Where Do Stories Come From? Examining the Exploration Process in Investigative Data Journalism

DILRUBA SHOWKAT, Northeastern University, USA
ERIC P. S. BAUMER, Lehigh University, USA

Investigative data journalists work with a variety of data sources to tell a story. Though prior work has indicated that there is a close relationship between journalists' data work practices and that of data scientists. However, these relationships and data work practices are not empirically examined, and understanding them is crucial to inform the design of tools that are used by different groups of people including data scientists and data journalists. Thus, to bridge this gap, we studied investigative reporters' data work practices with one non-profit investigative newsroom. Our study design includes two activities: 1) semi-structured interviews with journalists, and 2) a sketching activity allowing journalists to depict examples of their work practices. By analyzing these data and synthesizing them across related prior work, we propose the major phases in the data-driven investigative journalism story idea generation process. Our study findings show that the journalists employ a collection of multiple, iterative, cyclic processes to identify journalistically "interesting" story ideas. These processes both significantly resemble and show subtle nuanced differences with data science work practices identified in prior research. We further verified our proposal through a member check with key informants. This work offers three primary contributions. First, it provides a close glimpse into the main phases of investigative journalists' data-driven story idea generation technique. Second, it complements prior work studying formal data science practices by examining data-driven investigative journalists, whose primary expertise lies outside computing. Third, it identifies particular points in the data exploration processes that would benefit from design interventions and suggests future research directions.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in HCI**; **Computer supported cooperative work**.

Additional Key Words and Phrases: Investigative Data Journalism; Data Science; Transparent Journalism; Participatory Design; Inclusive Design

ACM Reference Format:
Dilruba Showkat and Eric P. S. Baumer. 2021. Where Do Stories Come From? Examining the Exploration Process in Investigative Data Journalism. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 390 (October 2021), 31 pages. https://doi.org/10.1145/3479534

1 INTRODUCTION

Investigative journalism represents a unique class of journalism that offers reporting based on a variety of data sources [22]. Unlike fast-paced news reporting organizations, investigative reporters often spend months or even years developing a single story [11]. The reporters are also very detail-oriented in their verification procedures [22, 48, 62]. As Kirkpatrick argues, "Investigative journalism plays an important role in preserving democracy; we need watchdog journalists to hold

Authors' addresses: Dilruba Showkat, Northeastern University, Boston, MA, USA, showkat.d@northeastern.edu; Eric P. S. Baumer, Lehigh University, Bethlehem, PA, USA, ericpsb@lehigh.edu.

2573-0142/2021/10-ART390 $15.00
https://doi.org/10.1145/3479534

institutions accountable, to shine a light on misdeeds" [59, p.16]. A few well-known investigative reporting organizations include The International Consortium of Investigative Journalists (ICIJ) [79], The Center for Public Intigrity (CPI) [36], ProPublica [84], and Mother Jones [49]. Their aim is to produce reporting that makes an impact. Notable examples of this impact in the US range from Sinclair's exposé on the Chicago meatpacking industry [94], to coverage of the Watergate scandal [90], to the analysis of racial bias in algorithmic risk assessment [50].

This most recent example [50] demonstrates how these news organizations are becoming more data-driven [8]. This data-driven evolution in journalism is occurring partly because of increased availability of varied types of data, including structured data (e.g., campaign contributions, government spending), unstructured text, audio/video recordings, testimony, hearings, and many others [21, 78]. As a result, news organizations are experiencing a growing demand for data journalists in the digital newsrooms [8, 14].

This increasing prominence of data-driven investigative journalism poses a variety of challenges. First, many investigative journalism organizations are non-profits, which can also face severe resource limitations, e.g., a limited technology budget, few staff members, and time constraints [7, 72, 74, 103]. Second, the professional practices of journalism differ from those of data science. Other research has closely examined work practices of how data scientists use techniques from machine learning (ML), natural language processing (NLP), or related areas [16, 57, 63, 75, 81, 105]. Even though the journalists use similar tools (e.g., Jupyter Notebook, RStudio) or engage in similar data analysis processes to that of a data science worker [41, 75, 85], less work has examined either the confluence of or differences between data science and data journalism [9, 42, 59]. Furthermore, while data scientists are proficient in these tools [75], many investigative journalists lack formal training in data science techniques [85]. Thus, although data science and data journalism may involve similar computational techniques, but they are applied within different professional contexts and likely involve different day-to-day practices.

As a third challenge, the available tools tend to focus on a specific, potentially limited aspect of journalistic practice [20]. For example, several prior researches have explored data-driven visual storytelling [17, 61, 64, 65], i.e., how best to tell a given story using data visually. To our knowledge, though, significantly less work closely examined *the processes by which journalists formulate ideas for what data-driven stories to tell* in the first place. For instance, given the dearth of tools that effectively search for a specific dataset, reporters have to rely on writing scrapers or manually downloading them, and sometimes finding a dataset can be challenging [85]. Even when they can obtain relevant data, reporters also lack tools specifically designed to enable easily sorting through multiple, heterogeneous data sources to identify story ideas that are interesting, relevant, and viable [85]. The few tools that are available are somewhat domain specific and not easy to adapt [11, 48], or they support developing creative angles on existing stories [69] rather than generating ideas for new stories. This gap – tools that effectively support generating journalistically "interesting" story ideas – may arise in part from a lack of prior work examining what makes an idea interesting to a journalist [37], either in general or specifically in a data-driven setting. While various work provides high level guidance about conducting data-journalism [9, 41, 86], there are fewer empirical investigations thereof.

These gaps in the prior literature give rise to the following two research questions, which guide the work presented here:

RQ1: What are the processes and practices involved in data-driven *story idea generation* in investigative journalism?

RQ2: How do investigative data journalism practices both *resemble and differ from data science*?

To address these RQs, this study examines the story idea generation process in the context of investigative data journalism work practices. We engaged members of the data team at an

investigative journalism non-profit in a qualitative study consisting of semi-structured interviews. We also conducted a sketching activity asking participants to depict their data-driven work practices. These data were analyzed in a qualitative, abductive fashion [15, 40, 67] to understand how ideas for data-driven stories are generated, as well as how they evolve and are shaped by the tools are practices that these journalists use.

The contributions of this research are threefold. In response to **RQ1**, this paper offers a high-level understanding of story idea generation through examination of the data work practices in investigative journalism. In response to **RQ2**, the paper then compares this model with prior work studying expert data scientists work practices. It identifies several points of similarities with that prior work, as well as some unique aspects that stem from the journalism professional context. Finally, this paper uses these results to articulate specific processes that could benefit from design interventions. It also notes which of these design opportunities have already been explored, which opportunities deserve additional attention, and which should perhaps be avoided. The interdisciplinary nature of this work has the potential to highlight new areas for future research at the confluence of data journalism, data science, and HCI with a specific focus towards story idea generation.

2 RELATED WORK

This paper draws on and synthesizes literature from two areas: data journalism, and studies of data scientists' practices. Although relevant to one another, prior work has done little to explore the resonances between these two areas, as described below.

2.1 Data Journalism

Most traditional news media outlets depend on the number of visitors, clicks, subscription methods, and advertisement to earn revenue [18, 72]. Whereas non-profit newsroom does not publish to be profitable, instead, they usually run through donation and grant funding [7, 72]. The key to non-profit organizations sustainability is through the practice of "technology adoption, technology use, and technology learning", [74, p.9]; non-profit investigative newsroom is no exception [11]. According to Kovach and Rosenstiel [62, p. 85], "The group calling itself Investigative Reporters and Editors, for instance, has tried to develop a methodology for how to use public records, read documents, and produce Freedom of Information Act requests," emphasizing the widespread use of public data among the investigative reporters [22, 41, 72]. Similarly, Gray et al. [41, p. 3], described data journalism as "journalism done with data. [...] Both "Data" and "Journalism" are troublesome terms." He quoted,

> "Data can be the source of data journalism, or it can be the tool with which the story is told – or it can be both. Like any source, it should be treated with skepticism; and like any tool, we should be conscious of how it can shape and restrict the stories that are created with it." [9, 41].

Investigative reporting can be original, interpretative investigative reporting or reporting about someone else's investigation in the analysis [62]. The investigative data journalists (in this study) applied a wide range of tools for data analysis [8, 41, 85]. They: i) work with various data formats such as CSV, shape files, Geo JSON, XML, delimited text, tables, various proprietary formats, ii) apply various programming languages for scrapping and web development including Python, Ruby, SQL, JavaScript, and R, iii) rely on different data analysis tools and API including Google suite (e.g., Google docs), Jupyter Notebook, QGIS, RStudio, PostgresSQL Database, Amazon AWS, Mapbox, HUD, Census API, CSPAN, and ArcGIS. They also applied Machine Learning (ML) algorithms (e.g., multiple linear regression, logistic regression, and libraries (e.g., pandas)).

Clearly, data and computational tools play a major role in data journalism. Some of the prior work described the data journalism practice as getting data, understanding data, and delivering data [41, 86]. While others described an inverted pyramid model of data journalism consisting of five phases: compile, clean, context, combine, communicate [9]. Several research proposed data-driven visual storytelling models and they are widely studied in the literature [61, 65, 85]. For example Kosara and Mackinlay [61] proposed a high level model of how the stories are developed based on how journalists work; they described that journalists collect data for exploration and presentation. In another work, Lee et al. [65] described a model consisting of three phases: i) explore data, ii) make a story, and iii) tell a story. In a recent study, Chevalier et al. [17] extended the model by incorporating different journalism roles such as data collector, scripter, presenter, etc., and external factors including ethics, time, cost, and legal issues. Study participants were journalists, designers, and researchers. Chevalier et al. [17] used thematic analysis [10].

Although data journalism [9] and data-driven visual storytelling models [17, 61, 65] are available, none of the prior work has closely examined how investigative reporters work with data to come up with a story idea, and how data science practices are tightly interwoven and interconnected with data journalism [42, 59] work practices; thus, this study will bridge this gap.

2.2 Data Scientists Working With Data

The conventional model described data scientists' work practices consisting of data discovery, data capture and cleaning followed by feature engineering [83]. A recent study proposed by Muller et al. [75] confirmed the conventional model and added "data as created" in the model; the model proposed increased human activity with data as the human move from data discovery phase to data creation phase in the data science worker pipeline using a qualitative interview-based study. There are other similar research exists in the literature that described data scientists work practices in different domains [42, 52, 57]; for example, Guo [42] described a data science workflow consisting of preparation, analysis, reflection, and dissemination through studying users from various domains (e.g, students, engineers, research scientists); Kandel et al. [52] proposed a five high-level task model consisting of discovery, wrangling, profiling, modeling, and reporting by studying data analyst in different enterprise organization (e.g., marketing, finance); Kim et al. [57] identified data scientists' roles in software engineering teams and discovered similar data activities described above. Researchers also examined how instructors teaches data science to novices in data science education [63].

Other work have examined collaborative processes, such as Zhang et al. [111] found that data scientists working in teams are highly collaborative working with different people, using various tools across all stages of data science workflow [83] using a survey based study. Wang et al. [105] reported the benefits and tensions as synchronous collaborative tools are adopted in data science teamwork. Passi and Jackson [81] studied how data science collaborative teams work around emergent tensions during algorithm-driven data analysis through skepticism, assessment, and credibility. Researchers also developed tools specifically to support computational notebooks [56, 106]. In a recent study Chattopadhyay et al. [16], suggested that professional data scientists face numerous challenges when working in the Notebook environment; these pain points extend from loading and cleaning data, explore and analyze the data, managing code, sharing data, replicating results among other problems.

Much of these prior research investigated professional trained data scientists work practices in the context of Computational Notebooks (e.g., Google Colab, Jupyter Notebook) [16, 56, 87, 105]. However, prior studies also indicate that untrained data science professionals such as the data journalists also use similar data science tools for data collection and data analysis [41, 85]. The significance of data science practices in data-driven journalism cannot be overlooked [42, 59].

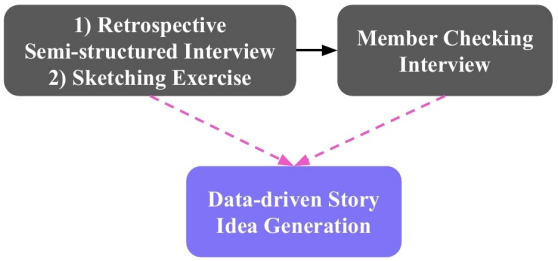

Fig. 1. The study design was composed of two major parts: first, a retrospective semi-structured interview followed by a sketching activity; second, a member-checking interview.

Essentially, data scientists and investigative journalists both depend on data, computational tools and programming languages; however, there is no study that examined the inter-relatedness between these two domains.

In this study, we bridge this gap through 1) uncovering how investigative journalists work with data to come with interesting story ideas, and 2) comparing data scientists and investigative journalists data driven work practices.

3 METHODOLOGY

3.1 Recruitment and Study Design

Participants were recruited via contacts at a non-profit investigative journalism organization based in the US called JustNews[1]. This study was part of a larger project working with journalists at JustNews to collaboratively design computational journalism tools. Since this study was carried out as the first step of this larger project, we did not seek to recruit journalists from other news organizations. We began with two key informant staff members, who introduced us to an additional four staff members. Of those six, a total of five staffers agreed to participate in our study. Although a relatively small number of participants, there are at least four reasons why these data are sufficient for the analysis presented here. First, our participants comprised over half of the staff members at JustNews who work on data-oriented reporting. Ideally, the entire population of data-oriented staff would be included, but practical and professional demands on staff members' time precluded this possibility. Second, this number of participants closely resembles the number of participants in some prior work [81]. Third, we found ourselves reaching theoretical saturation [40] around the fourth of the five interviews. Finally, the data are both copious and rich. We collected over 7 hours and 5 minutes (50,543 words) of detailed interview data, supplemented by the sketching activities and member-checking interviews described below (see Figure 1). Each participants was given a US$20 Amazon.com gift card as a token of thanks.

3.1.1 Semi-structured Interview. With each participant, we conducted retrospective semi-structured interviews [99]. We asked each participant a series of open-ended questions, including the participant's individual background and role at JustNews, and a project about which they were passionate, *"Please talk about a project that you have completed in the past and you were really passionate about."* We asked participants' to talk about a past project so that they can reflect upon the strategies they have used in the project [108]. We asked additional follow-up questions when required. Interviews were conducted between 7 Nov 2019 and 11 Dec 2019 via phone or video conferencing depending

[1] "JustNews" is a pseudonym referencing the organization's mission to bring justice through investigative reporting.

upon participants' preference. All interviews were audio-recorded and transcribed using Temi transcription tool [23].

3.1.2 Sketching. Each participant also completed a sketching exercise, based on the following prompt. *"We talked about various data sources, tasks, people, data analysis tools, methodologies, and output. Please sketch how all these activities go together in your reporting task"*. We used the virtual collaborative tool Sketchboard [98] to enabled participants to externalize their data work practices while talking through them [cf. 53]. The sketching was video recorded. At the end of the sketching task, we prompted the participant to described the sketch from start to finish. This helped us to understand their current work practices accurately and to find any steps that we were not able to capture during the interview.

3.1.3 Member-Checking Interview. After completing our analysis, the authors conducted a member check [3] with our initial two key informants. We presented the description of our proposal and discussed internal working procedures described below in the Results. The two key informants provided some minor corrections and additional details but mostly confirmed our proposal. This member check helped us address potential issues with the number of participants and increased our confidence that our proposed model accurately depicts the data-driven practices for story idea generation at JustNews.

3.2 Participants' Demographic Information

We interviewed a total of five journalists, all of whom lived in the US. Among the five participants, four identified as male and one identified as female. Two participants had a Masters' degree, and three had a Bachelor's degree. Their educational background was diverse, including liberal arts, journalism, and engineering. Most of them were self-taught in computational tools. This study was approved by the Lehigh University Institutional Review Board (IRB). Interviews were conducted in English. Interview data was anonymized by using pseudonyms across four different levels: 1) organization name, 2) participant's name, 3) project titles, and 4) role titles. In order to provide further anonymity and protection of our participants' identities, we anonymized several story related details that could accidentally identify them. We did this by iteratively sharing drafts of the paper with each of the participants and anonymizing according to their expectations. The various level of anonymization is essential to protect participants' from inadvertent identification [51, 71, 88]. We were really fortunate to have highly motivated participants, who tried their best to be good participants, helping us with providing story details and information by being receptive to researchers queries and cooperating with us as much as possible.

3.2.1 Investigative Journalism Roles. Our participants each held different data-related roles at JustNews, and each worked on different projects. Each of these roles involved working with data to varying degrees. Jakob and Drew are the *Data Journalist*, who perform data collection, data analysis, fact-checking, drafting, and story writing. Both worked on housing related stories. For example, Drew worked on the Tax Breaks in Housing story, which investigates Tax breaks in the low income housing projects developed for the poor people. Jakob worked on a similar story called Low Income Housing, where he worked with a local reporter to explore the connection between low income housing and economic areas. Jonas is a *Developer Journalist* and has several years of experience in traditional news reporting. Currently, his role mainly involves data analysis building and maintaining news web application interface [8, 85]. He worked on the Election Campaign project, which analyzed campaign contribution data. Emerson is a *Algorithmic Reporter*. In addition to performing data activities similar to that of a data journalist, this role requires the application of quantitative skills such as Machine Leaning (ML) in data analysis. Emerson worked on the Threat

Prediction story based on digital surveillance data, which identified demographic disparities in error rates of a device that attempted to infer threat from the data. Samantha is *an Editor*. As a senior member of JustNews, her current role involves less data analysis (though she has the skill). Her responsibilities include assigning stories to the team based on skill-set, checking whether a story is feasible or not, checking the data element for correctness, helping the team collaborate, and making sure that resources are available to the team. Samantha described the Health Care data project from when she was a Data Journalist at JustNews; the story revealed doctors who were charging Medicare at a higher rate than others for procedures they performed.

Each of these roles may have additional responsibilities at the organizational level. However, none of the four roles discussed in this paper have strictly predefined or fixed data-related tasks (e.g., data collection, data validation). Rather, all staff have the freedom to chase any ideas and collaborate; for instance, developer journalist could also work on stories. In every decision, though, an editor plays a significant role that everyone in the team looks up to. One of the key factors distinguishing these roles is in terms of journalism experience. In addition to having data analytic skills, each of these roles also required domain expertise for the stories. Oftentimes, collaboration and communication help bridge any knowledge gaps.

3.3 Data Analysis and Coding

We applied qualitative inductive analysis methods drawing on grounded theory [15, 40] and similar approaches [67]. The interview data were coded by the first author and discussed with the second author. Initially, we applied open coding, followed by axial and focused coding to discover higher-level categories. Themes emerged inductively from data [15, 40], and they were discussed, revised, and confirmed only after which it became part of our proposed model.

After analyzing the interview transcripts, we moved to analyzing the sketches. Informed by prior work [e.g., 53], each sketch was annotated by the first author using concepts and themes emerging from the analysis of the interview data. Annotations were based on the visual content of the sketch, as well as on the audio recording and screen capture of the participant creating their sketch. In this way, analysis of the sketches helped both to confirm and to refine these themes our analysis identified.

4 RESULTS: EXPLORATION IN DATA JOURNALISM

The exploration phase involves three main, tightly interwoven and interconnected processes: generating an *interesting story idea*, *data acquisition*, and *tinkering with data*. Staff at JustNews described that the exploration phase can begin with any of these three processes. Furthermore, this phase involves frequent oscillation among these three different processes. Thus, although described in a linear order below, in practice these processes unfold in an iterative, cyclic, and nonlinear fashion shown in Figure 2. After initial data exploration, the journalist move to the next phase of detailed verification process, which is not described in this paper (in great detail) for simplicity, we rather focused our model description about story idea generation.

4.1 Interesting Story Idea

"Journalism is storytelling with a purpose. The purpose is to provide people with information they need to understand the world. The first challenge is finding the information that people need to live their lives. The second is to make it meaningful, relevant and engaging" [62, p.189]. The following paragraphs describe where ideas for these stories come from, how the story ideas evolve and take shape over time, and the key aspects that determine whether an idea is interesting to be worth pursuing as a story. While some aspects of this phase resemble traditional investigative journalism [22, 62, 72], this subsection attends to how those journalistic practices manifest in data-driven work

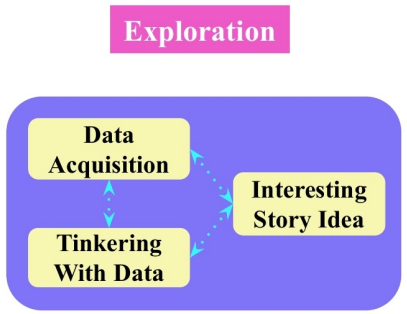

Fig. 2. Highly Iterative Nature of Interrelationships Among Data Acquisition, Data Exploration and Interesting Story Idea Phases.

[17, 41, 42]. This subsection focuses primarily on instances where the story idea was generated first. The latter subsections within the Results will draw out instances where the Exploration process begin either with Data Acquisition or with Tinkering.

4.1.1 Where Ideas Come From? Participants described different story ideas arising or originating from numerous channels. Examples include casual conversations with colleagues, domain experts (e.g., researcher, academics) recommendation [62], website forms for people to submit tips [79, 84], or during traditional reporting [85]. Often, individual reporters have personal interest or expertise [85] in a given area or in well-known projects that have existed for a long time, such as on housing [41] or on Medicare [11]. For example, Jakob described sketching (see Figure 3) out his work with a reporter, Sonya, in the Low-income Housing story, where he was trying to investigate if low-income housing gets built in locations that lacked access to opportunities. Jakob described his process of using data to scrutinize this hypothesis,

> "I'm kind of like looking at the process of how things like normally start. [...] Sonya has this thing she's interested in like affordable housing in [city]. She has an idea that's like, how can we use data to better understand affordable housing. [...] Kind of have Sonya doing traditional reporting [via phone] [...] I am able to use like databases and computers in order to kind of bolster the work of their reporting." - Jakob

This kind of collaboration practice is common at JustNews, where a reporter decides to help another reporter in a story based on mutual interest. Here, in this case Jakob was helping automate Sonya's (editor) data work who had carried out the data analysis by hand. In addition to in-house expertise, ideas also come from external domain experts in a relevant field [62, 70]. For example, Samantha described the origin of the story idea for the Healthcare Data project.

> "So, basically after talking to experts, you know, [Healthcare] programs (experts), what may be interesting to look at. A lot of people had said, you should look at the breakdown of, how things [expenses] are coded, these offices." - Samantha (Sketching).

Inspiration for story ideas also came from other sources such as a reporter visiting a place or an institution or having a spontaneous meeting. In other cases, story ideas emerged directly from data, as described further below in Section 4.3.

These excerpts speak to the variety of means by which story ideas are generated or originated, which essentially involves a lot of communication with inside and outside sources (see Figure 4). A

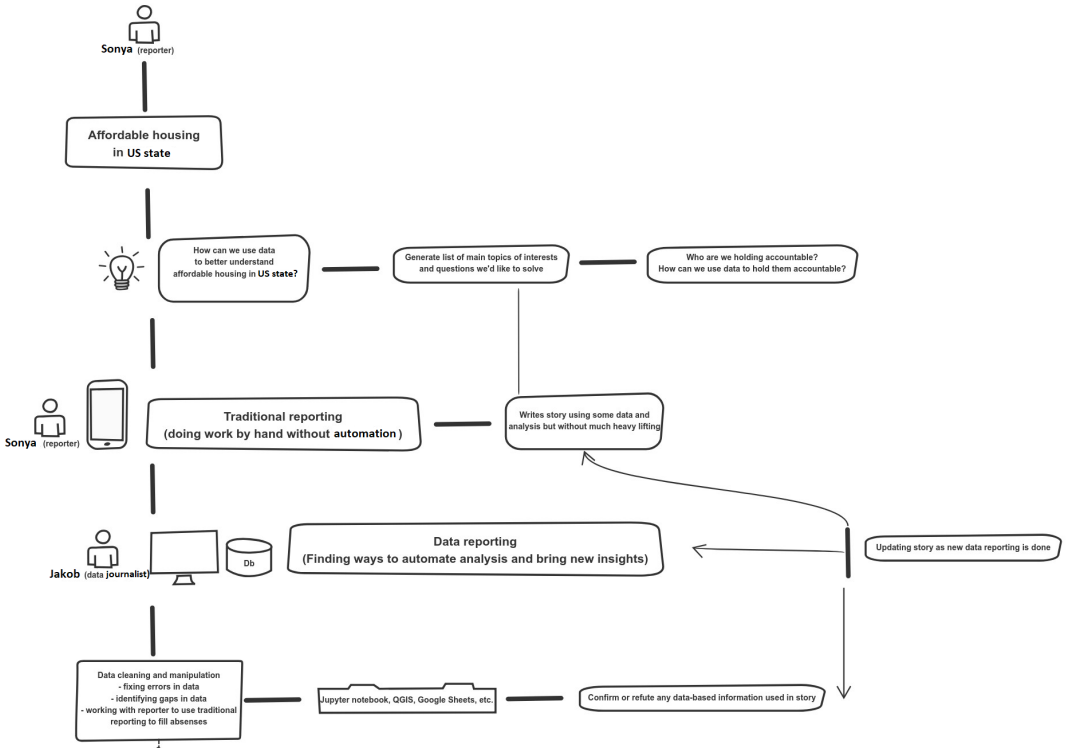

Fig. 3. Excerpt from a sketching exercise depicting the data journalist Jakob's activity on the Low-income Housing project. This sketch depicts Jakob's data activity to generate a story idea, highlighting his interactions with the reporter Sonya.

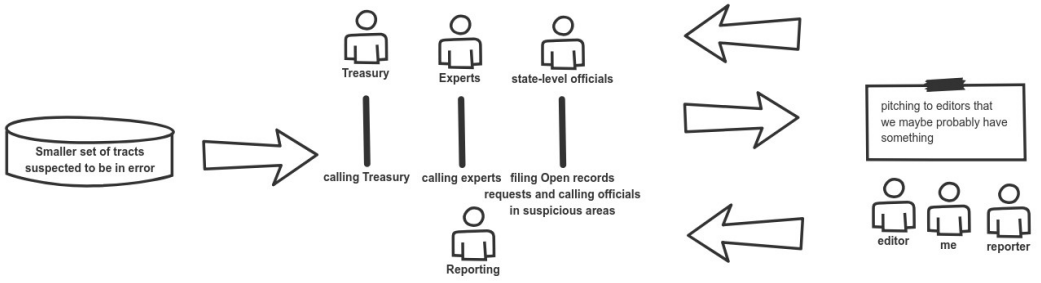

Fig. 4. Excerpt of the sketch illustrating Drew (near right, labeled "me") communicating with various internal and external sources once he had a hunch that they have a story.

similar variety of forces influence the subsequent shaping of a story idea.

4.1.2 How Ideas Are Shaped. Given a potential story idea, reporters evaluate the story idea based on their goals, their skill-set, and ethical considerations [85] and judgements [27]. While reporters' individual goals or factors that inspire them can vary, they are often aligned with more general motivations in investigative journalism. Primarily, reporters seek to provide readers with access to otherwise unavailable information. In describing the Low-income Housing project, Jakob said that the dataset listing the projects that the federal government had funded did not provide information about where these project are implemented in terms of economic opportunity information (e.g., employment access, job diversity, crime index). Thus, the reporters had to go and connect the two pieces of information (projects that received government funding with the location information for opportunity to tell the reader about it [89], to make them aware.

> "And *since the government themselves doesn't say* in that original database whether each of these projects were located, areas with high economic benefit, for example, we wanted to go and do that work ourselves." - Jakob (emphasis added).

Put differently, the focus of this story became shaped by a desire to provide the public with additional information about these housing projects and inform people about inaccessible resources.

Skills also play a major role in shaping story ideas. The journalists described evaluating their skill-set with respect to the story idea and a variety of other considerations. For instance, Drew, before devoting himself in any particular story, would ask himself several questions.

> "I feel like, there's some measures in my head of like, values that might be like, do I care about this as a topic? It's pretty high on the list. [...] am I interested in this for some reason? another thing will be like, can this be done without adding in? [...] is there an opportunity to really help? so, I would want to be involved [...] some degree there's like, has anybody done anything like this before? Can I really say something new if I do apply this skill-set to this problem?" - Drew.

Other journalists described asking themselves similar questions. These questions reveal the various factors that influence whether a story is pursued. An affirmative answer to one or more such questions increases the chance that they pursue the story or decide to collaborate in a story.

As suggested in the above quote from Drew, the journalists also mentioned that they want to work on "something not done before" or original [62, 89] stories rather than building on off someone else's reporting [22]. For instance, Jonas said he worked on the hate crime project in part because the government did not provide a systematic way to collect the information about hate crimes, so he worked on "connecting the dots" in that project to "fill that void."

Finally, the journalists make ethical judgements when deciding whether to work on a story at all. Jonas provided one example from a project about hate crimes. He described how, in such cases, the reporters need to be very careful about what data they are collecting and what they are going to publish.

> "Victims' information is pretty sensitive [...] We have to kind of take care not to, sort of re-victimize people [...], but sometimes like the whole practice or the process of collecting information itself will lead us to modify what we're doing, because we were like, 'Oh, well that's not a good thing.' [...] And it might be potentially harmful if we did or if we published it." - Jonas.

In other words, the reporters may discover that publishing or even collecting information would be harmful. In such instances, they may alter a story or even abandon it entirely to avoid causing harm [85].

This passage also highlights the non-linearity of the practices described here. That is, the "process of collecting information itself will lead [reporters] to modify" both the focus and the very

conceptualization of what a given story is actually doing.

4.1.3 What Makes an Idea Interesting. Reporters consistently described seeking out, in their words, "interesting" story ideas [85]. When asked to define or explain the term "interesting," though, reporters experienced some degree of difficulty. On the one hand, they couched their explanations in terms consistent with general values of journalism, such as informing the public, covering current events, documenting illegal or problematic activities, revealing ineffectiveness of systems or institutions, conveying information that is correct and accurate. Most importantly, the freedom and independence to work in areas nobody else is pursuing and oftentimes the reporters willingness "to take a risk" [62, 89]. On the other hand, there was a conscious admission about the difficulty of pinning down a prescriptive definition of what counts as interesting.

This difficulty became more apparent when talking about trying to articulate "in code" (e.g., Python, Ruby, SQL) means of identifying what counts as interesting. In any attempt to do so, Jonas said,

> "I am certain that we miss interesting things. It's just that the definition of what interesting is, [...] *is not necessarily easy to put in a code.* that's part of the issue. That's part of the challenge essentially." - Jonas (emphasis added).

An example came from Jonas's work on a project that involved tracking tweets from politicians. In one instance, a Democratic senator had been tweeting about a bill to protect the pensions of coal miners whose employers had gone bankrupt. That senator had been tweeting about the bill every single day for months. Later, Jonas noticed that a very similar bill had been introduced by Republicans. Upon inspecting their Twitter logs, that Democratic senator had stopped tweeting about the bill three weeks before.

Jonas uses this example to explain how identifying something "interesting" is quite complex. In this case, he notes,

> "That's a kind of a terrible process because like, it relies on me being aware of things that maybe I would miss and also like *it's not based on just simply the data itself.* [...] I do a lot of tracking of like, behavior that reoccurs over time, like just like this happens and it happens again or on a fairly regular schedule, but, [...] I don't have a great way of like, of notifying myself essentially when that stops." - Jonas (emphasis added).

We see here how the tension between wanting to rely "on just simply the data itself" and the intuitions of reporters "being aware of things" that could be interesting influences how story ideas are generated.

When asked about automated tools that might assist in this process, Drew went farther, saying,

> "I'm skeptical of any thing being applicable to all scenarios and I sort of prefer the idea that I should just be equipped with the skills that I would need and that like I prefer to operate like in really basic code. [...] Like I would rather just know how to run sequel [SQL] queries than have some system that tries to do something. So, that for me I would rather write code in Python than have something to scrape websites then use some tool that tries to coordinate that work for me [...] what if somebody comes up with something awesome someday? I'm like surely gonna be interested in evaluating and potentially using it, but I am not on myself really changing my tool kit." - Drew.

These examples reveal the situated manner in which reporters determine what is interesting. They also highlight the difficulty and occasional skepticism about applying general or universal general methods for outlier detection [11, 48]. Drew mentioned that for the Tax Breaks project initially he did not know what to look for. During casual exploration with a colleague, he happened

to find that a well known real estate development project was included in a certain census tract. As described further below, such "tinkering around" with the data can lead reporters, to "find anomalies that actually themselves [are] stories". However, before such tinkering occurs, journalists need to acquire relevant data [42].

4.2 Data Acquisition

All cases recounted by our participants involved acquiring data [42, 75]. While stories varied in the kinds of data and numbers of different data sets involved, the process of data acquisition was shaped according to two aspects. First, were data acquired as the initial step in the process of developing a story (i.e., before generating a story idea or tinkering with any data), or were they sought out later? Second, were the data public or private? These aspects shaped not only how data were acquired but also how that acquisition processes shaped the broader exploration of story ideas [9].

4.2.1 Public Data: Public data consists of data that are readily available, often from government sources [41, 85]. Examples include low income housing tax credit data, public portions of aggregate medical records, and others. Such data are usually acquired by downloading through API or scraping websites. For example, Drew described how near the beginning of a project,

> "I might be [...] writing scrapers, let's say for like scraping records off of websites. And there'll be a lot more programming. A lot of like, gathering stuff and putting it into like, Postgres [PostgreSQL] databases and trying to essentially see like, is there a way to get to make data useful?" - Drew.

In another example, Samantha described working on an existing Healthcare Data project. Her team was aware that the government medical source was releasing a second part of a dataset regarding the US doctors prescription data. So, they directly downloaded and used the data:

> "We have worked with different [Healthcare] dataset, which was the [second] part that was prescribing data [...] actual things that doctors do [...] So, we had known that they were getting ready to release, that data. [...] that came out. [...] you know, publicly available for download ..." - Samantha.

While the Healthcare Data story was built on a single data source, other stories require the reporters to integrate multiple datasets [41, 75]. For example, both Housing stories discussed in this paper were built on multiple datasets. Information about the geolocation of the construction sites was provided in one dataset, while information about the geographical boundaries of the economic area information (e.g., education, job benefits) were available in a different dataset. The datasets had different format (e.g., xls, json).

> "And, so, that's why we needed to connect, a geographic dataset that has the locations [of economic areas] with this other data set that has the locations of where the projects are that the government is funding. And by combining those two things, we're essentially making connections that otherwise aren't available." - Jakob.

In such cases, data acquisition already involves significant data "wrangling" work [41, 52, 85], as the data of interest need to be assembled into something coherent before analysis can even begin. Data can come in different format, for example, in the Low-income Housing story, a .CSV file containing housing addresses and GeoJSON file containing geographical coordinate information including low/high economic information, to make association between the two datasets, the reporter had to find situational workaround before creating a database to be able to run SQL queries. These acquisition processes are also sometimes more complex when the data of interest are not publicly available.

4.2.2 Other Data Sources: In other cases, reporters leverage data from non-governmental sources. In the Low-income Housing example described above, Jakob collected geographical coordinates (longitude and latitude) for low-income housing projects using the arcGIS API [41]. In his words,

> "[You] would enter in a project address and put it into the search field in the arcGIS website. And once you hit enter and it'll show you where that specific location is on the map." - Jakob.

Investigative reporters often use record requests in their reporting [22], and they generally face no problems when making record request from the government sources. However, our study found instances where the journalists were asked to pay for accessing certain datasets. Although such occurrence is undesirable, they are also not uncommon [17, 61]. This process of paying for data is generally more involved than the collection of public data [85]. Such context [32] associated with data acquisition often impacts story persuasion.

Journalists also receive data [75], sometimes, these data come from domain experts, researchers, or other collaborators. Receiving data from anonymous sources [62] is also fairly common, such as safety investigation reports received in a story about safety failures in manufacturing. While working with given or public data is comparatively easier, however, when working with annonymous datasets, the journalists have to be extra cautious, they usually consider asking themselves: "Who gathered it, when, and for what purpose? How was it gathered? (The methodology). What exactly do they mean by that?" [9] as a way to navigate around issues of trust relating to the data sources. As a result, working with such anonymous dataset involves rigorous verification, and sometimes the journalist may decide not to use the data [61] if the data is unreliable [9]. The journalists found that anonymous sources are usually motivated by "selflessness" to prevent some disasters from reoccurring.

In many cases, journalists' collect data through research, interview [61]. Reporters take consent from the story subjects (e.g., supervisor's consent when interviewing researchers) whenever possible. However, for sensitive stories, interviewing subjects is not easy, because of the constraints imposed on the subjects such as threats or job demotion. In such instances, the reporters usually go out of their way to engage with the subjects. For example, by prompting the subject to draw to convey information instead of speaking or going off-the-record conversation to provide freedom and respect subjects privacy. All these factors essentially influence whether a story gets completed or not.

Such issues of data access ease influenced both what stories were pursued and what reporters had available for exploration [61]. Oftentimes, during the course of the story, the reporters might require to collect more data or throw away incorrect or unusable data [9], as indicated by the cycle in Figure 2 among data acquisition, interesting story idea and tinkering with data described next.

4.3 Tinkering with Data

Reporters' data exploration practice was similar to Tinkering [101], which is described as "a playful and subjunctive modality – once often replete with the utterance and practice of what if, could be, maybe, perhaps, let's try it out etc." [104, p.4]. It "is one way of preserving what is valuable and reworking what it is not" [102, p.5]. As Jonas describes it,

> "... the other aspect of this is sort of me poking around in SQL for the most part or sometimes in code." - Jonas.

Some stories the journalists described actually began in this way – not by formulating an idea for a story, or by acquiring some new data set(s), but by tinkering with an existing dataset. During this

process, the reporter is likely unsure of the story idea. S/he engages in trial-and-error experimentation with the dataset to see what is there. For instance, Drew examined data to discover places in the data that warrant further inquiry.

> "I was using data to find places to go do more reporting and I wasn't using it in a way that a lot of data journalism is where like the story is like some big Sunday headline number that comes out of the dataset. So, like for me it was like, *what's the best way to find places that deserve scrutiny*? [...] it was very much like a kind of focusing the work, kind of like exercise in data analysis." - Drew (emphasis original).

In other cases, the story idea is already known for common stories covering Housing or Medicare [41] stories. However, the exact individual/institution that are "held accountable" are not discovered until tinkering with data. Journalists find them through discovering outliers in the data [11, 48]. For example, Samantha working in the Healthcare data story described how she found the doctor who charged medicare at a higher rate than peers through outlier detection,

> "I did the summary statistics showing, [...] separating, all the doctors in other specialties are so that we're only comparing apples, all the [doctors specialty] in [state] for example. And then just kind of did some exploratory data analysis, some kind of rough graphs to show *who are the outliers*, what does the typical breakdown of the scaling scale look like for an [doctors' specialty] in [state] and who are those people who don't fit that profile. So through doing that we were able to pretty quickly show, for each specialty and each date, both what the typical profile [...] was like, and we were able to pull out any people who didn't really fit that profile." - Samantha (emphasis original).

Out of the five complete stories participants described, three of them started with outliers or anomaly in the dataset during casual explorations. Thus, what an "outlier" means have different connotation for the journalists' [93]. Journalists' described an outlier or anomalies as interesting – a data point that is "different", "important", "useful", "controversial", "scandalous", "relevant", "valuable to the reader", mostly true and a story idea worth persuasion [37, 89]. We also want to note that, the practice of detecting outlier for story idea generation is now currently changing.

Journalists also ask questions such as "what would be the variables used?" in feature engineering [34, 75], as well as what metadata to use in the story. For example, in the Tax Breaks in Housing story, Drew was trying to understand how a particular project ended up in the Tax Breaks program. He had to replicate a previous analysis done by the government agency. Drew had to refer to additional source document and do additional research, and contact domain experts to further understand what variables might help him understand the agency's data analysis process and the variables used in the story [62]. In his words,

> "I think it was sort of three main data points [...] the [agency] was using. It would be the overall [...] the percentage of people in poverty and the median family income [...] were the main variables used. To determine eligibility for the program." - Drew

Participant's also examined data as a process of generating "a hypothesis" [57, 70] and testing. For instance, Emerson worked on the Threat prediction project. He wrote during the sketching activity about testing the threat detection device using standard datasets,

> "*Generate a hypothesis* based on the fact that we saw trends for specific demographic categories (a hunch, not statistical). Speak with editor, co-reporters (thought it was potentially interesting to pursue), experts in [surveillance data] analysis and other scientists (who brainstormed ways to test our hypothesis further); the algorithm reacted more strongly to [a particular user group] in our dataset. Is that a result of some characteristics of the algorithm?" - Emerson (Sketching), (emphasis original).

Again, though they may have "a hunch," the process of tinkering is what led to the specific idea about skewed results for a particular demographic (e.g., race, class, age).

The casual exploration to pursue a story idea may also enable the creation [61, 75] of an entirely new dataset. For example, the Threat Prediction project described above required Emerson and his team to collect ground truth data to test the device's algorithm. He wrote,

> "We went to [X institution], [...] to record dozens of [people] each time in solo and in group situations. [...] set up an experiment protocol, [...] just observe their interactions, record the whole time." - Emerson.

While experimenting with the data, the reporters would also check to see whether the data is complete, is consistent, has missing values, or has errors. This process also involves data transformation such as filtering, outlier removal [41]. In other words, data exploration, including dealing with formatting problems and mostly evaluating the data for their quality, helps reporters determine how usable the data are [65, 75].

Such testing involves a panoply of different tools, ranging from Google sheets, to hand-crafted MySQL queries, to generating exploratory visualizations, to other data transformation such as aggregation, to manually flagging problematic spots or something that seems like an "interesting" [85] story idea worth pursuing. As Jonas described in the context of the Election Campaign Data Analysis story,

> "Then we also derive aggregates [...] and put into tables. We create essentially new tables. So, at this level, like what we probably could or should be doing essentially is like, and we do a little bit of this is like *flagging interesting things*. [...] So we will do like booleans of like whether [...] simply something exists or not. [...] one example of this would be, does this filing have contributions over a hundred thousand dollars in it or something like that. " - Jonas (emphasis original).

This description shows how the process of playful experimentation with the data and identifying interesting stories can be quite interwoven in practice.

Data exploration might also begin with something as simple as sorting a spreadsheet, as Drew described his data exploratory experience in the Tax Breaks in Housing project.

> "Initially, I was asked to like look over a spreadsheet with a colleague and it would seem like it would just be some sort of routine activity [...] the spreadsheet was like all the data on these [geographic areas] that were designated [for economic development]. And we were basically sorting the Excel spreadsheet and going like, okay, let's look at like the richest ones." - Drew.

This examples illustrates how the exploration process is shaped by the affordances of data analysis tools. The JustNews staff did not begin with a plan to analyze the "richest" economic development areas [41]. Rather, the spreadsheet easily afforded this view, which subsequently shaped their analysis.

As described above, data exploration also involves checking how clean is the data and noting any missing fields. Such cleaning happens often, though not exclusively, while tinkering with the data. In the Election Campaign project, Jonas describes keeping track of each individual member of the US Congress.

> "There is a unique identifier for them, that the Congress actually assigns. [...] but this particular dataset doesn't include it for no good reason that I can see. And so what we have to do is, we have to try to match those files that are published. We have to try to match the names in there and the offices with our list of members. [...] we have a script that kind of goes through and tries to match those members to their ID" - Jonas.

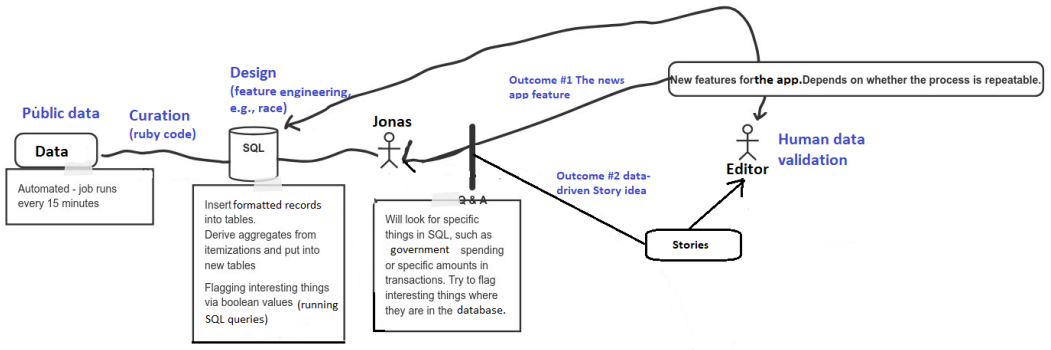

Fig. 5. Annotated sketch drawn by Jonas when describing the data work in the Election Campaign Data Analysis project. Blue text indicates researcher annotation.

Such missing or inconsistent aspects often only come to light during the exploration process. The exploration phase also involves determining how to address such issues, or whether the issue can be addressed at all. Jonas mentioned that they had to manually correct the names for member of the house to process the inconsistencies in the names [41]. The above paragraph demonstrates the various range of activities in which our reporters engaged during data tinkering.

4.4 Sketching Data Analysis

We collected 5 complete sketches, one from each of our participants, about the processes through which they go for data-driven story idea generation. After analyzing the interview data, we analyzed sketching data and found similar processes as they described during the interviews. In the interviews, some participants described more than one story, but during the sketching exercise participants sketched one story of their choice. Jonas sketched the Election Campaign Data story, Emerson chose to sketch the Threat Prediction story, Samantha drew the Healthcare Data story, and Jakob and Drew both made sketching illustrations of the Housing story they discussed during the interview. In the interest of space and to avoid redundancy, this section shows one annotated sketch as an example. An additional annotated sketch is presented in Appendix B. The sketches shown in this paper are anonymized (e.g., names of specific projects and government agencies are redacted) and partially cropped for simplicity and anonymity.

Figure 5 shows a sketch drawn by Jonas about a news application that he administers. Because this is a news web application, the initial process of data collection and data cleaning is automated before inserting into the database (e.g., using code). However, the experimentation (i.e., tinkering) depends on Jonas manually running queries (e.g., design [75]). This is depicted in the "Q & A" phase near the center of the sketch where Jonas himself appears. The sketch also shows the two possible outcomes of such tinkering: it can lead to new features for the news application, or it could lead to a story. Which of these outcomes is pursued depends in part on the editor, who is also responsible for validating the findings.

Thus, the analysis of the sketching data offers a form of methodological result triangulation [100], in that similar results were obtained both from the sketching and from the interviews. We have also shown through sketch annotation that the data analysis steps (e.g., data collection, capture, curation, engineering features, human inspection of data validation [75]) holds for different roles across different stories covered in this study, essentially indicating the richness of our data. That said, the collaboration and communication aspects were more evident in the sketches, as seen in

the coordination between Jakob and Sonya in Figure 3, or in the various interactions Drew had with individuals both within and outside of JustNews in Figure 5. This point aligns with prior findings that data work even in data-driven investigative journalism is a collaborative endeavor [e.g., 44, 70, 111].

4.5 Exploration Summary

To reiterate, the exploration phase involves three cyclic iterative processes: generating an *interesting story idea*, *acquiring dataset(s)*, and *tinkering with data*. Figure 2 depicts the iterative flow among these three processes. The exploration phase may begin at any one of these processes, as noted above. Of these three processes, generating an interesting story idea was the most crucial. At this stage the reporters had settled on what their story idea would be – what data they would be analyzing, what results they would be presenting, whom they would be holding accountable (see Figure 3), etc.

Essentially, once the reporter has their data in hand, they check to see what they can do with the data. This involves checking data quality, completeness, running queries, writing codes/scripts to flag interesting data points. These data points are manually (e.g., writing and saving queries) marked with the potential to become either a story or become part of the News Applications [41] or interactive tool [59]. They also decide what features [34] or variables to use in the story. Once the story idea is determined the team (e.g., an editor is always involved in a story) decide what technical and other resource needs are, and whether to work on the set of ideas immediately or work on the side. They also gather more data if needed, oftentimes essentially creating an entirely new datatset (e.g., consisting of multiple datasets or ground truth data) [59, 75]. They fetch various record request, contacting relevant outside sources etc. The journalists also engage in a very detailed verification process after exploration. The story is never published until the data that JustNews produced are thoroughly verified internally within the team and with outside experts. Subjects covered in a story are also told about the results and given the chance to respond, only after which the story is published [62]. In this study, we intentionally avoided describing the details of investigative journalist verification process for simplicity and primarily to focus on story idea generation.

5 DISCUSSION AND IMPLICATIONS

The results above suggest a model of how investigative reporters engage in data-driven work practices to generate story idea at a specific non-profit. This section compares the porposed model with related prior work studying data scientists as well as work on data journalism [9, 17, 25, 35, 42, 62, 75, 80, 81].

These findings complement prior research on data science practices, which has focused on experts: data scientists, software engineers, enterprise analysts, etc. [16, 35, 42, 52, 57, 75, 82, 111]. However, prior work has not provided a detailed comparison of the two distinct yet related areas of data science and data-driven investigative reporting [59]. The remainder of this section fills that gap by comparing our proposed model with prior work [42, 75, 81]. It also offers a brief comparison between our model and related work on models of data-driven visual storytelling [17, 65] and data journalism [9, 41], finally, we suggest implications for future design and research directions.

5.1 Comparison with Data Science Work Practice

As suggested by Kirkpatrick [59], we found some overarching similarities and subtle differences between data scientist's [57, 75, 81] and the investigative journalist's data work practices. We provide these comparisons using the empirically derived model of data scientists' work practices proposed by Muller et al. [75] – organized in terms of discovery, capture, curation, design, and creation –

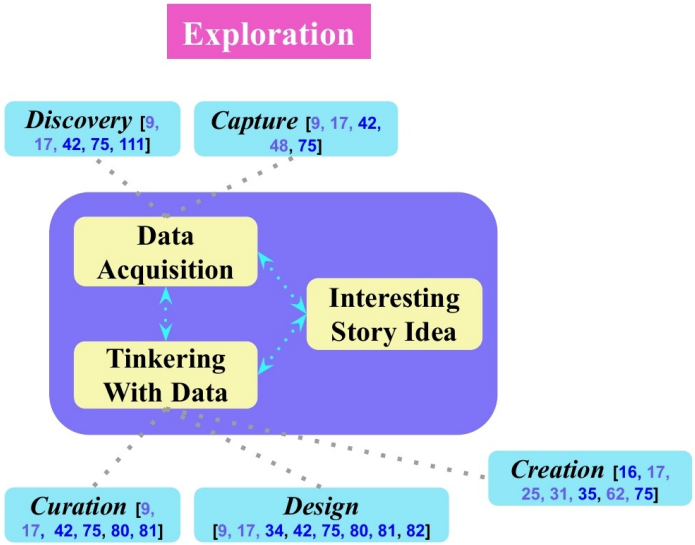

Fig. 6. Comparing our proposed model of investigative journalists' data work practices at JustNews with prior work studying data science practices (citations marked in blue) and on data journalism (citations marked in purple).

interleaved with our formal proposal of story idea generation (see Figure 6), and throughout, we also note implications for possible future design intervention and research directions.

In the data acquisition step, the journalist must first acquire data before performing any analysis [42], a process similar to the discovery phase in Muller et al. [75] model. Unless the dataset is given or publicly available for download, finding relevant datasets is not always easy for the journalist, similar to data scientist's [75]. For example, in one of the stories about a story about safety failures in manufacturing, the reporters knew what the story idea was, but they did not know how to obtain the data. After contacting the relevant government sources and waiting for several months they received an anonymous dataset [62] consisting of thousand of pages of documents about training reports. Processing through large volume of documents without much context requires extra situated work [63, 80]. The challenges that journalists face in working with such huge volumes of unstructured documents differs from the work of enterprise data analysts [52], who mostly analyze structured data. Working with unstructured data (e.g., emails records, court briefings) requires further experimentation with the data and sometimes pose uncertainty in the story persuasion [63, 80].

While some stories might based on single dataset (e.g., the healthcare data story), others may require the journalists to capture, combine, and integrate multiple datasets [52, 75]. The Tax Breaks in housing story presents one example. In such instances, journalists described the difficulty combining data obtained in different file format (e.g., PDFs, XML), which might require them to use different tools [16]. For example, Emerson had to use several advanced Python libraries to work with surveillance data. Thus, they faced various challenges when sharing files such as Jupyter Notebooks with each other, often raised due to teammates technical incompatibility, somewhat similar to problems when working in software engineering teams [96]. For example, one reporter who preferred working in Python mentioned that they cannot leave comments while collaborating with another reporter who prefers working in Excel Sheets [59], whereas Jupyter Notebooks easily

afforded this ability [16]. Thus, journalist wanted support for tools that allows for coding in more than one language.

Furthermore, the investigative journalist also often rely on other data sources, such as FOIA or record request [22], which might sometimes require them to adapt their request to receive the data. This is undesirable, because non-profit organization usually work under resource constrained circumstances [72, 74, 103] impacting what story gets to publish [72] and what information is made available to the public [72]. It also causes a delay in the publication of the investigative stories, part of the reason why some stories are never published due to the lack of raw data/numbers [75, 85], even though making the people aware of the story is necessary. The journalists also need to track the data sources to make sure that the data is up-to-date (e.g., static vs dynamic data [75]) similar to data scientist [42, 111]. Journalist' faced difficulty to track changes [16] when a data value changes due to manual update or update from code operation or source update [42], the lack of proper code comments or documentation [96, 111] was problematic for them. Though, one might suggest to use git or Jupyter Notebook's synchronous feature [105]; but, reporters non-technical background makes it infeasible. Currently, the journalists use manual process to keep track of such data update involves using google docs which they termed as "data diary" [58].

"Clean, complete, and consistent datasets – as every data analyst knows – are a theoretical fantasy" [80, p. 2443], our participant's also acknowledged this point similar to that of data scientist [42, 75]. Once the data is collected, the journalists starts initial experimentation with the data to check for its quality (e.g., completeness, missing data); he might engage in various data transformation processes (e.g., filtering, removing duplicates or human error [9]) [41], to curate and clean the data [16, 70, 75]. The journalists had to find situational workarounds to account for missing, ambiguous, inconsistent data [16, 80]. Though, each individual case might vary depending on the story. However, the journalist usually write code, run scripts, use different tools [42], derive manual heuristics to work with the data. For example, working with ambiguous data [81] was challenging for Jonas. Jonas worked on Congressional data story, where he had to classify why members of the congress missed votes using text data. He said, some of the explanations were obvious (e.g., daughters graduation, family emergency), while others were unclear or vague required careful reading and interpretation. Eventually, Jonas and his colleagues had to rely on manual heuristics similar to that of data scientist team reliance on low accuracy score given "extremely hard" [81] nature of the problem. Another way the journalist approach ambiguity is through directly contacting the agency or organization (part of traditional reporting process) that published the data to inquiry for the missing or incorrect information, and then manually updating the data point in the dataset [9, 85].

During the data tinkering process the journalists also look for outliers, anomaly, i.e., something different or surprising information [11, 48] in the data. They intentionally look for such surprising knowledge as opposed to the data scientist (individually or collaboratively) who might decide to either use or discard surprising information (e.g., outliers [75]) depending on various factors (e.g., business goals, computational ideals) described as "counter-intuitive knowledge" in [81]. Through discovering the anomaly in the data, the journalist decide whether there is a potential for an interesting story idea or not. Anomaly [11, 48, 59] (in addition to various tracking) plays a major role in investigative journalism. However, there are other criteria (described in 4.1) such as journalist's values [89], personal interest, skill-set collectively determines what it takes an story idea to be "interesting". The journalist engage in the design of meta-data or feature related information [34, 75]. For example, in the Tax Breaks in Housing story, Drew used several features (e.g., family median income, percentage of people in poverty) in the story. Drew ran several simplified queries (e.g., greater than 5%) [81] to determine the eligibility of a government project in certain geographic locations; a process very similar to that of an expert data scientist [81] trying to work with inconsistent data. The outcome from the tinkering phase is an interesting story idea.

The journalists along with other colleagues in the team check to see if the result makes sense or not by replicating results [87]; they take notes, hold meetings (e.g., with editors, reporters) to discuss resource, technical requirements and possible collaboration opportunities [42].

Oftentimes, the journalist might also gather more data [9] (indicated by a bidirectional edge in Figure 6) or end up effectively engaging in the creation [75] of their own (e.g., ground truth [16, 57]) dataset [cf. 24, 25, 61, 85].

In several cases, the journalist may or may not have direct access to the underlying algorithm. For example, Emerson worked on Threat Prediction story; he was able to access the threat prediction algorithm through the device; he verified the models underlying working using a three-fold test: first, using standard dataset; second, using synthetic dataset; third, in the real-world. This approach of model verification and building trust is slightly different than how expert data scientist work with large datasets proposed in [81], perhaps because the journalist was testing a device and worked with smaller dataset, and the nature of the problem domain and group dynamics.

In other cases, such as patent algorithms, the journalist have no way to investigate the internal workings or determine the input-output relationships. In such instances, traditionally, the journalists engage in a well-known technique called reverse engineering to replicate data analysis [24, 25, 31], a process similar to that of the data scientists or software engineering practice of reproducing [35] results by reusing code [16, 57]. This process was problematic when a journalist is trying to reproduce results using a tool (e.g., Notebook) that is different that the tool that the other reporter had used (e.g., Excel, SQL) raised from technical incompatibility among team members [96]. Code sharing doesn't help much either due to lack of proper code comments and documentation [111]. "Reverse engineering" is not always accurate [85], but it works [24]; relying on such approaches in also quite common among the journalists at JustNews.

To summarize, a lot of the technical challenges that our participants faced are well-known problems in software engineering [91, 96] and data scientist's working in different domains [57, 75, 80, 81], such as track changes (e.g., data value/row, provenance) [42, 87], collaborating in teams with incompatible technical skills, documentation problems (e.g., comment on data) [5, 96, 111], replication [87] or code reproducibility problems [16, 35], support for multiple programming language among others. While some of the subtle differences emerged primarily from journalistic professional context of data analysis for story idea generation of this study. These findings essentially provides interdisciplinary design implication [60] and calls for future investigation.

5.2 Comparison with Prior Work on Data Journalism

Significant work has considered the unique challenges and opportunities of data journalism [12, 18, 19, 25, 31, 45, 59, 95]. However, the comparison here focuses on the model of data-driven visual storytelling in journalism from Chevalier et al. [17], specifically the "exploring data" phase in the model, since they explicate the detailed processes that is most directly comparable with the one proposed here for story idea generation. We will also note design implications as applicable.

In the data-driven storytelling model proposed by Chevalier et al. [17] described a the initial phase of "exploring data". That phase's problems and pain points associated with data, computational tools and data analysis closely resembles to the challenges faced by the our participants. For example, our participant's faced problems of difficulty finding data, cleaning data, they had to work with incomplete/missing or incorrect data [9]. However, investigative journalists process of dealing with incorrect, ambiguous or missing data was thorough and detailed when compared to the verification for data visualization described in the data-driven visual storytelling [17, 61, 65] model. Even though we intentionally avoided discussing participant's data verification process for simplicity. Verification plays an important role for our participants and investigative journalism in general

[62]; thus, we want to provide a high level overview in contrast to data-driven visual storytelling [17, 65].

For instance, we found that our reporters were relentless in their verification process [35, 62], they rigorously check teammates data work, verified findings with external domain experts [73], they looked at data from different perspective and context [81], to prevent biases [9, 50, 72]. When checking teammates data work, the reporters would not doubt or question the other reporters work or capabilities, instead, the reporter would work through skepticism, asking questions and engage in effective communication [81] – a practice very similar to humility, i.e., the "journalists should be humble about their own skills", [62, p.101]; focusing towards making sure that the data that they put out there is true and accurate [62, 89].

For some complex data-driven stories, JustNews reporters publish the dataset used [95], data collection process [26], data analysis methodological details (e.g., summary statistics), and sometimes publish code on GitHub [31, 95] to remain transparent about their verification process. The reporters also share story draft, data source they analyze with the subjects [62] so that they are given the opportunity to respond before publishing. JustNews do not publish until all facts reported in the story is verified. Perhaps because JustNews staff were under less pressing time constraints than in the settings described by Chevalier et al. [17]. These difference may emerge from JustNews's exclusive focus on investigative reporting [22], wherein verification features perhaps more prominently than in other forms of journalism [41, 61, 62]. They might also arise from differences in the types of participants (e.g., roles) and organizations involved, since Chevalier et al. [17] drew their participants from multiple, diverse organizations, as opposed to our focus on a single news organization.

The notion of generating an "interesting" story idea was both more central to and more complex in our analysis. For Chevalier et al. [17] model, the story idea is already known before any data is collected and the origin of the story was not the primary focus. In data-driven visual storytelling model, sometimes the story is pre-written, and the reporters are simply determining how to recast the written story in visuals. In our model, though, reporters rarely began with a specific story idea in mind unless the story has existed for a long time or widely known (e.g., Election Campaign Data Analysis) [41]; in other words, just a high level story idea. However, the exact story idea or the individual/institution that are held accountable always emerged from the data (e.g., anonymous reports/documents, public dataset, or record requests) [62]. Thus, the journalists either knew what the story idea is (at a high level) and needs data to work with or they discover the story ideas directly from the data and generate hypothesis or ask questions that he would like to answer using the available data (see Figure 3) [9]. Furthermore, our reporters were inspired many different ways to work on a story. For example, journalists' values [89], ideology [109], personal interest (e.g., scandal [37]), news relevancy [37] and significance [62], and others play an important role in shaping story ideas (discussed in detail in section 4.1). When operating as a non-profit news reporting organization [74, 103], the source of funding and other resource constrains may also decide what story gets pursued [72].

In another point of contrast, alongside publishing stories JustNews also publishes News web application for relevant datasets (e.g., Election Campaign Data Analysis). They effectively engage in the data analysis with two goals in mind, i.e., whether certain feature(s) are useful for a story or publishable in the news application. Thus, depending on the data analysis (e.g., SQL/Python queries on demographics) outcome, data (e.g., a row/column, table) either becomes a story or part of the news application [41]. Essentially, data work involved in a story is rather manual as opposed to news applications where the data work is automated. For example, a particular row in the Healthcare dataset might become a story because it is interesting and if that row representing a doctor in the dataset is an outlier and all other evidence (e.g., data from his office, past records)

support the doctor held accountable. On the other hand, if the data analysis process is repeatable then it becomes part of the news application, for example, a table presenting the list of doctors prescription data in a given state/city over a year to show trends in the data. While the main objective of data-driven visual storytelling is to presenting and communicating data to persuade [61, 65, 85], the investigative reporters' goal is to make an impact such as law reformation, induct change [62, 72] and make people aware by providing valuable information [41]. Another important point to note that, the JustNews publishes investigative work so that other news reporter in other new media organizations could use them [22].

Thus, the proposed model of data-driven practices at JustNews helps identify similarities and differences between data scientists and investigative journalists data work practices. As Muller et al. [75] and colleagues note, "many people are doing the work of data science, across multiple job categories and titles, and that this diversity is likely to increase over time" [75, p.2]. The findings presented here help deepen our understanding about the nature of that diversity along with potential avenues for future research.

5.3 Implications for Design and More

Previous sections highlight the situated similarities and nuanced differences between data science and data journalism data work practices as they unfold in an investigative data-driven story idea generation process. Below we offer detailed descriptions for possible future design directions, starting with the most well-known yet unresolved challenges, and then moving towards unexplored research avenues. Throughout, we suggest specific areas that can benefit from interdisciplinary design, research, and development by synthesizing across human-centered design, social sciences, data science, software engineering, and related areas [16, 32, 60].

5.3.1 Previously Identified Challenges. Our results corroborate the importance of challenges identified in previous work. As an example, the journalists faced problems with name resolution, for instance, when a member of congress changed their last name. The journalists in our study had to manually track and update the database to reflect such changes [41]. Another related problem is to track when the value of a given field changes. These changes can occur either by their own operation or by changes in the source data they are working on (data provenance [cf. 42, 111]). Similarly, code provenance also poses unique challenges, for instance, while replicating another reporter's analysis. Currently, Git like tools are not workable due to technical incompatibility among the team members [44, 105]. Software design interventions to navigate some of these challenges will be valuable.

Furthermore, journalists usually receive their hands on experience in computational tools either through self teaching, from colleagues, by training, or from journalism coursework, and they have their own workflow. The training materials might cover only a selected few tools and programming languages [41]. Therefore, some collaborative training or joint research efforts at the intersection of data science education [42, 63] and data journalism to educate the journalists about the latest tools and technologies will be beneficial and provide encouragement. For instance, notebook style tools, such as Google Colab [6], can be learned within a week or two, and such tools implicitly support many of the capabilities that the journalists want. At the same time, applying these tools require the journalists to have some basic level of programming experience. This, again, highlights the potential impacts of interdisciplinary education approaches [60, 63].

5.3.2 Interesting Story Idea Generation. Some computational tools have been developed to assist journalists with news discovery [28, 68]. Tools to support journalists in finding creative angles (e.g., quirky, ramification) [69] on a story also exist. However, tools that will effectively support generating "interesting" story ideas that are "original" and "important" requires further attention.

While artificial intelligence (AI) and natural language processing (NLP) techniques to discover outliers from specific dataset exist [11, 48], these techniques are not necessarily designed to identify data points that are *journalistically* interesting. Varying degrees of experience among journalists can also limit their capabilities to make use of all the available data sources [111].

Journalists can be supported with tools that are specifically designed and developed for them to generate journalistically interesting data-driven story ideas. Such tools could be designed not to replace but rather to complement the existing journalistic processes of story idea generation. For instance, it may be possible to devise software that is able to detect outliers (e.g., text, numbers) from single and multiple datasets [11, 48] in a way that aligns with the characteristics of what journalists find interesting (e.g., original, important, make an impact). Doing so requires translating these definitions for the term "interesting" into text-based interactive tools by applying various ML/NLP capabilities [39], perhaps leveraging recent advances in design methods [e.g., 110]. The main challenge here is to move beyond the statistical approaches to outlier detection, instead shifting focus to a more qualitative characteristic meaning of an outlier or data point. We slightly touched upon the journalists interpretation of outlier and its contextual meaning for story idea generation. Thus, the various computational tools and dataset usage in several story idea generation context [32, 55] might provide areas for design implications in varied domains (e.g., software engineering, data science, data journalism) (discussed in 5.1) [35, 82] through opening up the doors for conversation among researchers coming from diverse disciplines [60].

Furthermore, existing tools (e.g., Rstudio, SQL [41]) also have a limited ability to capture social behavior, changes in behavior, and human emotions. Novel systems could take benefit of the well-known sociotechnical systems such as Twitter, Youtube, Reddit, to identify latest news trends [66]. Data travels a long journey of context [32, 55, 73] and situations [60] before becoming knowledge or a newsworthy interesting story idea. Such a path is influenced by journalists habit [109], tools they use (described in section 2.1) [82] and its affordances, journalists values [89], ethics and others factors. Understanding each of these data dimensions along with the context are crucial (e.g., [75]). The subjective decisions and judgments humans are able to make makes this a ripe area for technical exploration.

Additionally, our journalists described the data-driven story idea generation process from their individual perspective due to the interview nature of this study. However, one interesting observation is that during the sketching exercise the collaborative nature was more prominent (see Figure 4). Thus, we recommend that future researchers investigate more specifically into the collaboration aspects, such as interactions with the domain experts [76], in further detail. Doing so can help identify possible tensions that might arise, as well as the resolution of such conflicts [81, 111]. Stories which required the creation of ground truth data [75] and labeling [76], offer opportunities to examine such collaborations.

5.3.3 Inclusive Design for Journalism. Most of the journalists who participated in this study were proficient in data analysis and computational tools despite their lack of formal training in data science or relevant areas. Nonetheless, we found notable similarities (discussed in 5.1) about the challenges they faced to those of the data science work practices in other related domains [16, 52, 57, 63, 75, 96, 111]. This similarity may arise in part because the journalists currently use similar computational tools to those of expert data scientists. However, some of the challenges emerged directly from the journalism domain itself, suggesting a need for further examination of journalists' experience with various software tools (e.g., Notebooks) and programming languages (e.g., Python, R) [16, 41].

These findings do not necessarily suggest a need to design separate software tools (e.g., Notebooks) specifically for journalists that differ from tools designed for expert data scientists. However,

they do indicate that such tools should be tested further to detect software inclusiveness issues (e.g., inclusive HCI [46, 92, 97]), such that they support users with different computing expertise equally. Such testing can improve the tools' overall usability, not only for user groups such as data journalists, but also, potentially, for experts in data science work [16, 75].

5.3.4 Designing Transparent Algorithmic Interactive UI for Journalism. The design directions suggested above can be explored in combination with statistical or machine learning model interpretability and explainability capabilities [1, 77]. Doing so will not only help journalists be more confident in their data analysis, it will also support reproducible [35] and trustworthy journalism [62]. This approach will help in the practice of transparent journalism [2, 31, 62] by incorporating accountability in the data analysis process [111], both for journalists and also for the audience. It may be possible to adapt existing interpretability and explainability tools, such as InterpretML [77], into interfaces for journalistic data analysis needs. Similarly, transparency about the data sources such as transparent documentation [4] (e.g., data collection, who collected the data and why [32], transformation decisions applied [111]) can be achieved by adapting existing processes [e.g., 38] for journalistic practices. Tying both data analysis and data source transparency can be implemented at the user interface (UI) level for easier use [29]. The term "user" here could focus on UI designed for the journalists, on web interfaces designed for the audience, or for other stakeholders (e.g., direct and indirect [4]). Less technical approaches, such as fictional scenarios [30, 33], that take into account the situated context can be beneficial as well. Because, algorithmic transparency goes beyond the technical functioning of algorithmic systems. Such efforts would require further user studies with various groups of users including journalists to understand its usefulness, acceptance, and the intricacies that might be involved [43, 54, 92]. Again, researchers, designers, and practitioners from diverse domains can all play an important role in exploring these unexplored avenues [31, 60, 77].

5.3.5 Making Decisions and Judgement in the Journalistic Exploration Context Explicit. Passi and Jackson [81] asked an important related question with regards to human activity with data, "Do data science team members have a lot to lose if things do not work or are they allowed to experiment and make mistakes?" As our study suggests, there exists significant similarities between data scientist and investigative journalist data work practices [9, 42, 59]. However, since a journalist's first obligation is conveying truth to the citizen [62], journalists must make sure that the information that they put out in the world is accurate through engaging in detailed verification processes. Thus, in responding to Passi and Jackson [81], the investigative data journalists cannot make mistakes or ignore previous mistakes because they want to make sure that the story they tell is true. Despite the rigorous verification processes our participants used, mistakes still happen and often go unnoticed. When mistakes are noticed, the story gets updated after facts are corrected, and the correction is clearly specified on the news portal.

Moreover, our participants were often fearless, independent, risk takers and hold strong values in their journalistic practices [89]. However, humans do make decisions, and we were able to capture some essence of journalists decision making, but not all of it. Thus, one important future research area along these lines is the examination of how to make any assumptions, judgments [111], or power dynamics [32] during the exploration process more explicit or even explore journalists current approach to such decision making [2, 31, 47, 55]. We should also move our attention from a generalized focus of context to a more specific instance of context/situation (e.g., technical, organizational [33]).

That said, it is unlikely that a single computational pipeline could fit all journalistic needs because "a human investigative journalist can look at the facts, identify what's wrong with the situation, uncover the truth, and write a story that places the facts in context. A computer can't"

[59, p.2], [citing 11]. In other words, full algorithmic automation is not only impossible [27], but also undesirable [13, 25, 59, 107]. Thus, a more advantageous strategy likely involves designing tools targeted to support one specific phase or process, such as the exploration phase on which this paper focuses. Such intentionally focused design can help understand how those tools support, integrate with, and potentially alter existing practices.

6 LIMITATIONS AND FUTURE WORK

This paper's primary limitation is that it is based on a single U.S. based non-profit news organization with a small data team. As noted above, various strategies were used to address this limitation, including multimodal data collection (interviews and sketches), recruiting over half the data-oriented team at JustNews, and a member check interview where the final model was presented to key informants. Furthermore, the alignment between aspects of this model and aspects of related prior work increases our confidence that the proposed model may apply more broadly than only the JustNews organization. That said, future work should examine how these data practices unfold for other news organizations, national, and social context by addressing the following set of questions: How might the motives of profit-focused news production [cf. 18] differently shape the data-driven story idea generation process? Might the fluidity of responsibilities among different roles or the nature of collaborative interactions differ in organizations with larger data journalism teams? How might legal regulations or socio-cultural norms (e.g., in different countries) differently influence aspects such as access to or use of public records data? Lastly, the interviews were conducted over a period of three months, and member checking was completed after four months. Therefore, future study should investigate how the data-driven practice evolves or changes over time for story idea generation.

7 CONCLUSION

This paper proposes a high-level model of how investigative reporters' work with data to generate interesting story ideas based on a qualitative study. The proposed model consists of three highly iterative and cyclic processes: *data acquisition*, *tinkering with data*, and the discovery of *interesting story idea*. We described the situated manner in which the data traveled a journey of context, decision making, and judgement at both interpersonal and technical level before becoming a story. Our study findings suggests that JustNews's investigative reporters' work practices both resemble and differ from those of the work practices of expert data scientists in considerable ways. The proposed model also offers a means of organizing prior work on data journalism, both empirical and technical, and identifying that work's relationship with studies of data science practice. Furthermore, this model helps call out aspects of data-driven journalism that have received less attention and thus may offer opportunities for design intervention.

8 ACKNOWLEDGMENTS

We would like to thank all of our study participants for their time and dedication. We are thankful to the anonymous reviewers and AC's for their valuable comments and feedback. This material is based on work supported in part by the NSF under Grant No. IIS-1844901.

REFERENCES

[1] Microsoft 2020. 2020. What is responsible machine learning? (preview). https://docs.microsoft.com/en-us/azure/machine-learning/concept-responsible-ml Accessed: 2020-11-12.

[2] Tanja Aitamurto, Mike Ananny, Chris W Anderson, Larry Birnbaum, Nicholas Diakopoulos, Matilda Hanson, Jessica Hullman, and Nick Ritchie. 2019. HCI for accurate, impartial and transparent journalism: Challenges and solutions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[3] Eric P.S. Baumer, Xiaotong Xu, Christine Chu, Shion Guha, and Geri K Gay. 2017. When Subjects Interpret the Data: Social Media Non-use as a Case for Adapting the Delphi Method to CSCW. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1527–1543.
[4] Emily M Bender. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. In *Future of Artificial Intelligence: Language, Ethics, Technology Workshop*.
[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.
[6] Ekaba Bisong. 2019. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 59–64.
[7] Chris Bopp, Ellie Harmon, and Amy Voida. 2017. Disempowered by data: Nonprofits, social enterprises, and the consequences of data-driven work. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3608–3619.
[8] Jan Lauren Boyles and Eric Meyer. 2017. Newsrooms accommodate data-based news work. *Newspaper Research Journal* 38, 4 (2017), 428–438.
[9] Paul Bradshaw. JULY 07, 2011. The inverted pyramid of data journalism. https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/ Accessed: 2020-05-22.
[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[11] Meredith Broussard. 2015. Artificial intelligence for investigative reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories. *Digital Journalism* 3, 6 (2015), 814–831.
[12] Meredith Broussard. 2016. Big Data in Practice: Enabling computational journalism through code-sharing and reproducible research methods. *Digital Journalism* 4, 2 (2016), 266–279.
[13] Meredith Broussard, Nicholas Diakopoulos, Andrea L Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. 2019. Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly* 96, 3 (2019), 673–695.
[14] Pew Research Center. JULY 23, 2019. Digital News Fact Sheet. https://www.journalism.org/fact-sheet/digital-news/ Accessed: 2020-03-21.
[15] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.
[16] Souti Chattopadhyay, Ishita Prasad, Austin Z Henley, Anita Sarma, and Titus Barik. 2020. What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[17] Fanny Chevalier, Melanie Tory, Bongshin Lee, Jarke van Wijk, Giuseppe Santucci, Marian Dörk, and Jessica Hullman. 2018. From Analysis to Communication: Supporting the Lifecycle of a Story. In *Data-Driven Storytelling*. AK Peters/CRC Press, 169–202.
[18] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.
[19] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015).
[20] Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Commun. ACM* 54, 10 (2011), 66–71.
[21] DATA.GOV. 2020. The home of the U.S. Government's open data. https://www.data.gov/ Accessed: 2020-02-10.
[22] Hugo De Burgh. 2008. *Investigative journalism*. Routledge.
[23] Temi Developers. November 2019. Advanced speech recognition software. https://www.temi.com Accessed: 2019-11-12.
[24] Nicholas Diakopoulos. 2013. Sex, Violence, and Autocomplete Algorithms. *Slate, August. http://www. slate. com/articles/technology/future_tense/2013/08/words_banned_from_ bing_and_google_s_autocomplete_algorithms. html* (2013).
[25] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).
[26] Nicholas Diakopoulos. 2018. Ethics in Data-Driven Visual Storytelling. In *Data-Driven Storytelling*. AK Peters/CRC Press, 233–248.
[27] Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media*. Harvard University Press.
[28] Nicholas Diakopoulos. 2020. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digital Journalism* (2020), 1–23.
[29] Nicholas Diakopoulos. 2020. Transparency. In *The Oxford Handbook of Ethics of AI*.
[30] Nicholas Diakopoulos and Deborah Johnson. 2019. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society* (2019), 1461444820925811.

[31] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic transparency in the news media. *Digital Journalism* 5, 7 (2017), 809–828.

[32] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism.* MIT Press.

[33] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. *arXiv preprint arXiv:2101.04719* (2021).

[34] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2952–2963.

[35] Melanie Feinberg, Will Sutherland, Sarah Beth Nelson, Mohammad Hossein Jarrahi, and Arcot Rajasekar. 2020. The New Reality of Reproducibility: The Role of Data Work in Scientific Research. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–22.

[36] Center for Public Integrity. 1989. Center for Public Integrity. https://publicintegrity.org/ Accessed: 2020-03-22.

[37] Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of peace research* 2, 1 (1965), 64–90.

[38] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[39] Liqiang Geng and Howard J Hamilton. 2006. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)* 38, 3 (2006), 9–es.

[40] Barney G Glaser, Anselem L Strauss, and E Strutzel. 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research New York Aldine De Gruyter.

[41] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: how journalists can use data to improve the news.* " O'Reilly Media, Inc.".

[42] Philip Guo. OCTOBER 30, 2013. Data Science Workflow: Overview and Challenges. https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext Accessed: 2020-08-22.

[43] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[44] Youyang Hou and Dakuo Wang. 2017. Hacking with NPOs: collaborative analytics and broker roles in civic data hackathons. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–16.

[45] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics* 19, 12 (2013), 2406–2415.

[46] idrc. 1975. Inclusive Design Research Center. https://idrc.ocadu.ca/ Accessed: 2021-10-02.

[47] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric P. S. Baumer. 2019. Where is the human? Bridging the gap between AI and HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[48] Aniket Jain, Bhavya Sharma, Paridhi Choudhary, Rohan Sangave, and William Yang. 2018. Data-Driven Investigative Journalism For Connectas Dataset. *arXiv preprint arXiv:1804.08675* (2018).

[49] Mother Jones. February 1976. Mother Jones. https://www.motherjones.com/ Accessed: 2020-03-22.

[50] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2020-07-08.

[51] Karen Kaiser. 2009. Protecting Respondent Confidentiality in Qualitative Research. *Qualitative health research* 19, 11 (Nov. 2009), 1632–1641. https://doi.org/10.1177/1049732309350879

[52] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926.

[53] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*. 39–52.

[54] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[55] Jakko Kemper and Daan Kolkman. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society* 22, 14 (2019), 2081–2096.

[56] Mary Beth Kery, Bonnie E John, Patrick O'Flaherty, Amber Horvath, and Brad A Myers. 2019. Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[57] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 96–107.

[58] Yea-Seul Kim, Nathalie Henry Riche, Bongshin Lee, Matthew Brehmer, Michel Pahud, Ken Hinckley, and Jessica Hullman. 2019. Inking Your Insights: Investigating Digital Externalization Behaviors During Data Analysis. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. 255–267.
[59] Keith Kirkpatrick. 2015. Putting the data science into journalism.
[60] Marina Kogan, Aaron Halfaker, Shion Guha, Cecilia Aragon, Michael Muller, and Stuart Geiger. 2020. Mapping Out Human-Centered Data Science: Methods, Approaches, and Best Practices. In *Companion of the 2020 ACM International Conference on Supporting Group Work*. 151–156.
[61] Robert Kosara and Jock Mackinlay. 2013. Storytelling: The next step for visualization. *Computer* 46, 5 (2013), 44–50.
[62] Bill Kovach and Tom Rosenstiel. 2014. *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA).
[63] Sean Kross and Philip J Guo. 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[64] Bongshin Lee, Matthew Brehmer, Petra Isenberg, Eun Kyoung Choe, Ricardo Langner, and Raimund Dachselt. 2018. Data visualization on mobile devices. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–8.
[65] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications* 35, 5 (2015), 84–90.
[66] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1483–1493.
[67] John Lofland and Lyn H Lofland. 1971. Analyzing social settings. (1971).
[68] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. In *Computation+ Journalism Symposium*.
[69] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the news: Digital creativity support for journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.
[70] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.
[71] Annette Markham. 2012. Fabrication as Ethical Practice. *Information, Communication & Society* 15, 3 (April 2012), 334–353. https://doi.org/10.1080/1369118X.2011.641993
[72] Kelly McBride and Tom Rosenstiel. 2013. *The new ethics of journalism: Principles for the 21st century*. CQ Press.
[73] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.
[74] Cecelia Merkel, Umer Farooq, Lu Xiao, Craig Ganoe, Mary Beth Rosson, and John M Carroll. 2007. Managing technology use and learning in nonprofit community organizations: methodological challenges and opportunities. In *Proceedings of the 2007 symposium on Computer human interaction for the management of information technology*. 8–es.
[75] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 126.
[76] Michael Muller, Christine Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. *Submitted for publication to CHI* (2021).
[77] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).
[78] nostoros. 2018. CONNECTAS. https://www.connectas.org/sobre-nosotros/ Accessed: 2020-08-10.
[79] The International Consortium of Investigative Journalists. 1997. International Consortium of Investigative Journalists. https://www.icij.org/investigations/ Accessed: 2020-03-22.
[80] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2436–2447.
[81] 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 (2018), 1–28.
[82] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (2020), 2053951720939605.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[84] ProPublica. 2007. ProPublica. https://www.propublica.org/ Accessed: 2020-03-22.
[85] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. 2018. *Data-driven storytelling*. CRC Press.
[86] Simon Roger. July 2011. Data Journalism at The Guardian. https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism Accessed: 2020-03-22.
[87] Adam Rule, Aurélien Tabard, and James D Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[88] Benjamin Saunders, Jenny Kitzinger, and Celia Kitzinger. 2015. Participant Anonymity in the Internet Age: From Theory to Practice. *Qualitative Research in Psychology* 12, 2 (April 2015), 125–137. https://doi.org/10.1080/14780887.2014.948697
[89] Ivor Shapiro. 2010. Evaluating journalism: Towards an assessment framework for the practice of journalism. *Journalism Practice* 4, 2 (2010), 143–162.
[90] Alicia C Shepard. 2007. *Woodward and Bernstein: Life in the shadow of Watergate*. J. Wiley.
[91] Dilruba Showkat. 2018. Determining newcomers barrier in software development: An IT industry based investigation. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 165–168.
[92] Dilruba Showkat. 2021. Tinkering: A Way Towards Designing Transparent Algorithmic User Interfaces. In *Joint Proceedings of the ACM IUI 2021 Workshops*.
[93] Dilruba Showkat and Eric P. S. Baumer. 2020. Outliers: More Than Numbers?. In *Companion of ACM CSCW 2020 Workshops*.
[94] Upton Sinclair and Earl Lee. 2003. *Jungle: The Uncensored Original Edition*. See Sharp Press.
[95] Jennifer A Stark and Nicholas Diakopoulos. 2016. Towards editorial transparency in computational journalism. In *Computation+ Journalism Symposium*.
[96] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, and David Redmiles. 2015. Social barriers faced by newcomers placing their first contribution in open source software projects. In *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*. 1379–1392.
[97] Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2020. Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human–Computer Interaction* 13, 1 (2020), 1–69.
[98] Saiki Tanabe. 2017. Sketchboard. https://sketchboard.io/ Accessed: 2020-03-11.
[99] Steven J Taylor, Robert Bogdan, and Marjorie DeVault. 2015. *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons.
[100] Data Source Triangulation. 2014. The use of triangulation in qualitative research. In *Oncology nursing forum*, Vol. 41. 545.
[101] Sherry Turkle and Seymour Papert. 1990. Epistemological pluralism: Styles and voices within the computer culture. *Signs: Journal of women in culture and society* 16, 1 (1990), 128–157.
[102] David B Tyack and Larry Cuban. 1995. *Tinkering toward utopia*. Harvard University Press.
[103] Amy Voida, Ellie Harmon, and Ban Al-Ani. 2011. Homebrew databases: Complexities of everyday information management in nonprofit organizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 915–924.
[104] Shirin Vossoughi, Meg Escudé, Fan Kong, and Paula Hooper. 2013. Tinkering, learning & equity in the after-school setting. In *annual FabLearn conference. Palo Alto, CA: Stanford University*.
[105] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How data scientists use computational notebooks for real-time collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[106] April Yi Wang, Zihan Wu, Christopher Brooks, and Steve Oney. 2020. Callisto: Capturing the" Why" by Connecting Conversations with Computational Narratives. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[107] Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. 2021. How Much Automation Does a Data Scientist Want? *arXiv preprint arXiv:2101.03970* (2021).
[108] Robert S Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster.
[109] John Wihbey, Thalita Dias Coleman, Kenneth Joseph, and David Lazer. 2017. Exploring the Ideological Nature of Journalists' Social Networks on Twitter and Associations with News Story Content. *arXiv preprint arXiv:1708.06727* (2017).
[110] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on*

Human Factors in Computing Systems. 1–12.
[111] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

A BRIEF STORY DESCRIPTIONS

A.1 Housing Stories: Tax Breaks in Housing

These housing projects are government funded projects and meant to help the poor. The projects involved working with a public housing dataset and geo-location data. The housing dataset contains features such as housing project ID, name, address, state ID, and contact information. However, public housing datasets do not provide geo-location information specifying where these projects are actually built, such as economic information (e.g., job, education, crime rate) status. For example, economic information might say a lot about how new construction could impact the kinds of opportunities people might obtain by living in those areas. These geographical location information is collected via ArcGIS API. The reporters need to connect the two datasets to discover where these projects are physically located. In a related housing story, Jakob found that a low-income housing project was built in an area which lacked the proper infrastructure for economic growth. Thus, the gap between the poor and the rich never improved.

A.2 Threat Prediction Story

In this story, a threat detection device was installed at several places in the US. However, little information was available about the device underlying working, for instance, what algorithm was used or what dataset the device was trained on. The reporter tested the device in several steps. First, using standard dataset, the analysis showed skewed results for certain demographics. The other two tests (simulated test, and a real-world test) also showed similar results. Therefore, the device was unable to predict threat accurately.

A.3 Healthcare Data Story

This story used healthcare dataset, where each doctor code their patient visits in the ranging from low (0) to high (5), depending on how involved the patient assessment is. Examining this dataset revealed that a few doctors were charging Medicare to a higher rate, meaning they coded most of their patient visits at high as 5, compared to other doctors in the office for the same category (e.g., OBGYN).

A.4 Election Campaign Data News Application/Story

This project makes use of public election campaign finance datasets. The datasets are parsed and formatted before inserting into the database tables. The reporter then queries the database by running manual SQL queries to identify and detect interesting data points, such as specific transactions of a particular amount for a particular race or committee. For example, they make comparisons about the total amount spent this year compared to previous years for a specific time period by running aggregate SQL queries. The outcome from these activities could lead to either new features for the news application or could become a story (also shown in Figure 5). New features are added in the news application when the reporter finds that the process can be automated for displaying on the web.

B EXAMPLE ANNOTATED SKETCH

Figure 7 shows an excerpt from the sketch that Jakob drew for the Low-income Housing and Tax Breaks story, annotated with themes from the analysis. As described in the Results, Jakob was trying to automate the reporter Sonya's manual data work. The high level story idea was known in

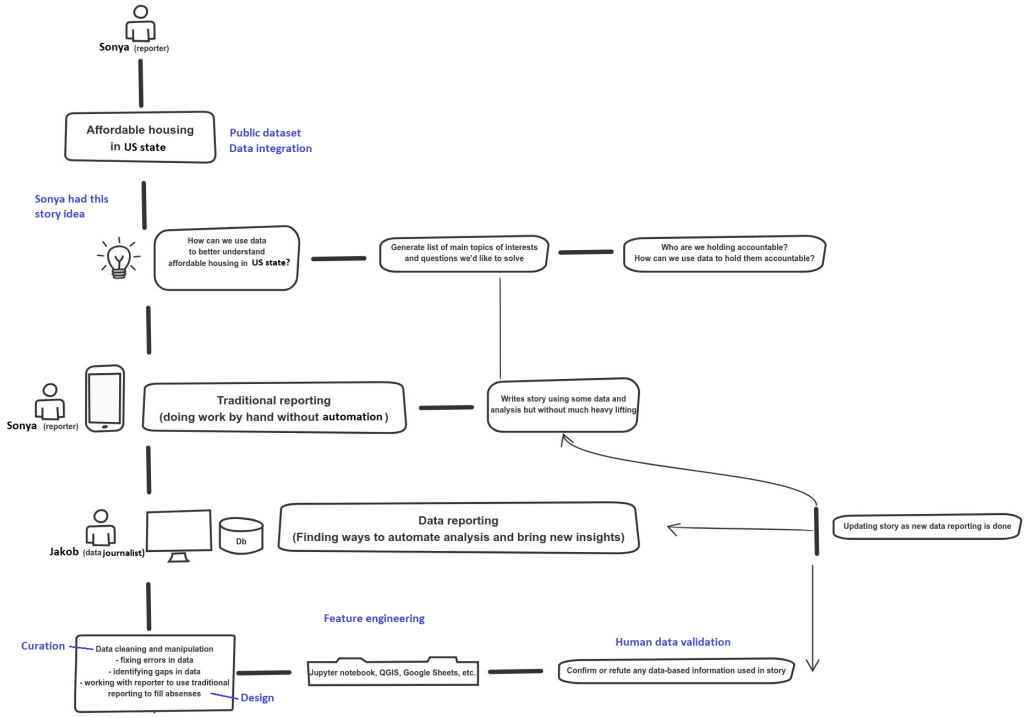

Fig. 7. Annotated sketch drawn by Jakob when describing the data work activity in the Low-income Housing story. Blue text indicates researcher annotation.

advance. Jakob was trying to check the data work of Sonya by running database queries, to make sure whether he gets the same data points (who was held accountable in the story) as the reporter Sonya. The story line is similar to the story described in A.1.

Received October 2020; revised April 2021; accepted July 2021