The New Reality of Reproducibility: The Role of Data Work in Scientific Research

MELANIE FEINBERG, University of North Carolina at Chapel Hill, USA
WILL SUTHERLAND, University of Washington, USA
SARAH BETH NELSON, University of Wisconsin-Whitewater, USA
MOHAMMAD HOSSEIN JARRAHI, University of North Carolina at Chapel Hill, USA
ARCOT RAJASEKAR, University of North Carolina at Chapel Hill, USA

Although reproducibility–the idea that a valid scientific experiment can be repeated with similar results–is integral to our understanding of good scientific practice, it has remained a difficult concept to define precisely. Across scientific disciplines, the increasing prevalence of large datasets, and the computational techniques necessary to manage and analyze those datasets, has prompted new ways of thinking about reproducibility. We present findings from a qualitative study of a NSF–funded two-week workshop developed to introduce an interdisciplinary group of domain scientists to data-management techniques for data-intensive computing, with a focus on reproducible science. Our findings suggest that the introduction of data-related activities promotes a new understanding of reproducibility as a mechanism for local knowledge transfer and collaboration, particularly as regards efficient software reuse.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; **Ethnographic studies**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: data work; reproducibility; replicability; scientific software development

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 35 (May 2020), 22 pages. https://doi.org/10.1145/3392840

1 INTRODUCTION

In September, 2018, the Kaggle data science community surged to two million users worldwide. At the same time, a Kaggle blog post noted that only 100,000 LinkedIn users have a current job title of "data scientist" [45]. As this discrepancy suggests, many people who perform data-related activities are not full-time data scientists. Accordingly, as the CSCW community seeks to understand and support data work, it is important to consider the experiences of people in a variety of professions who are integrating data-intensive tasks into their daily work practices [34, 40, 44]. What happens when work evolves to encompass a new array of data-oriented activities? Do people understand their work–and what it means to do excellent work–differently? In this study, we explore this

Authors' addresses: Melanie Feinberg, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Will Sutherland, University of Washington, Seattle, Washington, USA; Sarah Beth Nelson, University of Wisconsin-Whitewater, Whitewater, Wisconsin, USA; Mohammad Hossein Jarrahi, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Arcot Rajasekar, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

2573-0142/2020/5-ART35 $15.00
https://doi.org/10.1145/3392840

general question through the professional domain of scientific research, in which data-focused tools and tasks have been increasingly incorporated into research practice. In particular, we examine how the adoption of data tools affects scientists' understanding of a core professional concern–reproducibility.

Researchers throughout the natural and social sciences have needed to adapt their traditional methods to accommodate increasingly large datasets, which often require computational techniques for analysis [23, 26]. Often, these scientific domain experts lack formal training in computer or data sciences; instead, they cobble together skills and techniques for data management and analysis as necessary to accomplish their domain-specific research goals [1]. Because these new data work practices are evolving opportunistically and informally, many researchers may not consider data-related activities to be core components of the scientific enterprise. Nonetheless, new affordances associated with computational techniques challenge previous ways of understanding how science might–and should–be conducted, and the role of data in scientific work [25].

One area of scientific practice that has undergone significant change as a result of evolving computational techniques is collaboration. The ability to mobilize data from one situation as evidence in another situation is a core promise of this transformation [37], and, as such, it has significant implications for the way researchers share and reuse code or data, and build upon each others' work [9]. However, effective collaboration around data and code has been difficult to implement because of the extra work involved in sharing code and data, the proliferation of data formats, and the problem of communicating contextual information [4, 56].

Another key area of scientific practices impacted by emerging data and computational techniques is reproducibility. Data tools have been held up as a solution to the problem of reproducibility in science. The notion of reproducibility has long been integral to an understanding of good scientific practice [22]. In theory, by documenting data collection and analysis protocols in published papers, scientists enable others to replicate their experiments and verify their results [55]. In practical terms, what does it mean for science to be reproducible–to repeat an analysis with the original data? to redo an experiment by collecting new data?–has always been fundamentally ambiguous. Accordingly, the information that researchers need to provide in order to achieve reproducibility has also been uncertain [14]. Within a particular field, scientists have managed this ambiguity by relying on consensus understanding of shared work practices and associated publication conventions [60]. If a scientist within a particular field follows generally agreed-upon practices, then the work that scientist conducts is, according to the standards of that community, sufficiently reproducible [55].

Across fields of scientific endeavor, the increasing adoption of data-analysis techniques associated with large-scale datasets has put pressure on established conventions regarding reproducibility and good scientific practice [5]. For instance, previously common, accepted practices for data management, such as the reliance on a researcher's individual judgment to make case-by-case decisions regarding data cleaning, have been questioned [36]. At the same time, adoptees of new data management practices, such as the use of automated scripts to ensure consistent cleaning operations across entire datasets, may not recognize the extent to which these new techniques might necessitate changes to other aspects of their work [41]. For example, to enable the repetition of analysis techniques that may involve custom code and extensive, unknown software dependencies, the way that work processes are documented for scientific publication may also need to change.

[1]For instance, in 2017-2018, the Cyber Training initiative of the National Science Foundation has funded projects to provide supplementary training in data science techiques to researchers in atmospheric sciences, astrononomical sciences, high-energy physics, materials science, and social science and public services. As an example of the rationale for such projects, the project Web site for the Large Synoptic Survey Telescope (LSST) fellowship in astronomy describes its program as providing skills "not easily addressed by current astrophysics graduate programs." (See https://astrodatascience.org/) The workshop that we describe in this paper was also funded under the Cyber Training program.

More fundamentally, the idea of reproducibility itself may be transformed as work practices evolve to accommodate the new importance of data work in science.

In this project, we seek to understand what reproducibility means to scientists who are beginning to engage in data work. We present findings from an NSF-funded two-week workshop developed to introduce an interdisciplinary group of domain scientists (from fields such as ecology, bioinformatics, chemistry, and mechanical engineering) to data-management techniques for data-intensive computing, with a focus on reproducible science. In doing so, we concentrate on two research questions:

- How do data tools and practices shape workshop participants‘ understanding of reproducibility?
- How do workshop participants articulate the benefits of computational reproducibility?

Our findings show that reproducibility is very much a concept in flux, in terms of what it means, how it is achieved, and who it benefits. When asked to define reproducibility, workshop participants–both instructors and attendees alike–tend to associate reproducibility with the replication of data analysis, in accordance with an understanding of reproducibility as a means to verify results and ensure scientific integrity. However, when participants were asked about the utility of reproducibility, this global dimension of reproducibility, understood as promoting scholarly integrity and upholding methodological rigor, is seldom mentioned. Instead, participants invoke a local understanding of reproducibility as a mechanism for knowledge transfer and collaboration, particularly as regards efficient software reuse.

In our view, this multiplicity of meanings around the concept of reproducibility shows how the assimilation of data-science tasks might cause a cascade of effects on scientists' understanding of their work. On the one hand, tools and techniques associated with data science enable a more exacting and meticulous way of performing consistent operations on a dataset, both preparatory to and during analysis. The use of such tools and techniques then becomes associated with the concept of replicability. Concurrently, domain scientists–for whom computation is generally a means, rather than an end–see how adopting these techniques can, more immediately make their work easier by facilitating the reuse of code across studies.

Our findings reinforce Feger at al.'s [18] observation that the collaborative benefits of data tools and practices may increase scientists' motivations to make both code and data available to others for the purposes of reproducibility. Here, we extend this observation to make a stronger connection between the problem of collaborating around data or code and researchers' understandings of reproducibility. Our findings suggest that the adoption of data tools and practices for collaborating around those tools comes with a shift in the practical criteria for reproducing findings. Moreover, these criteria reflect a particular way of doing reproducibility, which, in the case of this workshop, was centered around automation. More generally, in providing a case study that shows how the adoption of data-science tasks can affect a domain's understanding of its own fundamental concepts, we contribute to CSCW's emerging understanding of data-science work practices [1, 40, 44].

Our paper proceeds as follows. First, we situate our work within the literature of reproducibility in science, of scientists' data sharing, and of data work more broadly. Next, we describe the workshop where we collected our data, followed by a summary of our findings. We conclude by discussing the subtle refinements in how our participants understand the importance of reproducibility, and in how reproducibility should be implemented. We connect these conceptual adjustments to participants' integration of data-science activities into scientific practice.

2 RELATED WORK

In this section, we briefly review related work in three areas:

- The concept of reproducibility in science, and challenges in understanding what reproducibility means for data-intensive, code-dependent research.
- The role of data sharing and code sharing as a mechanism for facilitating reproducibility in science.
- The work practices of data scientists and others who perform data work.

2.1 Reproducibility and Scientific Credibility

In recent years there has been a renewed focus on reproducibility as a criterion for credible science. Researchers from a variety of fields have raised concerns about the replicability of published findings (Ioannidis, 2005; Pashler and Wagenmakers, 2012; Collins and Tabak, 2014). For instance, a widely publicized effort to replicate findings from prominent social and cognitive psychology studies showed fewer significant findings and smaller effect sizes that the original studies (Open Science Collaboration, 2015). This rising concern with reproducibility, which has been termed a "crisis" in the field of psychology, has led to a reevaluation of certain aspects of scientific practice, such as the publication of research data and materials and the preferential publication of positive results [29, 58].

While reproducibility has long been an important principle of good science, the recent emphasis on carrying out reproducibility in practice is new. Traditionally, reproducibility has been seen as critical to the progress of science because it lends consistency to scientists' understanding of nature and is an expression of a scientific culture of organized skepticism [19]. However, actually replicating the results of a study requires time and resources, and it is seldom done (Baker, 2016). The recent impetus towards reproducibility therefore comes as a broad cultural change, with the goal of altering existing institutional and incentive structures, and changing the way science is done at the level of practice [2, 42]. Publication venues, especially in computationally intensive sciences, have placed new emphasis on making code and data available in order to support reproducibility and reuse [52]. Furthermore, researchers have attempted to establish best practices for computationally intensive research, which include not only the publication of data, but also an array of data management, project organization, and collaboration activities [53, 59].

Despite a general sentiment that reproducibility is good for science, no strong consensus has emerged concerning how reproducibility is done and what the precise goal of reproducing something might be. The term replication is often taken to mean that an experiment is performed again with the same materials and procedures, with the purpose of verifying the results of that experiment. However, the utility of exact replications is questionable. Schmidt [50] points out that if all aspects of the replication are the same as the original experiment then they are the same experiment and the replication does not serve as a confirmation. Instead, replication involves similarity with difference. The findings of an experiment are more convincing if they can be replicated with different samples or populations, or with different instruments, or by different researchers. Schmidt calls this "the proof that the experiment reflects knowledge that can be separated from the specific circumstances (such as time, place, or persons) under which it was gained" [50, p. 90]. This idea, which Schmidt calls conceptual replication, is similar to Cartwright's [7] definition of reproducibility. For Cartwright, replication involves the same procedures as the original experiment, whereas reproducibility involves variation (and therefore extension) of the original experimental setup.

As Collins [11] observes, instructions, recipes, or diagrams of the original experiment are not enough for replication without the tacit knowledge of the experimenter who carried them out. This leaves a fuzzy boundary between replication and reproduction, as replications are necessarily instantiated in new circumstances, and it becomes unclear whether the replicator is rehashing the old experiment, or performing a new one. What is accepted as a successful replication, then, is necessarily situated. This point directs us to investigate specifically what kind of confirmation is

being sought in the ongoing dust-up over reproducibility. Freese and Peterson [20], for instance, have characterized the emergence of a statistical metascience as one basis for scientific integrity. In connection with this notion, efforts in psychology have focused on reproducing findings with new groups of participants [10], whereas others have focused on exact replications of findings using the same code and data [46].

2.2 Sharing and Reproducibility in Scientific Work

In order for scientific work to be reproducible, research processes and outcomes must be shared [43]. Accordingly, data sharing is intertwined with the concept of reproducibility, and data reuse is promoted as a primary objective of data sharing [54]. As science has become more data-intensive, data sharing has been identified as a foundational component of scientific collaboration [57]. Data that has been preserved and made available for public access facilitates reproducibility by enabling re-analysis and replication studies, and enables novel combinations of data to emerge [12]. Data reuse can have economic impacts as it lowers the cost of redundant data collection by making collected data available to other researchers [46], providing opportunities for collaboration and co-authorship [48].

Because of the long-term benefits associated with data sharing, institutional players such as funding agencies, journal publishers, academic institutions urge scientists to make their data available to others [14]. Increasingly, these institutions are enacting policies that require data sharing for the projects that they fund, the articles that they publish, or the researchers that they employ. These policies do have some effect. For example, a survey sponsored by the European Commission cites requirements attached to public funding as the most important reason for preserving and sharing research data [28]. Internationally, significant public investments have been made in projects to support data sharing across disciplines, universities, and national boundaries. For example, the DataONE project was funded by the U.S. National Science Foundation to "ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data" [15].

However, despite institutional support for this type of cyberinfrastructure, data sharing among scientists–particularly across disciplinary boundaries–remains a challenging undertaking, even where policies require it [17, 46]. Many researchers recognize they lack the basic skills to use the metadata standards that facilitate public data sharing and reuse [24]. Furthermore, merely uploading one's data to a public repository seldom provides sufficient context to enable others to understand and reuse it [41]. To properly share data, scientists must explain how their data is structured (that is, explain their metadata) and document the operations performed on raw collected data to make it ready for analysis (in data science, these operations are often called "cleaning" or "wrangling") [55]. Preferably, scientists must also document–or provide–any software they used to perform data cleaning and analysis. In many scientific domains, this software may involve custom applications or scripts [27]. Savage and Vickers [49] argue that many researchers fail to develop reusable datasets (that is, datasets that are accompanied with sufficient documentation, such as metadata) because such efforts often involve too much work and too little incentive. All of these issues considered, the ideal of a seamless, interoperable infrastructure is not just a technical feat, but rather it is a tremendous collaborative and organizational task requiring collaborators to bring shifting technical, institutional and scientific requirements into alignment [3, 13]. It is precisely in focusing on these collaborative dynamics, and on practice, that empirical work in CSCW makes valuable contributions to our understanding of the shifting social conditions of science [31].

Given the tension between the labor involved in data sharing and its incentives, researchers have turned towards the core reasons that researchers have for sharing data, and how to better encourage or enforce reproducible practices. Researchers have identified the core practices of data sharing, such as communicating methods and other metadata, as essential to facilitating reproducibility

[21]. Making reproducible science a standard thus involves what Stodden et al. [53] calls a "culture change" and a shift in the policies of central scientific institutions, such as journals and funding agencies. Sholler et al. ([51], for instance, outline ways that researchers resist adopting the best practices of reproducible research and offer suggestions for overcoming these obstacles. Feger et al. [18] point to a number of ancillary benefits of reproducible practices, such as facilitating communication, automation, and collaboration. Feger et al argue that systems designed to support these ancillary benefits–potentially using gamification patterns to encourage engagement–will help to offset the required effort of reproducible practice.

2.3 Data Work and Practice

Research within CSCW has described tools and practice as integral to the way that researchers understand their vocation, and how remediations in tools or practice can alter these understandings [30]. A recent area of research has focused on the practices of data science and data scientists specifically, and how data scientists come to understand the work that they do. For example, Kandel et al [32] looked at the everyday processes of 35 data analysts in corporate environments, identifying five primary tasks associated with data work: discovery, wrangling, profiling, modeling, and reporting. Kim et al [35] examined the education and training of data scientists in a large software company, looking at the problems that data scientists engage with and how data scientists describe their work to others. Carter and Sholler [6] surveyed day-to-day work practices of data scientists and how these professionals may connect or diverge from the dominant discourses around data science and big data.

One focus in these studies has been to understand how data scientists interpret and shape the data that they work with [40]. Muller and colleagues [40], who synthesize the experiences of 21 data science workers at a large, international company, describe a continuum of active interventions on the part of data workers toward their data. In their study, few data scientists receive a clean, complete dataset from a client, which they can proceed to model. Instead, data scientists variously transform the data that they work with, via processes of capturing, curating, design, and creation. Notably, these transformation processes occur for "ground truth" data–data used as the standard for comparison, or the dependent variable for predictive analyses–as well as for experimental data. Although ground truth is commonly described as authoritative, objective, and factual, the data scientists that Muller and colleagues studied sometimes had to transform ground truth, rather than merely discovering and using it without intervention. As reported by Muller and colleagues, when data scientists needed to work actively with ground truth data and transform it, they talked about the idea of ground truth differently and understood their role regarding ground truth differently. In other words, as the data scientists in Muller et al's study reappropriated the tasks that they performed with experimental data for ground truth data–as they adjusted their practices with ground truth–they also adjusted their understanding of the concept of ground truth. We observe a similar phenomenon in our study. As domain scientists integrate data–science tasks into their activities, they adjust their practices with reproducibility, and they also adjust their understanding of the concept of reproducibility.

Similarly, Passi and Jackson [44] relate how data scientists and business analysts in another large corporation negotiate trust in data science systems. In Passi and Jackson's account, one element of trust involves the interpretation of quantitative metrics produced by data science models. The data scientists, for instance, found low, but improved, accuracy scores in relation to a hard problem (churn prediction) to validate trust in their models, whereas the business analysts were disappointed with these results. Here, also, the need to actively engage in the practice of trust building across roles (data scientists to business analysts) leads to an adjustment in the concept of trust–what it means to "trust the numbers" for both groups. In the churn prediction problem reported by Passi

and Jackson, the data scientists adapt their presentation of results from a binary label (a customer is likely or not likely to stop using a product, or churn) to a ranked list of probabilities (these customers are in decreasing order of their likelihood to churn). As the data scientists adjust their practices according to the sense of trust negotiated with the business analysts, they also adjust their understanding of trusted numbers. Our study demonstrates a similar pattern of conceptual adjustment; however, we observe this pattern with scientific researchers who are not data scientists, but who are integrating some aspects of data science work into their own practices.

3 METHODS

In this section, we describe

- The workshop that constituted our research setting.
- Our data collection and analysis methods.

3.1 Research Setting: Two-Week Workshop on Data-Intensive Computing for Domain Scientists

Our findings are based on an qualitative study of a two-week workshop conducted to train domain scientists to work and collaborate in data-intensive research settings. This workshop was the initial activity of a multi-stage project, funded through the CyberTraining program of the U.S. National Science Foundation (NSF), to educate domain scientists in management and analysis techniques for large-scale datasets. The workshop project team comprised faculty, research scientists, and technologists from multiple universities and research institutes. The interdisciplinary team included members with expertise in distributed computing, data management, and cyberinfrastructure.

The first four authors were involved in the workshop solely as program evaluators. (The fifth author was the main PI on the project and provided high-level supervision of all activities.) The evaluation team did not participate in curriculum development, nor did we provide instruction. We developed and conducted our data collection and analysis activities separately from the instructional team. The evaluation team was invited to attend project team meetings where curriculum planning was discussed, but our participation in these meetings was limited to observation.

This data management workshop provided a strategic site to investigate how researchers understand methods and purposes of reproducibility. In order to capture these understandings, we relied on ethnographic methods of observation and engagement. Despite the short time frame of the workshop the benefits of a rapid ethnographic engagement were essential for capturing the context of use [39]. Rendering broad notions or movements, such as the so-called reproducibility crisis, can be difficult using traditional ethnographic methods. This problem mirrors difficulties in studying large-scale infrastructures that researchers in CSCW have grappled with for a long time [33]. However, the workshop context made the problem of capturing broad understandings more tractable. In a strategically-situated ethnography, the ethnographer positions themselves in a site where broader understandings and systems are actively under construction or repair. This allows them to render something broader by following the subject's own sensibilities about other locales and happenings [38]. The workshop was a site in which proper applications of data-intensive tools and methods were under construction: students were learning a technical trade of reproducible science, and they were attempting to reconcile broad notions of reproducibility with their own day-to-day work. Here, we use this process as a scalar device, following the subjects' own management of the broader field of reproducible research ([47]. It is important to note that our object of study is not the learning process, nor how best to teach reproducibility; rather we are using the learning process as a methodological way in for examining how researchers come to understand what reproducibility means, and how it relates to notions of good scientific work.

3.1.1 Workshop Curriculum and Structure. The workshop curriculum was developed to provide domain scientists with a holistic introduction to computationally-intensive methods for managing and analyzing large datasets, according to current best practices for achieving reproducibility. The curriculum was designed to provide students with an array of complementary skills, including technical experience with data management and analysis tools, and project organization and collaboration skills for data-intensive projects.

The workshop was conducted over 10 days in July, 2018. All students were resident at the workshop, which took place at a university in the southeastern United States. Students stayed together in university lodgings, and ate breakfast and lunch together at the site of the workshop. Each day, workshop sessions were scheduled for 8 or 9 hours. Optional early morning sessions were also included on most days.

Structurally, the workshop was divided into two parts. The first 6 days of the workshop comprised traditional instructional sessions, which combined lectures with hands-on exercises. The technical portion of the curriculum included topics such as:

- Hands-on work with high-performance computing resources, such as JetStream and Amazon Web Services.
- Profile creation for a number of scientific data portals, such as Dataverse and HydroShare.
- Instruction on tools used for containerizing scientific workflows, such as Docker and Singularity.
- Overviews of basic principles of data management, including an introduction to metadata standards.
- An introduction to a selection of tools for data manipulation and analytics, such as Knime and R Shiny.

The portions of the curriculum that focused on organizational and communication skills introduced students to the fundamentals of basic software development and collaboration practices, such as stand-up meetings, pair programming, code walkthroughs, and project management tools. Students were encouraged to practice these organizational and communication skills throughout the course of the workshop.

The last 4 days of the workshop were devoted to group projects. In these projects, students were tasked with first replicating the data analysis in a published scientific paper and then with making the analysis in that paper more easily reproducible (for instance, by using container software to ensure that software dependencies did not cause problems for future reproducers). Students were assigned to work on one of three projects. One project was in the field of hydrology, one was in bioinformatics, and one was social science. There were two teams assigned to the hydrology and social science projects, and one to the bioinformatics project. The hydrology and bioinformatics projects were selected by instructors, and the social science project was proposed by one of the workshop attendees.

3.1.2 Workshop Participants. Our participants included both attendees and instructors. Attendees: Twenty-one workshop attendees were selected through a competitive application process, and 20 ultimately attended the workshop. All attendees were advanced doctoral students or postdoctoral researchers at various institutions in the United States. Although we may refer to attendees by using the term "students," the workshop attendees were practicing researchers with at least three years of doctoral training at the time they participated in the workshop. The majority of attendees were in the final stages of their doctoral studies and working on their dissertation projects. They came from a number of disciplinary backgrounds, including bioinformatics, mechanical engineering, ecology, chemical engineering, computer science, public health, materials science, neuroscience, and organizational psychology. Table 1 gives more information on the attendees.

Students' prior experience with data management practices and tools varied. Some students were comfortable with shell scripting and programming, data manipulation, and working on a command line, whereas others were just learning these skills. Similarly, some students had been working on data-intensive projects for some time, whereas others were seeking these skills in anticipation of working on such projects. Additionally, students varied in the kind and size of the data they typically used in their own work. For instance, bioinformatics researchers working with gene sequences needed to manage terabytes of data and required distributed computing resources to conduct their analyses. Other students were social scientists, working with social media data, electronic health records, or survey responses; for them, large datasets could easily be stored on a personal laptop.

Although our attendees ranged across a wide variety of fields, the emphasis in their own everyday practice tended to focus on analysis and modeling rather than data collection. For instance, the biologists in the workshop mostly worked with gene sequences, often from publicly available datasets, rather than directly collected observational data, and the social scientists worked with already existing electronic health records or social media data, rather than directly collected surveys.

Instructors: The workshop instructors likewise came from an array of disciplines. (Table 2 lists information for the instructors.) Although the group of instructors were all experts in their fields, they represented different aspects of data management and analysis and were not necessarily well-versed in other areas: for instance, the metadata expert might not be familiar with distributed computing, and the workflow expert might not be familiar with visualization.

Table 1. Workshop Attendees

Gender	Male: 7 Female: 13
Area of Expertise	Life Sciences: 7 Informatics and Information Science: 4 Social Sciences: 2 Computer Science: 1 Engineering: 4 Physical Sciences: 2
Professional Status	Postdoctoral Researchers: 3 Doctoral Students: 17
Total Attendees	20

Table 2. Workshop Instructors

Gender	Male: 10 Female: 2
Area of Expertise	Informatics: 2 Computing and Computer Science: 5 Engineering: 1 Data and Data Science: 4
Total Instructors	12

3.2 Data Collection and Analysis

Here, we summarize our data collection and analysis processes, and discuss the limitations of our study design.

3.2.1 Data Collection. We collected data through multiple methods throughout the duration of the workshop. These included:

- Participant observation.
- Interviews with students and instructors.

Two of the authors participated in direct observations of the workshop. One participated as a student, sitting through the lectures, working through class assignments, participating in group work, and in general attempting to absorb the skills of data-intensive science. The other sat through each class session and observed the students' participation, how they asked questions or asked for help, how they collaborated, and how they moved between groups. Both of these authors also participated in workshop discussions over the Slack messaging service. These discussions included interactions between students and instructors, and instances of students helping each other with technical problems or posting resources. Both of these authors generated field notes based on their participation and observation.

During the last few days of the workshop, we conducted semi-structured interviews with 19 students and 12 instructors (one student declined to be interviewed). Tables 1 and 2 provide more information on the interviewees. Interviews lasted around 30 minutes each, and were recorded and transcribed for later analysis. The first four authors conducted interviews, with interviews distributed equally across the team.

Both instructors and students were informed that the goal of the interviews was to understand participants' experiences with and thoughts about the workshop, partly to inform the development of subsequent workshops. In the interviews, all participants–instructors and students–were asked about what reproducibility meant to them in the context of their own research. Subsequently, interviews varied somewhat for instructors and students. Instructors were asked about their pedagogical goals and whether the workshop had achieved those goals. Students were asked about their everyday activities with data, the kinds of data-related issues that were currently important for their own research, and what they had hoped to learn from the workshop. Students were then asked about what they had learned from the workshop, and what they felt like they still needed to learn. We also asked students how the workshop had affected their understanding of reproducibility, if at all. (We continued by asking students more specific questions about the effectiveness of the workshop structure and logistics, but that data was not relevant to the concerns of this paper.)

Our position as neutral evaluators and our goal of improving future curricula encouraged all participants–instructors and students–to be forthright in their responses to our questions, as we had no stake in any particular outcome. Instructors and students alike were generally quite direct in telling us about programmatic strengths and weaknesses. While most of these assessments are not relevant to the concerns of this paper, the honesty of these critiques gives us confidence in the interviews as data sources. (As another output of our process, we created an evaluation report where such comments were synthesized.)

3.2.2 Data Analysis. Following a variation of the classical approach to grounded theory [8], the authors developed open codes through group discussion. These discussions surfaced observations, ambiguities, and experiences of the two observing authors, who then took these sensitizing concepts back into the field site for further observation and engagement. Following Charmaz [8], the early analytical frame stayed close to the observations of the two authors in the field, emerging around the notion and importance of reproducibility, and the use of data tools in accomplishing it. The goal

of developing the initial concepts through discussion, rather than through textual word-by-word or line-by-line coding, was both a way of establishing an iterative engagement with the field site in the short amount of time available, but also a way of relating the observations of the two authors in the field. Subsequently, we went back to the interviews and coded them according to the categories that we had collaboratively developed. As we applied our initial codes to the interviews, we iterated and refined our understanding of these categories and their relationships.

4 FINDINGS

In this section, we summarize findings related to our two research questions:

- How do workshop participants develop their understanding of reproducibility around data tools and practices?
- How do workshop participants articulate the benefits of computational reproducibility?

Our findings indicate that participants (both instructors and students) view reproducibility in two dimensions:

- A global dimension, in which reproducibility is a means to ensure scientific integrity through the verification of results. In the global dimension, reproducibility emphasizes data sharing. The global dimension is more explicit: when participants were asked to define reproducibility, this is the dimension that they were likely to articulate.
- A local dimension, in which reproducibility is a means to facilitate collaboration. In the local dimension, reproducibility emphasizes code sharing. The local dimension is more implicit: this dimension emerged when participants were asked about the utility of what they learned in the workshop and how what they learned might be productively applied to their own work.

4.1 Participants' Explicit Understanding of Reproducibility: The Global Dimension

We began our interviews by asking participants to describe their own research areas. As part of this discussion, we asked participants to describe what reproducibility meant to them, in the context of their own work. In response to this direct question, participants invoked two primary definitions of reproducibility:

(1) Reproducibility is when someone else can repeat your data analysis and verify your results. Sometimes, participants expressed this definition as a matter of proper implementation, or using the correct state-of-the-art tools to achieve this goal (e.g., reproducibility is when you create a Docker or Singularity container for your data and code).
(2) Reproducibility is when someone else can reuse something that you generate–your methods, your data, your code–and achieve similar results.

According to the notions of reproducibility that we outlined in our related work section, Definition 1 is similar to what Cartwright calls "replication" and Schmidt calls "exact replication" [7, 50]. Definition 2 is similar to what Cartwright calls "reproduction" and Schmidt calls "conceptual replication." As we note in the related work section, the distinctions between these definitions can be subtle. With replication (Definition 1), however, there is more emphasis on verifying the results of the initial experiment by repeating it as closely as possible. With reproduction (Definition 2), there is more emphasis on extending the results of the initial experiment by reusing some of the original components (methods, data, code) in a new experiment. Across scientific fields, most participants invoked some variation of Definition 1 when explicitly defining reproducibility. For instance, participant T6, an instructor and computer scientist, defined reproducibility as

> taking what somebody else has done or what you have done and having someone else be able to perform the same analysis and obtain the same or similar results

and P2, a workshop attendee in health behavioral sciences, said

> if somebody wants to replicate our results they can just look at the data and run whatever analysis you need, and check if the results match theirs, that's reproducibility

whereas P12, a bioinformatics researcher said

> I think it should be fairly straightforward for another researcher, not necessarily like a high school student, but another researcher in my field to be able to run through all the analyses, take my data which is open source and publicly available, run through all the analyses and produce the same outputs

The workshop itself did emphasize this definition of reproducibility: the goal of the projects in the second week of the workshop was for the participants to take a published paper, reproduce the data analysis, and then make it easier for others to replicate the data analysis. The workshop curriculum, which introduced students to a range of data-management tools, probably contributed to the responses of some participants in associating the concept of reproducibility directly with its implementation using particular tools, such as container software. For instance, when asked about reproducibility, P8, another bioinformatics researcher, said

> at least for the near future containers are probably one of the best ways to make things accessible and reproducible

However, although the workshop may have reinforced an association between the concept of reproducibility and the use of certain tools for implementing it, this mode of understanding reproducibility was already familiar to many attendees. Participants would, for instance, refer to the tools that they had been using before the workshop and compare them to the tools that they had learned about during the workshop. As one example, when asked to define reproducibility for her own work, P4, whose research area is in consumer health informatics, began talking about the tools she had used before attending the workshop:

> Before this conference I used Jupyter notebooks and I can send a Jupyter notebook of Python files... being able to do that without having to spend hours on end of trying to replicate exactly what someone else has done, that saves lots of time.

When asked by the interviewer to confirm that this was what reproducibility meant to her, P4 continued by connecting the implementation of reproducibility with particular tools to the notion of verifying results:

> Being able to replicate, being able to do exactly what someone else has done, I think that is very important, and then being able to reproduce it, so having some kind of value after you're able to replicate it, that's kind of what I mean

Notably, although P4 refers first to reproducibility as "replication," or doing "exactly what someone else has done," in line with Definition 1, she continues by alluding to Definition 2, which she describes somewhat ambiguously as "having some kind of value after you're able to replicate it" and labels with the term "reproduce." P4 was not alone in moving back and forth between different concepts of reproducibility, often with minimal recognition of doing so. For instance, participant T1, one of the instructors who was a hydrologist, initially described reproducibility as making sure that another researcher can repeat what you did and verify your results:

> Reproducibility is very important... you have to make it reproducible and make sure that people could easily execute your work.

But as T1 continues to discuss reproducibility, it evolves to encompass Definition 2 as well:

> You want to share your data, you want to share your code, you want to make it reproducible... you want a platform that does all that, makes it easy for you to do...

As with P4 and T1, participants tended to offer Definition 1 first. This occurred even when they actually did not find Definition 1 particularly relevant for their own context. For instance, when the interviewer asked P13, a materials scientist studying molecular dynamics, what reproducibility meant to him, he responded this way:

> In my field, I guess not many people do the science over, at least not the simulations because those take a really long time...But you should...be able to get the same approximate results from running the same system under the same conditions.

But when asked by the interviewer how often he thought about reproducibility, P13 said that he thought about it "probably every day" in terms of setting and documenting simulation parameters–using other people's data and making his data easier for other people to use. P13 commented in this context that

> If somebody were to give me like a trajectory through our main data file, I couldn't figure that information out. A lot of the tweaks there aren't in the file itself; you have to rely on them giving you the log files that they have, which are not always in the same place. So...

And when asked about how what he learned in the workshop might benefit his work, P13 emphasized both data and code reuse:

> Especially having other people, giving them my script and saying you can do whatever with this, or to run this specific build or analysis.

The value of making code and data portable was therefore not just for the purpose of exact replication, but also for making one's analyses available to collaborators or to the field, so that it might be used on different data. This value was often captured under the term reproducibility. P3, for instance, describes the importance of making analyses available to others in her field:

> I have had experience trying to use other people's work and it not working and with the field I'm in the methods move very quickly so if a paper comes out and you want to try to use it, so having the reproducibility is very important.

4.2 Participants' Implicit Understanding of Reproducibility: The Local Dimension

The trajectory followed by P13 was not uncommon. When discussing the importance of reproducibility generally, participants tended to articulate the global dimension of reproducibility, focusing on the replication of an analysis to verify results. But when asked about what they found valuable about the workshop, they described benefits of reproducibility on a local level, for themselves and their immediate colleagues, rather than for the scientific community in general. This involved not only being able to give code or data to a collaborator and have them reuse it, but also to be able to pick up their own code and data at some point in the future and reuse it.

Sometimes, these local benefits involved compliance with policies for data sharing mandated by particular scientific journals within their discipline. For instance, P7, who studies engineering and economics, commented that

> Sometimes there are high-end journals that require you to give them the datasets, sometimes the codes

More often, however, participants were not directly affected by such institutional policies; some of them only learned about publisher mandates for data sharing at the workshop. P15, an environmental scientist, noted that

> In my research. . .they haven't required me to upload the code. . .but I learned from this workshop that there are more and more publishers that are trying to require their authors to submit the code.

Participants were more likely to describe the benefits of reproducibility in terms of collaboration and knowledge transfer. Both attendees and instructors related stories of needing to continue a project begun by a researcher who had left the lab, and not being able to reproduce that colleague's work. T1, for instance, shared a story during the workshop about her own experiences as a graduate student attempting to reuse data that was left to her by a predecessor. The difficulty of understanding how to reuse somebody else's data inspired her initial interest in reproducibility. P11, a bioinformatics researcher, observed that

> We have difficulty reproducing people's work, and that's a very standard [problem], I think, that goes across fields. It's challenging in my lab to have knowledge transfer between generations of students and post docs because of the size of my lab and because of the structure of it. So I was hoping with the containers to be able to compartmentalize workflows and compartmentalize data analysis so it's more easily integrated for a new user and so that when you go from generation to generation there's...you can try and bridge the gap and knowledge between the two groups.

Participants focused on being able to reuse code in order to perform subsequent analyses. If they could not replicate an initial analysis, then they would not be able to reuse scripts for subsequent work. For instance, P17, a mechanical engineer, recalled attempting to continue with and extend a senior colleague's work, but not being able to replicate that colleague's results first, which affected his own ability to make progress:

> First, I had to go through reproducing his research, so when I saw that his output file and my output files were not exactly the same...[I could not] proceed ahead with my research advancing what he was doing.

Others, such as P14, a computer scientist, focused on the time lost in rewriting software that could not be reused:

> The previous student of my professor wrote a huge code for processing and doing simulations for passengers' movement in flight. . . we tried to just improve the code, but unfortunately it was not well documented and written. At the end we just wrote new code, completely new code, so [the previous student's work] was useless for us...

Although P14 was a computer scientist by training, few of the attendees considered themselves coding experts, which may have contributed to their interest in reusing code. The structure of the social science group project, which was brought to the workshop by one of the students, provides an additional example of this. This project focused on reproducing a scientific tool, rather than replicating findings from a paper. Although most of the attendees were already using computational techniques to analyze their data, they identified their expertise as data analysts, not as software developers, and most of them considered writing or debugging software–as opposed to customizing or using software–to be extraneous to their primary goals.

This distinction between acceptable and excessive forms of coding caused friction for another of the group projects, the bioinformatics project. For this project, the study that students were supposed to replicate was five years old, and, despite the efforts of the original authors to provide data, code, and documentation, some of the necessary code to replicate the original analysis could not be installed properly and needed to be debugged. Although this sort of problem was very realistic, it meant that the members of this project group needed to spend unanticipated time performing software engineering tasks, rather than replicating the analysis and containerizing the

workflow with current tools–the skills that they had learned in the workshop. One of the group members, P12, a bioinformatics researcher herself, found this situation very frustrating, and she abandoned this project and joined another group instead. P12 observed that

> we were so busy troubleshooting software that wasn't ours that we weren't familiar with and rewriting code...[it] was really frustrating.

However, others on this project team, while finding this situation to be unexpected, did not find it to be an inappropriate use of their time and expertise. For P11, another bioinformatics researcher on this project team, the level of software engineering required was reasonable:

> It was challenging certainly, but a challenge is a good thing to a certain extent... Everybody has different levels of motivation, or perseverance I think is a better word. I reached my limit a couple times, but then I just took a break and come back and started working on it again.

Unlike P12, P11 saw the kinds of programming tasks required by the workshop project as a necessary component of data analysis. P11 described how, in his lab, he has to learn new software tools through a process of trial and error, which has required him to build up his own coding skills:

> my PI doesn't pretty much know programming so I'm pretty much all self taught. I'm used to having to read when I'm using a new tool, reading a lot of documentation, looking through the options. It's a hard way to learn but it's a good way to learn. . .

The different attitudes that P12 and P11 (both bioinformatics researchers working with genomics data in their own work) exhibited towards the level and type of coding seen as necessary for data analysis are similar to the differences noted by Kandel et al. (2012) in their study of enterprise data scientists. In Kandel et al.'s study, some analysts ("hackers") employed multiple kinds of programming tasks in their work, whereas others ("scripters") primarily used analysis software packages such as R, and others ("application users") used applications like Excel or SPSS (Kandel et al., 2012). This significance here is that participants with greater and lesser levels of skill and interest in coding tasks (hackers, scripters, and application users alike) considered the reuse of code as a key aspect of reproducibility, in terms of personal benefit.

Notably, when participants described the benefits of reproducible code for collaboration and knowledge transfer, they were less likely to articulate similar benefits for sharing data and making data reproducible (for instance, in using metadata standards or documenting a dataset's metadata). Most of the participants did not collect data themselves: they analyzed public data or data produced by others (for instance, genomics data sequenced by an external facility, or electronic health records made available by a healthcare provider). Within their local teams (for instance, within a single lab), they were dealing with data from similar sources (or the same source), with similar structure; if they were interpreting the data differently from their local colleagues, they were not aware of it, and did not see this as a significant barrier to working together. Although a few participants described existing challenges in harmonizing data from multiple sources or in managing data for different clients–for instance, P1, a plant pathologist, talked about needing to ensure that metadata was appropriately recorded for his data, which comprised samples sent from around the U.S.–these were not generally significant concerns, in terms of day-to-day activities with colleagues.

5 DISCUSSION

Our findings suggest that domain scientists' integration of data-science activities into their work practices affects their understanding of reproducibility in several ways:

- In the global dimension, domain scientists are adjusting how they understand the implementation of reproducibility: how it is to be achieved.

- In the local dimension, domain scientists are expanding how they understand the motivation for reproducibility: why it is to be achieved.

In the following sections, we discuss these conceptual adjustments to the how (implementation) and why (motivation) of reproducibility.

5.1 The Implementation of Reproducibility: From Documentation to Automation

As summarized by Fidler and Wilcox, a key challenge regarding reproducibility has been what Collins called "the experimenter's regress": if the results of a replicated experiment do not confirm the results of the original experiment, is the problem with the original experiment or with the replicated one (Fidler and Wilcox, 2018)? Perhaps the replication did not sufficiently reenact the conditions of the original experiment, and the flaw is with the replicator, not the originator. The experimenter's regress underscores the conceptual fragility of reproducibility: there will always be some doubt as to whether a replication was sufficient, because a complete replication is impossible, no matter how extensively an experiment was documented. From an absolute perspective, this problem seems irresolvable. Just as one can never step into the same river twice, one can never truly replicate anything.

However, as computational techniques for data analysis have been assimilated into many scientific domains, there is a rising sense that using computation to automate operations upon data may limit the uncertainty of the experimenter's regress. In Kitzes, Turek, and Deniz's introduction to their 2018 handbook on the practice of reproducible research for data-intensive sciences, automation–the ability to produce all of a study's calculations and visualizations with "a single button press"–is the first strategy that they introduce for implementing reproducibility. If automation is sufficiently enabled, then documentation (which Kitzes, Turek, and Deniz call "provenance tracking") "can be instantiated and executed with reasonably minimal effort" [36]. Our participants, likewise, associated increased automation with increased reproducibility, to the extent that state-of-the-art technologies for facilitating automation–in particular, software containerization–were occasionally seen as a shorthand for reproducibility in general (e.g., using Docker to containerize your workflows "is" reproducibility).

The reorientation to automation, rather than documentation, as a principal strategy for implementing reproducibility is subtle, but significant. It narrows the scope of reproducibility as a concept, focusing it more closely on the verification of results. Additionally, it transfers the primary articulation of the researchers' understanding of their data to the details of the code used to manipulate it; documentation is oriented around the code, and describes the data through the code. The tendency of some of our participants to identify reproducibility with the use of particular technologies further limits and reorients the conceptual scope of reproducibility. Reproducibility therefore takes on a specific meaning in relation to a set of data tools, and this understanding is quite different from (though not incompatible with) statistical and metascience responses to the reproducibility crisis in other communities [20].

The workshop bioinformatics project, which students struggled to reproduce, provides a revealing situation here. As described in the previous section, the study that students were supposed to replicate was five years old; the study's authors had published data, code, and documentation, but the students had difficulties even installing the software that they were supposed to use. As it turned out, the instructors who selected this study for the workshop had not anticipated the level of difficulty that students experienced in attempting to replicate the study; they hadn't tested the tasks that they asked students to perform.

Among the students, there was a sense that their project had so many challenges because appropriate technologies to facilitate reproducibility–particularly in terms of alleviating issues with

unknown software dependencies–had not been developed when the original study was published. Accordingly, the original study was documented, but it was not sufficiently automated. As P11 explained, although the original study "was intentionally made to be reproducible" and "there was a lot of documentation," it was five years old, and "five years of computer space is like a century."

It would have been possible to perceive the situation with the bioinformatics project as a warning not to rely on current technologies to encapsulate the notion of reproducibility in a stable manner. Given the effort that they had taken, the authors of the original study had clearly imagined that the data and code that they had carefully documented and made available would remain viable for future research. Indeed, the workshop instructors themselves were caught somewhat unawares that the study would be so difficult to replicate. Instead, though, this experience seemed to paradoxically reinforce the association of reproducibility as a concept with its implementation via current tools. Participant P12, who switched project teams because of her frustration with so much time being spent debugging software, provides an illustrative example of this. P12 noted that, despite her dissatisfaction with the project, it was true to her everyday experience in the lab, commenting that

> I laughed about it with one of the professors, it's a real life example. This is exactly how it normally goes.

Despite this daily experience of encountering studies that do not accurately anticipate the conditions of future researchers, P12 continued by imagining that the difficulties posed by the project were due to a lack of appropriate technologies, hypothesizing that today's tools would obviate such problems in the future:

> [the bioinformatics project] highlights the importance of using containers and using virtual environments because once you get it into that format you can reproduce it...

5.2 The Motivation for Reproducibility: From Abstract Value to Concrete Benefit

The concept of reproducibility is deeply integrated with normative values regarding good scientific practice. One way of understanding the renewed interest in reproducibility that we describe in our related work section is as a reassessment of community norms regarding standard procedures for data collection and analysis. It is not, for example, that social psychologists whose studies could not be replicated in the Open Science Collaboration 2015 review were practicing poor science, but that the Open Science Collaboration was, through their review, suggesting that the scientific community should adapt its understanding of reproducibility and of appropriate mechanisms to achieve it [10].

However, although the recent conversation around reproducibility has created new ideas about what reproducibility involves and how to implement it, there has been less discussion about why reproducibility is important. Fundamentally, reproducibility continues to be associated with abstract notions of scientific integrity and epistemic confidence. The specific advantages of reproducibility remain vague and future-oriented, promising long-term value for the scientific enterprise as a whole. Although there may be increasing incentives in some scientific domains to adopt newly sanctioned processes regarding reproducible research, especially in fields like social psychology, where the results of previous studies have been questioned, these incentives tend to be punitive, rather than generative: not following these practices may lead to rejection by journals, or, more catastrophically, not following these practices might lead to public shaming, should one's study not be successfully replicated. Writing about the reproducibility "revolution" in social psychology for the New York Times Magazine, journalist Susan Dominus quotes Jay van Bavel, a social psychologist, on having his own work unsuccessfully replicated: "It is terrifying, even if it's fair and within normal scientific bounds" [16].

Unsurprisingly, our participants did not find these future-oriented motivations for reproducibility to be compelling. However, they did locate more immediate, concrete benefits to adopting techniques associated with reproducibility, focused around collaboration and knowledge transfer within local teams. This focus on local and concrete motivations, rather than global and abstract motivations, narrows and reorients the conceptual scope of reproducibility in a manner similar to that described in the previous section (although to different effect, as we will discuss in the following section). From a motivational perspective, reproducibility becomes understood more immediately as a set of practices for effective collaboration on continuing projects where key team members might be rotated on a regular basis–as is common with research lab situations–rather than a set of practices for opening one's work to general scrutiny from the overall scientific community. To return to the comments of P13, the materials scientist, it is unusual in his discipline to "do the science over" in alignment with the more global, abstract notion of reproducibility; nonetheless, P13 thinks about reproducibility "probably every day" in terms of making data and code reusable for others to pursue their own, similar projects.

5.3 Conceptual Contradictions and Pragmatic Associations: Relating the Motivation for Reproducibility (Local Dimension) and Its Implementation (Global Dimension)

In the related work section, we describe how the term "reproducibility" has been associated with a number of concepts that are closely related but not equivalent. Although some scholars have attempted to systematically distinguish these related concepts with different names (such as Cartwright's "replication" and "reproducibility" or Schmidt's "exact replication" and "conceptual replication"), most people tend to mix these different concepts and terms without realizing it. In our findings, we observed that our participants, too, often referred to "reproducibility" in multiple senses, typically without recognizing that they were using the same name for different (albeit very similar) concepts.

In our study, participants seem to understand reproducibility along two distinct dimensions: the global dimension, in which reproducibility is a means to ensure scientific integrity through the verification of results, and a local dimension, in which reproducibility is a means of collaboration and knowledge transfer. Conceptually, there is some friction between these two dimensions. In the global dimension, the influence of data science has manifested in a specific way of understanding how reproducibility is implemented, where reproducibility is best achieved through automation, as with containerization technologies. In the local dimension, the influence of data science tools and practices has manifested in a specific way of understanding the motivation for reproducibility, where a key benefit of reproducibility is code reuse on local teams. Some similar distinctions are drawn in the literature. The emphasis on automation in the global dimension implies an understanding of reproducibility as an exact replication, where the same data and tools are employed to verify a result. But the emphasis on code reuse in the local dimension implies an understanding of reproducibility as a conceptual replication, where the same tools are employed to pursue a new project that extends the original results. This observation goes beyond just an analytical distinction between reproduction and replication; it suggests that if we frame the problem of designing tools and practices for reproducibility around the notion of a reproducibility crisis we may be misunderstanding the motivations and sensibilities of many researchers on the ground.

Pragmatically the two dimensions are aligned in their association with current state-of-the art technologies (particularly container software and computational notebooks). Our findings here support and extend other recent studies. Feger et al. [18], for instance, has pointed to the collaborative benefits of reproducible computing as a way of incentivizing reproducible practices, suggesting that both purposes are tied to the same practices and tools. In a similar vein, Rule (2018) defines two different uses of computational notebooks: exploratory uses, in which researchers make

messy notes and code in the course of their personal explorations, and explanatory uses, in which the notebook is ordered and polished for replicating or presenting. Participants in the workshop conflated both the global and local understandings we outlined under the term reproducibility, and particularly under the notion of using computational methods for reproducibility. Rather than interpreting this as a misunderstanding or confusion, we point to this conflation of local and global as an indicator of researchers' understandings of underlying changes around the adoption of data tools. Some part of the crisis of reproducibility, especially in the context of data-intensive science, is not so much an epistemological crisis of scientific verification, but rather a more prosaic crisis of the shift towards collaborating around large data sets and tangled codebases. The shift to data science tools and methods implicates a shift not only in the way researchers verify a particular analysis, but more broadly the way they define good scientific work. If we understand the crisis as one of exact replication only, then the entanglement of replication and collaboration seems like a conflation, contradiction, or coincidence. If, however, we look at data science tools as being integral in the way researchers share and evidence their work, then the two understandings of reproducibility are sensibly connected.

6 CONCLUSION AND FUTURE WORK

In this paper, we've shown how the integration of data-science tools and techniques into the work practices of domain scientists coincides with a reorientation of the concept of reproducibility. As our participants assimilate data-science activities into their work, they begin to understand reproducibility more strongly along a local dimension–as a means to collaborate and facilitate knowledge transfer, via the implementation of state-of-the-art technologies, such as containerization software–and less strongly along a global dimension, associated with abstract values of scientific integrity.

Our goal here is not to generalize these specific findings to all disciplines in all locales. Rather, we use these findings to sensitize the research community to a number of useful approaches to understanding the proliferation of data tools in scientific work. Firstly, it suggests a way of examining the reproducibility movement not as a monolithic crisis, but as a highly variegated movement across different research cultures. Secondly, while the effort that has been put into constructing analytical distinctions between concepts like replication and reproducibility is valuable, there is also value in following the researchers' own understandings of reproducibility in practice. As argued by Jackson, Steinhardt, and Buyuktur [31], such meanings are crucial to policy and design of computational tools. Finally, we can avoid taking data tools as a neutral substrate for scientific investigation, and instead examine how such tools embody shifting standards of excellence in scientific work. Changes in tools have real consequences for the way researchers evidence their findings and evaluate the findings of others, and are at the center of the "culture change" occurring around reproducibility. This observation lends new recognition and new responsibility to the work of research software engineers, technicians, developers, and designers in science.

As the availability of large datasets continues to increase, and as the tools and techniques to manage and analyze these datasets become more widely accessible, we can expect that more people in diverse fields and professions will begin to integrate data-related tasks into their pursuits. These other domains of practice are likely to alter around the adoption of data work. In this paper, we provide an illustrative example of how the uptake of data-science activities might affect how people understand the fundamental concepts that inform their work. Still, because of the nature of our study, our findings are necessarily localized. While the workshop setting provides a strategic point of observation, it also presents some limitations. Extending the study longitudinally, and looking at the use of reproducibility tools in everyday scientific practice, rather than in the educational workshop setting would be valuable. Furthermore, as shown by the distinction

researchers made in the workshop between acceptable and excessive forms of coding, the kinds of tasks that constitute science are shifting in response to the adoption of data practices. Another area of fruitful study might explore the evolving divisions between science and often-sidelined types of technical work, such as research software engineering. Finally, there are strong affinities between the concerns of researchers in doing reproducible science and the concerns of the broader software development community in developing software effectively through documentation, adhering to development methods, and avoiding technical debt. These affinities are well worth investigating as further characteristics that distinguish computational reproducibility from broader notions of reproducibility, and as developing connections between scientific software development and industrial cultures of software development.

7 ACKNOWLEDGEMENTS

This work was funded by the U.S. National Science Foundation grant 1730390.

REFERENCES

[1] Karen Schepeler Baker. 2017. *Data work configurations in the field-based natural sciences: mesoscale infrastructures, project collectives, and data gateways.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
[2] C Glenn Begley, Alastair M Buchan, and Ulrich Dirnagl. 2015. Robust research: Institutions must do their part for reproducibility. *Nature* 525, 7567 (2015), 25–27.
[3] Matthew J Bietz, Eric PS Baumer, and Charlotte P Lee. 2010. Synergizing in cyberinfrastructure development. *Computer Supported Cooperative Work (CSCW)* 19, 3-4 (2010), 245–281.
[4] Jeremy P Birnholtz and Matthew J Bietz. 2003. Data at work: supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*. 339–348.
[5] Christine L Borgman. 2015. *Big data, little data, no data: Scholarship in the networked world.* MIT press.
[6] Daniel Carter and Dan Sholler. 2016. Data science on the ground: Hype, criticism, and everyday work. *Journal of the Association for Information Science and Technology* 67, 10 (2016), 2309–2319.
[7] Nancy Cartwright. 1991. Replicability, reproducibility, and robustness: comments on Harry Collins. *History of Political Economy* 23, 1 (1991), 143–155.
[8] Kathy Charmaz. 2014. *Constructing grounded theory.* sage.
[9] Lyra J Colfer and Carliss Y Baldwin. 2016. The mirroring hypothesis: theory, evidence, and exceptions. *Industrial and Corporate Change* 25, 5 (2016), 709–738.
[10] Open Science Collaboration et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.
[11] Harry M Collins. 1985. Replicating the TEA-Laser: Maintaining scientific knowledge. *Ders.: Changing order. Replication and induction in scientific practice. London/Beverly Hills: Sage Publications* (1985).
[12] Mark J Costello. 2009. Motivating online publication of data. *BioScience* 59, 5 (2009), 418–427.
[13] Jonathon N Cummings and Sara Kiesler. 2008. Who collaborates successfully? Prior experience reduces collaboration barriers in distributed interdisciplinary research. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 437–446.
[14] Renata Gonçalves Curty, Kevin Crowston, Alison Specht, Bruce W Grant, and Elizabeth D Dalton. 2017. Attitudes and norms affecting scientists' data reuse. *PloS one* 12, 12 (2017).
[15] DataONE. 2018. DataONE. https://www.dataone.org/what-dataone. Accessed: 2019-03-05.
[16] Susan Dominus. 2017. When the revolution came for Amy Cuddy. *The New York Times* (2017), 29.
[17] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. 2015. What drives academic data sharing? *PloS one* 10, 2 (2015).
[18] Sebastian S Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W Woźniak. 2019. Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[19] Jeremy Freese and David Peterson. 2017. Replication in social science. *Annual Review of Sociology* 43 (2017), 147–165.
[20] Jeremy Freese and David Peterson. 2018. The emergence of statistical objectivity: changing ideas of epistemic vice and virtue in science. *Sociological theory* 36, 3 (2018), 289–313.
[21] Carole Goble, David De Roure, and Sean Bechhofer. 2011. Accelerating scientists' knowledge turns. In *International joint conference on knowledge discovery, knowledge engineering, and knowledge management*. Springer, 3–25.
[22] Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. 2016. What does research reproducibility mean? *Science translational medicine* 8, 341 (2016), 341ps12–341ps12.

[23] Jim Gray, David T Liu, Maria Nieto-Santisteban, Alex Szalay, David J DeWitt, and Gerd Heber. 2005. Scientific data management in the coming decade. *Acm Sigmod Record* 34, 4 (2005), 34–41.

[24] Melissa A Haendel, Nicole A Vasilevsky, and Jacqueline A Wirz. 2012. Dealing with data: A case study on information and data management literacy. *PLoS biology* 10, 5 (2012).

[25] Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11, 3 (2013), 156–162.

[26] Tony Hey, Stewart Tansley, Kristin Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.

[27] James Howison and Julia Bullard. 2016. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology* 67, 9 (2016), 2137–2155.

[28] PARSE Insight. 2009. Insight into digital preservation of research output in Europe: Survey report.

[29] John PA Ioannidis, Marcus R Munafo, Paolo Fusar-Poli, Brian A Nosek, and Sean P David. 2014. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences* 18, 5 (2014), 235–241.

[30] Steven J Jackson and Sarah Barbrow. 2013. Infrastructure and vocation: field, calling and computation in ecology. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2873–2882.

[31] Steven J Jackson, Stephanie B Steinhardt, and Ayse Buyuktur. 2013. Why CSCW needs science policy (and vice versa). In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1113–1124.

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

[33] Helena Karasti and Jeanette Blomberg. 2018. Studying infrastructuring ethnographically. *Computer Supported Cooperative Work (CSCW)* 27, 2 (2018), 233–265.

[34] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

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

[36] Justin Kitzes, Daniel Turek, and Fatma Deniz. 2017. *The practice of reproducible research: case studies and lessons from the data-intensive sciences*. Univ of California Press.

[37] Sabina Leonelli. 2016. *Data-centric biology: A philosophical study*. University of Chicago Press.

[38] George E Marcus. 1998. *Ethnography through thick and thin*. Princeton University Press.

[39] David R Millen. 2000. Rapid ethnography: time deepening strategies for HCI field research. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. 280–286.

[40] 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*. 1–15.

[41] 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 (2017), 85–97.

[42] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.

[43] Irene V Pasquetto, Bernadette M Randles, and Christine L Borgman. 2017. On the reuse of scientific data. (2017).

[44] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[45] Gregory Piatetsky and Preet Ghandi. 2018. How many Data Scientists are there and is there a Shortage. *KDNuggets* (2018).

[46] Heather A Piwowar. 2011. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one* 6, 7 (2011).

[47] David Ribes. 2014. Ethnography of scaling, or, how to a fit a national research infrastructure in the room. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 158–170.

[48] Dominique G Roche, Robert Lanfear, Sandra A Binning, Tonya M Haff, Lisa E Schwanz, Kristal E Cain, Hanna Kokko, Michael D Jennions, and Loeske EB Kruuk. 2014. Troubleshooting public data archiving: suggestions to increase participation. *PLoS biology* 12, 1 (2014).

[49] Caroline J Savage and Andrew J Vickers. 2009. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one* 4, 9 (2009), e7078.
[50] Stefan Schmidt. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of general psychology* 13, 2 (2009), 90–100.
[51] Dan Sholler, Sara Stoudt, Chris Kennedy, Fernando Hoces de la Guardia, Francois Lanusse, Karthik Ram, Kellie Ottoboni, Marla Stuart, Maryam Vareth, Nelle Varoquaux, et al. 2019. Resistance to Adoption of Best Practices. (2019).
[52] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PloS one* 8, 6 (2013).
[53] Victoria Stodden and Sheila Miguez. 2013. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Available at SSRN 2322276* (2013).
[54] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. Data sharing by scientists: practices and perceptions. *PloS one* 6, 6 (2011), e21101.
[55] Carol Tenopir, Elizabeth D Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one* 10, 8 (2015).
[56] Erik H Trainer, Chalalai Chaihirunkarn, Arun Kalyanasundaram, and James D Herbsleb. 2015. From personal tool to community resource: What's the extra work and who will do it?. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 417–430.
[57] Jillian C Wallis, Elizabeth Rolando, and Christine L Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS one* 8, 7 (2013).
[58] Jelte M Wicherts, Marjan Bakker, and Dylan Molenaar. 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one* 6, 11 (2011).
[59] Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K Teal. 2017. Good enough practices in scientific computing. *PLoS computational biology* 13, 6 (2017).
[60] Ann S Zimmerman. 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values* 33, 5 (2008), 631–652.

Received October 2019; revised January 2020; accepted March 2020