al. [108].
- *Foster community around shared experiences of civic issues.* In our analysis, the largest theme in data-centered talk was using data to share situated knowledge, e.g., personal stories. Data journalism about a civic issue can elicit deep feelings for people who live with these issues in their every day. A newsroom may view its role in society as providing an online public square for people to find each other and to form community through data journalism. CSCW systems such as *BudgetMap* [61] and *Hollaback* [26] offer compelling examples of how a thoughtful approach to aggregating personal experiences from a community can influence public decision-making and can offer people steps toward recovery from a trauma.

Data visualizations can help readers to develop insights about a civic issue that are harder to express in text, and that process of discovery can motivate readers to read further [27, 51]. In this paper, we examine the online discussions about data that can follow from data journalism. As presented, data-centered talk is an example of the rare, yet valuable instances of social interaction that can emerge even amidst noisy online discussions at news sites. While we have offered several possible ways that news sites might promote data-centered talk through system design and facilitation, our analysis highlights key differences among news sites that stem from newsroom choices about how to present data [32, 107] and whether to host an online discussion [24, 25, 46].

7.3 Limitations

The research has several key limitations. We opted to focus on specific topics related to climate change, diving deeply into the commenting and context surrounding twelve articles (6,525 comments in total), rather than sampling broadly from a range of topics (e.g., [52]). While we selected the articles to reflect common discussion topics (e.g., CO^2, Energy, Weather), the study did not deploy a controlled experiment. For this reason, throughout the article we note important similarities and differences among the news sites, discussion topics, and data visualizations included in the collection as a case study.

We also have limited information about the people involved with the discussion; those who posted, shared, and recommended comments, but also those who moderate the discussion. For example, moderators play a pronounced role in what content is prioritized in an online discussion. Each of the news sites employ different moderation practices and systems, but we do not know the content and volume of comments actually removed by moderators at the news site. For that reason,

[7] Global Carbon Project data archive: https://www.globalcarbonproject.org/carbonbudget/archive.htm

our analysis is limited to the comments that were published on the collection date, some of which may have since been removed.

There are also political operators who wage strategic information operations against online discussions about civic issues [92], perhaps especially around issues of climate change [66]. Recent CSCW research about strategic information operations demonstrate how political operators collaborate to promote specific narratives by raising uncertainty about presented facts and discrediting information providers (e.g., journalists, moderators) [92, 94]. Political operators also create content to contest evidence about civic issues [93]. In our case, comments that denounce journalistic decisions or that raise uncertainty about an analysis, may reflect such political operations, but we do not know to what extent any such operations may have influenced our findings.

8 CONCLUSION

We analyzed 6,525 comments across three different news sites for the presence of on-topic comments that reference embedded data and data visualizations. Such data-centered talk (DCT) was rare, only occurring in 10.9% of comments, but potentially quite valuable. Eleven different themes across four main categories in situated and domain knowledge, data provenance, visual representation and data narratives, reveal how people discuss data visualizations. These instances of DCT can serve as design inspiration for discussion system designers, as they point to directions for data-centered moderation tools and discussion prompts, and offer a source of feedback for data journalists. We demonstrate how automated text classification methods can help to surface comments that contain DCT (90.36% accuracy), which could be used in moderation systems to help facilitate meaningful conversation about data. Finally, we discuss potential ways that newsrooms might apply this analysis to promote data literacy, data science, and to foster community around shared experiences.

ACKNOWLEDGMENTS

We want to thank Yiwen Hou, Aneesh Kashalikar, and Ziwen Zeng for valuable assistance throughout the research process. The paper improved tremendously through the review cycle and is substantially stronger thanks in no small part to the thoughtful contributions of expert reviewers. This research was funded by the National Science Foundation, grants #1122206 and #2009003. Additionally, we want to acknowledge Cliff Lampe for asking a thoughtful question at CSCW 2018 about “literacy barriers in civic discussion” that prompted our direction into this research.

REFERENCES

[1] Jessica S. Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B. Starren. 2006. Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association* 13, 6 (11 2006), 608–618. https://doi.org/10.1197/jamia.M2115

[2] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 959–968.

[3] W. Lance Bennett and Steven Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication* 33, 2 (4 2018), 122–139. https://doi.org/10.1177/0267323118760317

[4] Rahul Bhargava and Catherine D'Ignazio. 2015. Designing Tools and Activities for Data Literacy Learners. In *Web Science: Data Literacy Workshop*. MIT Media Lab, Oxford, UK. https://www.media.mit.edu/publications/designing-tools-and-activities-for-data-literacy-learners/

[5] Laura W. Black. 2008. Deliberation, Storytelling, and Dialogic Moments. *Communication Theory* 18, 1 (1 2008), 93–116. https://doi.org/10.1111/j.1468-2885.2007.00315.x

[6] Tanja Blascheck, Lindsay Macdonald Vermeulen, Jo Vermeulen, Charles Perin, Wesley Willett, Thomas Ertl, and Sheelagh Carpendale. 2019. Exploration strategies for discovery of interactivity in visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 2 (2 2019), 1407–1420. https://doi.org/10.1109/TVCG.2018.2802520

[7] Robin Blom, Serena Carpenter, Brian J. Bowe, and Ryan Lange. 2014. Frequent Contributors Within U.S. Newspaper Comment Forums. *American Behavioral Scientist* 58, 10 (9 2014), 1314–1328. https://doi.org/10.1177/0002764214527094

[8] Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean Daniel Fekete. 2014. A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (12 2014), 1963–1972. https://doi.org/10.1109/TVCG.2014.2346984

[9] Cameron Brick and Sander van der Linden. 2018. Yawning at the apocalypse. *The Psychologist* 5, 1 (2018), 30–35.

[10] Moira Burke and Robert E. Kraut. 2008. Mind your Ps and Qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 281–284.

[11] Moira Burke, Robert E. Kraut, and Elisabeth Joyce. 2010. Membership Claims and Requests: Conversation-Level Newcomer Socialization Strategies in Online Groups. *Small Group Research* 41, 1 (2 2010), 4–40. https://doi.org/10.1177/1046496409351936

[12] Sean Captain. 2017. Disqus Grapples With Hosting Toxic Comments On Breitbart And Extreme-Right Sites. https://www.fastcompany.com/3068698/disqus-grapples-with-hosting-toxic-comments-on-breitbart-and-extreme-right-sites

[13] Damian Carrington. 2018. Huge reduction in meat-eating essential to avoid climate breakdown. https://www.theguardian.com/environment/2018/oct/10/huge-reduction-in-meat-eating-essential-to-avoid-climate-breakdown

[14] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. *arXiv preprint arXiv:1504.00680* (2015).

[15] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4 (8 2014), 658–679. https://doi.org/10.1111/jcom.12104

[16] Stephen Coleman and John Gøtze. 2001. *Bowling together: Online public engagement in policy deliberation.* Hansard Society London.

[17] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: Using intelligent task routing to help people find work in wikipedia. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. 32–41. https://doi.org/10.1145/1216295.1216309

[18] Lisa Cox. 2018. Australia experiencing more heat, longer fire seasons and rising oceans. https://www.theguardian.com/environment/2018/dec/20/australia-experiencing-more-heat-longer-fire-seasons-and-rising-oceans

[19] CSIRO and Bureau of Meteorology. 2018. *State of the Climate 2018.* Technical Report. CSIRO, Canberra, Australia. 24 pages. https://doi.org/978-1-925315-97-4

[20] Richard Davis. 1999. *The web of politics: The Internet's impact on the American political system.* Oxford University Press.

[21] James Delingpole. 2018. Delingpole: No, Trump's Red State Base Is Not 'Suffering Most' from Climate Change. https://www.breitbart.com/politics/2018/09/03/delingpole-no-trumps-red-state-base-is-not-suffering-most-from-climate-change/

[22] James Delingpole. 2018. DELINGPOLE: No, Vegetarianism Won't Save the World from 'Climate Change'. https://www.breitbart.com/europe/2018/11/03/delingpole-vegetarianism-wont-save-the-world-from-climate-change/

[23] James Delingpole. 2019. Delingpole: May's Fake Tory Government Caves to Anti-Fracking Loons. https://www.breitbart.com/europe/2019/04/28/delingpole-mays-fake-tory-government-caves-anti-fracking-loons/

[24] Mark Deuze, Axel Bruns, and Christoph Neuberger. 2007. Preparing for an age of participatory news. *Journalism Practice* 1, 3 (2007), 322–338. https://doi.org/10.1080/17512780701504864

[25] Nicholas Diakopoulos and Mor Naaman. 2011. Topicality, time, and sentiment in online news comments. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 1405–1410.

[26] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. 2013. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 477–490.

[27] Graham Dove and Sara Jones. 2012. Narrative Visualization: Sharing Insights into Complex Data. In *Interfaces and Human Computer Interaction (IHCI)*. Lisbon, Portugal. http://openaccess.city.ac.uk/1134/

[28] Steve M. Easterbrook, Eevi E. Beck, James S. Goodlet, Lydia Plowman, Mike Sharples, and Charles C. Wood. 1993. A survey of empirical studies of conflict. In *CSCW: Cooperation or Conflict?* Springer, 1–68.

[29] O. Edenhofer, R. Pichs-Madruga, Y. Sokona, S. Agrawala, I.A. Bashmakov, G. Blanco, J. Broome, T. Bruckner, S. Brunner, M. Bustamante, L. Clarke, F. Creutzig, S. Dhakal, N.K. Dubash, P. Eickemeier, E. Farahani, M. Fischedick, M. Fleurbaey, R. Gerlagh, L. Gomez Echeverri, S. Gupta, J. Harnisch, K. Jiang, S. Kadner, S. Kartha, S. Klasen, C. Kolstad, V. Krey, H. C. Kunreuther, O. Lucon, O. Masera, J. Minx, Y. Mulugetta, T. Patt, N.H. Ravindranath, K. Riahi, J. Roy, R. Schaeffer, S. Schlömer, K. Seto, K. Seyboth, R. Sims, J. Skea, P. Smith, E. Somanathan, R. Stavins, C. von Stechow, T. Sterner, T. Sugiyama, S. Suh, K.C. Urama, D. Ürge-Vorsatz, D. Victor, D. Zhou, J. Zou, and T. Zwickel. 2014. *Summary for policymakers.* Technical Report. Intergovernmental Panel on Climate Change.

[30] Tristan Edis. 2019. While the government is in denial, the states are making staggering progress on renewable energy.

[31] Justin Ellis. 2015. What happened after 7 news sites got rid of reader comments » Nieman Journalism Lab.

[32] Martin Engebretsen, Helen Kennedy, and Wibke Weber. 2018. Data Visualization in Scandinavian Newsrooms Emerging Trends in Journalistic Visualization Practices. *Emerging trends in journalistic visualization practice* 39 (2018), 3–18. https://doi.org/10.2478/nor-2018-0007

[33] Dima Epstein, Cynthia Farina, and Josiah Heidt. 2014. The value of words: Narrative as evidence in policy making. *Evidence & Policy: A Journal of Research, Debate, and Practice* 10, 2 (2014), 243–258.

[34] Dima Epstein and Gilly Leshed. 2016. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Public Deliberation* 12, 1 (2016).

[35] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Fassett, and Sarah Duda. 2016. Storytelling with Data: Examining the use of data by Non-Profit organizations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, Vol. 27. Association for Computing Machinery, 1273–1283. https://doi.org/10.1145/2818048.2820068

[36] Bassey Etim. 2017. The Times Sharply Increases Articles Open for Comments, Using Google's Technology. https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

[37] Claudia Flores-Saviaga, Brian C. Keegan, and Saiph Savage. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, Palo Alto, California.

[38] Kristin Fontichiaro, Jo Angela Oehrli, and Amy Lennex. 2017. *Creating Data Literate Students.*

[39] Andrea Forte, Nazanin Andalibi, Thomas Park, and Heather Willever-Farr. 2014. Designing information savvy societies: An introduction to assessability. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2471–2480. https://doi.org/10.1145/2556288.2557072

[40] Rolf Fredheim, Alfred Moore, and John Naughton. 2015. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference*. ACM, Oxford, United Kingdom, 11. https://doi.org/10.1145/2786451.2786459

[41] Deen Freelon. 2015. Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society* 17, 5 (2015), 772–791.

[42] Pierre Friedlingstein, Matthew W. Jones, Michael O'Sullivan, Robbie M. Andrew, Judith Hauck, Glen P. Peters, Wouter Peters, Julia Pongratz, Stephen Sitch, Corinne Le Quéré, Orothee C.E. DBakker, Josep G. Canadell1, Philippe Ciais1, Robert B. Jackson, Peter Anthoni1, Leticia Barbero, Ana Bastos, Vladislav Bastrikov, Meike Becker, Laurent Bopp, Erik Buitenhuis, Naveen Chandra, Frédéric Chevallier, Louise P. Chini, Kim I. Currie, Richard A. Feely, Marion Gehlen, Dennis Gilfillan, Thanos Gkritzalis, Daniel S. Goll, Nicolas Gruber, Sören Gutekunst, Ian Harris, Vanessa Haverd, Richard A. Houghton, George Hurtt, Tatiana Ilyina, Atul K. Jain, Emilie Joetzjer, Jed O. Kaplan, Etsushi Kato, Kees Klein Goldewijk, Jan Ivar Korsbakken, Peter Landschützer, Siv K. Lauvset, Nathalie Lefèvre, Andrew Lenton, Sebastian Lienert, Danica Lombardozzi, Gregg Marland, Patrick C. McGuire, Joe R. Melton, Nicolas Metzl, David R. Munro, Julia E.M.S. Nabel, Shin Ichiro Nakaoka, Craig Neill, Abdirahman M. Omar, Tsuneo Ono, Anna Peregon, Denis Pierrot, Benjamin Poulter, Gregor Rehder, Laure Resplandy, Eddy Robertson, Christian Rödenbeck, Roland Séférian, Jörg Schwinger, Naomi Smith, Pieter P. Tans, Hanqin Tian, Bronte Tilbrook, Francesco N. Tubiello, Guido R. Van Der Werf, Andrew J. Wiltshire, and Sönke Zaehle. 2019. Global carbon budget 2019. *Earth System Science Data* 11, 4 (12 2019), 1783–1838. https://doi.org/10.5194/essd-11-1783-2019

[43] Michael Gonchar and Katherine Schulten. 2017. Announcing a New Monthly Feature: What's Going On in This Graph? https://www.nytimes.com/2017/09/06/learning/announcing-a-new-monthly-feature-whats-going-on-in-this-graph.html

[44] Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* 29, 3 (2013), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

[45] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. 2009. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Commun. ACM* 52, 1 (1 2009), 87–97. https://doi.org/10.1145/1435417.1435439

[46] Alfred Hermida and Neil Thurman. 2008. A Clash of Cultures. *Journalism Practice* 2, 3 (10 2008), 343–356. https://doi.org/10.1080/17512780802054538

[47] Rosemary Lucy Hill, Helen Kennedy, and Ysabel Gerrard. 2016. Visualizing Junk: Big Data Visualizations and the Need for Feminist Data Studies. *Journal of Communication Inquiry* 40, 4 (10 2016), 331–350. https://doi.org/10.1177/0196859916666041

[48] Sanne Hille and Piet Bakker. 2014. Engaging the Social News User. *Journalism Practice* 8, 5 (9 2014), 563–572. https://doi.org/10.1080/17512786.2014.899758

[49] Simon Holmes à Court. 2018. In 2018 the Australian government chased its energy tail. Here's a more hopeful story. https://www.theguardian.com/commentisfree/2018/dec/31/2018-australian-government-energy-more-hopeful-story

[50] David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied logistic regression* (third ed.). John Wiley & Sons, Inc.

[51] Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

[52] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. 2015. Content, context, and critique: Commenting on a data visualization Blog. In *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, Inc, 1170–1175. https://doi.org/10.1145/2675133.2675207

[53] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (1 2019), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[54] Luca Iandoli, Ivana Quinto, Paolo Spada, Mark Klein, and Raffaele Calabretta. 2017. Supporting argumentation in online political debate: Evidence from an experiment of collective deliberation. *New Media & Society* 20, 4 (2017), 1320–1341. https://doi.org/1461444817691509

[55] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing. In *Proceedings of the ACM conference on Computer supported cooperative work & social computing (CSCW)*. ACM, Baltimore, Maryland, 588–602. https://doi.org/10.1145/2531602.2531674

[56] Elizabeth Jensen. 2016. NPR Website To Get Rid Of Comments. https://www.npr.org/sections/publiceditor/2016/08/17/489516952/npr-website-to-get-rid-of-comments

[57] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (7 2019). https://doi.org/10.1145/3338243

[58] Dan M. Kahan, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. 2012. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change* 2, 10 (10 2012), 732–735. https://doi.org/10.1038/nclimate1547

[59] Helen Kennedy and Rosemary Lucy Hill. 2018. The Feeling of Numbers: Emotions in Everyday Engagements with Data and Their Visualisation. *Sociology* 52, 4 (8 2018), 830–848. https://doi.org/10.1177/0038038516674675

[60] Joy Kim and Andres Monroy-Hernandez. 2016. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1018–1027.

[61] Nam Wook Kim, Jonghyuk Jung, Eun-Young Ko, Songyi Han, Chang Won Lee, Juho Kim, and Jihee Kim. 2016. Budgetmap: Engaging taxpayers in the issue-driven classification of a government budget. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1028–1039.

[62] Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, and Jeffrey Heer. 2017. GraphScape: A model for automated reasoning about visualization similarity and sequencing. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2017-May. Association for Computing Machinery, 2628–2638. https://doi.org/10.1145/3025453.3025866

[63] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, New York, New York, USA, 1188–1199. https://doi.org/10.1145/2531602.2531677

[64] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (4 2014), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

[65] Dave Levitan. 2015. Nothing False About Temperature Data. (2 2015). https://www.factcheck.org/2015/02/nothing-false-about-temperature-data/

[66] Stephan Lewandowsky and Klaus Oberauer. 2016. Motivated Rejection of Science. *Current Directions in Psychological Science* 25, 4 (8 2016), 217–222. https://doi.org/10.1177/0963721416654436

[67] Kat Long. 2017. Meet The New York Times's Super-Commenters. https://www.nytimes.com/2017/11/25/insider/new-york-times-top-commenters-profile.html

[68] Anders Sundnes Løvlie, Karoline Andrea Ihlebæk, and Anders Olof Larsson. 2017. User Experiences with Editorial Control in Online Newspaper Comment Fields. *Journalism Practice* (3 2017), 1–20. https://doi.org/10.1080/17512786.2017.1293490

[69] Thomas W Malone and Mark Klein. 2007. Harnessing collective intelligence to address global climate change. *innovations* 2, 3 (2007), 15–26.

[70] Brian McInnis, Dan Cosley, Eric Baumer, and Gilly Leshed. 2018. Effects of Comment Curation and Opposition on Coherence in Online Policy Discussion. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*. ACM, Sanibel Island, Florida, 347–358.

[71] Brian McInnis, Gilly Leshed, and Dan Cosley. 2018. Crafting Policy Discussion Prompts as a Task for Newcomers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018), 1–23. https://doi.org/10.1145/3274390

[72] Alfred Moore. 2012. Following from the front: Theorizing deliberative facilitation. *Critical Policy Studies* 6, 2 (7 2012), 146–162. https://doi.org/10.1080/19460171.2012.689735

[73] Julia Moskin, Brad Plumer, Rebecca Lieberman, and Eden Weingart. 2019. Your Questions About Food and Climate Change, Answered How to shop, cook and eat in a warming world. https://www.nytimes.com/interactive/2019/04/30/dining/climate-change-food-eating-habits.html

[74] Walter C. Parker. 2006. Public Discourses in Schools: Purposes, Problems, Possibilities. *Educational Researcher* 35, 8 (11 2006), 11–18. https://doi.org/10.3102/0013189X035008011

[75] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 39–48. https://doi.org/10.1145/3287560.3287567

[76] Samir Passi and Steven J. Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Association for Computing Machinery, 2436–2447. https://doi.org/10.1145/2998181.2998331

[77] Samir Passi and Steven J. Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018). https://doi.org/10.1145/3274405

[78] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr. 2019. Data is personal: Attitudes and perceptions of data visualization in rural Pennsylvania. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300474

[79] Kendra Pierre-Louis. 2018. Greenhouse Gas Emissions Accelerate Like a 'Speeding Freight Train' in 2018. https://www.nytimes.com/2018/12/05/climate/greenhouse-gas-emissions-2018.html

[80] Kendra Pierre-Louis. 2019. Ocean Warming Is Accelerating Faster Than Thought, New Research Finds. https://www.nytimes.com/2019/01/10/climate/ocean-warming-climate-change.html

[81] Brad Plumer. 2019. As Coal Fades in the U.S., Natural Gas Becomes the Climate Battleground. https://www.nytimes.com/2019/06/26/climate/natural-gas-renewables-fight.html

[82] Nadja Popovich. 2018. How Does Your State Make Electricity? https://www.nytimes.com/interactive/2018/12/24/climate/how-electricity-generation-changed-in-your-state.html

[83] Cynthia R. Farina, Dima Epstein, Josiah B. Heidt, and Mary J. Newhart. 2013. RegulationRoom: Getting more, better civic participation in complex government policymaking. *Transforming Government: People, Process and Policy* 7, 4 (2013), 501–516.

[84] Susan Roth. 1995. Visual literacy and the design of digital media. *ACM SIGGRAPH Computer Graphics* 29, 4 (11 1995), 45–47. https://doi.org/10.1145/216876.216889

[85] Ian Rowe. 2015. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media* 59, 4 (10 2015), 539–555. https://doi.org/10.1080/08838151.2015.1093482

[86] David M Ryfe. 2006. Narrative and deliberation in small group forums. *Journal of Applied Communication Research* 34, 1 (2006), 72–93.

[87] Bahador Saket, Alex Endert, and Cagatay Demiralp. 2019. Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 7 (7 2019), 2505–2512. https://doi.org/10.1109/TVCG.2018.2829750

[88] Lynn M. Sanders. 1997. Against deliberation. *Political theory* 25, 3 (1997), 347–376.

[89] Priti Shah and Eric G. Freedman. 2011. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science* 3, 3 (7 2011), 560–578. https://doi.org/10.1111/j.1756-8765.2009.01066.x

[90] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing uncertainty about the future. , 1393–1400 pages. https://doi.org/10.1126/science.1191181

[91] Nina Springer, Ines Engelmann, and Christian Pfaffinger. 2015. User comments: motives and inhibitors to write and read. *Information, Communication & Society* 18, 7 (7 2015), 798–815. https://doi.org/10.1080/1369118X.2014.997268

[92] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019). https://doi.org/10.1145/3359229

[93] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. *Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains*. Technical Report. http://syriacivildefense.org

[94] Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindu Narasimhan. 2016. Could this be true? I think so! Expressed uncertainty in online rumoring. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 360–371. https://doi.org/10.1145/2858036.2858551

[95] John Stasko. 2014. Value-driven evaluation of visualizations. In *ACM International Conference Proceeding Series*, Vol. 10-November-2015. Association for Computing Machinery, 46–53. https://doi.org/10.1145/2669557.2669579

[96] Kim Strandberg and Janne Berg. 2013. Online Newspapers' Readers' Comments - Democratic Conversation Platforms or Virtual Soapboxes? *Comunicação e Sociedade* 23, 1 (2013), 132. https://doi.org/10.17231/comsoc.23(2013).1618

[97] Jennifer Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of public deliberation* 3, 1 (2007).

[98] Jennifer Stromer-Galley and Anna M. Martinson. 2009. Coherence in political computer-mediated communication: analyzing topic relevance and drift in chat. *Discourse & Communication* 3, 2 (5 2009), 195–216. https://doi.org/10.1177/1750481309102452

[99] Natalie Jomini Stroud, Joshua M. Scacco, Ashley Muddiman, and Alexander L. Curry. 2015. Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication* 20, 2 (3 2015), 188–203. https://doi.org/10.1111/jcc4.12104

[100] Lu Sun, Robert E. Kraut, and Diyi Yang. 2019. Multi-level modeling of social roles in online micro-lending platforms. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–25. https://doi.org/10.1145/3359235

[101] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (3 2010), 24–54. https://doi.org/10.1177/0261927X09351676

[102] W. Ben Towne, Aniket Kittur, Peter Kinnaird, and James Herbsleb. 2013. Your process is showing. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. ACM Press, New York, New York, USA, 1059. https://doi.org/10.1145/2441776.2441896

[103] Janet Vertesi and Paul Dourish. 2011. The value of data: Considering the context of production in data economies. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 533–542. https://doi.org/10.1145/1958824.1958906

[104] Fernanda B. Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (11 2007), 1121–1128. https://doi.org/10.1109/TVCG.2007.70577

[105] Nai Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring trade-offs between learning and productivity in crowdsourced history. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018). https://doi.org/10.1145/3274447

[106] Elke U. Weber. 2006. Experience-Based and Description-Based Perceptions of Long-Term Risk: Why Global Warming does not Scare us (Yet). *Climatic Change* 77, 1-2 (8 2006), 103–120. https://doi.org/10.1007/s10584-006-9060-3

[107] Wibke Weber, Martin Engebretsen, and Helen Kennedy. 2018. Data stories. Rethinking journalistic storytelling in the context of data journalism. *Studies in Communication Sciences* 18, 1 (11 2018), 191–206. https://doi.org/10.24434/j.scoms.2018.01.013

[108] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *Conference on Human Factors in Computing Systems - Proceedings*. 227–236. https://doi.org/10.1145/2207676.2207709

[109] Wesley Willett, Jeffrey Heer, Joseph M. Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: Structured support for collaborative visual analysis. In *Conference on Human Factors in Computing Systems - Proceedings*. 3131–3140. https://doi.org/10.1145/1978942.1979407

[110] Scott Wright and John Street. 2007. Democracy, deliberation and design: the case of online discussion forums. *New media & society* 9, 5 (2007), 849–869.

[111] Marc Ziegele, Timo Breiner, and Oliver Quiring. 2014. What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication* 64, 6 (12 2014), 1111–1138. https://doi.org/10.1111/jcom.12123

Received January 2020; revised June 2020; accepted July 2020