Guidelines For Pursuing and Revealing Data Abstractions

Alex Bigelow, Katy Williams, and Katherine E. Isaacs

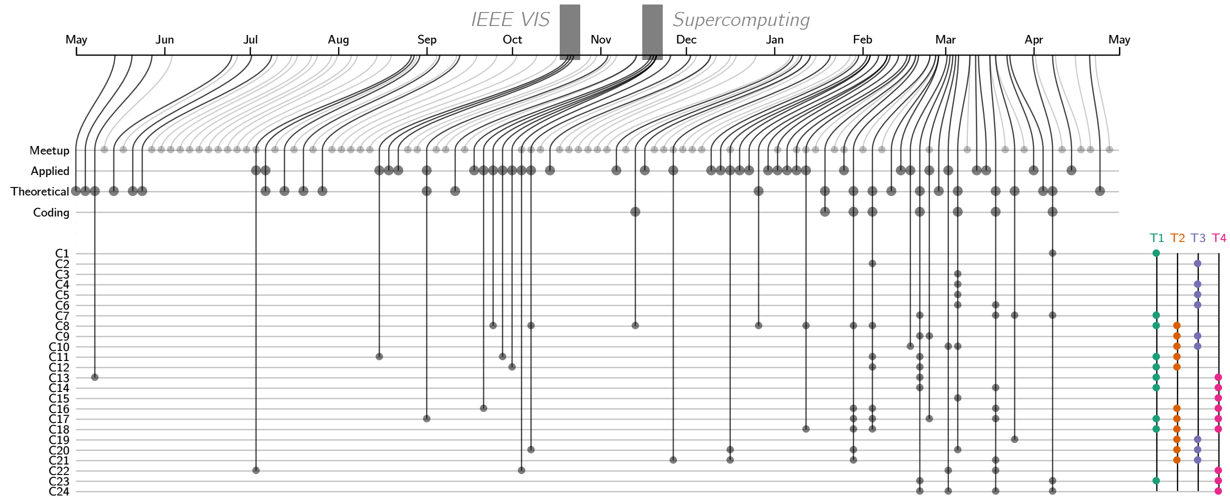

Fig. 1. A summary of memos, their relationships with codes, and code relationships with themes. The timeline at the top shows when individual memos were created, as well as the timing of two major deployments of the survey at conferences. The four rows below the timeline indicate the nature of the meeting, including when Meetup attendance occurred, when data workers discussed their applied datasets, when the authors engaged in theoretical discussions, and when the authors engaged in open coding. Rows C1-C24 show which memos directly informed the development of codes. Columns T1–T4 show which codes directly inform which themes.

Abstract— Many data abstraction types, such as networks or set relationships, remain unfamiliar to data workers beyond the visualization research community. We conduct a survey and series of interviews about how people describe their data, either directly or indirectly. We refer to the latter as latent data abstractions. We conduct a grounded theory analysis and develop a substantive theory that (1) interprets the extent to which latent data abstractions exist, (2) reveals the far-reaching effects that the interventionist pursuit of such abstractions can have on data workers, (3) predicts why and when data workers may resist such explorations, and (4) suggests how to take advantage of opportunities and mitigate risks through transparency about visualization research perspectives and agendas. To continue the discussion, we make our dataset open along with a visual interface for further exploration.

1 Introduction

Data abstractions are fundamental to a wide set of visualization activities, from performing and documenting the provenance of data wrangling operations, to understanding the mental models of domain experts in design study research, to justifying design decisions in technique- or systems-focused research, and to reasoning about the role of data abstraction in theoretical visualization research. Difficulties in reasoning about and communicating data abstractions therefore have far-reaching implications: effective communication about data abstractions is critically important to the way researchers justify design decisions in technique- or systems-focused research. A poor understanding of the mental models of domain experts in design study research is a significant threat that risks creating solutions and systems that do not address real needs [27]. Too much focus on a single data abstraction has been observed to limit creativity [2] and to warp scientific analysis [1]. However, the extent to which these effects apply, in terms of specific abstractions, is poorly understood.

• *Alex Bigelow, Katy Williams, and Katherine E. Isaacs are with the University of Arizona. E-mail: alexrbigelow@email — kawilliams@email — kisaacs@cs.arizona.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

We set out to understand how malleable a data abstraction is, and to better understand the process of pursuing **latent data abstractions**. We define a latent data abstraction to be a data abstraction that is meaningful and useful, yet undiscovered. It has yet to be fully elucidated, communicated, documented, and formatted. A data abstraction becomes less latent as coherent details are identified, as its details are spoken or written, and as its artifacts in a computer are actualized into relevant forms.

Because there were blind spots in the questions that we should even ask, we chose to conduct a Grounded Theory Method investigation seeking to discover how a diverse range of **data workers**, from spreadsheet users to programmers, across different disciplines, consider different data abstractions. This investigation analyzes memos from conversations, meetings, and interviews and the results of a deployed survey.

The result is an evidence-based, interpretivist substantive theory regarding data abstractions with implications for how visualization project teams and individuals discover, wrangle, manage, and report their data abstractions. In particular, we find that introducing a **data abstraction typology**—a model that describes the space of possible data abstractions and/or data wrangling operations—can elicit rich communication and reflection about data and uncover latent data abstractions, even when such a typology is imperfect. We show how visualization researchers can increase actionable communication with data workers by introducing and critiquing a typology together, as a visualization design activity [22].

The substantive theory in this paper also adds to existing literature by explaining some of the reasons why communicating about data abstractions can be so challenging, and it provides guidelines for communicating with data workers about data abstractions. These guidelines also have applications for more crisp communication about data abstractions in design study, technique, systems, and theoretical research papers.

We have further made the raw data collected in our survey available through an open interactive visual interface[1] along with an archive of codes, themes, an audit trail [5], and their revision history[2] that summarize observations recorded in memos from a year of interviews and meetings with diverse data workers, as well as observations from the visualized survey responses.

In summary, our contributions are:

1. An interpretivist substantive theory of data abstractions and their implications for visualization design (Sect. 5),
2. Guidelines regarding the development of data abstractions (Sect. 6.1),
3. The design of an open survey regarding the description of data and the malleability of data abstractions (Sect. 3.3, Sect. 6.4), and
4. An open dataset of survey results with a corresponding interactive visualization.

We begin by discussing necessary background and a review of related work (Sect. 2). We then further discuss our methodology (Sect. 3). We present the codes derived from our study (Sect. 4) and use them to describe the substantive theory. We follow with our guidelines and reflections.

2 Background and Related Work

We discuss the theoretical underpinnings of our work (Sect. 2.1), related background in thinking and communicating about data in analysis and design projects (Sect. 2.2), the importance of documenting real-world wrangling needs (Sect. 2.3), and the context in which this work fits into research into creativity (Sect. 2.4).

2.1 Theoretical Underpinnings

Interpretivist research has a different philosophical goal than the *positivist* research that we typically see in the visualization research community [23]. Interpretivist research aims to *describe* phenomena and *generate* hypotheses. This is in contrast to the positivist approach used in the scientific method that aims to *test* hypotheses.

We present a *substantive* theory, that has different objectives than a *formal* theory. Substantive theories are transferable, in contrast to formal theories that are generalizable. Both kinds of theory require systematic analysis of evidence, but the nature of supporting data and the ways that data are collected and analyzed are different. This study employs a team-based [45], interpretivist form [44] of Grounded Theory Methodology as a systematic way to develop and refine a substantive theory.

Grounded Theory Methodology is an appropriate fit for beginning this investigation because our initial theory—that non-tabular data abstractions may be comprehensible, useful, and under-utilized among the broad population of data workers—is very general and based on a small number of surprising observations [1, 2]. The overarching theory and the nature of the questions that we should pursue were prone to rapid revision and refinement as additional, surprising observations arose. Grounded Theory is an approach that is uniquely suited for investigating and describing phenomena in which questions evolve rapidly, as the theory is constructed by the data as it is being collected.

Consequently, we used *surprise* as a principled way to collect data [25]; the extent to which we pursued interactions with data workers and adapted and deployed a survey were motivated by identifying gaps in our own knowledge and unanticipated findings. In contrast, we also used our lack of surprise as a qualitative indicator to know when *codes*—or concepts that describe phenomena—had reached saturation, and needed no further investigation.

In presenting qualitative research, we are careful of pitfalls [37] in our reporting of numbers and counts: we include the full visualized corpus of survey responses to maximize available context. Our numeric statements and visualizations are meant to be interpretivist descriptions of the phenomena associated with how data workers think and communicate about data abstractions, not positivist statements of statistical significance.

Although this is not a visualization design study, Meyer and Dykes' six categories for judging and reporting rigor [23] are relevant for the kind of interpretivist research that we present. This research is *informed* by our relevant prior research experiences; *reflexive* in our efforts to *constantly compare* [6] collected data and gaps in our understanding; *abundant* through the number of survey participants and diversity of interview and Meetup participants; *plausible* through documented connections from memos and survey responses, to codes, and to themes; *resonant* in that the substantive theory has broad implications for how visualization research is conducted and reported; and *transparent* through the public release of the survey, its responses, and the revision history of the evolution of our codes, themes, and relevant metadata.

2.2 Thinking and Communicating About Data

We build on other efforts to understand how data workers think and communicate about data. From the beginning of our research, our main focus has been to expand understanding of one specific approach identified by Muller *et al.* [26]: how data workers approach the *design* of their data, as opposed to *discovery*, *capture*, *curation*, and *creation*.

Many authors have noted the designed nature of data abstractions [24], such as the handcrafted nature of many cybersecurity datasets [17]. Feinberg observes that the mere use of a dataset makes the user a designer of its abstraction [9], even if many data workers may not be aware of their inherent flexibility. Consequently, there is a need to learn to develop a "data vision" to exercise discretion and creativity in designing abstractions [31]. This is especially important in light of a data designer's ethical responsibility to structure data effectively [8], as the design of what is measured and how it is stored can be overtly political acts [34].

The responsibility to design effective abstractions does not always fall upon data workers in isolation. In the context of visualization design studies that involve individuals with diverse roles and expertise, designing effective data abstractions [27] and communicating effectively about those abstractions as they evolve [38], are critical to the success of a project.

However, difficulties arise in effectively communicating about data abstractions [35]. There are myriad aspects to data abstractions in design projects, such as adapting to data changes, anticipating edge cases, understanding technical constraints, articulating data-dependent interactions, communicating data mappings, and preserving data mapping integrity across iterations [43]. These difficulties are consistent with reports of there being surprisingly little documentation about the design of abstractions [47]. The lack of documentation makes human decisions invisible and threatens future analysis. In strictly machine-learning contexts, some authors have gone as far as suggesting that "deemphasizing the need to understand algorithms and models" [32] may be an effective way to increase trust in model predictions. We show that the inverse is also true: that education and transparency can foster healthy skepticism of data models and abstractions, which can be important for fairness and provenance. We argue that transparency about data abstractions can be especially important for data wrangling and visualization, in which data workers need to "interact not only with the interface but *with the data*" [43].

To facilitate communication about a particular project's specific data abstraction, the visualization research community often relies extensively upon data abstraction typologies [7, 10, 28]. Currently, the main purposes of such typologies are to guide a researcher in the selection of appropriate visual encodings, and to support transferability across

[1] osf.io archive of survey responses: `https://osf.io/s2wmp/`
[2] osf.io archive of codes, themes, and audit: `https://osf.io/382fn/`

different design studies. However, aside from highly contextual design study research itself, there is little data that reveals the extent to which the visualization research community's typologies are compatible with data workers' perspectives and language, and, although the interventionist nature of design study research is known [20], the effects of introducing foreign data concepts have yet to be described in detail.

2.3 Data about Applied Wrangling Needs

Little applied data wrangling work has been published in the visualization community, even though novel algorithms, data structures, and infrastructure need to be implemented in ways that correctly address nuanced worker needs. Such efforts often consume the bulk of the labor involved in applied visualization research [13, 15, 26], and can include rich refinements in terms of task clarity and data location that advance science and constitute important visualization research contributions in their own right [38], yet, until very recently, such work lacked a clear venue for publication without also engineering a polished visualization system.

The lack of such work leaves a major gap in needed visualization research. Although our work includes a preliminary qualitative dataset and a substantive theory that begin to fill this gap, the extent to which data workers use or even consider different data abstractions is still difficult to analyze or test quantitatively, as data wrangling decisions are rarely documented in research or in practice [47]. When such decisions are documented in research, they typically only exist as justification for a visualization design; resulting in limited information about the data abstraction, its provenance, and important documentation about how and why it was reshaped.

It can consequently be difficult to justify technique-driven or systems-focused research into general-purpose data wrangling software systems [3, 13–15, 19, 39, 41], as such efforts often lack grounding in real user needs. Instead, they are forced to rely upon past researcher experience, scant hints about real-world data wrangling precedents that exist in design study literature, and speculation about how data workers might think and what operations they might find useful. This study, and future standalone publications that are focused on data transformations, can help to better inform the design of such systems.

2.4 Creativity and Creative Roles

Discovering a latent data abstraction can have powerful creative benefits, such as inspiring radical visual innovations [21, 30]. Although the work that we present has implications for visualization researchers and their interactions with the broader population of data workers, our primary objective is to compare and contrast sets of creative objectives that can be held by any kind of data worker—including visualization researchers themselves. Consequently, we identify the role of an **abstraction theorist** that seeks to discover useful latent data abstractions, and contrast that objective against the broad set of all other concerns that a data worker may need to consider, such as data wrangling, data ownership, workflow management, the design and implementation of visualizations, evaluating a visualization, and reporting on visualization research.

Contrasting these roles is similar in spirit to Von Oech's popularized "explorer, artist, judge, warrior" creative roles [42]: "theorist" and "worker" may refer to distinct individuals in a collaborative environment, such as a visualization researcher and domain expert, or they could refer to different priorities that a single individual is considering on their own. Therefore, we describe differences through a pragmatic lens, instead of attempting to analyze different populations' creative styles or cognition [40].

We add to precedents for pragmatic guidance for creativity in visualization design, including the design of creativity workshops [12, 16] and creativity exercises [22]—we propose the pursuit of latent data abstractions as an additional creativity exercise, specific to the design of data itself.

3 Methodology

The evidence upon which we base our findings comes from two sources: memos and a deployed survey about data abstraction perspectives. Here we discuss both sources, and the way that they both influenced, and were influenced by, our internal data abstraction typology.

3.1 Memos and Timeline

We wrote memos in four contexts: 1) regular attendance at data-centered community Meetups, 2) applied conversations with data workers in diverse contexts about their perspective on their data, 3) theoretical discussions about data abstractions among the authors, and 4) collaborative open coding sessions. A summary of all memos, their relationships with codes, and code relationships with themes, are shown in Fig. 1, and an associated audit trail [5] is available in the supplemental material[2].

This project began with theoretical conversations about the nature of data abstractions between the authors, that arose occasionally as part of regular meetings. Early on, we decided to engage with an existing local Meetup group that regularly met to seek or provide help with data: a core group of regular members met twice per week at a coffee shop or bar, and continued to meet remotely beginning in March due to social distancing measures. Members and visitors frequently brought laptops to show data and code that they were working with, to solicit advice or help with debugging in a casual context. The core group and its frequent visitors included a diverse array of researchers, administrators, and data scientists from the local university and surrounding community. A selected subset of these community members—those that provided specific information that informed the development of a code—are shown in Table 1.

Later, as our survey was developed, it was deployed among this group, as well as at the IEEE VIS and Supercomputing conferences. Deployments of the survey often presented additional opportunities to discuss applied data problems with data workers outside of the context of the survey itself.

As concepts and patterns began to be less surprising, the authors began to identify codes from supporting evidence, in a collaborative open coding environment similar to the one described by Wiener [45]. We separately explored the survey responses, following our own curiosity and trying to answer our own questions, before periodically meeting and discussing our findings. As we discussed different patterns in the data, each author brought supporting personal experience or memos from a related interview to support or contest the proposed code.

3.2 Data Abstraction Typology Evolution

We began our investigation by adapting the data abstraction typology described by Tamara Munzner [28] to a data wrangling context: our initial objective was to describe a design space of possible data wrangling operations, so we modeled operations as edges in a complete graph,

Table 1. Informants

Informant	Role	Domain
I1	Professor	Medicine / Bioengineering
I2	Research Assistant	Linguistics
I3	Postdoctoral Researcher	Biology
I4	Research Director	Information technology
I5	Professor	Mathematics
I6	Research Director	Interdisciplinary institute
I7	Professor	Interdisciplinary institute
I8	Postdoctoral Researcher	Information science
I9	Postdoctoral Researcher	Biology
I10	Postdoctoral Researcher	Biology
I11	Program Coordinator	Interdisciplinary institute
I12	Postdoctoral Researcher	Bioinformatics
I13	Professor	Public health
I14	Professor	Biology
I15	Professor	Computer Science
I16	Professor	Computer Science
I17	Research Assistant	Computer Science
I18	Engineer	Industry
I19	Data Scientist	Industry

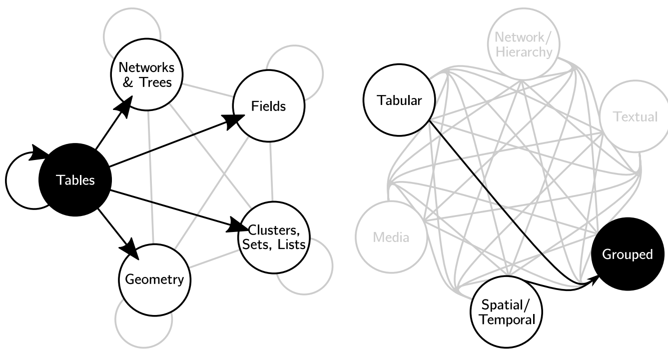

Fig. 2. The evolution of our data abstraction typology. Initially, we modeled abstractions as fitting into five specific data abstraction types, with every node in the complete graph representing a potential latent abstraction (left). Data wrangling operations, such as converting rows in a table as nodes in a network, or performing dimensionality reduction of tabular columns in a high-dimensional space, are modeled as directed edges that require changing to a non-tabular data abstraction. As we engaged in applied conversations with data workers and designed our survey, the specific categories in our typology evolved, as did its model. The final typology models the process of considering a latent abstraction as a hyperedge coming from a hybrid set of different categories to a new target latent abstraction (right), such as imagining ways to cluster rows in a table based on columns containing geographic information.

connecting each of five broad data abstraction types, as shown on the left in Fig. 2. Each edge represents a transition from one data type to another; for example, network modeling tools [3, 14, 19, 39] would largely support operations along an edge from "Tables" to "Networks & Trees." Self-edges describe wrangling operations that transform a dataset to a different form of the same type, such as transposing the rows and columns of a table.

We quickly discovered many weaknesses of this model, through our own theoretical discussions and applied conversations with data workers. Most datasets have elements or relevant metadata that could be described heterogeneously: with more than one type of abstraction. Furthermore, many of the dataset types that we initially selected were a poor fit for specific datasets, such as text corpora. These weaknesses caused us to reflect on the overall purpose of such a model: one that attempts to delineate all possible data wrangling operations may not be possible. However, adapting it to be more flexible could potentially aid in the process of exploring latent data abstractions.

Motivated by the weaknesses that we had discovered about our initial model, we began designing a survey for investigating how malleable a set of identified abstraction types can be in practice and the extent to which enforcing a different perspective on a real-world dataset can have creative benefits in its own right. Additionally, it could reveal unmet needs for data wrangling tools, although that objective was no longer prioritized. For this survey, we adapted our model to describe the act of theorizing about an alternative data abstraction, instead of performing a concrete data wrangling operation. We modeled these acts as hyperedges, also shown in the right side of Fig. 2, with the target alternative abstraction remaining singular but allowing for any combination of source abstractions. This makes it possible for survey participants to describe their dataset with any combination of abstraction types, and yet still explore an abstraction type that may be less familiar.

3.3 Open-ended Survey Design and Deployment

We developed and deployed our survey in the form of an interactive web page. It is designed in three phases, shown in Fig. 3: after the first introductory phase, the main phase of the survey invites the participant to describe a real or imagined dataset, in terms of our data abstraction typology's six broad data abstraction categories: tabular data, network / hierarchical data, spatial / temporal data, grouped data, textual data, or media. Further details are requested from participants where they indicate that they at least "rarely" interpret the data in terms of a particular abstraction category.

The final phase of the survey chooses randomly from the abstractions that a participant has indicated that they think about the *least*, and encourages them to try to think creatively about their data with that abstraction. It solicits qualitative, self-reported feedback on the extent to which the imagined transition is perceived to be useful, as well as which software tools participants would be likely to use to accomplish the transition.

Throughout all phases, the survey solicits meta-feedback about the survey itself. Participants can challenge the survey design, questions asked, our set of possible data abstraction categories, and the terminology that we used. The glossary is interactive, allowing participants to provide alternate terms or definitions. Participants can also skip sections of the survey that begin to ask questions that the participant feels have strayed from their perspective or use case.

We deployed the survey among three groups: attendees at regular community data Meetups, attendees at the 2019 IEEE VIS Conference, and attendees at the 2019 Supercomputing Conference. Although the latter two groups had a less diverse computing focus, we were aware of ongoing discussions about data abstractions within these communities, and suspected that these groups were particularly likely to offer direct critiques of our typology and approach.

4 Codes

Here we present the codes for phenomena that we identified in our collaborative open coding process, with selected supporting evidence. For more detailed supporting evidence for each code, see the supplemental archive. [2] As there are 24 codes, we present them as groups for readability, however, the analyzed themes presented in Sect. 5, and their relationships with codes, are more complex, as shown in Fig. 1.

Codes **C1–C6** are based mostly on patterns that we observed in the visualized corpus of survey responses.

C1. Compared to the diverse responses in how participants described thinking about their data, the way that they characterized how it is represented in a computer was disproportionately tabular. This disconnect between the mental model and physical computer representation indicates not only a possible need for new data storage or data wrangling tools but also a lack of awareness of other data storage options. Data workers may default to tabular data organization because it more easily fits into their current workflow and tools, or because they do not know of existing "unconventional," non-tabular tools.

C2. There was wide variation in reported dataset scales. Taken from the median response for each of the "Basic Dataset Characteristics" questions (e.g. "Approximately how large is this dataset?"), the median dataset was on the order of megabytes (close to gigabytes) in size, with thousands of items in the dataset and tens of attributes.

C3. Participants included broad techniques in their responses for wrangling tool support. When asked to actually transform their initial dataset into the alternative abstraction type, most participants listed software tools or programming languages but some listed techniques. These techniques included natural language processing ("NLP, Python", "Python, nlp techniques"), machine learning, and mathematical operations ("cluster into connected components", "Morse Smale Complex").

C4. Participants sometimes noted that they would need to ask a domain or visualization expert for help in order to change data abstractions. Along with techniques and software solutions appearing as answers to how the participant would actually transform the data abstraction, some participants acknowledged they either needed more information from a data theorist (e.g. "Could be displayed as a tree, I would hire someone") or from a domain expert ("...would need to discuss this in more detail with a domain expert...this data was not provided").

C5. Participants sometimes noted that more information would need to be collected and added to the data before transitioning to a different abstraction. To transform their data from one abstraction

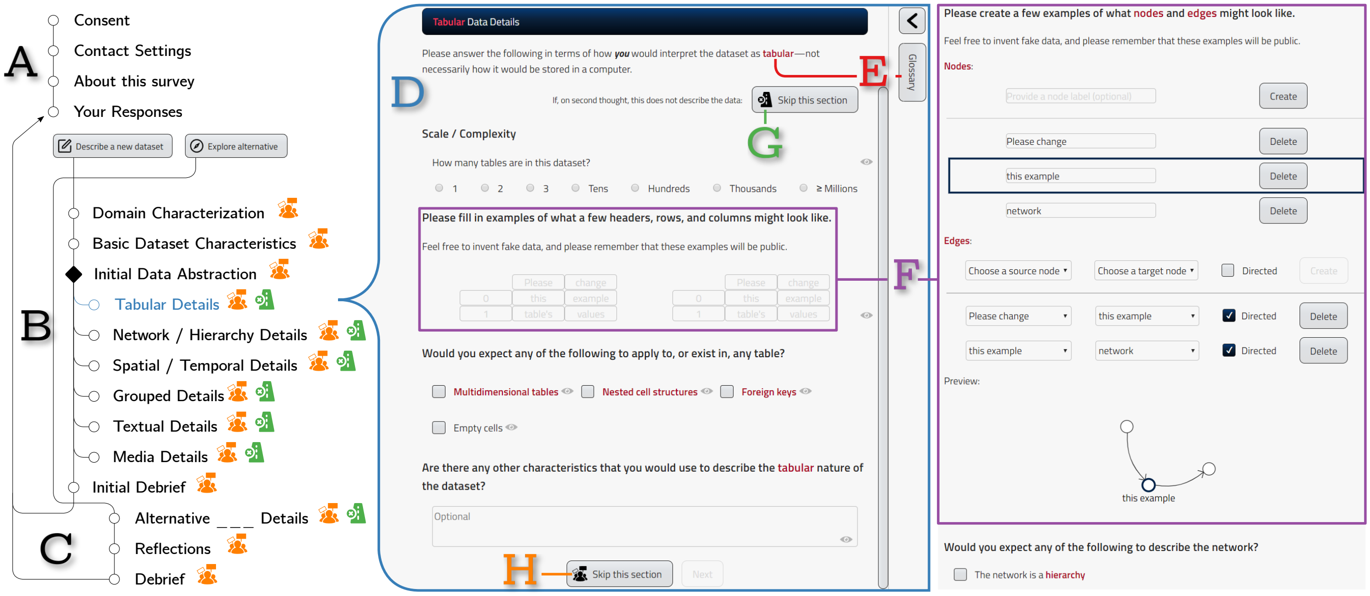

Fig. 3. An overview of the survey that we deployed. The survey is divided into three sections, shown here as a flow diagram. The first section (A) includes consent forms, contact settings, an introduction to the innovations in the survey, and a summary of responses that redirect to the other two survey portions. The main "Describe a new dataset" portion of the survey (B) invites participants to describe a real or imagined dataset, and asks them to reflect upon the extent to which they think about the dataset in terms of the six dataset types that we identified. Where participants reply that they at least "rarely" think of their data in terms of a given type, they are asked for more details in a specialized Details section of the survey. The final "Explore alternative" portion of the survey (C) invites participants to imagine their dataset as the type that they initially thought about the *least*, and fill in the associated Details portion of the survey with this new perspective. As an example, the Tabular Details interface is shown (D). Participants are encouraged throughout the survey to look up terminology highlighted in red, where participants can edit the terms and suggest alternative definitions in the glossary (E). In some Details sections, participants are asked for a small sample of what they imagine the data to look like, to help ground their thinking (F). At any point in a Details section (G), or at the end of most other sections (H), participants can choose to skip the section to provide targeted critique on the survey itself if the questions have strayed far enough from the participant's mental model.

to another, participants stated that they would need to collect additional data, such as images, speech transcripts, recordings, and labels.

C6. There was a wide distribution of the tools and techniques that data workers would use to wrangle data. Survey participants reported 54 different tools by name, with many tools being unique to a single participant. Tools that were mentioned by multiple participants tended to be programming languages.

Codes **C7–C12** are based on evidence from multiple sources, and are suggestive of unspoken perspectives, intuitions, and fears that may be common among data workers.

C7. Even before the survey guided participants to alternative abstractions, they discussed how they could see their data in other forms. This manifested both in conversations with participants before they took the survey, as well as in comments in the earliest sections of the survey before the question was asked.

C8. Many data workers did not feel that what they work with "counts as data." This comment was a common refrain while soliciting survey participation at both technical conferences, as well as through deployment across the university. However, outside of the survey, three informants (I1, I2, I4) independently made this observation while reflecting on their experiences working with people new to Data Science. For example, I2 often runs a data science workshop in the humanities but it tends to get very low attendance—often the same three participants. Seeing information as "data" may take a certain level of creativity and willingness to experiment and fail. One Supercomputing survey participant working on hardware design felt that treating circuit diagrams as "data" would be very strange, and perhaps inappropriate.

C9. Thinking about alternative data abstractions can provoke fears of scope creep. During a discussion with informants I6–I12, there was a consensus that exploring alternative abstractions can be very beneficial for the success of a project, however, it was also cautioned that it would have the potential to cause misalignments in the vision of a collaboration—usually termed "scope creep." Data workers are often cognizant of the impacts that changes to the design of their abstraction will have, including considerations and costs that they may or may not be able to articulate in detail.

C10. Data abstractions are often personal in nature to a data worker. Based on prior experiences, such as designing a visualization with I6–I12, the authors recognized that abstractions can be personal, subjective, and contextual. Wrapped in an existing data abstraction are a data worker's personal preferences, prior data science knowledge, and domain knowledge. Thus, suggestions to change this abstraction are often met with feelings of confusion and resistance. Some of these emotions stem from concerns about additional work overhead, such as those identified by **(C9)**. Other times, these emotions stem from the ecosystem of how the data was created, the people it may impact, and the subjects of the data—all things that a data worker may understand but a theorist may be unaware of.

C11. Data workers often have "gut feelings" or intuition about their data as networks. Data workers, regardless of whether their data is known to be network data or not, tended to have some intuition about the existence of networks within their data, even if specifics such as the meaning of a node or edge were unknown. Special types of networks, such as DAGs and trees, were also mentioned.

C12. Data workers often have "gut feelings" or intuition about their data as clusters, sets, or groups. Similar to **(C11)**, data workers also had intuition about the existence of groups in their data. They sometimes referred to hierarchies existing in and among these groups, and also intuited patterns and clusters in their data.

Codes **C13–C18** highlight informative weaknesses of our typology.

C13. There is wide variation in how data workers describe hierarchies. There was some initial difficulty designing the survey when deciding where hierarchies should fall. Even among the authors, we recognized that one could describe hierarchies as spatial, as networks, as nested sets. We questioned whether a tree and a hierarchy are the same thing, but concluded they have semantic differences. In the final survey, hierarchies were grouped with networks as a "Network/Hierarchy" abstraction type, with "Hierarchy" chosen deliberately to seek feedback. This diversity of perspectives was confirmed; one participant commented that they more closely align hierarchies with groups: "I find the separation of hierarchies and groupings to be a bit problematic for this domain. Many codes, such as diagnosis codes, exist in a hierarchy (defined by metadata). However it is quite common to refer to areas of this hierarchy as groupings."

C14. Most datasets did not fit in one category, and participants talked about not just the raw data, but derived values, metadata, or even "multiple datasets." Participants often selected multiple data abstractions in response to the initial question of categorizing their dataset. Heterogeneous datasets are very common, such as when metadata takes a different form from the main dataset, or when one dataset is a nested "value" inside another of a different type.

C15. "Media" as a category had a less well-defined mental model, resulting in a space with too little structure for participants to map their data crisply when forced to think of their data as "media." When asked to consider media as an alternative abstraction, a common response was to imagine screen-capturing to record images and video of a visualization of the data. But thinking of their data in this way elicited feelings of discomfort from some participants; comments such as: "This is weird. I think of the data not as media but I'm actively trying to turn it into media" and "I have displayed this data by mapping some of it [to color channels in a heatmap], but I don't consider the data itself to 'be' media or 'have' media." Some data workers understand some sort of inherent visual quality of their data. For example, one response was "The data set itself does not include any media, but interpretations of it are visual in nature... The data could be illustrated by addition of multidimensional images or 3D meshes when interlinked with concepts in the graph."

C16. Even very technical data workers find some data abstraction concepts, language foreign. We noticed confusion and misunderstanding surrounding our abstraction terminology; notably, terminology surrounding tabular data (e.g. items, attributes) was unknown to one Supercomputing participant and needed to be related to the physical spreadsheet (e.g. rows, columns) to clarify. This difference in theory-based thinking and practice-based thinking shows that there is a disconnect between how visualization people talk about data, and how data workers in general talk about data.

C17. Many data workers consider functions to be data. One unexpected finding, after reviewing responses aligning with **(C8)**, was that a subset of participants recognize functions as data. These datasets include continuous models, functions like regression models from housing data, collections of partial differential equations, or constraint data for linear or integer programming, which I5 and one author did not consider to be "spatial" as defined in the survey.

C18. Many data workers consider code to be data. As part of a larger discussion about open science and data sharing, several informants noted that code should be considered data. At a minimum, code acts as "metadata" by providing provenance of where a given dataset came from. As I6 noted that, "one person's metadata is another person's data."

Codes **C19–C21** describe the different ways that it was difficult to focus conversations with data workers on the design of a data abstraction.

C19. The design of a data abstraction proved difficult to talk about in isolation from specific file formats. Related to **(C16)**, some survey participants misunderstood the connection between an abstraction and its implementation (e.g. a table vs. a spreadsheet). As a result, in response to our request for "Other Generalizations," they suggested file formats that were clear fits for our existing six abstractions such as: "directed graph represented in a format such as dot" instead of Network/Hierarchy, "CSV file" instead of Tabular, "a collection of free text" instead of Textual.

C20. The design of data abstraction proved difficult to talk about in isolation from software and programming language abstractions. One author noted difficulties in focusing conversations on how a person thinks about their data; informants frequently pivoted to talking about abstractions imposed by software that were often only loosely associated with the data model itself, such as git's model of remotes and branches, or Jupyter's statefulness.

C21. The design of a data abstraction proved difficult to talk about in isolation from discovery, capture, curation, and creation. [25] Discussions often detoured from data design to topics such as data provenance and other data wrangling concerns. Similarly, when prompted to transform their data from one abstraction to another, some participants suggested collecting entirely new datasets, rather than transforming the existing data.

Codes **C22–C24** describe things that appeared to aid reflection and communication about data abstractions.

C22. Showing real data, such as a spreadsheet, helps data workers and theorists communicate effectively about data abstractions. Many different interactions at community meetups, such as with I3 and I19, were enhanced by the culture of bringing laptops to show data and inspect it together.

C23. Data abstraction typologies help data workers discover latent data abstractions. Asking questions about a data abstraction and how it fit, or did not fit, into a typology helped expand data workers' view of their dataset. One participant noted: "The questions made me think more about 'the nature' of this dataset. I had always considered it to be 'just tabular' but I realize that there is a hierarchy and geographic data (and a geographic hierarchy) which I hadn't really considered before. As I type this, we could layer in time and sets when considering multiple elections." Data abstraction typologies can help data workers discover underlying latent abstractions, like hierarchies, or how visualizing their data with additional data abstractions may augment understanding, like adding images to patient records.

C24. Data abstraction typologies help data workers communicate at a sufficient level of detail to design a visualization system. We observed this directly with I6–I12. A survey participant also noted that the mental exercise of the survey "prodded me into thinking about my annotations as more of a central player in the overall visualization as opposed to a secondary thought or supporting contextual element." Discussing abstraction typologies helps create a common data design language and reinforces the value that both sides (the data worker and visualization designer) bring to the data problem.

5 Substantive Theory

Together, these codes form a substantive theory with four overarching themes, including the prevalence of latent data abstractions, the interventionist impacts that pursuing latent abstractions can have, why many data workers may express hesitancy to pursue latent abstractions, and the benefits that transparency about data typologies can have for the latent abstraction discovery process. Here, we enumerate the evidence that supports each theme.

T1**: Latent data abstractions are very common.** At least initially, raw data formats are not designed in such a way as to anticipate all abstractions that may be needed or useful, yet even though these abstractions may not be fully actualized in a computer, data workers are often aware of meaningful, useful abstractions that they can communicate about without specific prompts **(C7)**. Some of these abstractions, particularly networks **(C11)** and groups **(C12)**, are very intuitive to many data workers.

This theme validates a known [2, 38] phenomenon that data rarely has a "correct" abstraction, even where predominant file formats exist; we observed that discrepancies between raw file formats and the way that a data worker thinks about their data are common **(C1)**. Instead, data abstractions have a complex and evolving form **(C14)** that must be explicitly designed.

The designed nature of data abstractions makes it important to note that neither data workers nor theorists possess comprehensive knowledge of all possible latent abstractions, and open-minded communication is necessary for meaningful, useful abstractions to be discovered. This is true for both parties: theorists are often aware of abstractions that data workers might not consider to "count" as data **(C8)**. Similarly, data workers may be aware of abstractions that theorists do not consider to "count" as data **(C17) (C18)**. Data workers and theorists may also think about the details of the same abstraction differently **(C13)**. Introducing a typology of data abstractions can expose abstractions that neither party has considered, in that a typology can contain new abstractions that data workers may not be aware of, or they may lack new abstractions that theorists have not considered **(C23)**.

T2: The visualization community identifies data abstractions for its own transferability needs, but the process of identifying an abstraction is an intervention with far-reaching effects. Collaborations with data workers beyond the visualization research community stand to benefit—and can be harmed—by the way that both parties introduce, articulate, and explore data abstractions.

Our data validates that visualization researchers, as theorists, are not operating in a vacuum; some abstractions that are common in the research community are intuitive to many data workers **(C11) (C12)**.

However, although these commonalities may be good news for the validity of the work that visualization researchers perform, there are also areas in which the culture of visualization research clashes with data workers at large: there is a often a disconnect between what theorists consider to be data and what data workers consider to be data **(C8) (C17) (C18)**. Disconnects also occur between the language that theorists use to describe data, and the language that data workers use **(C16)**. These differences in culture risk miscommunication at best, but also may put a project at risk of devolving into a bad collaboration, where either the goals of the theorist or the goals of the data worker become subordinate.

Consequently, for better or worse, introducing a theoretical perspective is almost always an intervention, and the effects of such interventions can be profound. Because the design of data abstractions is so inextricably linked to the other concerns of data discovery, capture, curation, and creation **(C21)**, changes to the design of a dataset can result in changes to all of its other aspects. Similarly, influencing a data worker's mental model of their data can have far-reaching practical effects, including disruptions in workflows and changes to the file formats **(C19)** and software **(C20)** that data workers use.

Data workers are often cognizant of the impacts that changes to the design of their abstraction will have **(C9)**, even if they may not be able to fully articulate these impacts in detail **(C10)**.

This is why we predict that **T3: data workers are less willing to pursue latent data abstractions when the design of an existing abstraction is already fundamental to their workflow.** When there exists a direct mapping between familiar software and the format of the raw data, efforts to introduce a new abstraction will likely be met with resistance.

The costs of a changed data abstraction design can include a need to learn new file formats **(C19)** and new software **(C20)** that may come with the need to learn new software skills such as programming. The tight coupling between data abstractions, workflows, and software can be seen in the bespoke wrangling software needs that arise from the combinatoric expansion of diverse abstractions, diverse workflows, and diverse dataset scales **(C6) (C2)**. However, the added cost is reduced when software practices have not yet been established and investments in learning new skills have not been made. This cost can also be mitigated when theorists are willing and able to provide expert help **(C4)**, such as wrangling the data to its needed forms.

Similarly, the costs of pursuing latent data abstractions can propagate to other data concerns **(C21)**, such as the need to collect additional data **(C5)**. The fears that data workers often feel **(C10)** and voice **(C9)** are suggestive that data abstraction changes can spill over into task abstraction changes that may begin to depart from data workers' actual needs. This potential cost can actually be an opportunity if care is taken to solicit critique whenever theoretical perspectives are introduced, as such introductions often encourage data workers to provide detailed information about their mental models that they might not otherwise articulate.

Theorists need not wait for such impositions, however, to solicit this kind of targeted feedback. **T4: Like access to real data, introducing a data abstraction typology helps to focus reflection and communication about data abstractions at a level of detail that includes actionable information.**

Our data **(C22)** validates the known pitfall [38] in which the lack of access to real data can doom a design study collaboration, because visualization researchers are less likely to have enough actionable information to articulate an accurate data abstraction. It also validates that a culture of data review [46], that is careful to emphasize good communication and transparency about the data abstraction, can compensate for a lack of access to real data because the detailed abstraction is a joint objective that all parties have a stake in.

When theorists take the time to be transparent about their agenda, including the typology that they are attempting to fit a worker's data into, revealing the typology can have similar benefits in that it helps a data worker understand what a theorist is looking for **(C24)**. Introducing typologies can expose data workers to latent abstractions that they may not have considered **(C23)**, and provides an opportunity to provide detailed feedback that might otherwise be left unspoken **(C17) (C18) (C13)**. For example, introducing a typology that is a poor fit in how it subdivides data abstraction categories can serve as an aid to communication, in that it can highlight the detailed ways that a worker considers their data to fit or partially fit more than one abstraction category **(C14)**.

Not all shortcomings of a typology are equally beneficial, however. Data abstraction categories that are too general **(C15)** or rely too heavily upon jargon **(C16)** may have limited utility. These limits are highly contextual, for example, a typology that differentiates between partitions of an abstract mathematical space and regions of a physical three-dimensional space might be useful for a data worker with a rich mathematical background to reflect upon; however, for a worker with less mathematical training, the amount of unfamiliar jargon introduced could inhibit detailed feedback.

When introducing a data abstraction typology as an explicit design activity [22], care should be taken to choose a typology with an appropriate level of granularity and enough accessible concepts to encourage feedback and critique.

6 Discussion

The substantive theory that we present describes phenomena that are suggestive of guidelines for theorizing about data abstractions. Additionally, it has implications for reporting data abstractions in many kinds of visualization research. We also reflect on our experiences and their implications for the design of data abstraction typologies, and lessons learned from our innovations in survey design and deployment.

6.1 Guidelines for Pursuing (Latent) Data Abstractions

Reflecting on the presence of latent data abstractions **(T1)**, the interventionist nature of defining data abstractions **(T2)** and in some cases the resistance to it **(T3)**, and the focusing power of typologies **(T4)**, along with our coded findings in Sect. 4, we proffer the following guidelines:

Data owners and abstraction theorists should collaboratively probe raw data. A typical design workflow may have data owners *describe* their data synchronously and then *give* one or more data files to the abstraction theorists for later review. There are several surfaces of loss in this approach, in which latent information remains latent. Data owners may forget to review elements of their data. Abstraction theorists may make assumptions given the data file that are only revisited much later, if at all. Instead, we recommend that initial meetings with data owners involve the presentation and collaborative probing of at least one raw dataset.

Abstraction theorists should introduce the typology and process that they follow. Just as theorists can feel lost without exposure to the raw data, data workers can feel lost when theorists attempt to fit a worker's project into an opaque typology or framework. For example, if a worker does not understand, at least at a basic level, that a theorist is attempting to identify relevant data abstractions before considering visual encodings, workers are forced to second-guess the theorist's needs. In such a situation, discussing their data in terms of potential visualization designs may appear to be *helpful*. As theorists request that workers provide at least one raw dataset, theorists should also reciprocate by preparing and presenting sufficient background about what they are hoping to learn or observe.

Create artifacts that document and convey abstraction details and demonstrate possible permutations. We discovered that even in discussion among the authors, people who had a close working relationship and were operating from the same typology, there were times when we believed we were discussing the same abstraction of the data, only to discover we had completely different assumptions once drawings or classifications were made explicit. Explicitly stating ideas serves as not only a communication aid, but also as a method to explore the creative space of possible abstractions and as documentation for resulting abstractions. Furthermore, writing or drawing such low-level details can be an effective strategy to ground a derailed conversation and refocus it back on the design of the data.

Challenges are an effective means of probing. They require an artifact to be challenged. Throughout our interactions with data workers, we observed that suggesting a concrete abstraction, particularly one that was unlike how the data worker usually conceptualized their data, elicited rich feedback about their data and their thinking on it. Responses beginning with phrases like "That wouldn't work because..." or "That makes no sense" were precursors to valuable reflections on their data. Setting up such a response requires some form of artifact, verbal, pictorial, textual, or otherwise to be challenged. We recommend such situations be approached sincerely as an honest, creative exercise towards considering other forms.

Typologies can serve as a guide to elicit latent elements of the data abstraction from data workers. A given typology may not fit all elements of a particular problem and dataset. However, it provides a corpus of possible abstractions with which to consider the data. These possibilities can serve as a jumping point to discuss and challenge possible abstractions of the data. Through our survey and interviews, we observed that discussions of fitting the data to various forms evoked more detail about the data itself as well as provided structure to exploring possible alternative abstractions.

Document and share the provenance of datasets. It is appropriate that a visualization and analysis solution operates on a brief period of a dataset's lifecycle and often only a subset of all possible data available. However, it is beneficial to document the latent elements of the data beyond that directly used by that solution. The source of the data, the transformations it has gone through, and related data all provide context. This context can be used to better understand how the data worker, who is more familiar with all of these elements, conceptualizes the dataset.

Assess opportunities inherent in derailments. The space of (possibly latent) data abstractions is vast in comparison to the minimal data abstraction represented in a visualization project. In following these guidelines, it can be easy for both theorists and workers to feel that the discussion has become derailed: workers may begin to discuss other data concerns such as data discovery, capture, curation, or creation. Workers may also discuss specific software or even prematurely begin to volunteer visualization encodings and techniques. Similarly, theorists may appear to be exploring esoteric concepts that do not have a clear application to a worker's project, and their explorations may threaten to add unnecessary labor to a worker's workload.

These derailments can be an opportunity to gain insight: First, discussing the design of a dataset has a tendency to prompt communication of important low-level information—even if seemingly unrelated—that workers would not otherwise bring up. Second, workers may actually be speaking on topic, but using seemingly irrelevant language about formats, software, or visualization as proxies that can be revealing about domain conventions or language, as well as revealing a need for the theorist to be more transparent about what they are looking for. Third, seemingly irrelevant topics may be indicative of a high-level mismatch of objectives, differences in perspective, or other miscommunications that could otherwise go unnoticed.

Actively seeking critique from data workers can help to identify a theorist's own derailments. Once derailments are identified, ascertaining the extent to which any of these three opportunities exist can help guide a theorist as to whether, when, and how to attempt to recenter the conversation.

Document objectives and revisit them regularly. Collaborators often have different high-level expectations, ideas, agendas, and sub-goals/tasks. This is complicated by the potential for a latent abstraction—even considering one hypothetically—to change collaborator perspectives and goals in ways that may not be communicated immediately. We recommend documenting the objectives of the project, and revisiting those objectives, especially when derailments are indicative of high-level mismatches.

Schedule interventions to revisit data abstractions. The above guidelines discuss how to make the latent *apparent*, but require the latent exist in the minds of people or the artifacts (e.g., the raw data) available. However, over the course of the project, all people involved may discover new facets of the data or incorrect assumptions previously made. Sometimes these discoveries lead to immediate intervention, but sometimes they expand the latent space. We recommend scheduling time to revisit, challenge, and refine data abstractions, given possible discoveries that are latent.

6.2 Implications for Reporting Data Abstractions

Our data suggests that providing the expert help that many data workers need can make visualization researchers more effective collaborators. Until recently, as we discuss in Sect. 2.3, performing, documenting, and reporting on this kind of work may have been difficult to accomplish by itself, even though there is a great need for published guidance and experience to inform many different kinds of visualization research.

We expect performing and reporting on detailed, applied data wrangling work better equips visualization experts to collaborate effectively. Recent acknowledgements of "Data Transformation," [29] "Data Abstraction," and "Data Structure" [18] as potential standalone contribution areas may aid in these efforts. We also suggest that such reports may be able to help ground technique- and systems- focused research in more evidence-based user needs.

6.3 Implications for Designing Abstraction Typologies

Our experience in attempting to apply the same data abstraction typology to a diverse array of data workers and datasets revealed wide variability in the extent to which typologies are likely to fit a particular context—both the diversity of datasets and the diversity of data worker expertise and perspectives can risk a poor fit.

Our data shows that this is not necessarily problematic. It demonstrates how typologies can be useful in pursuing latent data abstractions despite—and, in some circumstances, *because* of—their limitations. In the spirit of the observation that "all models are wrong but some are useful" [4], shortcomings of a typology can create opportunities to aid in detailed communication and reflection that might be less likely if the typology were a perfect fit.

This also suggests that typologies may not scale well for purposes beyond the pursuit of latent data abstractions: typologies must generalize in order to be tractable and support comparison, however, generalizations fundamentally censor diverse, individual voices and risk stifling important exceptions and innovative thinking. Our corpus of survey responses demonstrates a way that a conversation about the nature of data abstractions can be conducted at scale, in a way that balances the need for generalizability, while giving priority to individual viewpoints and grounding discussion in the context of real-world applications. In the way that our survey explicitly sought critique on the typology that we presented, it allowed for enough organization to visualize, compare, and contrast hundreds of viewpoints, while giving wider freedom for participants to engage directly with the theoretical questions implicit in the design of the survey.

6.4 Reflections on Survey Innovations and Deployment

Unlike typical surveys that primarily collect quantitative information for well-defined questions, our main objective in deploying the survey was to probe for blind spots in our own understanding of what data abstractions exist, and how data workers think about them.

Consequently, we sought to create a survey that was as open-ended as possible. Closed questions are therefore least ideal, as they provide zero opportunities for a participant to signal to researchers that the researchers may have missed something—researchers have to anticipate every possible response [36].

Open-ended, free response questions at least make it possible for participants to submit critique, but because they're expensive to code and analyze, and because they introduce more survey fatigue, they often take the form of a single comment field at the end that are only used as an "outlet" for participants, rather than a prioritized source of data [11].

The extent to which participants freely made use of the ability to skip survey sections suggests that this approach has several benefits. Replacing a whole section of a survey with a single free response field appears to help mitigate survey fatigue. The free response field is at least as open-ended as regular free response questions, and consequently incurs no additional analysis cost. The act of stepping outside the normal flow of the survey appears to have encouraged participants to think about the design of the survey itself, and in some cases, engage at a theoretical level that more closely resembles a forum than a survey.

In contrast, our interactive glossary did not appear to have garnered as much attention or use—this may have been due to its placement outside the flow of the survey, and/or its position in the corner of the screen.

The survey innovations created opportunities to improve our understanding of what data abstractions exist, what terminology is actually used by diverse data workers, to refine the evolving theory, and we expect it will inform future iterations of the survey.

7 Limitations and Future Work

Here we document the limitations of the survey that we present, our intent to deploy it to a broader audience, and suggest future uses for the dataset that we have released.

7.1 Survey Design and Evaluation

The archive of survey responses that we present is not without typical technical difficulties. One major drawback of its design was that the length of the survey varied, depending on the difference between "rarely" thinking about a dataset as a certain type and "never." This resulted in some participants filling out lengthier surveys, who began to show signs of fatigue. Additionally, a question in several of the Details sections had a bug that failed to capture the data completely. Finally, as participants almost always took the survey on their own devices, connectivity and browser incompatibility issues, especially for specific iOS devices, occasionally arose. These challenges, together with a small number of responses in which participants appeared to abuse the ability to skip sections, etc., resulted in a set of questionable responses that we have flagged.

Rather than suppress these errata, they are included in the archive of the data, and documented in context in the visualized summary of each question. The set of questionable responses can also be interactively filtered out.

As the survey design itself is not our primary contribution, we have only evaluated the extent to which our innovations were effective in achieving our qualitative aims. We can not speak to whether they are effective ways to solicit critique from participants in general, nor engaging enough to encourage theoretical reflection at the levels that we observed.

7.2 Further Survey Deployment

The feedback that we collected may also have been influenced by the groups where we deployed the survey and wrote memos about our observations. The populations we engaged with during this study all had a high interest in computing: domain scientists who come to hacking-oriented meetups and attendees at computing conferences. Although the Supercomputing conference has thousands of attendees who are there for reasons other than the technical program, in some interactions, we had difficulty convincing those people that their data counts as "data." Thus, our data and subsequent findings are lacking representation among people who do not identify with data.

Effectively engaging people with less overt interest [33], that may not share the goals represented by our "data worker" persona, is an ongoing effort that we hope to pursue in future work. Subsequent survey deployment and memo writing will target more diverse data perspectives and skill sets, by networking with people from non-Computer Science backgrounds. For example, Meetup attendees have already referenced ongoing discussions about data abstractions in a paleontology community. They are considering how to best match and connect competing ontologies from different sources. Similarly, we have been connected with a group of vehicle mechanics that are adapting their tables of diagnostic metrics to changes introduced by increasing numbers of electric vehicles. Other potential domains include linguistics, sociology, bioinformatics, construction equipment and manufacturing, and athletics. We intend to advertise and deploy our survey to more diverse groups of data workers, through academic and professional conferences, at relevant community Meetup events, and through word of mouth.

7.3 Data Reuse

We have released the public portions of the survey data in a visual, searchable format as a standalone research contribution, so that individual voices can be heard and reviewed by researchers studying similar phenomena, beyond our research aims. Such aims might include creating terminology maps across domains, using evidence in our survey responses to motivate and justify the design of general-purpose visualization and data wrangling tools, and other analyses.

8 Conclusion

Our grounded theory investigation into the malleability of data abstractions has resulted in a substantive theory about data abstractions and their implications for visualization design, guidelines for the development of data abstractions, the design and deployment of an open survey, and a corpus of survey responses that represent a discussion about the nature of data abstractions at scale. This work has implications for how data abstractions are reported, how typologies are designed and discussed, and may inform future efforts to seek critique through the deployment of surveys. Ultimately, this work sheds light on why thinking and communicating about data abstractions can be difficult, and shows how to best take advantage of opportunities inherent in that process, as well as mitigate its risks.

References

[1] A. Bigelow. *Driving Genetics with Experimental Visualization*. Undergraduate thesis, University of Utah, Salt Lake City, UT, USA, 2012.

[2] A. Bigelow, S. Drucker, D. Fisher, and M. Meyer. Reflections on how designers design with data. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, AVI '14, pp. 17–24. ACM, 2014. doi: 10.1145/2598153.2598175

[3] A. Bigelow, C. Nobre, M. D. Meyer, and A. Lex. Origraph: Interactive network wrangling. *IEEE VAST*, 2019.

[4] G. E. Box. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pp. 201–236. Academic Press, 1979. https://linkinghub.elsevier.com/retrieve/pii/B9780124381506500182. doi: 10.1016/B978-0-12-438150-6.50018-2

[5] M. Carcary. The research audit trial–enhancing trustworthiness in qualitative inquiry. *Electronic Journal of Business Research Methods*, 7(1), 2009.

[6] K. Charmaz. *Constructing Grounded Theory*. sage, 2014.

[7] E. Chi. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pp. 69–75, Oct. 2000. doi: 10.1109/INFVIS.2000.885092

[8] M. Da Gandra and M. Van Neck. *InformForm: Information Design: In Theory, an Informed Practice*. Mwmcreative Limited, July 2012.

[9] M. Feinberg. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 2952–2963. Association for Computing Machinery, Denver, Colorado, USA, May 2017. https://doi.org/10.1145/3025453.3025837. doi: 10.1145/3025453.3025837

[10] A. Figueiras. A typology for data visualization on the web. In *2013 17th International Conference on Information Visualisation*, pp. 351–358, July 2013. doi: 10.1109/IV.2013.45

[11] J. G. Geer. Do open-ended questions measure "salient" issues? *Public Opinion Quarterly*, 55(3):360–370, Jan. 1991. https://academic.oup.com/poq/article/55/3/360/1927045. doi: 10.1086/269268

[12] S. Goodwin, J. Dykes, S. Jones, I. Dillingham, G. Dove, A. Duffy, A. Kachkaev, A. Slingsby, and J. Wood. Creative user-centered visualization design for energy analysts and modelers. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2516–2525, Dec. 2013. doi: 10.1109/TVCG.2013.145

[13] P. J. Guo, S. Kandel, J. M. Hellerstein, and J. Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pp. 65–74. Association for Computing Machinery, Santa Barbara, California, USA, Oct. 2011. https://doi.org/10.1145/2047196.2047205. doi: 10.1145/2047196.2047205

[14] J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *IEEE Visual Analytics Science \& Technology (VAST)*, p. 10, 2011.

[15] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363–3372, 2011. doi: 10.1145/1978942.1979444

[16] E. Kerzner, S. Goodwin, J. Dykes, S. Jones, and M. Meyer. A framework for creative visualization-opportunities workshops. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):748–758, Jan. 2019. doi: 10.1109/TVCG.2018.2865241

[17] Á. Kiss and T. Szirányi. Evaluation of manually created ground truth for multi-view people localization. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, VIGTA '13, pp. 1–6. Association for Computing Machinery, St. Petersburg, Russia, July 2013. https://doi.org/10.1145/2501105.2501106. doi: 10.1145/2501105.2501106

[18] B. Lee, K. Isaacs, D. A. Szafir, G. E. Marai, C. Turkay, M. Tory, S. Carpendale, and A. Endert. Broadening intellectual diversity in visualization research papers. *IEEE Computer Graphics and Applications*, 39(4):78–85, July 2019. doi: 10.1109/MCG.2019.2914844

[19] Z. Liu, S. B. Navathe, and J. T. Stasko. Ploceus: Modeling, visualizing, and analyzing tabular data as networks. *Information Visualization*, 13(1):59–89, Jan. 2014. doi: 10.1177/1473871613488591

[20] N. McCurdy, J. Dykes, and M. Meyer. Action design research and visualization design. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pp. 10–18. Association for Computing Machinery, Baltimore, MD, USA, Oct. 2016. https://doi.org/10.1145/2993901.2993916. doi: 10.1145/2993901.2993916

[21] S. McKenna. *The Design Activity Framework: Investigating the Data Visualization Design Process*. PhD thesis, University of Utah, June 2017.

[22] S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design activity framework for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2191–2200, Dec. 2014. doi: 10.1109/TVCG.2014.2346331

[23] M. Meyer and J. Dykes. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019. doi: 10.1109/TVCG.2019.2934539

[24] M. Meyer, M. Sedlmair, P. S. Quinan, and T. Munzner. The nested blocks and guidelines model. *Information Visualization*, 14(3):234–249, July 2015. http://journals.sagepub.com/doi/10.1177/1473871613510429. doi: 10.1177/1473871613510429

[25] M. Muller. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In J. S. Olson and W. A. Kellogg, eds., *Ways of Knowing in HCI*, pp. 25–48. Springer, New York, NY, 2014. doi: 10.1007/978-1-4939-0378-8_2

[26] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, C. Dugan, and T. Erickson. 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*, CHI '19, pp. 126:1–126:15. ACM, 2019. doi: 10.1145/3290605.3300356

[27] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009. doi: 10.1109/TVCG.2009.111

[28] T. Munzner. What: Data abstraction. In *Visualization Analysis and Design*. CRC Press, Dec. 2014.

[29] T. Munzner, A. Endert, A. Lex, A. Ynnerman, C. Garth, M. Chen, P. Isenberg, and L. Shixia. Revise committee town hall. https://drive.google.com/drive/u/0/folders/1dqssldHbXLmAD9zeOqHCbfNTb8gjeHKS, Oct. 2019.

[30] C. Nielsen, S. Jackman, I. Birol, and S. Jones. Abyss-explorer: Visualizing genome sequence assemblies. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):881–888, Nov. 2009. http://ieeexplore.ieee.org/document/5290690/. doi: 10.1109/TVCG.2009.116

[31] S. Passi and S. Jackson. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pp. 2436–2447. ACM Press, Portland, Oregon, USA, 2017. http://dl.acm.org/citation.cfm?doid=2998181.2998331. doi: 10.1145/2998181.2998331

[32] S. Passi and S. J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):136:1–136:28, Nov. 2018. http://doi.org/10.1145/3274405. doi: 10.1145/3274405

[33] E. M. Peck, S. E. Ayuso, and O. El-Etr. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12. Association for Computing Machinery, Glasgow, Scotland Uk, May 2019. https://doi.org/10.1145/3290605.3300474. doi: 10.1145/3290605.3300474

[34] K. H. Pine and M. Liboiron. The politics of measurement and action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 3147–3156. Association for Computing Machinery, Seoul, Republic of Korea, Apr. 2015. https://doi.org/10.1145/2702123.2702298. doi: 10.1145/2702123.2702298

[35] A. J. Pretorius and J. J. Van Wijk. What does the user want to see? what do the data want to be? *Information Visualization*, 8(3):153–166, Sept. 2009. https://doi.org/10.1057/ivs.2009.13. doi: 10.1057/ivs.2009.13

[36] U. Reja, K. L. Manfreda, V. Hlebec, and V. Vehovar. Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, 19(1):159–177, 2003.

[37] M. Sandelowski. Real qualitative researchers do not count: The use of numbers in qualitative research. *Research in Nursing & Health*, 24(3):230–240, June 2001. doi: 10.1002/nur.1025

[38] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the statcks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[39] A. Srinivasan, H. Park, A. Endert, and R. C. Basole. Graphiti: Interactive specification of attribute-based edges for network modeling and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):226–235, Jan. 2018. doi: 10.1109/TVCG.2017.2744843

[40] R. J. Sternberg. *Handbook of Creativity*. Cambridge University Press, 1999.

[41] R. Verborgh and M. D. Wilde. *Using OpenRefine*. Packt Publishing Ltd, Sept. 2013.

[42] R. Von Oech and G. Willett. *A Kick in the Seat of the Pants: Using Your Explorer, Artist, Judge, & Warrior to Be More Creative*. Perennial Library, 1986.

[43] J. Walny, C. Frisson, M. West, D. Kosminsky, S. Knudsen, S. Carpendale, and W. Willett. Data changes everything: Challenges and opportunities in data visualization design handoff. *arXiv:1908.00192 [cs]*, July 2019. http://arxiv.org/abs/1908.00192.

[44] M. Weed. Capturing the essence of grounded theory: The importance of understanding commonalities and variants. *Qualitative Research in Sport, Exercise and Health*, 9(1):149–156, Jan. 2017. doi: 10.1080/2159676X.2016.1251701

[45] C. Wiener. *Making Teams Work in Conducting Grounded Theory*, pp. 292–310. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom, 2007. http://methods.sagepub.com/book/the-sage-handbook-of-grounded-theory/n14.xml. doi: 10.4135/9781848607941.n14

[46] K. Williams, A. Bigelow, and K. E. Isaacs. Visualizing a moving target: A design study on task parallel programs in the presence of evolving data and concerns. *To appear in IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis '19)*, Jan. 2020. doi: 10.1109/TVCG.2019.2934285

[47] A. X. Zhang, M. Muller, and D. Wang. How do data science workers collaborate? roles, workflows, and tools. *To appear in ACM Human-Computer Interaction*, 4(CSCW), May 2020. http://arxiv.org/abs/2001.06684.