# Interrogating Data Work as a Community of Practice

ANNABEL ROTHSCHILD, Georgia Institute of Technology, USA
AMANDA MENG, Georgia Institute of Technology, USA
CARL DISALVO, Georgia Institute of Technology, USA
BRITNEY JOHNSON, Georgia Institute of Technology, USA
BEN RYDAL SHAPIRO, Georgia State University, USA
BETSY DISALVO, Georgia Institute of Technology, USA



We apply Lave & Wenger's construct of a community of practice to identify and position members of the data work community of practice, focusing on members on the periphery who have received less attention – as compared to full practitioners (e.g., data scientists). Reporting on results of interviews with 19 civic workers who perform data work as their main task, we identify an atypical relationship between subject-domain experts (such as our interviewees) and full members of the data work community. Our interviewees may have less computational skill in data work, but they have extensive and varied practices to engage in data contextualization that data scientists and other full community members could learn from. In identifying the attributes of data workers on the periphery, we also hope to call attention to the challenges they face in performing data work in low resources institutions (e.g., governmental, non-profit). Our findings contribute to the larger conversations in human-centered data science about who performs data work and how they go about it, in order to addresses questions of power, fairness, and bias in data-intensive systems.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**.

Additional Key Words and Phrases: data work, data science, community of practice, human-centered data science

**ACM Reference Format:**
Annabel Rothschild, Amanda Meng, Carl DiSalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Interrogating Data Work as a Community of Practice. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 307 (November 2022), 28 pages. https://doi.org/10.1145/3555198

## 1 INTRODUCTION

We know that participants in data work are not limited to data scientists – like any professional community, there is a diverse bevy of practitioners; there are data wranglers [32, 33] or domain experts who engage in data work to facilitate advances in their subject area [41], alongside data scientists [7, 42]. There have also been extensive explorations of how individuals in other settings, such as nonprofits [3] and healthcare [59] engage with data; however, these works describe how data work – as a secondary task – affects mission-drive work and organizational agents rather than understanding relevant individuals as practitioners of data work and addressing their needs and experiences as such. We propose adopting the community of practice framework to identify

Authors' addresses: Annabel Rothschild, arothschild@gatech.edu, Georgia Institute of Technology, Atlanta, USA; Amanda Meng, a.meng@gatech.edu, Georgia Institute of Technology, Atlanta, USA; Carl DiSalvo, cdisalvo@gatech.edu, Georgia Institute of Technology, Atlanta, USA; Britney Johnson, bjohnson324@gatech.edu, Georgia Institute of Technology, Atlanta, USA; Ben Rydal Shapiro, bshapiro@gsu.edu, Georgia State University, Atlanta, USA; Betsy DiSalvo, bdisalvo@cc.gatech.edu, Georgia Institute of Technology, Atlanta, USA.


2573-0142/2022/11-ART307
https://doi.org/10.1145/3555198

and think about the various groups engaging in data work [38], as the framework yields a way to understand the intricacies of a given community of practitioners. Community of practice dictates that rather than using a central and exterior binary to think of data work, we see membership in the community as a gradient. Groups like data scientists and data analysts can be considered full (i.e. experienced practitioners) members and groups like data wranglers as peripheral members, or novice participants who, upon gaining more skill and experience, may or may not join one of the groups of full practitioners. Given the variety of titles used for computationally-intensive data practitioners, we isolate only these two terms to suggest one role that has, theoretically, experience in working on datasets with diverse subjects (data scientists) and one role that focuses on a subset of data specific to a given organization or subject area but is still computationally intensive (data analyst). When examining data work from this perspective, there is a disproportionate amount of scholarship related to full participants in comparison to peripheral members.

Why does our relative lack of knowledge about peripheral data workers matter? In the last few years, the CSCW community has come to understand data work as a human-centered process subject to discretion and interpretation [20, 48, 50, 53, 56]. There are even new subsets of data science based on the need for nuance when working with data about individuals, such as "human-centered data science" [37]. Moving the focus from the textual data exclusively to the way that human data workers interact with and process that data has key implications for fairness and transparency in data-driven systems; as a research community, we know that data is never “raw” (and to describe it as such "is both an oxymoron and a bad idea", in the words of Bowker [4]) and must be understood in context [25]. Engaging in active efforts to both document and preserve the context of datasets is an important first step towards designing equitable and just data-intensive systems; however contextualization often falls to the wayside in favor of other facts such as ”universality”, as Klaus Scheuerman et al. describe in their study of computer vision datasets [63]. Contextualization contributes to what Monroe-White describes as *emancipatory data science*, or "data work that frees members of marginalized communities from being the ‘object’ to the ‘subject’ of data science framings" [49]. Monroe-White describes that these practices will help eventual dataset users understand the "decisions regarding why, how, what, when, and where data are collected, managed, analyzed, interpreted and communicated" and that contextualization should involve those who are the subject of the data being used [49]. Loukisass, too, questions the role of "locality" in data [39] in this manner, while D’Iganzio and Klein suggest a feminist-informed handling of data [13].

Even the most minor attributes of a dataset’s "texture", or the relationship between the dataset’s infrastructure and environment, has weight on the correct (or well-informed) contextualization [21]. Only by understanding the motivations and actions of the human actors can we begin to reason about the functionality of such systems in their entirety. As there is an increasing call for XAI (explainable artificial intelligence) which requires uncovering the social context of these "human-AI assemblages" [16], the need for tools and process evaluations to support such work is only growing.

To kick-start this investigation of human-centered data work, Muller et al. established several open, key questions about the role of humans in data science [52]. Figure 1 shows the original text of these questions, with our summary category tags; throughout this paper we address these questions by their shorthand tags. Answering these questions pushes us both towards a scholarly understanding of these work practices and the development of tools and processes to support a broader array of data work. Our research attends to those who perform data work, but are not data scientists in the common use of that term, or other fairly-full practitioners (e.g., data analyst, data engineer, etc.). These workers are experts in other domains, but they collect, clean, analyze, and share data as an essential part of their work and identify data as a principle task of their role. In

Table 1. Guiding questions for studies of data workers, from Muller et al. (2019):

| Question number | Question, as posed by Muller et al. | Question tag |
|---|---|---|
| 1 | How do data science workers approach their tasks? What are their strategies? What can we learn about their views of data and process? | How? |
| 2 | What (and who) is missed in our prevailing accounts and assumptions around data and data work? If "data is never raw", who does the cooking, and how is this work performed, recognized, and accounted for? | Who? |
| 3 | Data science tools are generally designed for one user at-a-time. Other complex tasks have benefited greatly from collaboration practices and collaboration technologies. Are there opportunities to bring some of these lessons to data science? What are the distinctive collaboration needs and constraints in data science? | Tools for collaboration |
| 4 | What methods are valid and tractable to assess the usefulness and usability of tools for data scientists in service of iterative design? How can learnability be assessed when domain experts, not programmers, need to learn new languages and tools simultaneously, the learning of which takes far longer than typical usability studies? How can efficiency for skilled users be assessed when the uses of the tools are so diverse that benchmark tasks have little face validity? | User-based tool assessment |
| 5 | Who are the consumers of data science work and what are their needs? How can data science processes and tools address these needs, e.g., the level of transparency and comprehension consumers desire? | Implications of data work |

other words, they engage in activities that are familiar to data science but they are not, generally speaking, data scientists. We believe attending to these workers and practices is important for fostering a diverse approach to what counts as data work, which in turn aligns with commitments to equitable and just principles of labor and broadening participation in computing.

In this paper we describe the results of interviews with 19 individuals engaged in civic data work. We discuss both the specific needs and assets of this group of data workers and why their experiences and methodologies can both inform the tools we build for other data worker populations (such as data scientists) and call for professional tools and systems that address their needs. All work in the civic sector – either directly for local and state government or for non-profits in large cities in the United States. They work on "civic data" which Sinders describes as data that reflects "people who live in the communities and cities where the data was gathered" [64] and we extended to include datasets that describe the infrastructure and municipal services associated with those cities. Most of are participants first and foremost domain experts in another field, for example, the main cause of the non-profit they work for, and perform data work that amplifies and extends that mission. Within their organization they would be described as 'the data person', or the individual with the most computational data savvy; they are the primary data workers within their respective

organizations, showing the outsize impact of their work on their data ecosystems. When civic data ecosystems function correctly, especially when they increase transparency through open-access initiatives, they have myriad beneficial social implications [43, 45]. The contrapositive also holds: as Irani and Marx describe, "[t]he withholding of public information obscures and obstructs the democratic process by denying ordinary people the right to know what's being done in their name." [30]. Further, given the nature of their work, the clients of these organizations and governmental entities do not have much of a choice when engaging with these services, making them important sites of transparency investigations. Citizens' private data, too, is increasingly becoming public data in the era of revolutionary civic tech – as described by Boehner & DiSalvo [2] – which further compounds the need to understand how these data workers go about their work. Understanding this group of data workers, too, lends us insight about what data work in low and limited resource environments looks like, in contrast to the large technical organizations many participants in past literature were employed by.

Our results inform several of Muller et al.'s questions, with respect to civic data workers specifically, but extend to larger group of data workers more generally. First, our participants shed light on how data work is done in the civic contexts, which compared with high-tech data science efforts are relatively low-resourced (Question 1). We also discuss their applied and function-oriented perspectives on data. Second, we highlight these workers as "peripheral data workers" and demonstrate their role in major data ecosystems (Question 2). We use the term "peripheral" in a technical sense to call attention to how the tools and processes of these workers are at the margins of what is commonly considered as the data science community of practice. While peripheral data workers have been missing from much of the scholarship, we argue that they play as consequential a role as their full professional peers, namely data scientists at large technical organizations. Third, we describe their collaboration practices that are uniquely shaped by their comparatively low-resource work environments, replete with outdated tools and little ability to acquire new ones (Question 3). Fourth, we describe the tool acquisition practices of this group of domain experts, rather than formally trained technical workers; namely, that given their work environments and educational background, they are limited in choice of, and training for, new tools (Question 4). Fifth, and perhaps most importantly, we describe the ways in which this group of peripheral data workers goes about contextualizing the datasets they work with and the novel formal and ad-hoc systems they develop to do so (Question 5). These practices are informed by the proximity the data workers have had built into their workflows or have actively sought and or developed with the data.

We close with two insights for the CSCW community. First, a need to continue to expand our definition of who is a data worker and, subsequently, how the needs of data workers might vary between subgroups. We identify and position various subgroups of data workers via the community of practice mapping, which highlights a more complex relationship between full and peripheral practitioners than professional vs aspiring. We further suggest that rather than fetishizing the advanced computational skills of data scientists as model community members (and subsequently positioning them as the community members to emulate), we consider at the ways in which data scientists can learn from other members of the data work community, particularly with regards to data contextualization. Second, by examining one of the "peripheral" groups of data workers, we describe specific limitations hindering the use of tools and systems by data workers in low and limited resource environments and propose ways to address them.

## 2 RELATED WORK

The CSCW research community has paid increasing attention to how data scientists go about their work in recent years. Our work furthers that effort, introducing the domain-experts-as-data-workers and their contextualization practices, which hold promise for other, more full participants

in data work. Accordingly, our work builds off three areas of work: understanding the role of humans in data work, identifying who participates in data work, and finally, how the currently studied data workers go about their work.

### 2.1 Human-centered data science

While data science – and data work more broadly – has traditionally been seen as a "rational 'data-driven' process of 'discovery' that reveals the underlying nature of a domain" [53], this is no longer the case. Increasingly, we understand the way that humans performing data work shape the understanding and dissemination of the data they work with. Feinberg describes individual data workers as "designers" of data, rather than objective appropriators, highlighting the role the that human perception, background, and motivation play in the way dataset contents come to be understood [18]. Even the act of reading a database (making sense of how information is organized within it) is an act of "awareness, reflection, and control" [19]. Data work is also a collaborative practice subject to social interaction and dynamics: Koesten et al. explore how individuals use a series of patterns of activities in order to engage in data sensemaking, describing the cognitive and verbal practices that take place [36]. Miceli et al. expand on the power dynamics that take place in data work, highlighting the case of data annotation in computer visualization datasets; they call it a "sensemaking practice" that is frequently influenced by the labeler's higher-ups, whose decisions sometimes conflict with the labeler's factual or logical inclinations [46].

However, much more attention is paid to improving models than improving datasets, even though data provides an upper bound for quality in machine learning and algorithmic systems [31]. One of the biggest issues with data is the lack of proper contextualization. The implications go beyond reduced accuracy and poor model fitting: a lack of contextualization leaves room for issues of bias and fairness. When issues arise in datasets they tend to result in data cascades, or compounding, negative events that compromise the quality of the data as a whole, as termed by Sambasivan et al. [62]. D'Ignazio & Klein underscore the importance of contextualization practices, introducing reflexivity as a necessity for restoring context in data work. They identify "social, cultural, historical, institutional, ... or material," and who participated in the creation and curation of that dataset as starting points [13]. Loukissas, too, argues that contextualization of data requires an intimate understanding of the environment from which it originates [39]. When the associated origins and context of a dataset become separated from the textual data, it is easier to misuse or misapply a dataset. For example, consider a model to predict increases in home value in urban areas that fails to account for decades of housing discrimination and inequality in parts of those cities.

Miceli et al. identify how this context restoration praxis might take place specifically for large computer vision datasets and the individuals involved in labeling them [47]. There are other contextualization practices proposed for large-scale datasets: Bender & Friedman proposed data statements, essentially a short biography, to accompany NLP datasets [1] and Gebru et al. propose datasheets for datasets which include, among other information, "motivation, composition, collection process, recommended uses" to help facilitate conversation between dataset creators and consumers in the style of fact sheets that accompany electronic equipment [24].

### 2.2 But who does the (data) work?

One limitation to the innovative contextualization practices described above is who they target. While large-scale machine learning datasets are undoubtedly important places for contextualization practices to take place, they are not the only place. Further, these practices are aimed at researchers and practitioners in large technical organizations, who have both access to a community of skilled peers and ample opportunities for continued training. If we look at the data work as a *community of practice* including, but not limited to, individuals who meet this criteria – such as data scientists

who work with large scale datasets in high-tech organizations – it becomes clear there are other groups who could benefit from these contextualization initiatives. They demonstrate new modes of data contextualization and help us refine the contextualization tools and practices we develop for data scientists.

Lave & Wenger define a community of practice as the participants in a shared activity, at varying levels of experience and skill [38]. The community is defined not by the kind of work being practiced – many kinds of knowledge-building communities can be understood as such, formal or informal – but by the "structure and character of [the] community [that] emerges" [6]. Critically, a community of practice has full participants (experienced, highly knowledgeable members of the community) and peripheral ones (those who are learning the craft). Lave & Wenger give examples of communities of practice based on physical craft, e.g. tailors and midwives, who have more and less formal apprenticeship practices respectively. The community of practice construct has been applied in the computing space previously, for example to understand Wikipedia editors [6] or to grow the global HCI practice [66].

When we examine data work as a community of practice, we can make sense of data scientists – and similar data roles that prioritise computational skills over work in the subject domain – as full practitioners. It is these full practitioners who have been the subject of much of the CSCW community's focus, rather than members who may be on the periphery of data science but are full member of a more inclusive data work community of practice. Data science pulls from several closely-aligned fields, including information science, statistics, computing, and knowledge domains [42] and peripheral data workers are likely to reside at the intersection between one of these aligned fields and data science proper. Peripheral data workers engage in some subset of the full member's practice, while constantly becoming closer to full members as they hone the skills prized by the data work community. Data wranglers are an example of another group of the periphery of data work, as they focus on the early cleaning and organizational steps of data work [32, 33], while full practitioners – e.g., data scientists – have experience with the entire data lifecycle. Previous research has shown how individuals outside the data work community could join it. For example, young children have become data science apprentices, learning the skills of data production, analysis, and consumption [11, 27, 28, 72]. We do note that these are not professional practitioners, rather they are hobbyists or early-stage learners. Content moderators are another such group: while some do so at a professional level and have little agency over what happens to the data after they pass it off [61], others are volunteer moderators, such as [34] have more agency over the platforms they work with, but are usually not receiving compensation or professional evaluation for their work. Feinberg et al.'s reproducibility workshops for scientists from domains besides computing demonstrates a more professional apprenticeship experience [22]. Other professionals encounter data work as part of their profession, though it is not their main role, e.g., health care workers and the challenges they incur while doing so [59]. The data workers we discuss are unlike these previous groups in that they are not necessarily trained in data work, but perform it as a part of their compensated, professional role. Further, they consider data work to be a primary component of their role, but they rely heavily on their subject area expertise.

Building a broader understanding of data work as a community of practice may also address issues of collaboration. The work of Feinberg et al. also highlights the division between domain experts and technically skilled workers; those with extensive domain expertise have spent their careers learning the ins and outs of their subject, while groups like data scientists have spent theirs expanding their technical skill. Bringing the two groups together is not easy. For example, data work produces friction in the roles of health care providers [59]. According to Crisan et al., full practitioners of data work, who may have a variety of job titles, have varying levels of domain expertise with which to perform contextualization work [9]. When external domain experts are

brought in to consult on data work projects, however, the process is challenging and often fraught with technical and time limitations, as described by Mao et al. [41]. Therefore, the experiences and needs of those performing data work on the periphery differ from those of full community practitioners, including data scientists.

Most insights we have from ethnographic or human-centered studies about data workers, which inform this kind of contextualization work, focus on these full practitioners. For example, Pereira et al. study data workers in diverse fields and find that they have similar tool needs and challenges despite different domains, but their participant population all have focused degrees in data work and all but one have titles that explicitly include the word "data" (e.g., data scientist, data analyst, or data architect) [57]. Wang et al. provide insight on the way data workers use computational notebooks, but their survey population is comprised of students in data science or computer science degree programs, or professionals with an extensive technical data science or computer science background [69]. Zhang et al. uncovered the depth of collaboration in data science, but their respondents were a self-selected population of data workers at IBM [73]. Bopp et al. are an exception to this; they explore the impacts of data work on non-profits [3]. However, they find that the focus on data disempowers organizations, while our respondents are part of organizations that have long embraced data and do not struggle with its role in their organization.

This oversized focus on highly skilled technical workers as the targets of data work ethnographies and tools is not surprising. Data science and its practitioners have demanded attention in recent years from all corners of society, let alone this research community. However, these highly skilled technical workers are both a limited portion of a much larger population of data workers, and they have been only recently subjects of contextualization inquiry. We theorize that this is the result of data science being used interchangeably with data work in common parlance, as well common definitions of a data scientist's required capabilities listing domain expertise as a final add-on or one skill amongst many others [7, 40]. This deprioritization could be due, in part, to variability between domain expertise required by different roles and difficult defining it. For example, the EDISON framework developed to create a common definition of data science skills and competencies prioritizes technical skills over domain knowledge [12].

We postulate that many of those engaging in data work – perhaps even the majority – would not be considered full practitioners of data science. Data wranglers, for example, would be seen as peripheral at most to the data science practice, however they have a much fuller role in the data work community. All groups that are peripheral to data science are part of the data work community, in which they may still be peripheral or full participants. Peripheral data workers are to full data workers as paramedics are to physicians; paramedics perform applied, acute care for a subset of maladies, where physicians are responsible for a broader array of conditions and conducting further research into the state of the art. In many cases, these groups of peripheral data workers are first and foremost domain experts in their area of expertise and secondarily practice data work as it pertains to their domain. They are more likely to work in low or limited resource environments and, corresponding to their domain expertise, are often in close conversation with those involved in the data's origin, or they are directly involved in that process. This leaves us with two questions, which our work will begin to answer: 1) how well are other groups of data workers, such as those on the periphery of data work, engaging in contextualization practices, and 2) do they have contextualization practices that could be adopted by more full participation in data work (such as data scientists)? Further, given our focus on this group of peripheral data workers, we highlight the tool needs of this specific group and those in similar low or limited resource technical environments through design suggestions. These are pertinent questions especially in the face of the increasing automation of data work, for example, the AutoDS project, and the likelihood that they be widely adopted [14, 70, 71].

## 3 METHODS

To understand how those engaging in informal data work approach and understand their role, we interviewed 19 individuals between October 2018 and July 2019. These participants work in urban areas in the United States either for local or state government or other civic organizations. The interview protocol was approved by our Institutional Review Board and all participants consented to the study. Our process was based on the grounded theory approach [67].

*3.0.1 Researcher Positionality.* The authors on this paper are academic researchers who have worked with civic and non-governmental organizational data through assorted research projects and industry positions. We approached this project as researchers on the DataWorks project, a work-training program to broaden participation in middle-skill data work; our intention was to understand how data is collected, organized, and processed in civic and non-governmental organizational contexts – or, environments outside of data-centric or data-focused technical organizations.

*3.0.2 Participant selection.* The criteria for participant selection included individuals who described their job as data work, or who would refer to themselves as "the data person" within their organization but did not work for a for-profit company. Specifically, we sought individuals those working for local governmental and civically aligned non-for-profit organizations. Potential participants were identified through the research team's local connections and network. Additional participants were identified through snowball sampling [26], with early participants suggesting others. Recruitment happened through email and word of mouth. Participants were not compensated, as many were local or state government employees and could not accept compensation for interactions that took place while they were in their professional capacity.

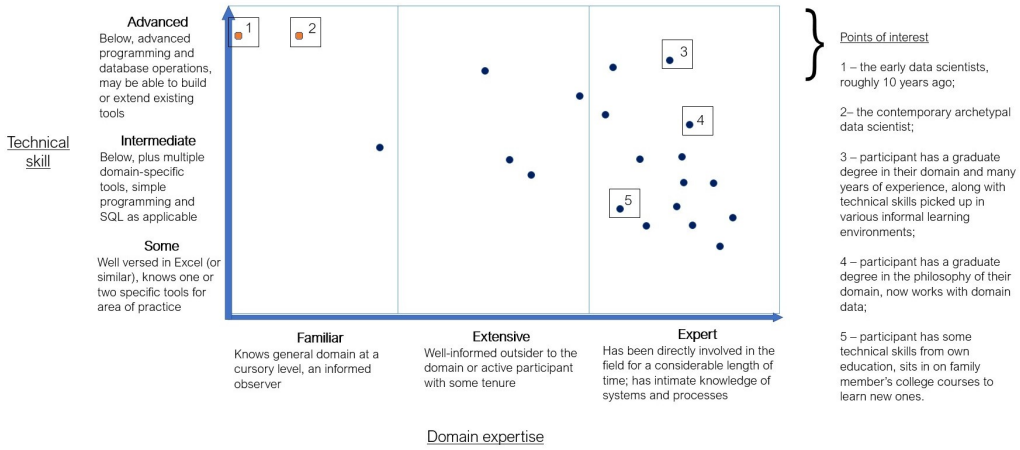
*3.0.3 Interview structure.* The second author lead all interviews. During interviews participants were asked a series of questions following a semi-structured interview style [15] and asked to describe their desk set up and asked to draw a diagram of example data flow within their organization, with these last two tasks styled as situated observation [65]. Interviews were recorded and transcribed. Interviews lasted between 28 min and 89 minutes with a total of 14.4 hours of interview recordings analyzed. Four of the interviews were conducted in pairs with two individuals who work on the same team; P8 and P9 were interviewed together, as were P10 and P11.

*3.0.4 Data analysis.* Transcripts were analyzed pursuant to the open coding framework [23] by the research team through multiple reviews and revisions. First, members of the research team reviewed different transcripts and discussed themes that emerged to develop the initial code book. Second, the first author then reviewed all of the transcriptions, began coding them with the initial code book and returned to discuss with the research team potential modification and nuances to change the code book based upon findings.

Third, after themes were thus refined, one researcher the first author coded all of the transcripts into four themes. Fourth, then two researchers the first and last authors reviewed each excerpt that was identified for each code and discussed how it applied. If there was disagreement as to which codes were applicable researchers discussed and refined codes to help with clarification. Fifth, the first author then applied the final version of the code book to the full corpus, with the last author providing peer review on random excerpts from each code to ensure continuity and reviewed excerpts that were difficult to code. Sixth, the research team then collaboratively identified themes within these the set of excerpts for each code, following thematic analysis [5].

Given the open-ended nature of many of the interview questions, responses went in multiple directions. In recognition of this variance, we generally avoid reporting quantitative metrics about participant responses and instead highlight reoccurring themes within categories of responses.

Fig. 1. Participants' backgrounds, in terms of technical skill – as defined in traditional data science verbiage – and domain expertise. Points 1 and 2 are red to denote that they are comparisons to our participants, where all participants points are dark blue.




## 4 FINDINGS

We separate our analysis into two sections, namely participants' backgrounds and then observations of how they go about their work. This division roughly corresponds to the dual purpose of this paper. First, we identify this group of dataset domain experts and their ability to contextualize datasets. These contextualization practices can be used to work against bias in automated systems by documenting limitations of the dataset. Second, we observe their work environments and experiences with civic data and provide insight about how this community of tool designers and ecosystem ethnographers might best serve these underserved members of the data science community.

When discussing participants, we purposely choose to obscure their identities as much as possible, given both the sensitive nature of their work and to allow them to be critical of the environments in which they operate. Therefore, individuals' backgrounds, education, and demographics are reported in aggregate and we use the gender-neutral "they" pronoun throughout.

### 4.1 Participants

*4.1.1 Participant backgrounds.* Our participants vary in gender and race from national averages in computer and information technology workforce participation [55]. Roughly 30% of our participants are female-identifying, 70% male-identifying, with the former slightly higher than national averages. Participants identify as Black or African American (n=7), Asian American (n=1), and White (n=11); this varies greatly from national averages. They mostly ( 80%) hold graduate degrees in their domain areas – for example, a Masters in Public Administration for someone who works in government – and the remainder hold Bachelors degrees in their domain area.

In Figure 1, we plot the relative domain expertise and technical skill of our participants, in comparison to one another. The plotting methodology is not precise; rather we aim to demonstrate the relative grouping of participants, taking into account their educational background, demonstrated and self-described technical skill, and any previous work experience or history mentioned. Domain

expertise relates to formal knowledge of a space gained through explicit educational experience, or self-described history within a specific organization or overarching body (for example, the municipal government). Technical experience is comprised of both formal training, such as a BA in computer science, alongside any skills learned on-the-job, in this specific role or a previous one. In our ranking of technical skills, we rate the technical complexity of tools used, too – for example, performing data cleaning and visualization with GIS tools is considered more skilled than collecting participant responses through Google Forms surveys. While both types of skills are important and require much practice and experience, we refer to the commonly accepted definitions of technical skill for traditional data scientists as our measuring stick, which prioritizes a particular suite of tools. Therefore, our metric measures technical skill and does not necessarily reflect professional experience and capabilities in a broader sense, so much as they do the metrics by which data scientists are traditionally judged.

As shown by Figure 1, our participants tend to be domain experts who have secondarily acquired the technical skills necessary to do data work in their primary domain of expertise. None of the participants are employed by primarily technical organizations. Their employers range from local and state government to small nonprofits and public utilities. Notably, seven of the participants have titles that include "data", "analyst", or "analysis". Only one participant describes their work even indirectly as "data science"; this participant is also the most technically-skilled member of the participant pool. Some work on primarily technical teams; for example, P19 manages a small team (less than 5 other individuals) of GIS practitioners in an organization that conducts regional policy analysis. Others are the primary data person within their organization; P18, for example, moved into their role as the organization's data manager after having gained a reputation within the organization as an "Excel nerd".

*4.1.2 Participants' perspectives on data work.* We asked participants a series of questions about their perspective on data work, to capture what similarities and differences they have with the full participant peers, namely data scientists. As participants focus on different parts of the data lifecycle, it is not surprising their responses are varied – from the logistics around each activity (e.g., designing questionnaires and writing guides for focus groups) to analyzing the dataset. The two most frequently mentioned tasks were data collection and cleaning, which were mentioned by five and four participants, respectively. Several participants mentioned the challenges of data collection, as they had to find and integrate data sources from several offices or departments. P13 describes this process when answering what part of their job takes up most of their time:

> P13: *Collecting. So like pulling data together from different departments. Yeah, I would say collecting.*
>
> Interviewer: *Is that mostly what you mean by that? Just figuring out where I get this data set.*
>
> P13: *Exactly.*
>
> Interviewer: *And asking folks for it?*
>
> P13: *And coordinating with people on to get needs met. So it's kind of like a project manager function. Pulling the people together and getting it in that useful format is probably where the bulk of the time goes.*

For P13, collecting the data encompasses not only obtaining the dataset, but coordinating cleaning that dataset to render it usable. P19 goes into more detail about the all-encompassing nature of cleaning: *"I kind of consider collecting like intrinsic to the job, so I don't think about it being an action. It's just what we do. I don't count that time really. Cleaning is a pretty big deal."* P8 elaborates on the hurdle that messy data presents to project progression:

> Interviewer: *You said you spend most of your time cleaning. Is there something that you would rather spend more time on?*
>
> P8: *I'd like to spend my time more analyzing. But see, ... we don't have great information for a lot of things, so I'm always trying to figure out to the extent of how much information is user providable and how much I may have to ask you go and clean it up... I want to spend more time analyzing.*

Here, cleaning again gets in the way of performing other data tasks the participant would prefer to engage in. P4 sums up their feeling on the labor to reward ratio of data cleaning: *"it's kind of like finding a needle in a haystack."* Cleaning comes up, too, as a task participants find particularly tedious. In the words of P7: *"It just takes so long to sort and making sure it's all formulated right and it's all coded properly. And it's just annoying. It's very tedious ... And it doesn't actually provide you with the information you really need. ... It's not like the analysis portion where you actually run the numbers and come to conclusions. It's like literally just like making sure that everything is like coded exactly."* Like P8, P7 finds data cleaning a distraction from the kind of labor (here, analysis) that is more central to their role.

However, participants were more uniform when it came to important qualities in a data worker. Close to half of participants described the most important feature of a good data worker as attention to detail and/or interest in the datasets they worked with. P17 identifies the need to be detail oriented in data work: *"it's real easy to miss things."* P4 has a more elaborate description of the tension between needing to be detail oriented, but being confronted with the monotony of data work:

> *I think that anybody can really clean data or learn, you know, about how to analyze, properly analyze data. But, I think, if you're not interested in it, you know, you're probably going to get tired of cleaning it, tired of analyzing it because a lot of the job is just looking at the same data sets over and over again and trying to figure out how you can merge data sets or, you know, what those data sets can actually tell you. So, it's a lot of repetition. So, you've got to actually have some vested interest in what you're looking at.*

Collection and cleaning can intersect, too, to cause extra frustration when multiple data streams are involved. P1 describes: *"And I think the ability to see themes across the data. Right? Especially when we're getting it from all different sources, and to pick kind of how you're going to format it and kind of structure the data..."*

However, the importance of good data stewardship takes on extra importance in light of the domain (public service work). P18 responds to a question about the most important quality of someone doing their job:

> *Integrity. I want to say something more technological, like you know, attention to detail, or ... like an ability to plan or something like that, but I think that when it comes down to the very end of things, when I took my stats class ... I realized that you can make numbers say pretty much anything you want them to say and I could easily like omit data from a dataset to make the story that we're trying to tell look more positive for us. But I think that in my heart of hearts, it's important to make sure that you're not doing anything unethical with the data, that you're presenting it as it is, letting people see the good but also letting them see the bad so that if there's a problem, they can fix it. If I hide it by you know, kind of selecting a certain subset that's only going to make us look good, then I'm doing disservice to the agency and G[*]d knows to the field at large.*

P6 stated similar sentiments: *"...From a government standpoint, I think the biggest quality is looking for accuracy... I think integrity's the biggest thing where you're working with, with data. I think the*

*numbers are what they are."* Here, the position of P6's role – as a government employee – adds an additional weight to the need to both properly prepare and analyze data from a moral standpoint.

Summary: especially when our participants are the primary "data person" in their organization, they are solely responsible for multiple tasks in the data lifecycle, meaning they spend much more time on early steps like data collection and cleaning than other data workers (particularly those in high resources environments) who may either purchase or obtain pre-cleaned datasets or be able to offload some of the monotonous or tedious steps of data cleaning. This is unique to this group of data workers, but the gist of their work – their perspective on the trials and tribulations of data work – demonstrates that they share a domain with their full participant peers.

### 4.2 How data work gets done in the civic sector

While our participants work in numerous organizations and have different day to day job duties, there are several themes that appear consistently in our interviews.

*4.2.1 Proximity to data.* The participants work with data that they are subject experts in, meaning that they understated the who/what/where of the data stream that results in the datasets they work with. In some cases, they themselves collect the data and engage in the analysis – e.g., P7's role involves developing policy decisions and they describes deploying a survey to relevant stakeholders. Similarly, P14 engages with offices and public service providers, such as the city code enforcement division, police and fire departments, and tax assessor's office, to negotiate data reporting procedures to collect the data the participant needs to answer questions of interest to their office.

For other interviewees, their proximity to the context of the data's origination is facilitated by their previous experience either in the organization, or within the specific sub-sector of work, say, affordable housing organizations. P19 is an example of the first type of proximity. Having worked in GIS for several branches of the local government, including utilities and transportation, they know that the local utility companies – all of whom would have reasons to dig up section of road – lack a shared database. Therefore, a proposed moratorium system to prevent digging up portions of newly laid roads is not implementable until such a shared database exists between these governmental and extra-governmental entities. As an example of the second kind of contextual experience: P12's job involves tracking use of their city's recreational facilities and the participant is in constant contact with facility managers. Together with one of the managers, they found that annual membership cards were resulting in visits to recreation centers were not being counted properly; this had direct implications for the amount of funding the center got.

Other interviewees rely on a data liaison for contextualization. The liaison is either directly involved with the collection process and can provide the data worker with more information, or their sole role might be to function as a go-between the individuals producing the dataset and the interviewee who performs analysis on it. An example of the first kind of liaison is the informal institutional knowledge that P8 and P9, who work at public utility service, source from their colleagues who work in the field (maintaining utility infrastructure):

> P8: *We've got quite a few gentlemen [field workers] who've lived there for two decades. They know our problem areas, so they'll say that's a problem area.*
> P9: *Yeah, they know our problem areas, so they'll say that's a problem area, it's back over where [old hospital building] used to be. 'Oh, yeah, I know exactly what it is.'*

Other participants have liaisons employed by their organization. P14, whose work involves budget analysis, describes how their liaison helps them track down points of confusion or concern: *"[The oversight office] they actually get the bills. And so I work with a liaison who, if I have a question on something, [the liaison] would work with me and the vendor, whoever it [of the contracting utility*

*companies] may be, annd be able to get an actual bill".* Another participant – P7, who deals with civic infrastructure performance – describes how their data liaison (a specially hired consultant) helps the participant make sense of unexpected data points in often-messy infrastructure datasets:

> *And the street address is that's associated with the [infrastructure landmark] but sometimes those are just made up addresses, depending on where the [landmark] is ... it's an address where if you look it up on Google Maps, it's just like a patch of grass or something... we send it to our consultant ... and then [the consultant] would just essentially go through and remove anything for which there's low confidence of accuracy.*

Here, P14's data liaison can make sense proxy data that would otherwise be considered "mislabeled" or even "incorrect". When P15 embarked on a housing data project and wanted to explore property tax assessments, they sourced their own liaison by bringing on a collaborator: an academic who specializes in the housing policy and was familiar with the data from their own work, underscoring the role of a personal network of collaborators. Another participant, P16, handles data for their city's public schools and explains how the school system has designated liaisons (roughly one per ten schools) to help interface between the tool builders (such as the participant) and the end users (teachers and administrators with questions). These liaisons, formally "specialists" for the information system, clarify data questions and manage feature requests that are passed back to the development team.

Summary: our participants have a unique proximity to the data they work with. Either they are directly involved in the dataset's collection, or they find ways manufacture proximity through a formal data liaison or ad-hoc interpersonal connections. This infrastructure allows them to capture the context of the data they work with and address anomalies or points of confusion in the data. Critically, for those removed from collection stage of the data they work, they might recognize they might misunderstand the data – despite their expertise in the domain – without the lack of (in)formal data liaisons who can restore that original context. The participants gladly welcome the data liaisons and consider them an important part of their workflow.

#### 4.2.2 *Collaboration.*

> *"Collaboration hot; Separation old; We will give it all we've got; Together we have told!"* An idealized description of collaboration in data work, excerpted from "In the Data Kitchen" [51].

The importance of collaboration comes up repeatedly for participants. First, given the limited number of data workers within their respective organizations and offices, participants describe the role that collaboration with others has. P19, for example, describes how talking to other GIS workers throughout the city helps them and the rest of their team stay up to date on tools: *"Which is good because I need to keep fresh on the tools and the data...And then also be able to speak out to others with it."* Peer expertise provides support that P19 could not otherwise access. P6 illustrates how cooperation across organizations and offices both betters data flow and the projects involved:

> *Yeah. We were, they [offices sharing data] were here in person. Yeah. So, we were actually sharing that as we were gathering, they were gathering. I was looking at it but we would also go back to the departments that shared the different data. So, we checked them in the loop. That even made them a lot more excited about sharing it, even more. So, they were able to say, hey, what about this? Hey, what if we could layer this on top of that? ... So, what it began to do was it began to have some of the departments think a little bit more strategically in how they use the data be able to see it. Because I think one of the challenges is people gather the data and they go away and no one says what was that data used for? I have no idea. But when you go back to, continue to go back and show*

> *them the output of it. That's when you start to see some more creativity for the folks who are actually using it. So, that was kind of the case with that project."*

For P6, not only does collaboration provide them with access to needed data flows, but it also engenders projects that integrate the expertise of the two departments.

There is another theme that P6 hits on: given the civic nature of the projects that the participants work on, they rarely work with a single data stream. Rather, most of the participants pool data from several sources. Particularly for those working in domain policy positions, they must collate data collected by different branches of the municipal government or various extra-governmental organizations. For example, P13 describes how developing housing policy requires the cooperation of several municipal agencies: *"So where we had to get the three agencies together and hash out all the details of how our programs work, how we measure things, and how it should be reported. And that has, it takes, it's kind of like you can't just have data people there. You have to [have the] domain person."*

The coalescing of multiple data streams that P13 describes has implications for interoperability between datasets as well. As P4 states, *"... that's the biggest impediment I think in, you know, organizational partnerships... all our data's in different, you know, ... stages. Some of it's... still in paper format. I believe, if you start looking at nonprofits, you're going to see a lot more organizations that are still very paper heavy."* Not only can the nuance of the subject domain present hurdles, but so can the format the data is recorded and stored in. P1, too, describes the challenges of first obtaining data from another department and then attempting to work with it once received:

> *So, when I put out the call to each of the public agencies for the data, and I basically just requested what they had, right? However, you're tracking it or producing a report, send it. And so, trying to be as flexible as possible... very slowly, we're seeing responses. And it all came in – I'd say it all came in, in Excel, but in very different levels of detail...Like some, like the [housing] stuff is super detailed, but it's also because of how they're required to put it into their system. And then the other stuff was like two columns ... it was like very different, and then everybody had slightly different fields. And it's like important information... I attempted you know, to put it into one Excel document, but I'm sure I [messed] some things up when I did that... I attempted then to start to merge it so that we could de-duplicate because a lot of it is the same... A lot of their money is in the same properties.*

P1 is describing multiple public agencies providing funding for government-sponsored housing around the city; despite the overlap in properties, there exists no centralized system and P1 was responsible for matching relevant information about a given property from the records of the various agencies. However, when working with multiple offices or agencies within a municipal government, data doesn't necessarily exist in the same format given different standards (both prescribed by use and unintentional) as P13 describes: *"... the biggest problem is the data were in different formats, so police, if I recall correctly, used addresses and then fire had latitude and longitude."*

Participants frequently highlighted times that a lack of cooperation either halted their projects or presented major obstacles. Along with issues of data format, some organizations and municipal and state agencies fail to share data point blank. P13 their project lacked key data due to inter-office politics: *"We tried to get some data from the [relevant municipal office] and they weren't cooperative ... We had to leave it out."* P6 describes the problem as being endemic to the city in which they operate at an existential scale:

> *Well, in the city we when you say collaboration, first thing comes to mind: silos. Silos. I think the team that's here in this office, right here, are probably the most collaborative in the city... we cut across horizontal... any collaboration that takes place in the city, a*

> *lot of times, that relates to the data takes place with us. In terms of collaborating from department one collaborating with department two, there's not a whole lot of collaboration that takes place, I think. That's one of the areas I think city could benefit from.*

P14 also describes the challenges of working in a state with an outdated data storage system and how this is compounded by the challenge of limited technical personnel:

> *It's just more time consuming maybe because we, you don't really have the resources as of yet to kind of just have the application like some states actually have it to where it's [on the] web. So they just go online. [A coworker] do[es] it and then it like sends it all the way to us. So everything is electronic ... [that coworkers] kind of gets it on one spreadsheet or you do that now but it's a little bit more automated?...[the same coworker] does it so how he has the application coded, he basically has the codes on the back end. And unfortunately, he likes to keep his stuff together so no one knows the codes and no one can actually unlock the spreadsheets to do it themselves.*

Critical work in the state, as described by P4, is reliant on the personality and process of a single staff member.

Data work in the civic sector also requires collaboration and support from individuals both superior and subordinate. P12 describes the limitations of data collection when it requires inputs from individuals completing field work for their office:

> *We would love to be able to make it as where we can streamline it to where when they go out they do everything digitally, but the only thing that would speed that up is being able to change everybody's job description where they'd have to be able to work. Because believe or not, we still have people who don't want to work with electronics...If it's not in their job description, you know, you'll find out real quick that they don't want to do it. And so that's part of the difficult thing with some of the people in the field...Like, you might have [an individual]who – and some of these metrics you'll see – whose job is to [perform infrastructure maintenance work]. You know, [this individual's] not going to want to be inputting [their] data [on electronic devices].*

P12's attempts at datafication are hindered by a lack of corresponding enthusiasm for data by those in charge of the field workers, who have not incorporated extra allowances for data collection into their subordinate's schedules. The field workers are thus discouraged from engaging in data collection as it falls outside of the duties on which they will be evaluated and compensated.

Summary: collaboration – and the challenges it presents – are of elevated importance to participants. Unlike professional data scientists, they often need to communicate their findings to both those above and below them in institutional hierarchy. Having buy in from participants at all steps of the data lifecycle is critical for our participants; they often work on shared datasets that require extra technical and logistical support to obtain and use. Personal relationships between the individuals we interviewed and those in their professional networks are critical components of performing their role.

*4.2.3 Communication.* Communication is likely an important topic for any data worker, but our interviewees placed particular emphasis on its role in the civic sector. Many of our interviewees are responsible for formulating policy that directly impacts residents of the city, or they present their findings to elected officials who will take that action. Hence, this group of data workers plays an important role in shaping how the contents of the datasets they work with come to be understood.

P5 describes specifically augment policy with data that will appeal to a board of elected municipal decision makers: *"We usually will do some research specifically on the district and its demographics and if we have them. I work on a lot of projects that involve construction in one way or another,*

*on multiple assets, so I'll put together a map identifying them city-wide and then concentrations within districts and all that."* Here, P5 uses data as a way to personalize policy to a given decision-maker. P15 describes an incident where a higher-up incorrectly described a data dashboard to the organization's board of overseers, so the data workers in the organization removed the dashboard – which made information available to the public – as a result. We cannot share the actual quote in more detail without compromising the participant's identity. The accessibility of data here hinged on the behavior of a single individual higher in the professional hierarchy. Similarly, P17 describes trying to run intra-organization focus groups, which required the buy in of upper-level management: *"...they had trouble recruiting, and so I spent a lot of time trying to get leadership in the areas we were recruiting – these were operations employees, folks in operations – to really get the word out and encourage people to come to the focus groups and to work with the research firm. So it was a lot more hands on than I thought."* The project could not have continued without the support and cooperation of the participant's supervisors. P3 also describes how they had to seek buy-in from high-level officials to begin a project:

> *so we talked with the [Department of Education]... they actually do a really good job of providing data at the school level. We wanted it at the neighborhood level. So, we first discussed the need with the higher ups. So, you have to have some sort of access to even be able to have these conversations. And so, we've built up enough cache that we can go straight to the [second-in-command], the person in charge of data collection for all [the state] and say hey, can we come and talk to you about this.*

However, unlike many organizations and organizational settings, our participants also require the work of colleagues in roles lower in the chain of command. P8 and P9 describe how they need to effectively communicate with the field workers in their organization, as well as city council members:

> P8: *We do have conversations with them both directly and through their leadership as a conduit. Last February we had a series of special meetings on Saturdays to have a conversation about their frustrations ... So they called us in on a Saturday and we made our best attempt to show them... When we see 120 [infrastructure problems], this is what it looks like to us. I know to you it looks like chaos and, you know, you had one [occurrence] that it took three days to get resolved and another one, well, that was hardly a [big issue], it took like to an hour to clamp, but this is what it looks like to us. And also to say, hey, if you do a job and you don't close it out there, I have no idea that you did that. So when you're making the claim that you did this much work over the course of the day and I come back to you and say maybe, but I don't see that, and I've got hundreds of other work orders open, why haven't you done that, this is where we're missing each other... we had a slightly different format for trying to show them [the field works vs high level officials in the organization] that ... But admittedly, that's difficult ... We've got gentlemen and ladies ... who have no statistical background to be able to engage with you in a conversation about meetings or anything. It's just trying to communicate on this is what-matters-to-you information basis ... So that's a very different conversation.*

> P9: *We focus mostly on the council members because the difficulty is – well, for the mayor's office, like for our whole agency monthly reports and we do share information on what's going on. They're actually invited to our [intra-organizational] meetings as well and do attend, so they can see kind of the nuts and bolts, but as far as local officials, as far as elected officials, generally we try to shy away from a lot of that kind of stuff, ... overall larger aspects, information about the system because... you get to a lot of situations where*

> *people might not understand, well, your area has a lot more development so you might see a lot more things in your area.*

The buy-in necessary to complete their roles was, for P8 and P9, dependent on correctly tailoring their work to two different audiences. P12, similarly, details delivering safety metrics to field workers in their organization in order to increase productivity after a workplace accident; we cannot describe the incident in more detail without compromising the participant's identity. P12's role dependent on the field workers continuing their work and therefore, they were stuck until the field workers resumed work.

Given that the civic sector often answers to the public living in its jurisdictions, our participants also describe challenges around making data available to the public. In some cases, the public is mostly unaware that the data can be accessed, as P14 describes:

> *I guess for us, I don't think people realize that our data is essentially available to anybody. Because we are a state entity, so we really don't own our data. The people own our data. So I think people don't realize that they can actually request a lot of our data from us. I think the issue sometimes is maybe a cost factor. You know, if it does require me to come up with new data that we may already have in a different format, they do charge people... So I think if more people knew that, it will be, we probably have more like databases on more data kind of circulating around if people knew how easy it was. Or even that a lot of the data is on our website in spreadsheets.*

P4 details the desire to make their office's data more available to the public, so that city residents can see what's actually being done: *"We also definitely want it to be more front-facing. So, to actually, you know, say what [a colleague] has said before, we want to be able to have something that's a front end where citizens can actually see what our office is doing in them and, you know, actually engage a little bit more. Because, we're really trying to up engage that with the community, for sure, which, you know, is why we engage in [an affordable housing] study."* Here, data dissemination is not only a component of P4's job, but an important part of their self-identified duties as a civic worker.

Summary: communication takes on an added weight for data work in the civic context. Participants describe a need to communicate both to decision makers and the general public who will be the recipients of the policies and data driven systems our participants design and/or implement. These goals share some overlap with those of data intermediaries, as we will discuss further in 5.2.

*4.2.4 Tools.* All of our participants explicitly reference working with Excel, though they do so to varying degrees; for example, P18 describes their Excel work as one of the more technical parts of their job and how they use *"buttons, macros, and user forms"* to augment basic spreadsheets. In contrast, P16 uses Excel only when they receive datasets in Excel's custom format. Beyond Excel, tool use is more varied between participants, owing to the diversity in precise role highlighted in Figure 1. For example, participants with more emphasis on backend systems referenced setting up and managing Oracle and CRM datasets, while others more focused on collection described using tools like Google Forms and Survey Monkey.

However, a unifying theme for participants is that working in low or limited resource technical environments hinders their work. Specifically, their environments limit the tools they have access to and makes it difficult to obtain new ones. P6 succinctly explains why they are unable to acquire new tools when they find faults with (or major limitations to) existing ones: *"I mean, far as in the private sector, there'd probably be a yes to it [getting a better tool]. But in government, it's not yes."* Participants expand on the nuances of limited tool maintenance and acquisition in their work, describing the following:

- Limited funding. P7 describes how governmental workplaces rely on outdated software due to limited funding:

  *I think the way we put our metrics into a manual process is very tedious. It's time-consuming. And then, that will go away when we automate it ... Now, I think probably for some of the departments, a lot of the information from the field is inputted into some of those systems, manually, as well, too. And I think if we had like more type of that, I mean if it was able to be sped up electronically, I think that would help out a whole lot... I think it's just a lot of different manual processes throughout government, in general. So, I think if I was looking at ways of improving government in this. Removing a lot of different manual processes that are, kind of, typical standard automated in a lot of different private sectors.*

  Here, there is a clear optimal solution that would better facilitate the work of P7 and their peers, but the solution is inaccessible due to a lack of budget. The same is the case for P16, who describes a similar experience working with a local system that runs off a 25 year-old operating system version:

  *So it's more so kind of getting up to today's dates with software... it [the old operating system] has a lot of glitches... you have to like click stuff in order to go to the next page. So you would have to like click and then it pops up another one. You need to click that. Pop up another window. So it's really, it's really difficult to use. Most people won't even bother to learn how to use it. ... I mean, I had to learn how to use it because a lot of the information that I need is in there. And so, I learned how to use it but most people, they look at it and they're like, "I can't do it." Because if you, if you don't really use it all the time, you can forget and you won't be able to really get to where you need to get to because it's not really like you look at it and you're like, "Okay, I know how to click here." And like, okay, where do I put it?*

  Again, there is a simple fix (updating the software system to a modern one), but there is not the funding to make that investment.
- High turnover. Another issue that plagues civic organizations is that of turnover. P4 describes how data initiatives spiral out of control in part due to the high turnover resulting from administration changes and how those initiatives are quickly abandoned:

  *Particularly for this department, just the huge amount of reporting we're now required to do. I think there's an interest in really showing what the [the city] and other agencies are doing... just for this office to accurately report on the information that we're required to with the frequency that we have to, we definitely need better formats for storing and recovering data. We can't, every time, there's a report have to scramble around for various spreadsheets that, you know, we might not even know exists because there has been some turnover in the office. And, when that happens, that information gets lost. So, it's important to have these structures where, you know, if somebody leaves, their information doesn't go with them.*

  Information that is collected is both rendered unusable by personnel turnover, as well as changes in priorities of elected officials. In some cases, after an initiative falls from priority, the data collection and reporting system is still actively staffed, even though the data is no longer needed. Sometimes the design of a system itself and the training provided to users, however, is responsible for a lack of use. P15 describes such a situation:

  *I think it's becoming a more like widely valued thing here, but we had sort of the traumatic experience – we got a new [customer resource management system (CRM)]. No one ever got trained on it, but it was mandated that everyone had to log in every day. I know we had some really bad data, like we still don't know go through [that CRM], like every department ended up going rogue and just either doing Excel or like we use our own project management software ... Every department does its own thing now, and we're, we're rolling out a new*

> *Salesforce program that's going to be pretty cool, but it's really hard to buy into that because like our organizational memory is like [non-existent].*

Not only was the new system a waste of resources because it was never used, but it caused institutional fear of new software systems.

Subsequently, participant workflow is obstructed by a lack of reliable, affordable tools. Collaboration is hindered; P7 illustrates the ad-hoc nature of collaboration on a hand-coding effort with file type requirements:

> *[One team member] in particular who actually like created the macro that input all the data from the Google Sheets into the Excel, into much more usable format. ... Basically I had an Excel file which all the data from the Google spreadsheet – well, the Google Survey dumped into a Google Sheet, which we then made into an Excel file in order to use the macros and then like organize all the data. Then we put it back onto a Google Sheet to like assign everything to everybody once the data was in a more usable format. But because of the macros that [the team member] had installed didn't work on Google Sheets a lot of the people had to redownload Google Sheets and give it to me in Excel format again once they were done with it. And I plugged it back in to Google Sheets. It was more complicated than it should've been.*

In very few workplaces would this be considered an ideal process. However, workflows in civic organizations are defined by what tools are available, not what is optimal. Reliance on outdated tools and systems also increases organizational security risks. This is particularly true for governmental offices and agencies, which are already the subject of cyberattacks and foreign interests [58]. One participant describes the process of manually updating changes to a large database while their computer system was held hostage in an organization-wide ransomware attack.

Summary: government offices and agencies, as well as independent non-governmental organizations, often have limited budgets that are reliant on public funding or private grants, rather than data workers' needs. Further, governmental entities are subject to high turnover stemming from frequent administration changes. Participants frequently mentioned the impacts of limited funding and high turnover on their work, noting that it required them to work with outdated and frustrating tools. Beyond worker experience, there are large institutional security concerns – not to mention citizen and customer data privacy – posed by these systems.

## 5 DISCUSSION

Our analysis is framed by the questions Muller et al. present, as we seek to understand how peripheral data workers compare and contrast to their full participant peers and the subsequent implications for the future design of tools and processes. But more so than any specific implications for design, what emerges from our findings is the need to continually expand and refine our concept of data workers to include individuals on the periphery of the data work community of practice. While our participants comprise a particular subgroup of peripheral data workers – those in the civic sector – we theorize their concerns and practices are applicable to peripheral data workers in other sectors and professional settings. Namely, the civic sector is not the only one plagued by funding constraints and high personnel turnover; difficulties related to performing data work in this setting can thus be extrapolated. The one sector likely not touched by these constraints – the high tech industry – is already well-studied by members of this community, as discussed earlier in this work. Part of motivation for this work, in fact, is to identify, describe, and theorize about a broader range of data work beyond data science as it is commonly construed within and in relation to the high tech industry. Further, civic data workers, as an example of a peripheral group of data

workers, may engage in practices that could be beneficially adopted by their full peers, such as data scientists.

### 5.1 Understanding data work as a community of practice

In order to address the question of who performs data work, we suggest that our conception of data work needs to be reformatted – as others have suggested – into a more inclusive community of practice. When we treat data work as a community of practice with multiple options for participation, we, as a research community, can begin to identify the breadth of individuals and groups contributing to and subsequently defining our data ecosystems. Until we do so, any efforts we make to understand the contextualization practices of data workers writ large will lack key participants in the various data ecosystems. Further, using the community of practice framework, we can better ensure that the tools and systems we build are applicable for the full range of users; in other words, that we are developing tools that meet the needs of specific groups of users. We should, thus, pay attention to individuals on the periphery of data work who are full participants in their domain of expertise, and, similarly, understand data scientists as peripheral participants in the domains of the subjects they analyze.

**Re: Who performs data work?** We propose careful identification and examination of the workflows and processes of peripheral members of the data work community, who may be full members of another domain-specific community of practice. In order to understand who contributes to our data intensive systems, we suggest studying the needs and challenges of this group along with data scientists, data analysts, novices to the data work [11, 22], data wranglers [32, 33]. By arranging these subgroups of data workers in relation to one another, we can begin to understand common and unique experiences between them. Critically, we position the current conception of data scientist as full participants, with our interviewees as peripheral members, since, in many cases, they belong and identify with **another** community of practice as their primary affiliation. Similarly, data scientists with some expertise in a given subject area are peripheral members of that subject's community of practice. By broadening our understanding of how data work is completed across different fields, we can begin to make sense of how we, as a community, can both help transfer beneficial practices between different groups and also understand how data work is practiced outside of primarily technical organizations, such as large tech companies, and the difficulties therein.

In acknowledgement of a primary appointment or position in their domain of expertise, we actively avoid naming or labelling our interviewees as a class of data worker. Instead, we group them in relation to their peers who practice data work as their main professional practice. We have two reasons for this: first, our participants have their own professional identities that spans multiple domains. For this reason, we hesitate to label them as data scientists directly, in acknowledgement of their expertise in their respective subject domains. Second, given the range of our interviewees and the variety between their day-to-day work and specific organizations, we risk minimizing the nuanced experiences of our individual participants. "Peripheral data workers" serves not as a static classifier, but is meant to demonstrate relative position to a group already known to this community – full, professional data workers such as data scientists.

There is one immediate caveat to our proposition that data work can be understood as a community of practice that highlights an oversimplified view of the relationship between full and peripheral participants. In the traditional conception of community of practice [38], members on the outer rings of the community are often actively seeking to become more integrated into the community; for these participants, their role is much like that of an apprentice. When this apprenticeship process is well constructed, Lave and Wenger term it "legitimate peripheral participation", such that members on the periphery are engaging in meaningful sub-units of work in the domain under the

supervision of a more experienced member, who both guides and legitimizes them. However, we do not mean to suggest that in all cases peripheral data workers *want* to become more full members of the community, nor does there often exist an explicit mentor-mentee relationship between the full and novice members of the data work community of practice. Besides acquiring more data skill that would help them with their present role, few indicated interest in becoming full time members of, say, the data science community, nor are do they describe being in frequent contact with full-time data work professionals. However, this distinction presents an opportunity for **collaboration** between primary subject matter professionals and primary data work professionals, given the extensive contextualization skills of the former and computing skills of the latter. Such a link would be more akin to a *partnership* than *mentor-mentee* relationship, because each individual would serve in both roles, as both a teacher (in their domain) and learner (in their partner's). This bidirectional relationship represents a caveat to the work of full participants of the data work community of practice, we note, in that it indicates that groups like data scientists may not have the necessary domain expertise to complete their tasks and actually rely on adjacent subject matter communities for assistance and contextual knowledge.

Given that contextualization requires extensive time spent in, or understanding, the ecosystem that produced that data, it is these domain experts – or peripheral data workers – who may provide insight about how that contextualization might be done. As they are close to the origins of the data, they may be best positioned to answer Loukissas' questions about the originating environment [39]. They can also make sense of factors that impact the data as it travels from collection to a summary or report handed to a decision maker. Having this knowledge makes them well-equipped to address contextualization throughout the creation *and* curation stages, as D'Ignazio and Klein discuss [13] and bring us closer to Monroe-White's concept of *emancipatory data science*, which supports the conscientious and consent-concerned use of data created by marginalized communities [49].

### 5.2 Understanding the practices of peripheral data workers

Muller et al. [52] pose the questions of how data workers go about their tasks, how they collaborate, what the implications of their work is, and what their needs regarding collaborative tools are. We answer these questions in turn, with our responses focused on the group of peripheral data workers identified earlier in this work, who belong to an understudied demographic of data workers.

**Re: How is data work performed?** As discussed earlier in this work, most of our knowledge about the way data work happens uses data scientists as the study population. While data scientists are an important and sizeable community of data workers, they receive an oversized amount of attention in comparison to groups more on the periphery. In particular, many efforts to address bias and fairness in algorithmic systems involve warning labels – or, said more mundanely, a kind of nutrition label – about who and what is represented in that dataset, along with clear statements about the situation and context from which the data arose. The solutions are geared towards the community of highly technically skilled data scientists, without specifying how data scientists should source this information, or practice conscientious consumption of it. Meanwhile, many peripheral data workers work in immediately public-facing setting and have found ways to meaningfully practice these conscientious consumption and reflection practices. Our interviewees are a good example of this: the data work they engage in has a direct effect on city residents and various vulnerable populations (those served by the non-governmental organizations). As shared in our findings, our interviewees have ample access to the data they work with, either having collected it themselves or via a data liaison. If we want to implement documentation procedures to address issues of representation in datasets, these data workers are uniquely positioned to show us how those practices might be established in other groups within the data work community. Further, given their public facing role – or implications of their work – our participants share some

overlap in characterization with "data intermediaries" (DIs), or individuals with data skills who connect members of the interested public, particularly members of marginalized communities, with data related to their interests or concerns [44]. However, their are two notable differences between our participants and DIs: first, the latter do so as avocation while our participants are salaried employees of public-facing organizations and institutions, and, second, our participants are more valuable given their level of contextual knowledge, while DIs are sought as partners for their technical skills.

**Re: How does collaboration happen in data work?** This group of peripheral data workers have developed systems, both formal and ad-hoc, to answer their questions about a given dataset. Consider the aforementioned data liaisons, as well as the institutional knowledge that P8 and P9 rely on to make sense of errors in their dataset, respectively. This corpus of interviews also highlights the crucial role of cross-hierarchy communication; ideally all data workers could be able to communicate the contents of, and concerns about, a dataset they work with to both those in their technical setting, but also to the laypeople who are affected by those systems. As there are increasing concerns about the fairness of algorithmic decision making in all facets of life, it is worth noting that our interviewees, all of whom work in the civic sector, play a part in systems that city residents usually have no option to disengage with. For example, city residents are limited in their choice of municipal services and utilities. Those who receive aid from the various non-governmental organizations represented amongst our participants very rarely enter the partnership in a position of power, which would enable them to ask how and why decisions were made and what their data was used for. In the more general case, it is important for citizens to be able to understand the open data the respective governments and organizations seek to publicize; civic data may be open for access, but, as Boehner & DiSalvo describe, closed by lack of access to context [2]. As a community, we should aim to both recreate and further extend these contextualization and cross-hierarchy communication practices.

**Re: How can we design tools for collaboration?** Beginning the work we found that discussions of both data science and civic data often suggest that idealized contexts and practices would be seamless, with a free flow of interoperable data across a common toolchain and coherent process [8, 17, 35, 60]. Whether or not that is desirable or achievable, it is not the case. We found these peripheral data contexts and practices are seamful: characterized by misalignments, inoperability, and the ongoing labor of translation and negotiation [68]. As Inman and Ribes argue, neither seamlessness nor seamfulness should be considered inherently good or bad [29]. Seams mark strategies of revealing and concealing how a system or process works. As such, seamlessness and seamfulness afford different means of access and control. Much of the labor of these peripheral data workers is tactical, developing work-arounds for the seams in their context and practices. In some cases these seams are particular to the data, e.g. P8 & P9 utilizing their personal experience (and that of the field technicians) to correct faulty addresses. In other cases, these seams relate to tools – consider P7 and their coworkers' issues with analysis software, requiring switches between Google Sheets and Excel – and in still other cases these seam are socio-technical, e.g., P12 struggling to get their field technicians to input data when it seemed like an unnecessary imposition without payoff. These practices echo the labor of work with seams that Vertesi describes in her ethnography of space science. As Vertesi describes, even well-funded scientific endeavors—such as sending a robot to mars—require negotiating, or what she refers to as suturing, seams. It is not surprising, then, that such seams and tactics of suturing exist in domains with fewer resources, such as local government and civil society. These seams and the tactics to suture them become opportunities for joining and expressing data in novel and useful ways. For instance, Dailey and Starbird have studied and described how residents, government workers, and journalists sutured the seams of social media platforms to create needed and responsive communication channels in response to a

natural disaster [10]. We argue that describing and attending to such resourcefulness in seamful data work can lead to both a better appreciation of this labor and inform the design for data tools and processes for domain experts.

### 5.3 Designing data work tools for domain experts

The answer to the question of tools needed for data workers should be the result of analysis of different constituent groups in the community of practice. Our findings illustrate several challenges faced by civic data workers, as subject domain experts, but peripheral data workers. Namely, there is often a lack of funding, or funding is provided by mercurial grants. This results in limitations on what new tools can be acquired and an overreliance on legacy systems that produce inflexible data streams. For those of our interviewees who learned their data work skills on the job, they had less exposure to broader groups of tools and tend to stick to those tools with which they were already familiar, even in cases where they would like new ones. For those who work for local and state government, there are also issues with documentation and archival practices, as new administrations present new data collection priorities and practices and personnel turnover is high. As local and state governments are also increasingly targeted by ransomware attacks, there are also specific security concerns both for tools used and data storage methods.

However, these limitations are likely not unique to the civic sector. Other data workers on the periphery are also embedded in organizations with non-technical foci; many of data workers potentially experience the same obstructions. Critically, these issues are much less likely to bother data scientists, who often operate within large technical organizations. In contrast, peripheral data workers need tools that 1) create flexible, more interoperable, and sustainable data streams and 2) provide accessible entry points for users with less technical backgrounds. In building these tools, we should incorporate space for the contextualization these data workers practice, both concretizing their observations and sensemaking practices, as well as facilitating documentation in roles with naturally high-turnover (such as those occupied by political appointees). By limiting our conception of who performs data work, we ignore a specific and documented need for data work tools that meet these requirements; this is something our community can and should address in our tool designs and user experience assessments.

**Re: Implications for stakeholders and affected population?** In designing these tools there is also the opportunity to bake in some of the contextualization or reflection practices already suggested by this community. If we understand contextualization as one type of collaboration (namely, conjoining context and the textual data), some of these guides can also be adapted to the specific working context of peripheral data workers. For example, what would it look like to build data analysis tools for peripheral data workers that helped them concretize those contextualization practices which they already engage in, and also add new ones? We can imagine something like space for Gebru et al.'s datasheets for datasets [24] made in these tools, that helps convey context between collaborators. The datasheets could be extended, too, to the civic data work space. Data workers could be asked to record not only recommended uses, but to explicitly warn what imagined uses would be inappropriate or reason for concern, given the naturally transient staffing of governmental organizations and the likelihood that a given dataset may be reused or revisited in the future when those most familiar with it are not around.

**Re: Tool Evaluations**. While we do not directly answer how to shape evaluations of tool learning-curves and usability, our findings suggest some preliminary ideas. First, tool uptake is significantly shaped by institutional resources, meaning that evaluations must take place in situ and in a variety of environments and account for the price of access to the tool. Second, tools should be graded for their usefulness – once learned – for given application areas, to help individual data workers in low-resource environments decide if they are worth the payoff. Third, tool evaluations

should be made available and accessible to non-researchers, to provide a centralized place for peripheral community members to search for new tools.

### 5.4 Future work

In order to acknowledge and potentially even mitigate issues of bias in data-driven systems, we propose an extension of Muller et al.'s questions to understand the practices and experiences of data workers (Figure 1). Specifically, we should take the perspective that data work is a community of practice, with participants at varying levels of participation; further, in many cases, those peripheral participants are full professionals in their own primary community of practice. There appears, too, to be a relationship between position in the community of practice and tools and resources available. Subsequently, we should augment the existing questions to encourage answers both specific to one group in the community of practice, as well as the larger community.

Building off Neff et al. who describe how critiques of data science might improve "with an approach that considers the day-to-day practices of data science" [54], we should add one new question to the list: what contextualization practices are incorporated in a given data worker's workflow? And, who else's help – direct or indirect – is needed to facilitate these practices?

### 5.5 Limitations

We note that all of our interviewees within this sample work in the civic sector in the United States of America. This choice was intentional for two reasons. First, while language around, and specifics of, municipal services varies by community and national setting, common concepts such as water treatment (or lack thereof) are more accessible than a specialized discipline, such as a natural science. This allows us to focus on the way in which data is handled by our participants, rather than the background of their work. Second, within the United States, civic organizations – either explicitly governmental or not – are typically sites of limited financial resources. However, we acknowledge that findings from this sample population describe technical challenges faced by a one group of individuals working in an applied field of data work and is limited, too, by the social and cultural aspects related to the singular nation of focus. However, we believe that their technical challenges, or balancing a need for better tools with a limited budget and high turnover, is not unique to the civic sector (or the United States) and can be found in many organizations, particularly those that are not large tech companies, where only a small portion of the population of all data workers are employed.

Finally, our interviewees' demographics, as noted in the results section, also vary some from national averages in computing and information technology. While many of our participants are not employed in roles that would necessarily be counted in these categories, they are the categories that many full participants of the data work community would be counted in. We theorize that while our survey population may not thus be representative of the computing and information technology workforce, it may be more accurately representative of the demographics of those on the periphery of data work, who are spread out across multiple fields, and may have slightly lessened gender and racial disparities compared to tech fields. Given that our participants worked in various organization types, this is difficult to state with certainty.

## 6 CONCLUSION

In this work we interrogate data work as a community of practice, using Lave & Wenger's framework [38]. We identify the unusual role of data workers on the periphery of the community of practice. They are often full professionals in a different community of practice but engage in more robust efforts to contextualize the data they work with, as compared with more full practitioners of data work (e.g., data scientists). We report the findings of our 19 interviews with civic employees in

large cities in the United States who engage with data, but who are peripheral members of the larger data work community. Pulling from these interviews, we suggest how other data work practitioners could learn from these contextualization methods. Further, we aim to raise the profile of data workers on the periphery and suggest that as a research community, we consider the unique challenges of performing data work in low resource technical environments. Our goal is to contribute to both better contextualization practices and systems for full data workers and promote the needs and concerns of data workers beyond data scientists. Our findings address open questions in the burgeoning human-centered data science community [53], including highlighting a wider range of data workers, along with their particular data work methods and workflows, along with their role in the greater data work community.

## 7 ACKNOWLEDGMENTS

We appreciate the time and perspective shared by our participants. Our anonymous reviewers provided much helpful critique. This work was funded by the National Science Foundation (US) under grant number #1951818.

## REFERENCES

[1] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (12 2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[2] Kirsten Boehner and Carl DiSalvo. 2016. Data, Design and Civics: An Exploratory Study of Civic Tech. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2970–2981. https://doi.org/10.1145/2858036.2858326

[3] Chris Bopp, Ellie Harmon, and Amy Voida. 2017. *Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work.* Association for Computing Machinery, New York, NY, USA, 3608–3619. https://doi.org/10.1145/3025453.3025694

[4] Geoffrey C. Bowker. 2005. *Memory practices in the sciences.* MIT Press, Cambridge, Mass. OCLC: ocm60776866.

[5] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of Research Methods in Health Social Sciences*, Pranee Liamputtong (Ed.). Springer Singapore, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[6] Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1099203.1099205

[7] Longbing Cao. 2017. Data Science: A Comprehensive Overview. *Comput. Surveys* 50, 3 (Oct. 2017), 1–42. https://doi.org/10.1145/3076253

[8] Silvia Cazacu, Nicolai Brodersen Hansen, and Ben Schouten. 2020. Empowerment Approaches in Digital Civics. In *32nd Australian Conference on Human-Computer Interaction* (Sydney, NSW, Australia) *(OzCHI '20)*. Association for Computing Machinery, New York, NY, USA, 692–699. https://doi.org/10.1145/3441000.3441069

[9] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2021. Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2 2021), 1860–1870. https://doi.org/10.1109/TVCG.2020.3030340

[10] Dharma Dailey and Kate Starbird. 2017. Social Media Seamsters: Stitching Platforms & Audiences into Local Crisis Infrastructure. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1277–1289. https://doi.org/10.1145/2998181.2998290

[11] Sayamindu Dasgupta and Benjamin Mako Hill. 2017. Scratch Community Blocks: Supporting Children as Data Scientists. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3620–3631. https://doi.org/10.1145/3025453.3025847 [Online; accessed 2020-10-18].

[12] Yuri Demchenko, Luca Comminiello, and Gianluca Reali. 2019. Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge. In *Proceedings of the 2019 International Conference on Big Data and Education - ICBDE'19*. ACM Press, London, United Kingdom, 124–128. https://doi.org/10.1145/3322134.3322143

[13] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism.* The MIT Press, Cambridge, Massachusetts.

[14] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 297–307. https://doi.org/10.1145/3377325.3377501

[15] Rosalind Edwards and Janet Holland. 2013. *What is qualitative interviewing?* Bloomsbury, London : New Delhi. OCLC: ocn855705441.

[16] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. *arXiv:2101.04719 [cs]* (Jan. 2021). https://doi.org/10.1145/3411764.3445188 arXiv: 2101.04719.

[17] Sheena Erete and Jennifer O. Burrell. 2017. Empowered Participation: How Citizens Use Technology in Local Governance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2307–2319. https://doi.org/10.1145/3025453.3025996

[18] Melanie Feinberg. 2017. A Design Perspective on Data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2952–2963. https://doi.org/10.1145/3025453.3025837 [Online; accessed 2020-08-17].

[19] Melanie Feinberg. 2017. Reading databases: slow information interactions beyond the retrieval paradigm. *Journal of Documentation* 73, 2 (March 2017), 336–356. https://doi.org/10.1108/JD-03-2016-0030

[20] Melanie Feinberg. 2017. The value of discernment: making use of interpretive flexibility in metadata generation and aggregation. *Information Research-an International Electronic Journal* 22, 1 (2017), 22.

[21] Melanie Feinberg, Daniel Carter, Julia Bullard, and Ayse Gursoy. 2017. Translating Texture: Design as Integration. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, Edinburgh United Kingdom, 297–307. https://doi.org/10.1145/3064663.3064730

[22] 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 (May 2020), 1–22. https://doi.org/10.1145/3392840

[23] Uwe Flick. 2009. *An introduction to qualitative research.* https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc_100025409254.0x000001 OCLC: 1052103480.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (19 3 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[25] Lisa Gitelman (Ed.). 2013. *"Raw data" is an oxymoron.* The MIT Press, Cambridge, Massachusetts ; London, England.

[26] Lisa M. Given (Ed.). 2008. *The Sage encyclopedia of qualitative research methods.* Sage Publications, Los Angeles, Calif.

[27] Lisa Hardy, Colin Dixon, and Sherry Hsi. 2020. From Data Collectors to Data Producers: Shifting Students' Relationship to Data. *Journal of the Learning Sciences* 29, 1 (Jan. 2020), 104–126. https://doi.org/10.1080/10508406.2019.1678164

[28] Samantha Hautea, Sayamindu Dasgupta, and Benjamin Mako Hill. 2017. Youth Perspectives on Critical Data Literacies. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 919–930. https://doi.org/10.1145/3025453.3025823 [Online; accessed 2020-09-15].

[29] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300508

[30] Lilly Irani and Jesse Marx. 2021. *Redacted.* Taller California, San Diego, CA.

[31] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562. https://doi.org/10.1145/3394486.3406477 [Online; accessed 2021-02-21].

[32] Britney Johnson, Ben Rydal Shapiro, Betsy DiSalvo, Annabel Rothschild, and Carl DiSalvo. 2021. Exploring Approaches to Data Literacy Through a Critical Race Theory Perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445141

[33] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, Vancouver, BC, Canada, 3363. https://doi.org/10.1145/1978942.1979444

[34] Charles Kiene, Kenny Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun "Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, and Bryan Dosono. 2019. Volunteer Work: Mapping the Future of Moderation Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 492–497. https://doi.org/10.1145/3311957.3359443

[35] Antti Knutas, Victoria Palacin, Giovanni Maccani, and Markus Helfert. 2019. Software Engineering in Civic Tech a Case Study about Code for Ireland. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society* (Montreal, Quebec, Canada) *(ICSE-SEIS '19)*. IEEE Press, 41–50. https://doi.org/10.1109/ICSE-SEIS.2019.14

[36] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking datasets – Understanding data sensemaking behaviours. *International Journal of Human-Computer Studies* 146 (2 2021), 102562. https://doi.org/10.1016/j.ijhcs.2020.102562

[37] 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*. ACM, Sanibel Island Florida USA, 151–156. https://doi.org/10.1145/3323994.3369898

[38] Jean Lave and Etienne Wenger. 1991. *Situated learning: legitimate peripheral participation*. Cambridge University Press, Cambridge [England] ; New York.

[39] Yanni A. Loukissas. 2019. *All data are local: thinking critically in a data-driven society*. The MIT Press, Cambridge, Massachusetts.

[40] Luis Felipe Luna-Reyes. 2018. The search for the data scientist: creating value from data. *ACM SIGCAS Computers and Society* 47, 4 (July 2018), 12–16. https://doi.org/10.1145/3243141.3243145

[41] 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 (5 12 2019), 1–23. https://doi.org/10.1145/3361118

[42] Gary Marchionini. 2017. Information Science Roles in the Emerging Field of Data Science. *Journal of Data and Information Science* 1, 2 (Sept. 2017), 1–6. https://doi.org/10.20309/jdis.201609

[43] Amanda Meng. 2014. Investigating the Roots of Open Data's Social Impact. *JeDEM - eJournal of eDemocracy and Open Government* 6, 1 (Oct. 2014), 1–13. https://doi.org/10.29379/jedem.v6i1.288

[44] Amanda Meng, Carl DiSalvo, Lokman Tsui, and Michael Best. 2019. The social impact of open government data in Hong Kong: Umbrella Movement protests and adversarial politics. *The Information Society* 35, 4 (Aug. 2019), 216–228. https://doi.org/10.1080/01972243.2019.1613464

[45] Amanda Meng, Carl DiSalvo, and Ellen Zegura. 2019. Collaborative Data Work Towards a Caring Democracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 42 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359144

[46] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (14 10 2020), 1–25. https://doi.org/10.1145/3415186

[47] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 161–172. https://doi.org/10.1145/3442188.3445880 [Online; accessed 2021-03-10].

[48] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (Dec. 2016), 205395171667967. https://doi.org/10.1177/2053951716679679

[49] Thema Monroe-White. 2021. Emancipatory Data Science: A Liberatory Framework for Mitigating Data Harms and Fostering Social Transformation. In *Proceedings of the 2021 on Computers and People Research Conference* (Virtual Event, Germany) *(SIGMIS-CPR'21)*. Association for Computing Machinery, New York, NY, USA, 23–30. https://doi.org/10.1145/3458026.3462161

[50] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating Data Science. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 467–473. https://doi.org/10.1145/3406865.3418584 [Online; accessed 2021-02-21].

[51] Michael Muller and Thomas Erickson. 2018. In the Data Kitchen: A Review (a design fiction on data science). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–10. https://doi.org/10.1145/3170427.3188407

[52] Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299018

[53] 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, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[54] Gina Neff, Anissa Tanweer, Brittany Fiore-Gartland, and Laura Osburn. 2017. Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data* 5, 2 (June 2017), 85–97. https://doi.org/10.1089/big.2016.0050

[55] US Department of Labor Statistics. 2021. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/cpsaat11.htm [Online; accessed 2021-03-30].

[56] 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, Portland Oregon USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

[57] Paula Pereira, Jacome Cunha, and Joao Paulo Fernandes. 2020. On Understanding Data Scientists. *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 1–5. https://doi.org/10.1109/VL/HCC50065.2020.9127269 [Online; accessed 2021-02-21].

[58] Nicole Perlroth. 2021. *This is how they tell me the world ends: the cyberweapons arms race.* Bloomsbury Publishing, New York.

[59] Kathleen H. Pine, Claus Bossen, Yunan Chen, Gunnar Ellingsen, Miria Grisot, Melissa Mazmanian, and Naja Holten Møller. 2018. Data Work in Healthcare: Challenges for Patients, Clinicians and Administrators. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 433–439. https://doi.org/10.1145/3272973.3273017

[60] David Ribes. 2017. Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1514–1526. https://doi.org/10.1145/2998181.2998344

[61] Sarah T. Roberts. [n.d.]. *Behind the screen: content moderation in the shadows of social media.* Yale University Press. OCLC: on1055263168.

[62] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 15.

[63] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *CoRR* abs/2108.04308 (2021). arXiv:2108.04308 https://arxiv.org/abs/2108.04308

[64] Caroline Sinders. 2020. A Solution without a Problem? Seeking Questions to Ask and Problems to Solve within Open, Civic Data. *Interactions* 27, 5 (Sept. 2020), 46–49. https://doi.org/10.1145/3411292

[65] Barry Smart, Kay Peggs, and Joseph D. Burridge (Eds.). 2013. *Observation methods.* SAGE, Los Angeles. OCLC: ocn816163690.

[66] Olivier St-Cyr, Craig M. MacDonald, Elizabeth F. Churchill, Jenny J. Preece, and Anna Bowser. 2018. Developing a Community of Practice to Support Global HCI Education. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3170427.3170616

[67] Anselm L. Strauss and Juliet M. Corbin (Eds.). 1997. *Grounded theory in practice.* Sage Publications, Thousand Oaks.

[68] Janet Vertesi. 2014. Seamful Spaces: Heterogeneous Infrastructures in Interaction. *Science, Technology, & Human Values* 39, 2 (2014), 264–284. https://doi.org/10.1177/0162243913516012 arXiv:https://doi.org/10.1177/0162243913516012

[69] 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 (Nov. 2019), 1–30. https://doi.org/10.1145/3359141

[70] Dakuo Wang, Josh Andres, Justin Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. *arXiv:2101.05273 [cs]* (13 1 2021). https://doi.org/10.1145/3411764.3445526 arXiv: 2101.05273.

[71] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359313

[72] Michelle Hoda Wilkerson, Kathryn Lanouette, Rebecca L Shareff, Tim Erickson, Nicole Bulalacao, Joan I Heller, Natalya St Clair, William Finzer, and Frieda Reichsman. 2018. Data Transformations: Restructuring Data for Inquiry in a Simulation and Data Analysis Environment. *International Conference for the Learning Sciences* (2018), 2.

[73] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *arXiv:2001.06684 [cs, stat]* (16 4 2020). http://arxiv.org/abs/2001.06684 arXiv: 2001.06684.

Received April 2021; revised November 2021; accepted March 2022