How Data Science Workers Work with Data:

Discovery, Capture, Curation, Design, Creation

Michael Muller*, Ingrid Lange*, Dakuo Wang*, David Piorkowski*, Jason Tsay*, Q. Vera Liao*, Casey Dugan*, and Thomas Erickson**
*IBM Research, Cambridge, MA, USA **Minneapolis, MN, USA
*{michael_muller, cadugan}@us.ibm.com **snowfall@acm.org
*{ingrid.lange, dakuo.wang, david.piorkowski, jason.tsay, vera.liao}@ibm.com

ABSTRACT

With the rise of big data, there has been an increasing need for practitioners in this space and an increasing opportunity for researchers to understand their workflows and design new tools to improve it. Data science is often described as data-driven, comprising unambiguous data and proceeding through regularized steps of analysis. However, this view focuses more on abstract processes, pipelines, and workflows, and less on how data science workers engage with the data. In this paper, we build on the work of other CSCW and HCI researchers in describing the ways that scientists, scholars, engineers, and others work with their data, through analyses of interviews with 21 data science professionals. We set five approaches to data along a dimension of interventions: Data as given; as captured; as curated; as designed; and as created. Data science workers develop an intuitive sense of their data and processes, and actively shape their data. We propose new ways to apply these interventions analytically, to make sense of the complex activities around data practices.

ACM Reference format:
Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *CHI Conf. on Human Factors in Computing Systems Proceedings (CHI'19), May 4-9, 2019, Glasgow, Scotland UK.* ACM, New York, NY, USA. Paper 126, 14 pages. https://doi.org/10.1145/3290605.3300356

KEYWORDS: Data science; work-practices; data discovery, data capture; data curation; data design; data creation; grounded theory.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK.
© 2019 ACM. ACM ISBN 978-1-4503-5970-2/19/05...$15.00.
DOI: https://doi.org/10.1145/3290605.3300356

1 INTRODUCTION

Data science is often described as a rational "data-driven" process of "discovery" that reveals the underlying nature of a domain (e.g., [44], [47], [48], [53], [76], [91], [92]). A 2015 keynote address at an ACM conference asserted that, "Increasingly it is data, vast amounts of data, that drives scientific discovery" [27] and, two years later, a second keynote stated that "The value of data explodes when it is integrated" [65]. Similar claims about data science appear in editorials of journals:

> the computer becomes an active question asking machine as opposed to a pure analytic servant. By initiating interesting questions and refining them without active human intervention, it becomes capable of creating new knowledge and making discoveries on its own… [2]

The theme of autonomous data science technology, also appears in scholarly papers: "[W]e develop the Data Science Machine, which is able to derive predictive models from raw data automatically" [59]. While some of these claims are stated for the future [24], Agarwal and Dhar editorialize that "This is powerful… we are in principle already there" [2]

There do not appear to be any humans in this view of data science. Computations occur "automatically" and "without active human intervention." The computer "becomes capable of creating new knowledge… on its own." And yet, as Yang et al. have shown, "crafting these solutions generally requires knowledge that is possessed by only a few" [98], [99]. Shah et al. observe tartly, "Good data won't guarantee good decisions" [86]. Scholars in the earlier tradition of Science and Technology Studies (STS) have questioned claims of objectivity in data [8], [37], [51], [79], [82]. In this paper, we join other HCI and CSCW researchers [28], [71], [73] in applying those STS lessons to data science, and in re-centering an account of data science on the humans who do the work.

As we will show, human expertise intervenes between the raw data and the analysis, crucially shaping the data, the choice of analysis, and in some cases the truth claims associated with the analysis. We propose a classification of those human interventions, and we associate them with two interrelated HCI literatures.

In this paper, we provide a background to our project, emphasizing the people who do the work, the work itself, and a pair of theoretical frames that we used as part of the grounded theory method of sensitizing concepts, to focus our attention while coding data. We then describe our ground theory approach, followed by our investigation of how data science workers interact with their data through 20 in-depth interviews. We conclude with implications for design and for organizations, and proposals for future research.

2 BACKGROUND

2.1 Data Science Workers

If we talk about data science as a human activity, then of course we must talk about the people who do the work. There is little agreement about the role of "the data scientist" in terms of personal attributes, necessary skills, or work outcomes [17]. The Kaggle organization conducted a survey of 16,716 people who work in data science, enumerating 15 distinct job titles [41], [42], [47]. For the 63% of respondents who answered this question, "Data scientist" was the modal answer at 15%, followed by "Engineer" (14%), "Researcher" (10%), "Data analyst" (7%), and "Business analyst" (5%) (each of the remaining categories accounted for less than 5% of the sample, respectively).

There is a growing blog discussion about differences among some of these roles (e.g., [4], [57]). Miller discussed future needs for data scientists, and recommended to consider "other big data jobs [such as] information strategists, information system professionals, and data governance and ethics professionals" [66]. Kim et al. reported on 16 interviews of people in data science activities [53]: Out of eight distinct workplace roles, "Data scientist" accounted for only 33% of the sample, followed by "Director" and "Program Manager." We conclude that many people are doing the work of data science, across multiple job categories and titles, and that this diversity is likely to increase over time [14], [69]. Data scientists use diverse tools, ranging from simple programming languages through shared notebooks through graphical canvases for laying out a data pipeline [42], [47], and apply diverse skills from data-wrangling through modeling to visualization [42], [92]. On that basis, we have expanded our overall role-description from "data scientist" to "data science worker."

2.2 Data Practices in Data Science

In one perspective, data science work is a subset of larger concerns around "big data" as a reciprocal or interactional partner in shaping and being shaped in assemblages with humans, methods, and tools [87]. Scholars in STS have developed rigorous and urgent critiques of the semantics and politics of big data in society [8], [37], [82], and have made powerful arguments to problematize ways in which algorithmic practices fade into invisible infrastructures [51], [79].

By contrast, this paper focuses on data and data practices in data science endeavors. We are concerned with local applications of data science methods to proximate research questions (e.g., [7]). We contribute to opening the "[opaque] box" through which the dynamic human shaping of data [74] tends to erase the data's "epistemic charge" to become part of future, largely invisible infrastructures [79], [82]. We adapt Ruppert's proposed strategy for social science engagement with big data experts [81], and we begin a first update of that strategy to engage with data science workers.

For both practical and theoretical reasons, it is important to understand how data science workers engage with their data. In practical terms, data cleaning – or more broadly, data wrangling [39], [49], [78], [88] - has been described as requiring up to 80% of the time and effort in a data science project [36], [49], [78]. "Dirty data" was the most commonly-reported challenge in the Kaggle survey [41]. As Sutton et al. note, large datasets often present multiple problems, leading to "death by a thousand wranglings" [88]. Martin reports that data science workers tend to spend less time doing analyses, and more time preparing their data: "Weeks or months is a realistic time frame. Hours is not." [62]. Engaging with data takes a lot of time and effort.

Researchers in HCI and CSCW have developed intriguing theoretical insights into how data science workers approach their data. Passi and Jackson described an on-going tension over the use of algorithmic rules [71]. They propose that data science students can learn to practice a kind of *data vision* that treats rules more as guidance ("rules-based") than as formal constraint ("rules-bound"):

> Effective algorithmic analysis... demands mastery of the ways that worlds and tools are put together, and which worlds and tools are so combined (across the wide range of methods, tools, and objects amenable to representation). Taken together, these two seemingly contradictory features constitute what we call data vision:

> the ability to organize and manipulate the world with data and algorithms, while simultaneously mastering forms of discretion around why, how, and when to apply and improvise around established methods and tools in the wake of empirical diversity.

Their work provides insights into how humans work within supposedly determinate algorithmic systems, and how humans take responsibility to make their own meaning that is *informed* by algorithms, but not *governed* by those algorithms. In this paper, we begin a more extended view of human formative work practices in data science.

2.2.1 Discovery of Data

Several scholars have questioned the nature of "data" in the specific practices of data science and related endeavors. Working in the hybrid space of French and English literatures, Bilis contrasted two views of the analyst's relationship with data [10]. In one view, the analyst takes a relatively passive stance, and *receives* data as "given" by the environment ("donné"). The analyst may take responsibility for finding an appropriate dataset, but that history is quickly forgotten. As a result, the assumption becomes that the dataset itself is complete and objective.

2.2.2 Capture of Data

Bilis contrasted this view with a more active activity in which the analyst *captures* data ("capta"). In this second view, the analyst takes a more determinate role, choosing which data to capture, arranging how to capture the data, and deciding when the data are complete. The analyst is thus accountable for her or his choices. (See Charmaz's constructivist grounded theory method for a comparable argument about accountability in social science [16].)

Pine and Liboiron offer a related but more explicitly organizational account of big data practices [73] (emphases in the source text):

> Though often described as 'raw,' this data is produced by techniques of measurement that are imbued with judgments and values that dictate what is counted and what is not, what is considered the best unit of measurement, and how different things are grouped together and "made" into a measurable entity... [H]uman-computer interactions start *before* the data reaches the computer because various **measurement interfaces** are the invisible premise of data and databases.

Pine and Liboiron were particularly concerned with political assumptions and motivations that may shape the ways that data are captured (e.g., [10]), labeled, and stored. Scholars in the Value Sensitive Design tradition have studied older cases in which the design of a system influenced the kind of data that could be represented, and the reality that was consequently asserted by that system [31]. While we agree with these analyses, we take up only a part of their aggregate lesson in this paper: namely, the ways that people shape the data that they subsequently analyze.

2.2.3 Curation of Data

Mentis et al. described the use of visual evidence in medicine as a matter of "crafting the image" – i.e., selecting aspects of the data so that it contains those attributes that will be useful to the people who will view it [64]. Taylor and colleagues showed how a local community actively curated physical and digital data about its locality, and "enacted a multiplicity of 'small worlds.'" [89].

These acts of curation extend Bilis's concept of capture into an even more active and more principled organization of the data for a particular user or usage. In addition to the human users implied by Mentis and by Taylor, Kandel et al. wrote about making data "palatable to databases, statistics packages, and visualization tools" [49].

2.2.4 Design of Data

Feinberg et al. have extended this problematization of data practices [28], [29], [30]. In contrast with the conventional sense of data that Bilis described as "given" by the environment or more actively "captured" from the environment [10], Feinberg and colleagues describe data as *designed* [28]. The actions of designing data may be necessary in order to make the data tractable or analyzable. Kiss and Szirányi note that "data sets used for cyber-security research are most commonly handcrafted" [55]. In a trivial sense, we do this whenever we deal with missing data (e.g., through imputation or elimination), and we do this again when we consider how to use (or not use) outliers and other anomalous values. In the necessary practices of dealing with missing or anomalous values, we engage in active *curation* of our data.

Patel et al. showed that data science workers engage in creating features as inputs to formal modeling software [72], in contrast to the more conventional, "given" data concept [10] or of extracted features [48], [85]. In a particularly strong example, Kiss and Szirányi recommend the construction of "synthetic data" for certain problems that require large, difficult-to-acquire datasets [55]. In all of these research programs, data appear to be more *produced* than discovered or revealed. We emphasize that these acts of production are not cause for scandal, but are reported as necessary professional practice.

Feinberg and colleagues considered both simple, unitary datasets, and also more complex, integrated or combined datasets [28]. While "[t]he value of data explodes when it is integrated" [65], Feinberg et al. show that data science workers engage in complex translation practices during the integration of datasets [30]. One consequence of integration is the tendency to exclude non-conformant data [29]. The integrated dataset may thus amplify the types of elisions and transformations that concerned Bilis, Pine and Liboiron, and Taylor et al., and that we described above as a matter of curation.

2.2.5 Summary and a Look Forward

We have described four types of human interventions in relation to data: discovery, capture, curation, and design. Later in the paper, we will introduce a fifth type – namely, creation. As we move across these interventions, we move away from a naïve view of "raw" data (as critiqued by [32], [70]), and further into data as a human-influenced entity. Indeed, even the "discovery" of data [10] carries an implication of human agency: If data are discovered, then *someone* must be doing the discovering. These degrees of relationship become important when we consider where human interpretation enters into data science work practices, and hence data science processes. We will develop an analytic dimension of data interventions, and we will propose further research to test that dimension.

2.3 Skilled Work

Data science is generally considered to be a form of skilled work (e.g., [40]). We therefore look to the literature on expertise and skilled work for perspectives that can inform our analysis of the skilled work of data science. When a data science worker confronts a new project, they have to engage in sensemaking to understand what the project is, and what it requires. Russell et al. [83] characterized sensemaking as

> Sensemaking can be a core professional task in itself, as it is for researchers, designers, or intelligence analysts... It arises when new problems, opportunities, or tasks present themselves, or when old ones resurface. It involves finding the important structure in a seemingly unstructured situation...

In this section, we briefly review two lenses for considering skilled sensemaking in data science.

2.3.1 Expertise

Data science involves data, concepts, and methods, and thus requires expertise. Three decades ago, Chi et al. published an influential account of expertise in diverse human activities, which has been recently re-issued [19]. In their integrative overview, she and R. Glaser summarized seven attributes of expertise and experts [35]. Four of those attributes may serve as sensitizing concepts (i.e., similar to B. Glaser's coding families in ground theory [34]) in our analysis:

1. **Domain-specificity.** Experts excel mainly in their own domains [19]. Expertise occurs at the intersection of expert and topic.
2. **Time.** Experts spend a great deal of time analyzing a problem qualitatively [19].
3. **Patterns.** Experts see large meaningful patterns in their domain [19]. Posner described this phenomenon in terms of coding and chunking smaller concepts into larger ideas [75]. In a study of how people interpret annotations, Hong and colleagues described a search for useful representations [45]. Lesgold et al. made a similar observation, but in a more instrumental way: Experts look for a schema that can "control" their work – i.e., that can provide rules or procedures, and then specify how the work is to be carried out [58].
4. **Representation.** Experts represent problems in their domain in a more principled way ([19]; see also [45], [52], [58], [83]).

2.3.2 Craft

It seems obvious to say that data science involves skilled interaction with data. But in what ways? Pine and Leboiron showed evidence for a series of decisions that constrain how data are "'made' into a measurable entity" [73], which is then available to be discovered or captured (donné or capta, per [10]). As mentioned above, Feinberg and colleagues wrote about the "design" of data in data science [28]. Kiss and Szirányi referred to the "handcrafted" nature of data in cybersecurity research [55].

If data may be designed in these multiple ways, then we might find insights into data science practices from craftwork practices. Within HCI and CSCW, craft has been studied with several emphases. Raman and Hellerstein invoked craft as a guiding metaphor in Potter's Wheel, a tool for cleaning data [77]. Wiberg describes HCI design as borrowing concepts of materiality from craftwork [96]. Lingel found commonalities across craft and development related to Internet of Things, highlighting the materiality of media, embodied skills, flow of work, and collaboration [59]. Rosner et al. also focused on themes of materiality, finding commonalities in domains of software development and the craft of ceramics [80].

Schön characterized design as a conversation with design materials [84], "to discover a framework of meaning... through practical operations in the situation" [94]. Craft has been characterized in similar terms [97].

This generative insight has led to multiple research programs in CSCW and HCI, in which the "design materials" are now commonly understood to include digital components [19], [22], [38], [90].

From this literature, the following four aspects may serve as sensitizing concepts for our analysis:

1. **Conversation with materials.** Through the conversation with materials, there is often a sense of intimacy with materials and media [18], [59], [80], [97].
2. **Control.** Craft-workers labor at an intersection of control and unpredictability [80], [97].
3. **Tools and methods.** Tools are important to craft-workers [18], [60], [80]. Correspondingly, the work practice or method for using tools may be distinct from the tool itself [59], [60], [80], [97].
4. **Appropriation.** Craft-workers make creative use of their tools (similarly to [71]), appropriating them for new purposes, and boot-strapping intermediate tools in order to prepare a task-required tool ([60]; see also [15], [23]).

3 METHOD OF INQUIRY AND ANALYSIS

Data science projects can take a long time to accomplish. As we noted above, the operation of data cleaning can take "months" [62]. Sometimes data science workers have to try different approaches – including (in this paper) as many as three substantively different types of data – before they have sufficient data of sufficient quality. We wanted to become aware of those kinds of strategic reconsiderations. We therefore chose retrospective open-ended interviews as our data-collection method.

We conducted 20 interviews with 21 data science workers in IBM, a large international company (two people from the same project asked to be interviewed together, I-20a and I-20b). Informants were invited based on convenience and snowball sampling. Most of the informants were remote from us, working in both formally-chartered research divisions and also informal research groups within consulting divisions. 24% of informants were women, which compares favorably with recent estimates of 15% women in tenure-track faculty in computing [95], 20% women in data science positions worldwide [54], and 17% women in the Kaggle survey [47]. We did not inquire about gender identity. While IBM supports diverse gender identities, those data are considered sensitive personal information, because of differing and sometimes severe legal risk factors in some of the countries in which IBM's employees work.

Table 1. Informants

Informant	Sex	Role	Domain
I-01	M	AI-novice software engineer	Education
I-02	M	Team lead	Business analytics
I-03	M	Lead data scientist	Sales
I-04	M	Model builder	Transportation
I-05	M	Applied AI researcher	Education
I-06	M	Model builder	Healthcare
I-07	M	Applied AI researcher	Information technology
I-08	M	Applied AI researcher	Chatbot
I-09	M	Applied AI researcher	Information technology
I-10	M	Lead data scientist	Business analytics
I-11	M	Model builder	Healthcare
I-12	M	Model builder	Image classification
I-13	F	Domain expert	Speech
I-14	F	AI-novice software engineer	Education
I-15	M	Applied AI researcher	Retail
I-16	M	Model builder	Natural resources
I-17	F	AI-adept software engineer	Natural resources
I-18	M	Senior researcher / junior modeler	Text classification
I-19	M	Theoretical AI researcher	(not domain specific)
I-20a	F	Applied AI researcher	Chatbot
I-20b	F	Applied AI researcher	Chatbot

Informants worked in local or remote locations on several continents, on diverse projects, including seismic applications, large-scale transit applications with streaming data, medical imaging, remote monitoring of virtual machines, and text analytics. Interviews were semi-structured, with interviewers guiding the conversation to cover principal topics of workplace role, type of project, team configuration, and major challenges. In general, we asked interviewees to discuss a completed project retrospectively, so that they could draw conclusions about their work that might not have been possible when they were only partway through their projects (see also [97]).

Table 1 summarizes information about the informants and their projects. Data science workers at IBM worked on diverse problems that involved variations on prediction, classification, and reinforcement learning. As we will see, those diverse projects led to some common work practices, and some very diverse and project-specific work practices.

3.1 Grounded theory method

To understand the rich and complex topics of the interviews, we used a form of constructivist grounded theory method [16], adapted for use in HCI [68]. In general, grounded theory methods begin with data, and apply rigorous qualitative analysis principles such ***constant***

comparison, ***theoretical sampling***, and ***abductive logic*** ([16], [21], [68]). The goal is to build an understanding of the themes and concepts in the data through *open* (descriptive) *codes*, and then to combine those low-level concepts into more powerful configurations of codes called *axial codes* that represent multi-class ***categories*** (similar to an un-ordered list) or ***dimensions*** (ordered lists). More theoretically powerful, higher-level codes can be constructed from the axial codes. These are sometimes called *selective codes* [21], and they indicate deliberate interpretive choices by the analyst(s) [16], [68].

Interviews were scheduled for 60 minutes, and then were audio recorded and transcribed, for a total of 500 pages (range 17-43) and 183,244 words (range 6,321-11,992). During analytic iterations, we extracted what became 357 passages (mostly single sentences), and applied open coding and axial coding among those passages [21], [68]. All coding was done by one researcher, relying on frequent discussions with other researchers. We returned to the passages, and to the inter-views, repeatedly as necessary (constant comparison), to look for additional evidence and to test and revise our emer-gent understanding (theoretical sampling, abductive logic). These iterative analyses led to a core set of 19 axial codes, which we combined into the 5 selective codes of section 4.

There has been extensive debate about when to consult the research literature when doing grounded theory analysis. We briefly summarize here. Glaser [33] took a strong position against any use of the research literature until all data were collected and analyzed. More recently, grounded theory researchers have taken a more pluralistic view of the use of the research literature [68]. McGhee et al. [63] and Dunne [26] noted that researchers need to be able to discern new research questions, and to shape their questions based on what is already known. We made selective use of the research literature through the well-established practice of sensitizing concepts [12] focusing on expertise and craft.

3.1.1 Sensitizing concepts

Bowen reviewed the 50-year history of the use of sensitizing concepts in qualitative research [12]. He attributes the idea to Blumer, who distinguished between *definitive concepts* and *sensitizing concepts* [11]:

> A **definitive concept** refers precisely to... a clear definition in terms of attributes or fixed bench marks... A **sensitizing concept**... gives the user a general sense of... guidance in approaching empirical instances.

Bowen further explicates that "Social researchers now view sensitizing concepts as interpretive devices and as a starting point for a qualitative study." Ribes summarizes related arguments as, "sensitizing concepts tell the investigator where to look but not what to see" [79]. Glaser's own views evolved, and he came to offer similar advice, eventually accu-mulating 40 *coding families* to organize the initial phases of a grounded investigation [34]. For additional papers that used sensitizing concepts in HCI and CSCW, see [9], [56]. As discussed above, we used four sensitizing concepts from the domain of expertise, and four sensitizing concepts from HCI/CSCW studies of craft. In our report, we make reference to those sets of sensitizing concepts in **bold font**.

4 RESULTS

4.1 Data Practices: "This is usually the kind of art"

In general, data science workers need data! Twenty of the 21 informants discussed aspects of finding, refining, combining, pre-processing, and assessing data in 96 segments of our conversations.

4.1.1 Data Acquisition

Finding data for analysis can be easy, or the task may be challenging. We begin with informants who received their data easily.

For I-16's team, the data were sent by a seismology client who tightly controlled the curation of data and labels from a third party:

> *we have this database which was annotated, and which was acquired by the client... For each retrieved sample we would check if the sample belonged to the same seismic category as the query...*

I-19 was able to use "just... standard [image-labeling] datasets... very common in the community... nothing special..."

For these teams, the data seemed similar to Bilis's concept of "given" data (donné, [10]).

Other cases presented a pattern of capturing data ("capta", [10]) from multiple sources. In a project to optimize the routing of shipments, I-15's team had to align data from two very different temporal paradigms, and needed to be ready to perform initial modeling with only "half of that":

> *You need two types of data. You need the historical data to train all your models... even with a new customer you have half of that... you also need to setup the real time data that you're going to be regularly using to update things...*

I-04 faced a similar problem of static vs. dynamic data while trying to find a dataset that could meet the needs of a large transit-sensing application:

> *We were validating and evaluating [based] on historical data… And so I had to make sure that this is exactly the same as what is being replicated in a real time system....*

For I-04 and I-15, the problem of combining datasets was also a problem of combining data paradigms, where static data and dynamic data (e.g., streaming data) had very different properties, time-courses, storage requirements, and error-recovery possibilities. To make these determinations, I-04 and I-15 had to have highly **domain-specific** know-ledge, as well as the ability to understand sophisticated **re-presentations** of the data. These two explicitly useful data paradigms confronted I-04 and I-15 with the choices of both *capturing data* [10] and *designing data* (e.g., [28], [73]).

4.1.2 Cleaning and Integrating

Once the data have been found, the next step in the conventional model (e.g., [85]) is the **time**-consuming process of cleaning the data. Similarly, to other reports [40], our informants said that *"Dirty data is number one problem"* (I-12) and that *"90% of the time, we probably spend on data massaging and infrastructuring"* (I-10). I-02 saw **time**-based practical obstacles to acquiring more data: *"often getting... new sets of data that you can test is more time-consuming than actually testing it…"* and that data aggregation involved a **conversation with the data**:

> *[The] most time-consuming aspect is probably just understanding the data… all the quirks, finding the one person who's able to explain the data… it goes exponentially more difficult as soon as you have several data sources that need to be combined.*

Similarly, as experts, informants described a **time**-course of learning about the peculiarities of their data, while searching for **patterns** in the data. I-13's core task was "collecting coughs" as part of a project to diagnose disease through analysis of auditory signals. He said,

> *it was pretty easy to get… someone who was coughing a lot to be willing to cough while a phone was recording them. It was a little harder to get the metadata that we thought would help us classify these coughs... age, history of smoking…*

The real world of over-stretched medical services presented additional challenges – especially for a team that extended across continents. Analysts had to be vigilant against conceptual contradictions, designing their data [28] for consistency:

> *And so you get a report back from the data… that they were a smoker but they are 18 months old... [D]id something get checked off wrong?*

I-13 repeatedly engaged in a **conversation with the data**, based on deepening **domain-specific** knowledge.

Often, the data were incomplete, and data science workers had to take action on missing values. I-15 had to know both data **patterns** and **domains** to perform optimizations for a package-delivery client:

> *In order to update the cost, we needed to know the item weights - and actually the dimension of the package… But… They only had weights for about 50% of the items... we had to fill in that missing data to support computing shipping costs.*

I-08 reported similar difficulties as a *general condition* of a type of data, reflecting his **domain** knowledge and a **conversation with data:**

> *We are trying to do this time series… filling missing value with the closest values and then we are trying to explain that… it's a time series data so there's gaps in this data...*

Sometimes, analysts had to use their familiarity with the data and ultimately their intuition based on **conversations with the data** and **patterns** in the data. In these cases, intimate **domain** knowledge had to substitute for more formal measures:

> *I mean, there are some heuristic ways to… determine or… quantify - but it's going to be very imperfect and expensive to compute.*

I-13 reported similar difficulties which – in this particular case – involved subtle issues of the perspectives of other stakeholders:

> *Obviously, we don't want to manipulate the data… But we can't use data that... we can't vet... we did have to not use some of the data and that was frustrating… Looking at the data in different lenses but knowing what other people were looking at so we could identify inconsist-encies in someone else's treatment of the data...*

In general, we see broad support for Feinberg's contention that data science workers often need to design the data over which they compute models ([28] – see also [73]), aided by informants' **domain** knowledge and extensive and **time**-consuming **conversations with the data**. To detect anomalies and to impute missing values, informants also needed to exercise strong familiarity with **patterns** in their data, and they often had to spend additional **time** to insure usable **representations**.

4.1.3 Engineering Features

In the sequence of the conventional model [85], the step after cleaning is engineering features. The selection of appropriate features involves careful inspection for **patterns** and regularities through **conversations with the data**, critical conceptual evaluation of potential **representations**, and the **appropriation** of certain data features in novel combinations.

Several informants worked on projects to manage cloud computing resources. A major problem was to identify and close down the virtual machines (VMs) that were no longer being used. I-07 described conceptual struggles to derive features based on a deeper **domain** knowledge and discernment of **patterns** while using well-understood **tools** that were supposed to be sensitive to human activity:

> *we just kind of thought... previous work... tended to just focus on like monitoring rules... [H]ow much CPU does this VM use? How much memory does this VM use?... [W]e wanted to go a little bit deeper... [t]he networking traffic, and the number of user logins... can we derive more meaningful active-versus-inactive classifications?*

As I-07 continued, he discovered **patterns** of CPU activity that were false indicators, through further **conversation with the data**. Those patterns appeared to indicate intentional human usage for specific tasks. Further investigation led to a very different conclusion:

> *CPUs... go on [at] random times throughout the day and that was because different virus scans were running on the virtual machine... or... updates... And that would hog up CPU time... this VM might be active... at 2 o'clock in the morning... those tended to [be] actually inactive VMs that were just running virus scans and these updates.*

A subtler method was needed to detect intentional usage, through **appropriating** Linux tools that could be repurposed to tell *which* processes were running on each of the virtual machines, and hence whether a human was the likely initiator of the processes. In this example, **domain** knowledge of the nature of the logged events was crucial for I-07 to choose a more effective set of factors.

I-14 worked on an educational application for children. To enrich the somewhat semantically-sparse data, they augmented each word with related words - i.e., designing the data, as in [28], [30], [73], for enhanced semantics. However, this approach ceded **control** to rigorous but adult-originated **tools and methods** that could lead to the incorporation of age-inappropriate language and concepts. They wanted to filter-out child-inappropriate words:

> *we expand those words using concept expansion, and... we search [each word] in a... story corpus ... a child corpus.... And if an expanded word is not lying in a child corpus, then the probability... is very high that it is not... age-appropriate...*

Again, we see that I-14 used **representations** based on principled knowledge of the children's **domain**, and applied **patterns** from an **appropriated** dataset that had been repurposed from a different purpose (story-telling). In this example, we see that feature extraction requires an interaction of **domain** knowledge with practices of design-of-data [18], [28], [60], [80], [97].

I-11 faced a different challenge, which was medical-imaging data that were too rich to allow for efficient feature extraction. Based on extensive **domain** knowledge, I-11 and colleagues had principled **representations** of their targets (cancer micro-nodules), but the sheer volume of the data was defeating their **tools and methods** and was impairing their **conversations with the data**:

> *when you work with 3-D images, they're very, very heavy to process... Running 3D convolutional neural networks [is]computationally very, very consuming. And also memory use is also huge... you had to scale down the image, or you had to get pieces of this image... trying to find bi-dimensional, and then one-dimensional representations of the volumetric data...*

Later in the interview, I-11 returned to this theme, highlighting a need for trial-and-error experimentation (another form of **conversation with the data**), without a quantitative set of criteria regarding whether the representation was fit-to-purpose:

> *I always have to modify because these models, they're really sensitive to the image size... There's some trial and error... what you need to observe, is whether the model is converging or not.*

In this work, we can see an intersection of **control** and unpredictability, similar to Rosner et al.'s accounts of breakdown and repair in software and ceramic crafts [80], and as a repeating theme in the Wilson's interviews with craft-makers [97]. I-11 understood that he was engaged in curating [64], [89] data **patterns** into appropriately-scaled **representations**, which he summarized as:

> *This is usually the kind of art.*

4.1.5 Summary: Data Practices

The fortunate few data science workers receive well-structured, complete datasets (e.g., Bilis's "given" data [10]). Everyone else has to cope with data that require complex and sometimes inventive work to make those data whole – or perhaps we should write "whole enough" to support formal analysis (e.g., Bilis's "captured" data [10]. Often, data are not ready for analysis, and must be designed to meet the requirements of an algorithm [28], [73]. In some cases, data must be combined or even excluded in a form of curation [64], [89].

While data science workers engage with data in these ways, they apply **domain** knowledge in diverse activities that lead to **conversations with the data**. They look for **patterns** and establish **representations**, leading to **control** of the execution of data science activities by **tools and methods**. Upon need, they **appropriate** data or tools to construct new ways of working.

4.2 Ground Truth Practices: "I am the ground truth"

In the practice of data science, *ground truth* are a special type of data. In a predictive analysis, ground truth are typically the dependent variable(s). As the name implies, ground truth variables are traditionally considered to be authoritative and objective. We consider data science workers' practices in relation to ground truth in this separate section. We coded 51 quotations from 18 informants about data science workers' practices regarding ground truth.

4.2.1 Ground Truth as "Fact"

Several informants spoke of ground truth as unquestioned fact. In work to predict crowding in a large metropolitan transit system, I-04 told us that

> *All of this model building was done against ground [truth] surveys.*

I-14 worked with colleagues to develop a reliable set of ground truth labels:

> *Using annotators, we created a validation set, and using that validation set we were getting approximately 57% accuracy...*

Some projects had the advantage that they received well-labeled cases from their clients. I-08's team worked on natural language processing for a chatbot, with other teams who were responsible for ground truth:

> *We are relying on the other team to work with the ground truth*

I-16 worked on characterizing seismic images:

> *We do have a ground almost truth. So what we have is interpretations from the experts.*

These informants have engaged a more subtle form of **control**. By defining ground truth as fact, they are able to engage **tools and methods** to take advantage of that status. Less **conversation with data** is required in this view of ground truth,

4.2.2 Ground Truth as Curated Derivation

Data science workers sometimes perceive **patterns** in their data, in which they have to balance cost and practicality against precision of ground truth. I-04's project was solving a civic problem of transit crowding in a large city. Survey data provided the most direct **representations** of commuters' actual experiences. However, there were too many transit stations, and too few survey personnel, and the project needed data round the clock. The team engaged in a series of **conversations with different types of data** to find a less labor-intensive measure with better temporal coverage:

> *[W]e also worked on video data, which is cameras located at the train stations that take a continuous feed of the platform scenario... It was more like manual annotation to understand how much crowd is there at one point in time...*

The team **appropriated** the CCTV cameras' signals as crowd-estimation signals. These were an improvement over the surveys, but even this approach was not computationally sustainable. I-04's team negotiated to read a stream of anonymized cell-tower signals from mobile devices.

Now they had 24-hour data. The resulting **pattern** was complex, involving the *practical* ground truth data (mobile phone signals) as an approximation for the *derived* ground truth measure (cameras) which were an approximation of the *original* ground truth measure (surveys). I-04 and colleagues were acutely aware that the initial derivation lost the subjective data of the surveys, and the second derivation omitted anyone whose mobile was not actively communicating with a cell-tower (similar to integration issues in [25]). In this case, I-04 and colleagues were curating their ground truth estimates, and they were making trade-offs of feasibility against validity and accuracy.

4.2.3 Ground Truth in Context

However, ground truth labels were not always available. Earlier, we mentioned I-16's project with seismic data. I-16 went on to say,

> *this is the kind of ground truth we have.... for seismic... a process that takes months with a lot of people working together... sometimes we don't even have... the final ground truth. We have people's opinions.*

Like other types of data, ground truth data could also require some cleaning [28], [30], [73]. For a project involving users' texts, I-08 spoke of "grooming" the data to improve classification:

> *... we groomed it down because we realized that... some of them did not have enough classification... we use it to confirm the match of whatever the person is saying to the concept in the hierarchy.*

This grooming process involved assigning text to categories. I-08 and colleagues had to apply their **domain** knowledge in conjunction with a well-organized hierarchical representation, to assign ground truth. That **representation** also presented an opportunity to assign procedural **control** to the hierarchical classification scheme.

Even in well-labeled cases, context could be important for understanding the label (i.e., the ground truth indicator), requiring an intimate **conversation with the data** and with the context of the data, based on **domain** knowledge. [18], [60], [75], [97]. I-12 described an application for tagging the contents of photographs:

> *Let's say for example your data is already well-labeled... You have two well-labeled sources. Then you have to disambiguate. This data says "war spaceship", [but] that [data] says "warship"… A string match... fails, but it means the same thing, so you have to resolve those kinds of things.*

4.2.4 Ground Truth as Improvised

In other cases, our informants had to modify their criteria. In an education-support project to label photographs that were taken by children, I-01 told us that the team could not collect the quantity of photographs that they needed for stable modeling. They considered how to simulate a photograph by a child (e.g., [28], [73]), by exploring,

> *could we measure the motor skills of a child and how they take a picture and could we make a machine imitate... similar motor skills. So if your hand shakes this way and I can measure it with let's say a gyroscope could we imitate that?*

However, this approach would not provide the characteristic content or photographic framing that a child might choose. Could they crowdsource photographs by children?

> *I realize that we're not going to get a bunch of kids who take 50 images for every [category] unless we go find, you know, essentially pay people... the crowdsourcing approach.*

They realized that crowdsourced data presented different problems:

> *Just how do you verify the kids that took the picture?*

I-01 summarized the team's difficulties and breakdowns [80] as the outcome of a long **conversation with the data**:

> *you essentially have to get the data set trained with the data set. Does the data set work? And then realize no okay, this data set that I have, it works but it doesn't work for the users… Okay how can I generate data that makes it look like it's enough data?... then… test it with real users.*

For I-01's team, the data were insufficient, and they tried several strategies to supplement the data. Their **domain** knowledge caused them to doubt that each method would produce trustworthy ground truth.

4.2.5 Ground Truth as Created

In the preceding accounts, informants discussed their difficulties with finding reliable labels to represent ground truth. Two teams had to create their own labels. For a question-answering chatbot, the architecture called for a preliminary categorization of a user's query, which then assigned the query to one of a group of specialist service modules that contained algorithms and **representations** to answer questions of that category. Each of these two assignment operations constituted a type of **control** that determined subsequent processing. Regarding the preliminary categorization step, I-20a explained,

> *Look at the sentence that says, "How can I change my contribution?" This could be a very ambiguous sentence – contributions to what?... it's actually a grey area in terms of the (text) classification.*

There was no definitive label available for each user query. I-20a had to label them herself:

> *When you want to bootstrap your agent, you come up with a set of examples and a set of labels. [Interviewer question: "So the ground truth is really what you think up?"] That's right. Correct.*

I-18 had a similar situation. He was responsible for both labeling users' text submissions, and for predicting those labels on the basis of user attributes. He was predicting his own labels. He summarized by saying,

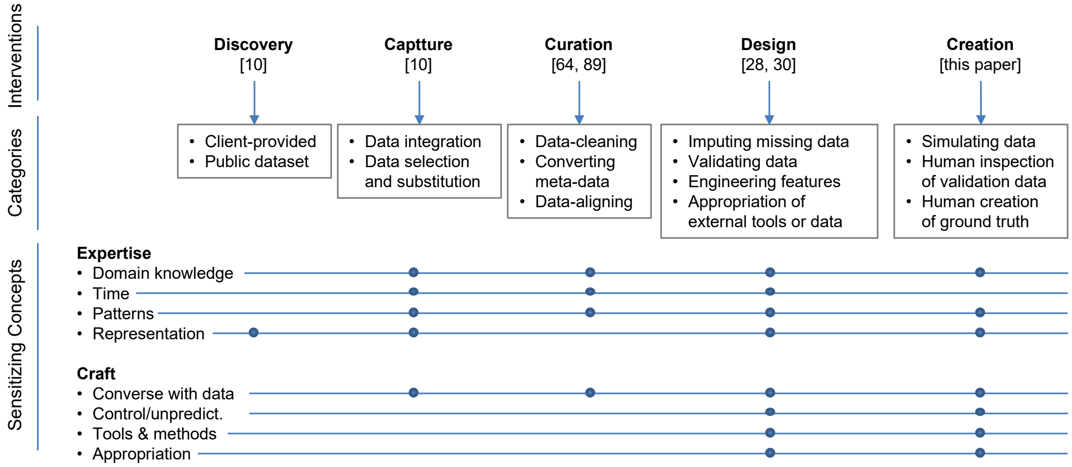
> *I am the ground truth.*

In these two cases, researchers had no ground truth. To perform their classifications, they had to create ground truth.

4.2.6 Summary: Crafting of Ground Truth

In many cases, there is explicit and reliable ground truth available for modeling. In these cases, the word "truth" appears to be a claim of *fact* (e.g., Bilis's "given" data [10]). In additional cases, it is relatively straightforward to resolve ambiguities through I-08's "grooming" – a form of excerpted or captured ground truth data (e.g., Bilis's "capta" data [10]), based in **domain** knowledge. In yet other cases, there is a clear derivational path of **representations** involving a **conversation with the data** from a primary form of ground truth, through intermediary, **appropriated** designed forms, resulting in a useful measure that has reasonable surface validity despite the indirect nature of the data (e.g., [18], [28], [73]).

However, other situations are more problematic, requiring various forms of improvisation. In the study of photographs, I-01's team simply did not have enough data, and they **conversed with the data** through various methods for simulating the data that they needed. Each method had strengths and weaknesses. I-18 and I-20a spoke candidly about the need to *create* the ground truth values



	Discovery	Captture	Curation	Design	Creation
Expertise					
• Domain knowledge		●	●	●	●
• Time		●	●	●	
• Patterns		●	●	●	●
• Representation	●	●		●	●
Craft					
• Converse with data		●	●	●	●
• Control/unpredict.				●	●
• Tools & methods				●	●
• Appropriation				●	●

Figure 1. Human interventions in data science work practices.

for their projects. In these cases, the data science workers had no practical choice other than to write their own labels, and then to learn or predict the labels that they had written (e.g., [28], [73]), and finally to document how they had created their ground truth.

5 DISCUSSION

5.1 Limitations

The informants in our study worked primarily in research organizations. The experiences of data science workers in other departments might be different (e.g., sales, consulting, human resources). While we interviewed across multiple locations, our sample was entirely English-capable, and may have inadvertently excluded some cultures. We sampled primarily from people who work in data science roles, and we might have learned about different practices from domain experts or business-process owners who engage in data science. Finally, our informants all worked for IBM, and data science might be done differently in other companies, universities and colleges, governments, and non-profits.

5.2 Implications for Data Science Practices

As we reviewed earlier, data wrangling can require much time and labor [36], [39], [41], [49], [62], [78], [88]. Based on the reports of [10], [28], [73] and the analysis in this paper, we propose to rethink aspects of data wrangling as a dimension of interventions (Figure 1).

In Figure 1, we summarize a monotonic scale of human *intervention* with data, from least ("discovery") to greatest ("creation") – i.e., a ***dimensionalized selective code,*** in the language of grounded theory. The simplest case – and the world implied by the optimistic views of [27], [65] – is the *discovery* of data [10]. Selective *capture* of data [10] is a more transformative process, followed by the more purposive operations of *curation* of audience-specific views [64] and micro-worlds [89]. Feinberg and colleagues describe a more intentional *design* of the data [28], [30]. Finally, in this paper we have documented the deliberate *creation* of data upon need, including the creation of ground truth data. In general. researchers reported this human work with data in a transparent manner, in internal reports and in conference publications. We note also that all of these transformations take place in the usually-invisible context of the human-informed measurement plan [73], [79] that determines what is considered as "data," and how those data are measured.

Each of these manipulations may be proper and necessary. We hope that the intervention view of Figure 1 will help researchers to write more explicitly about the degree to which they have needed to process and transform – i.e., to intervene - with their data. As researchers transit through human-influenced interventions from discovery to curation to creation, etc., they move farther and farther from what boyd and Crawford called the "mythology" of "the aura of truth, objectivity, and accuracy" in large scale analyses [13]. This mythology is particularly powerful with regard to ground truth data. It will be important to problematize this characterization when we inevitably intervene between "data" and "ground truth."

Janssen and Kuk called for enhanced algorithmic transparency in "unravelling the imperceptibility, material-ity and governmentality of how *algorithms work*" (em-phasis in original) [46]. We take a similar interest to make visible *how data work* in contemporary data science [51], [74], [79]. We noted that Bilis's data discovery [10] itself is a form of human intervention - i.e., *who* is discovering the data? We join with Kemper and Kolkman to ask, for *whom* are the data transparent [51]? We hope that, if data science workers document their interventions, then this kind of data provenance may eventually require less data science literacy [37], and may be made more "readable" for people with diverse backgrounds and expertises [81].

5.3 Implications for HCI of Data Science

In the preceding section, we advocated to take a more transparent approach in reporting data transformations in data science, emphasizing the transit away from raw data. We propose that these categories may also provide analytic value in characterizing and extending data wrangling approaches. Several approaches have explicated wrangling activities, including a line of development from Potter's Wheel [77] to Wrangler [39], [49], and then Trifacta [43], and a second line of development from Freebase Gridworks to OpenRefine [93]. Sutton et al. addressed a subset of data wrangling operations from the perspective of a collection of diff-like transformations that can reformat a dataset into a desired format ([88]; for similar propositions, see [39], [43]; for critique, see [74]).

We propose that wrangling operations might be mapped analytically in relation to the five interventions of Figure 1. Capture, for example, seems to involve simple filtering and deleting, while curation might involve a more rule-guided approach to those steps (but see [71] with regard to rules-based data science). Design is closer to feature extraction and feature engineering, which are often accomplished in wrangling through operations such as clustering and transforming. A more explicit and perhaps principled description of wrangling activities may help to keep them visible, rather than infrastructural [7], [37], [70], [79].

While not using this kind of vocabulary, data science workers engage in changing their assemblages [87] and dynamically re-working their information ecologies [6], [74]. Some aspects of data capture may take place before the data become visible during data wrangling, and may thus have become invisible by the time that data science workers begin to craft their data [51], [55], [64], [73], [79]. Some aspects of curation for specific audiences [61], may take place after data wrangling is complete – for example when choosing a model that can be explained to a particular client. We propose that further investigations into the relationship of the five interventions of this paper, with the common activities of data wrangling, may bring clarity to both of these complex activities.

5.4 Data Expertise and Data Craft

To organize our analysis, we used two sets of sensitizing concepts [11], [12], [34], [79]. The first set of four concepts was based on the expertise-in-HCI literature [19], [75], and the second set of four concepts was based on the craft-in-HCI literature [18], [55], [59], [60], [77], [80], [96]. This was an experiment, to explore how well those sets of concepts could inform our analysis.

These sets of concepts originated in two distinct strands of HCI literature. We were surprised to discover that we tended to use the two sets of concepts in an interleaved fashion, often coding the report of a data science worker in the vocabularies of both expertise and craft:

- Frequently, we interpreted an informant's report as involving a conversation with the data (from the *craft* literature), but this was also a form of *expertise* in the domain of the work.
- As a second example, we applied the *craft* concept of balancing control vs. serendipity, and found that we could apply the concept of control to the frequent *expertise*-based action among data science workers of identifying a pattern in the data that was suited to a method or a tool, and then ceding control to that method or tool.
- The opposite of control in *craft* is serendipity, and we repeatedly learned how data science workers encountered unexpected issues in their data and/or domain, how they applied domain *expertise* to detect new patterns, and how they thus transitioned from serendipity to control.

The interventions of Figure 1 imply a very active human shaping of the analysis, based on the data science workers' time-consuming conversations of the data, leading to discernment of patterns, and leading to the application or appropriation of appropriate methods and tools - supported by the data science workers' domain knowledge. Sometimes the informant had a principled reason to do this, but other times the informant based their next move on their sense of the data. Rosner et al. analyzed explicit and tacit knowledge in the HCI of craft [80], and our emerging understanding of data science activities is convergent with their analysis.

6 CONCLUSION

In this paper, we have analyzed the data practices of data science workers into a series of five increasingly assertive, creative, and formative interventions between data and algorithmic analysis. We attempted to show that each such action was necessary and appropriate. We proposed a future research agenda to relate these dimensioned interventions to other actions in data science, and particularly to the tools and methods of data wrangling. Finally, we used concepts from the expertise literature and the craft literature to make sense of these interventions.

In keeping with [32], Verborgh and De Wilde remarked [93]:

> Data is often dubbed the new gold, as it is of tremendous value for today's data-driven economy. However, we prefer to think of data as diamonds. At first they're raw, but through great skills, they can be polished to become the shiny assets that are so worthy to us.

We began this paper with the claims of Durrante-Whyte [27] and Miller [65] which implied a universality and objectivity of data in data science. We hope that the dimensioned interventions of Figure 1 can add epistemic rigor to the work of data wrangling. If humans passively "discover" data, then perhaps our analytic work really could reflect a kind of realist account of the data-world, and of our derivations of that world. To the extent that we have to intervene actively with our data – to capture or curate or design or create our data – then our interpretations and perhaps our intentions enter into the construction of our data-world, and into the actions that we take on the basis of that construction. We need to make these interventions, interpretations, and intentions visible for inspection, criticism, accountability, and perhaps reconciliation.

REFERENCES

[1] Sebastian Abt and Harold Baier (2014). A plea for utilizing synthetic data when performing machine learning based cyber-security experiments. *Proc. AISec 2014.*

[2] Ritu Agarwal and Vasant Dhar (2014). Editorial – Big data, data science, and analytics: The opportunity and challenge for IS research. Info. Sys. Res. 25(3), 443-448.

[3] Ashton Anderson, Jon Kleinberg, and Sendhil Mullainathan (2017). Assessing human error against a benchmark of perfection. *TKDD 11*(4), Art. 45.

[4] Jesse Anderson (2018). Data engineers vs. data scientists. O'Reilly. https://www.oreilly.com/ideas/data-engineers-vs-data-scientists .

[5] Lora Aroyo and Chris Welty (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *Proc. Web Science 2013.*

[6] Karen S. Baker and Geoffrey C. Bowker (2007). Information ecology: Open system environments for data, memories, and knowing. *J. Intell. Inf. Syst. 29*, 127-144.

[7] Karen S. Baker and Helena Karasti (2018). Data care and its politics: Designing for local collective data management as a neglected thing. *Proc. PDC 2018*, Art. 10.

[8] Jo Bates, Yu-Wei Lin, and Paula Goodale (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Soc. 3*(2), 1-12.

[9] Steve Benford, Gabriella Giannachi, Boriana Koleva, and Tom Rodden (2009). From interaction to trajectories: Designing coherent journeys through user experiences. *Proc. CHI 2009*, 709-718,

[10] Hélène Bilis (2018). Mapping fiction: Social networks and the novel. Presentation at *Shifting (the) boundaries* conference, Wellesley College.

[11] Herbert Blumer (1954). What is wrong with social theory? *American Sociological Review 18*, 3-1.

[12] Glenn A. Bowen (2006). Grounded theory and sensitizing concepts. *Int. J. Qual. Meth. 5*(3), 12-23.

[13] danah boyd and Kate Crawford (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenonomenon. *Info. Comm. Soc. 15*(5), 662-679.

[14] Ciara Byrne (2013). The rise of the DIY data scientist. *Fast Company.* https://www.fastcompany.com/3014018/the-rise-of-the-diy-data-scientist .

[15] Jennie Carroll. 2004. Completing design in use: closing the appropriation cycle. In *Proceedings of the 12th European Conference on Information Systems* (ECIS 2004). 337–347.

[16] Kathy Charmaz (2015). *Constructing grounded theory.* Sage.

[17] Akemi T. Chatfield, Vivian N. Shlemoon, Wilbur Redublado, and Faizur Rahman (2014). Data scientists as game changers in big data environments. *Proc. ACIS 2014*, 1-11.

[18] Amy Cheatle and Steven J. Jackson (2015) "Digital entanglements: Craft, computation and collaboration in fine art furniture production. *Proc. CSCW 2015*, 958-968.

[19] Madeline T.H. Chi, Robert Glaser, and Marshall Farr (1988/2014) (eds.). *The nature of expertise.* Psychology Press.

[20] Søren Christensen, Jens Bæk Jørgensen, and Kim Halskov Madsen (1997). Design as interaction with computer based materials. *Proc. DIS 1997*, 65-71.

[21] Juliet Corbin and Anselm L. Strauss (2007). Basics of qualitative research: Techniques and procedures for developing grounded theory. 3rd edition. Newbury Park, CA, USA: Sage.

[22] Andrew Dearden (2006). Design as a conversation with digital materials. *Des. Stud. 27*(3), 399-421.

[23] Alan Dix (2007). Designing for appropriation. *Proc. BCS-HCI 2007*, 27-30.

[24] C. Dobre and F. Xhafa (2014), Intelligent services for big data science. *Fut. Gen. Comp. Sys. 37*, 267-291.

[25] Anca Dumitrache, Lora Aroyo, and Chris Welty (2018). Crowdsourcing ground truth for medical relation extraction. *TIIS 8*(2), art. 11.

[26] Ciarán Dunne (2011). The place of literature review in grounded theory research. *Int.J. Soc. Res. Meth. 14*(2), 111-124.

[27] Hugh Durrante-Whyte (2015), Data, knowledge and discovery: Machine learning meets natural science. *Proc. KDD 2015*, 7.

[28] Melanie Feinberg (2017a). A design perspective on data. *Proc. CHI 2017*, 2952-2963.

[29] Melanie Feinberg, Daniel Carter, and Julia Bullard (2014b). A story without end: Writing the residual into descriptive infrastructure. *Proc. DIS 2014*, 385-394.

[30] Melanie Feinberg, Daniel Carter, Julia Bullard, and Ayse Gursoy (2017b). Translating texture: Design as integration. *Proc. DIS 2017*, 297-307.

[31] Batya Friedman, Peter H. Kahn, and Alan Borning (2006). Value sensitive design and information systems. In P. Zhang and D. Galletta (eds.), *Human-Computer Interaction and Management Information Systems: Foundations.* M.E. Sharpe.

[32] Lisa Gitelman (2013) (ed.), *"Raw data" is an oxymoron.* MIT Press.

[33] Barney G. Glaser (1998). *Doing grounded theory: Issues and discussions.* Mill Valley, CA: Sociology Press.

[34] Barney G. Glaser (2005). *The grounded theory perspective III: Theoretical coding.* Mill Valley, CA, USA: Sociology Press.

[35] Robert Glaser and Micheline T.H. Chi (1988/2014). Overview. In Michelene T.H. Chi, Robert Glaser, and Marshall J. Farr (eds). (1988/2014). *The nature of expertise.* Taylor and Francis.

[36] Michele Goetz (2015). 3 ways data preparation tools help you get ahead of big data. Forrester. https://go.forrester.com/blogs/15-02-17-3_ways_data_preparation_ tools_help_you_get_ahead_of_big_data/ .

[37] Jonathan Gray, Carolyn Gerlitz, and Liliana Bounegru (2018). Data infrastructure literacy. *Big Data & Soc. 5*(2), 1-13.

[38] Shad Gross, Jeoffrey Bardzell and Shaowen Bardzell (2014). Structures, forms, and stuff: The materiality and medium of interaction. *Pers. Ubiquit. Comput. 18*(3), 637-649.

[39] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer (2011). Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. *Proc. UIST 2011*, 65-74.

[40] Bob Hayes (2018a). A majority of data scientists lack competency in advanced machine learning areas and techniques. BusinessOverBroadway, http://businessoverbroadway.com/a-majority-of-data-scientists-lack-competency-in-advanced-machine-learning-areas-and-techniques .

[41] Bob Hayes (2018c). Most used data science tools and technologies in 2017 and what to expect for 2018. *BusinessOverBroadway.* http://businessoverbroadway.com/most-used-data-science-tools-and-technologies-in-2017-and-what-to-expect-for-2018 .

[42] Bob Hayes (2018b). Top 10 challenges to practicing data science at work. *BusinessOverBroadway.* http://businessoverbroadway .com/top-10-challenges-to-practicing-data-science-at-work .

[43] Jeffrey Heer, Joseph M. Hellerstein, and Sean Kandel (2015). Predictive interaction for data transformation. *Proc. CIDR 2015.*

[44] Tony Hey, Stewart Tansley, and Kristin Tolle (2009). *The fourth paradigm: Data-intensive scientific discovery.* Microsoft Research.

[45] Ming-Tung Hong and Claudia Müller-Birn (2017). Conceptualization of computer-supported collaborative sensemaking. *CSCW 2017 Companion*, 199-202.

[46] Marjin Janssen and George Kuk (2016). The challenges and limits of big data algorithms in technocratic governance. Gov. Info. Quart., 33(3), 371–377.

[47] Kaggle (2017). Kaggle ML and data science survey, 2017: A big picture view of the state of data science and machine learning. Kaggle. https://www.kaggle.com/kaggle/kaggle-survey-2017 .

[48] KDNuggets. (2018). Doing data science: A Kaggle walkthrough – Cleaning data. https://www.kdnuggets.com/2016/03/doing-data-science-kagglea-walkthrough-cleaning-data.html .

[49] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer (2011). Wrangler: Interactive visual specification of data transformation scripts. *Proc. CHI 2011*, 3363-3372.

[50] James Max Kanter and Kalyan Veeramachaneri (2015). Deep feature synthesis: Towards automating data science endeavors. *Proc. DSAA 2015*, 1-10.

[51] Jakko Kemper and Daan Kolkman (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Info. Comm. & Soc.* DOI: 10.1080/1369118X.2018.1477967

[52] Allison Kidd (1994). The marks are on the knowledge worker. *Proc. CHI 1994*, 186-191.

[53] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel (2016). The emerging role of data scientists on software development teams. *Proc. IEEE CSE 2016*, 96-107.

[54] John King and Roger Magoulas (2015). *2015 data science salary survey: Tools, trends, what pays (and what doesn't) data professionals.* O,Reilly. http://www.eli.sdsu.edu/courses/fall16/cs696/2015-data-science-salary-survey.pdf .

[55] Ákos Kiss and Tamás Szirányi (2013). Evaluation of manually created ground truth for multi-view people localization. Proc. VIGTA 2013.

[56] Peter Gall Krogh, Marianne Graves Petersen, Kenton O'Hara, and Jens Emil Grønbæk (2017). Sensitizing concepts for socio-spatial literacy in HCI. *Proc. CHI 2017*, 6449-6460.

[57] Cheng Han Lee (2014). Data career paths: Data analyst vs. data scientist vs. data engineer: 3 data careers decoded and what it means for you. Udacity. https://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html

[58] Alan Lesgold, Harriet Rubinson, Paul Feltovich, Robert Glaser, Dale Klopfer, and Yen Wang (1988/2014). Expertise in a complex skill: Diagnosing X-ray pictures. In Michelene T.H. Chi, Robert Glaser, and Marshall J. Farr (eds). (1988/2014). *The nature of expertise.* Taylor and Francis.

[59] Jessica Lingel (2016). The poetics of socio-technical space: Evaluating the Internet of things through craft. *Proc. CHI 2016*, 815-826.

[60] Jessica Lingel and Tim Regan (2014). "it's in your spinal cord, it's in your fingertips: practices of tools and craft in building software." *Proc. CSCW 2014*, 295-304.

[61] Paul Luo Li, Andrew J. Ko, and Jiamin Zhu (2015). What makes a great software engineer? *Proc. ICSE 2015*, 700-710.

[62] Karen Grace Martin (2018). Preparing data for analysis is (more than) half the battle. Analysis Factor. https://www.theanalysisfactor.com/preparing-data-analysis/ .

[63] Gerry McGhee, Glenn R. Marland, and Jacqueline Atkinson (2007). Grounded theory research: Literature reviewing and reflexivity. *J. Adv. Nurs. 60*(3), 334-342.

[64] Helena M. Mentis, Ahmed Rahim, and Pierre Theodore (2016). Crafting the image in surgical telemedicine. *Proc. CSCW 2016*, 744-755.

[65] Renée J. Miller (2017). The future of data integration. *Proc. KDD 2017*, 3.

[66] Steven Miller (2014). Collaborative approaches needed to close the big data knowledge and skills gap. *J. Org. Des. 3*(1), 26-30.

[67] Julia Moehrmann and Gunther Heidemann (2012). Efficient annotation of image data sets for computer vision applications. *Proc. VIGTA 12.*

[68] Michael Muller (2014). Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In Judith Olson and Wendy A. Kellogg (eds.), *Ways of knowing in HCI.* Springer.

[69] Syed Sadat Nazrul (2018). DevOps for data scientists: Taming the unicorn. *Medium: Towards Data Science*, https://towardsdatascience.com/devops-for-data-scientists-taming-the-unicorn-6410843990de .

[70] Gina Neff, Ahissa 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), 85-97.

[71] Samir Passi and Steven J. Jackson (2017). Data vision: Learning to see through algorithmic abstraction. *Proc. CSCW 2017*, 2436-2447.

[72] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison (2008). Investigating statistical machine learning as a tool for software development. *Proc. CHI 2008*, 667-676.

[73] Kathleen H. Pine and Max Liboiron (2015). The politics of measurement and action. *Proc. CHI 2015*, 3147-3156.

[74] Sarah Pink, Minna Ruckenstein, Robert Willim, and Melisa Duque (2018). Broken data: Conceptualizing data in an emerging world. *Big Data & Soc. 5*(1), 1-13.

[75] Michael I. Posner (1988/2014). Introduction. In Michelene T.H. Chi, Robert Glaser, and Marshall J. Farr (eds). (1988/2014). *The nature of expertise.* Taylor and Francis.

[76] Krishna Rajan (2013). Informatics for materials science and engineering: Data-driven discovery for materials science and engineering. Elsevier.

[77] Vijayshankar Raman and Joseph M. Hellerstein (2001). Potter's wheel: An interactive data cleaning system. *Proc. VLDB 2001.*

[78] Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras (2017). *Principles of data wrangling: Practical techniques for data preparation.* O'Reilly.

[79] David Ribes (2017). Notes on the concept of data interoperability: Cases from an ecology of AIDS research infrastructures. *Proc. CSCW 2017*, 1514-1526.

[80] Daniela K. Rosner, Miwa Ikemiya, and Tim Regan (2015). Resisting alignment: Code and clay. *Proc. TEI 2015*, 181-188.

[81] Evelyn Ruppert (2013). Rethinking empirical social sciences. *Dial. Hum. Geo. 3*(3), 268-273.

[82] Evelyn Ruppert, Penny Harvey, Celia Lury, Adrian Mackenzie, Ruth McNally, Stephanie Alice Baker, Yannis Kallianos, and Camilla Lewis (2015). *Socializing big data: From concept to practice.* CRESC, U. Manchester, Open U.

[83] Daniel M. Russell, George Furnas, Mark Stefik, Stuart Card, and Peter Pirolli (2008). Sensemaking workshop 2008. CHI EA 2008, 4751-4754.

[84] Donald Schön (1983). *The reflective practitioner. How professionals think in action.* Basic Books.

[85] Scikit-Learn (2017). scikit-learn Tutorials. http://scikit-learn.org/stable/tutorial/index.html .

[86] Shventank Shah, Andrew Horne, and Jaime Capella (2012). Good data won't guarantee good decisions. *Harv Bus Rev,* Apr 2012.

[87] Susan Elliott Sim, Marisa Levitt Cohn, and Kavita Philip. (2009). The work of software development as an assemblage of computing practices. *Proc. CHASE 2009*, 92-95.

[88] Charles Sutton, Timothy Hobson, James Geddes, and Rich Caruana (2018). Data diff: Interpretable, executable summaries of changes in distributions for datq wrangling. *Proc. KDD 2018.*

[89] Alex S. Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasilis Vlachokyriakos, Lillie Grainger, and Jessa Lingel (2015). Data-in-place: Thinking through relations between data and community. *Proc. CHI 2015*, 2863-2872.

[90] Jakob Tholander, Maria Normack, and Chiara Rossitto (2012). Understanding agency in interaction design materials. *Proc. CHI 2012*, 2499-2508.

[91] Paul F. Uhlir and Peter Schoder (2007). Open data for global science. *Data Sci. J. 6*, 36-53.

[92] Wil M.P. van der Aalst (2014). Data scientist: The engineer of the future. *Proc. I-ESA 7*, 13-26.

[93] Ruben Verborgh and Max De Wilde (2013). *Using OpenRefine.* Packt.

[94] Leonard J. Waks (2001). Donald Schon's [sic] philosophy of design and design education. *Int. J. Tech. Des. Educ. 11*, 37-51.

[95] Samuel F. Way, Daniel B. Larremore, and Aaron Clauset (2016). Gender, productivity, and prestige in computer science faculty hiring networks. *Proc. WWW 2016*, 1169-1179.

[96] Mikael Wiberg (2014). Methodology for materiality: Interaction design through a material lens. *Pers. Ubiquit. Comput. 18*(3), 625-636.

[97] Fo Wilson (2010). The new materiality: Digital dialogues at the boundaries of contemporary craft. *Cultura Visual 1*(14), 83-88.

[98] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld (2018a). Investigating how experienced UX designers effectively work with machine learning. *Proc. DIS 2018*,

[99] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos (2018b). Grounding interactive machine learning tool design in how non-experts actually build models. *Proc. DIS 2018*, 573-584.