Uncovering Data Landscapes through Data Reconnaissance and Task Wrangling

Anamaria Crisan*
Dept. of Computer Science
University of British Columbia

Tamara Munzner†
Dept. of Computer Science
University of British Columbia

ABSTRACT

Domain experts are inundated with new and heterogeneous types of data and require better and more specific types of data visualization systems to help them. In this paper, we consider the data landscape that domain experts seek to understand, namely the set of datasets that are either currently available or could be obtained. Experts need to understand this landscape to triage which data analysis projects might be viable, out of the many possible research questions that they could pursue. We identify data reconnaissance and task wrangling as processes that experts undertake to discover and identify sources of data that could be valuable for some specific analysis goal. These processes have thus far not been formally named or defined by the research community. We provide formal definitions of data reconnaissance and task wrangling and describe how they relate to the data landscape that domain experts must uncover. We propose a conceptual framework with a four-phase cycle of acquire, view, assess, and pursue that occurs within three distinct chronological stages, which we call fog and friction, informed data ideation, and demarcation of final data. Collectively, these four phases embedded within three temporal stages delineate an expert's progressively evolving understanding of the data landscape. We describe and provide concrete examples of these processes within the visualization community through an initial systematic analysis of previous design studies, identifying situations where there is evidence that they were at play. We also comment on the response of domain experts to this framework, and suggest design implications stemming from these processes to motivate future research directions. As technological changes will only keep adding unknown terrain to the data landscape, data reconnaissance and task wrangling are important processes that need to be more widely understood and supported by the data visualization tools. By articulating a concrete understanding of this challenge and its implications, our work impacts the design and evaluation of data visualization systems.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Technological advances are changing the volume and variety of data that can be collected and analyzed for scientific inquires. For example, a researcher studying the evolution of deadly disease outbreaks can now draw upon data from electronic health records, spatial polygons, social networks, genomics, and sensors embedded in the local environment. The visualization research literature assumes that experts have an understanding of these data and intend to derive actionable insights through exploratory visual analyses (EVA) [2]. However, domain experts who need to integrate and analyze heterogeneous data are becoming increasingly overwhelmed by the complexity and heterogeneity of their data, in addition to its volume. Data visualization systems that support EVA aim to help domain experts explore a dataset in depth, a process we define as **investigative exploration**. However, we have also observed that domain experts frequently need help to identify and triage viable datasets within a complex and unfamiliar **data landscape**: a large space of datasets that are either available to them now, or that they could gather, or that they could request from some gatekeeper who controls access.

This latter challenge of uncovering data landscapes is overlooked by the visualization research community. While we were made aware of this challenge through our collaborations with domain experts, we have since identified a similar challenge described in other visualization studies. Through our observations and analysis of the literature, we have defined two co-ordinated processes of **data reconnaissance** and **task wrangling** that domain experts undertake, often in an *ad hoc manner*, to uncover unfamiliar data landscapes in order identify viable datasets for analysis.

Our work aims to characterize the challenges that domain experts face when analyzing unfamiliar heterogeneous data landscapes and the design of visualization systems that may help them. Our first contribution is a concrete definition of data reconnaissance and task wrangling processes and to contrast these processes to investigative exploration. While other studies have described analogous challenges and processes, we are the first to reify these concepts. Our second contribution is a conceptual framework that illustrates how these processes serve to evolve an understand of data and tasks over time. Our third contribution is an initial retrospective analysis of visualization design studies to assess the prevalence of data reconnaissance and task wrangling processes in prior art. Finally, we provide a discussion of the effects of using this framework with domain experts as well as concrete implications for design.

2 RELATED WORK

We situate our contributions within branches of visualization research that concern methods for eliciting information from domain experts, existing theories for exploratory visual analysis, and finally to design study methodologies. We also comment on the limitations of prior work relative to the challenges attending complex data landscapes.

Elicitation Methods. The difficulty of understanding the needs of domain experts is well known in the visualization research community, which has long advocated human-centered design methods for data and task elicitation ([3], [11], [12]). These methods assume that domain experts have some understanding of their data landscape and also that data to be visualized are immediately available. However, this assumption does not hold in many situations. In our own prior research, we have documented how regulatory and organization constraints introduce overhead to visualization design and analysis by imposing limitations on data access [5]. We have found that domain experts who are sufficiently unfamiliar with a data landscape typically cannot articulate their analysis needs without further investigation of it. They may have many possible questions, but determining which of them have a reasonable chance of being answered depends on what data is available to them.

*e-mail: acrisan@cs.ubc.ca
†e-mail: tmm@cs.ubc.ca

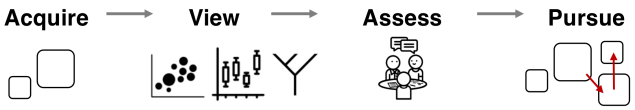

Figure 1: **Conceptual framework for data reconnaissance and task wrangling.** The four phases of acquire, view, assess, pursue are repeated across multiple cycles of a data reconnaissance and task wrangling process. Squares represent individual datasets, and the red arrows indicate domain experts obtaining new data informed by their assessment of existing data.

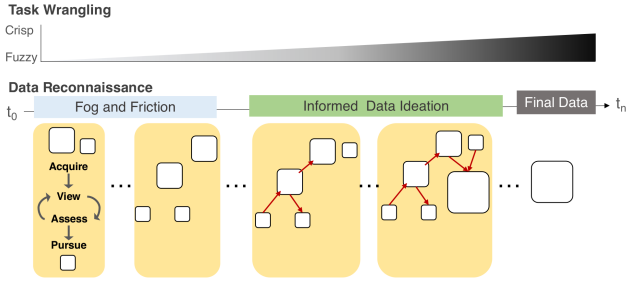

Figure 2: **Data reconnaissance and task wrangling phases over time.** Within a single cycle (yellow square) a domain expert will perform the acquire, view, and assess phases in order to familiarize themselves with their data and determine what to further pursue. Over a period of time ($t_0 ... t_n$) a domain expert will gather more datasets for analysis and form crisper ideas of the tasks these data could be used for. We distinguish three stages of this process, where the familiarity with the data landscape grows over time: the fog and friction, informed data ideation, and the demarcation of the final data.

General Purpose EVA Tools. Data visualization tools that support general exploratory visual analysis are unlikely to address the difficulties of navigating an unfamiliar data landscape. Previous exploration systems such as Tableau [13] or Keshif [15] have been architected around the assumption that domain experts are ready for a deep dive into an existing specific dataset to conduct exploratory searches [2]. General EVA tools may be used to support data reconnaissance and task wrangling processes but are not explicitly designed for this purpose. We posit that this lack of support leaves considerable burdens on the shoulders of the users.

Design Study Methodologies. An alternative approach to general exploratory systems is to develop bespoke data visualization systems using approaches such as the design study methodology (DSM) [12]. The DSM employs iterative rounds of task elicitation and prototype development to uncover data and tasks from an uncertain data landscape. The two major limitations of this approach are the need to invest many months of time into the process, and the dependence on a very specific data configuration that weighs heavily in the design of the eventual solution. It is thus a poor match for assessing unfamiliar data landscapes during data reconnaissance.

3 Data Reconnaissance and Task Wrangling

We provide a general framing of data reconnaissance and task wrangling as an abstraction of the processes undertaken by domain experts attempting to explore an unfamiliar data landscape.

3.1 Operational Definitions

Exploration is a general term that is broadly applied and consequently, captures many different complex processes [2, 10]. Here we distinguish **data reconnaissance** as the process of exploring an

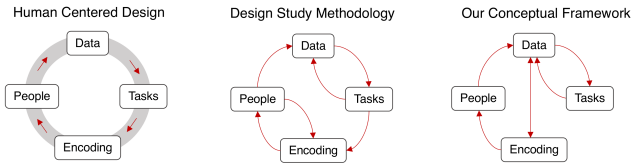

Figure 3: **Comparing different approaches to human centered design.** (a) The steps of a human centered design process [11] (b) A widely used design study methodology [12] (c) Our conceptual framework.

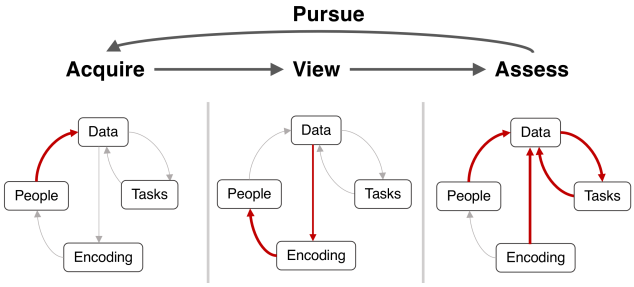

Figure 4: **Data reconnaissance and task wrangling phase breakdown.** We indicate the acquire, view, assess, and pursue phases of our conceptual framework on the human centered design loop. In our conceptual framework, the visual encoding is not informed by a specific task, contrary to the central dogma of design methodologies. However, we argue that domain experts may not have a clear notion of tasks, but through the process that we call task wrangling their ideas may evolve over time as they view and analyze data.

unfamiliar **data landscape**; that is, the very large space of existing heterogeneous and multidimensional datasets that are not yet understood by a specific person. It includes datasets that already exist but have not been assessed, or that do not yet exist but could be gathered, or that do exist but have barriers to access (e.g. due to regulatory or organizational constraints [5]). Data reconnaissance differs from *data wrangling* and *investigative exploration* which are aimed at a specific existing dataset that is transformed and analyzed in depth to generate new insights.

Task wrangling is the process of progressively forming a crisper notion of both what tasks a domain expert needs to address and whether available data is suitable for them. We follow the DSM [12] definition of a clarity axis with crispness in contrast to fuzziness on its two ends, where a crisp task has a clearly defined goal with a known set of steps. The DSM also states that task crispness should evolve over time, but assumes a clearly demarcated dataset. Task wrangling describes situations where the data itself is also fuzzy.

Data reconnaissance and task wrangling are related but distinct. A better understanding of the data landscape can help to improve the clarity of tasks, and clearer tasks provide further information about which areas of the data landscape to next pursue. We posit a chronological ordering where data reconnaissance and task wrangling come before investigative exploration and data wrangling. The objective of data reconnaissance is to identify relevant data, and the objective of task wrangling is to identify relevant visual analysis tasks. The outcome of these processes can feed subsequent exploration and data wrangling activities. Although we reify and name these processes explicitly for the first time, there is clear evidence that previous visualization researchers have indeed faced these challenges. For example, Ghani *et al.* [7] state that the challenges of forming crisper tasks for their study are exacerbated because "multimodal social networks are not a well established concept even in social science". Wood *et al.* [14] discuss the complex trajectories of different do-

mains that do not follow a "a linear progression" toward crisper tasks, but involve a more dynamic relationship between data, tasks, and understanding.

3.2 Conceptual Framework

We have devised a conceptual framework, shown in Fig. 1, that is composed of four phases that delineate data reconnaissance and task wrangling processes: acquire, view, assess, pursue. In the first phase, domain experts **acquire** some initial data in the form of one or more heterogeneous datasets. Domain experts may have some pre-determined research questions and are attempting to assess the suitability of available data to address this question [1]. Alternatively, domain experts may not have any research question in mind, but instead may wish to derive a hypothesis [2]. For example, domain experts may conduct initial pilot studies to explore an unfamiliar phenomenon and their findings can inform the design of a more robust and hypothesis driven analysis. Even in this latter scenario, domain experts will still attempt to ascertain whether data are suitable and sufficient or whether new data must be acquired. In the second phase of data reconnaissance and task wrangling, domain experts will **view** these data to gain a sense of what these datasets are, how they may be related, and a high-level overview of what they show. Here, we use the term *view* to imply that they visualize their data in order to generate a rapid and high level assessment of its utility. The key point is that viewing to quickly assess utility does not require the sophistication of viewing to address the ultimate task. In the third phase, domain experts use the resulting views to **assess** whether these data meet any of their needs, whether more data may be collected, and what tasks these data could be used for. As part of their assessment, domain experts can begin to determine whether these data serve particular analytic and domain specific tasks. Finally, in the fourth phase, domain experts can then opt to **pursue** additional data sources. Additional data are pursued when domain experts determine that the present dataset is insufficient to robustly support a research question. If access is restricted by regulatory and organizational constraints, experts could also use the these assessment-oriented views of the data to build a case for accessing other restricted data. By building on the results of the initial three phases, domain experts pursue data in a principled manner to further expand their knowledge of the data landscape through a process of progressive elaboration.

These cycles of *acquire*, *view*, *assess*, and *pursue* are repeated in order to form a final dataset for analysis and crisper notion of the visual analysis tasks. Domain experts' understanding and mental model of data, goals, and tasks evolves over time, and we have defined three stages that delineate this progression, shown in Fig. 2. During the initial phase, which we call **fog and friction**, an expert attempts to familiarize themselves with the data landscape, specifically what data are available and what data may still need to be accessed or collected. Viewing and assessing the data over repeated cycles provides a clearer sense of the tasks, that produces a targeted pursuit of datasets, or even field attributes that link datasets - an analysis phase we refer to as **informed ideation**. Eventually, experts can conclude with **final data**: a demarcated dataset that can be used for some specific analysis goal (e.g. [4, 10]). This final data may be used for a design study if a clear need for a bespoke solution emerges, or an existing general purpose EVA system might suffice for investigation.

Fig. 3 illustrates the differences between data reconnaissance and task wrangling and two previous approaches: human-centered design generally [11] and design study methodology specifically [12]. Broadly, all three of these approaches involve people, data, tasks, and visual encodings. The prior approaches require that people provide some context of creation, data, and some initial set of tasks that are further refined over time and that directly inform the design of visual encodings. Our conceptual framework permits a loosely defined creation context and advocates using the data itself to inform the creation of visual encodings in order to help experts progressively identify relevant tasks and motivate the pursuit and acquisition of new data. In Fig. 4, we overlay the phases of our conceptual framework on the human-centered design loop to indicate the goals of each phase.

3.3 Example Scenario

We conclude this section by presenting an example scenario of data reconnaissance and task wrangling that was informed by our collaboration with public health experts working with many heterogeneous datasets of different sizes and collected through multiple modalities, including electronic health records, social networks from community investigations, laboratory data, and spatial data.

The overarching goal of domain experts is to arrive at a finalized dataset that may be analyzed and visualized to inform practice standards and policy making. Experts begin in *fog and friction* where they attempt to acquire, view, and assess available data. They perform some initial statistical and visual analyses to make a quick assessment to identify data needs and pursue additional datasets, some that are available, and others that may need to generated or requested from gate keepers. Over time, experts will move to *informed ideation*, where dataset linkages are explored and solidified. Data visualizations will also become more complex as more data are made available and linked together. Visualizations developed in this phase may be considered pilot studies that provide initial evidence for some hypothesis or reveal gaps and issues in the available data; initially exploratory studies may also be conducted to delineate analytic and visual tasks more clearly. These studies will grant further access to other data, or funds to generate new data. Finally, experts can arrive at their *final data* with a crisper notion of their tasks. Traditional design study methods or existing EVA tools can be suitably employed, depending upon expert needs.

4 Analysis of Prior Design Studies

We detected suggestive references to data reconnaissance and tasks wrangling procedures in at least two instances [7, 14], and we sought to identify the extent that these procedures may have occurred in prior visualization design studies. To do so, we collected and analyzed data types and tasks reported in previously published design studies.

4.1 Methodology

We assessed 100 design studies published between 2010 and 2018. We primarily source these studies from vispubdata [9] and further augmented that data by web scraping session titles of prior VIS and EuroVis conferences to identify other design studies and also by conducting an analysis of article title and abstract text for the terms `(design AND stud*)` or `application`. We analyzed full papers and keywords for data types and tasks. We generously considered a task to be anything that proceeded the term 'task', 'design requirements', or 'design requirements' in the text; we did not include instances when the term task was used in a general sense (e.g. as in the phrase "*users performed different tasks for analysis.*"). We classified data and tasks according to their specialization: general, analytic, and specialized. The 'general' classification, refers to data and tasks that could be broadly interpreted, such 'exploring', 'browsing', or 'communicating' data. 'Analytic' classification indicate a more specific analytic intent for available data, such as 'comparing','summarizing', or 'aggregating' data, but that do not indicate a specific data type. Finally, specialized tasks have a crisp delineation of a data type and task, such as 'compare the structure of RNA isoforms'. We consider general tasks to potentially indicate instances of data reconnaissance and task wrangling. We have made our analyses and its results available online at https://github.com/amcrisan/datarecon_taskwrangle_analysis.

4.2 Results & Interpretation

Of the 100 articles, 42 articles did not indicate any tasks at all and were discarded from further analysis; we were able to extract and analyze 235 tasks from 58 unique articles. Interestingly, of the 58 articles that had tasks listed, many reported evaluative tasks, which assess the performance of a system, but do not list user specified tasks, which articulate needs of experts and motivate the design of the system. In this analysis, we do not differentiate between user and evaluative tasks, but we note that evaluative tasks may be driven more by visualization designers, rather than the needs of experts to understand their data landscapes.

Next, we derived high-level categories of tasks (general, analytic, and specialized) by examining whether the task made reference to a specific data type. We classified a total of 52 (52/235; 20%) tasks as falling into either general tasks (14/52; 27%) or analytic tasks (38/52; 73%) that are communicated without a specific reference to a domain or data type. We interpreted the presence of general tasks to suggest that data reconnaissance and task wrangling processes may have been a factor in those prior studies because a more refined relationship between data and tasks was not indicated. There are evident limitations to this retrospective analysis approach as it is not possible to go back in time and truly tease out the intentions of visualization researchers or their targeted domain experts. However, examining the dichotomy between sets of very general and specific tasks is a useful indicator of potential instances of data reconnaissance and task wrangling. Further study with additional data could serve to better delineate the boundary between data reconnaissance behavior from experts seeking to find viable datasets, and investigative exploration conducted by experts who seek to deeply probe a final dataset.

5 Discussion

Domain experts need help understanding their complex data landscapes before they can begin to deeply explore a specific dataset to derive new and actionable insights. Existing data visualization systems are primarily developed with the assumption that experts wish to explore a specific dataset, a process we call *investigative exploration*, and do not help domain experts understand broader data landscapes, including data that are presently inaccessible to them. We have introduced *data reconnaissance* and *task wrangling* as higher-order processes used by experts to uncover a data landscape. We have also defined a conceptual framework with four phases, acquire, view, assess, and pursue, to concretely describe actions driving these processes.

Data reconnaissance and task wrangling processes may be captured through application of the DSM, and indeed we have seen potential evidence of this in our examination of prior research. However, the DSM and other similar methodologies require more concrete data and tasks and are also labor intensive, often requiring multiple rounds of qualitative investigations, prototyping, and evaluation to inform visual encoding design. It would be useful to speed up the data reconnaissance and task wrangling processes. Furthermore, domain experts and visualization researchers may not have access to all available data for a study, but may use such a tool to articulate a clear need for data access.

5.1 Response of Domain Experts

We were motivated by our collaboration with public health stakeholders to create concrete definitions for data reconnaissance and task wrangling, along with the conceptual framework. Our insights arose through discussions and interviews intended to elicit goals, data, and tasks from these domain experts. In this process, one expert articulated an interesting point that he could not begin to understand what a data visualization, or even analytic, system should do because he himself was learning about this new complex data. We found that this perspective is, in fact, common and has been observed by others working with domain experts that are increasingly inundated with new and unfamiliar types of data and access constraints [8].

We have previously published on the constraints surrounding data available [5] and confidences of domain experts to interpret emerging types of data derived from new genomic technologies [6]. However, we have only recently characterized the challenges we uncovered in this prior work as resulting from data reconnaissance and task wrangling. Once we began to use the term *data reconnaissance* in lieu of the more general *exploration*, the tone of our conversation with experts changed. They felt they could associate much more strongly with the idea of reconnaissance and it led to more concrete discussions of what a data visualization system should do.

5.2 Implications for Design

Data reconnaissance and task wrangling processes introduce new challenges for developing data visualization systems because they acknowledge an expert's incomplete and evolving understanding of data, goals, and tasks. This has implications for data visualization systems and models for investigative EVA that assume a specific configuration of data and crisper goals and tasks . Our conceptual framework suggests that there would be value in developing new systems that are able to quickly and even automatically develop visual encodings from data for the purposes of exploring these data landscapes. We have proposed that these new systems should be:

- **Responsive** to different types of data that exist within a heterogeneous data landscape
- **Adaptable** to the visual encoding, interactions, and constraints arising from these different data types
- **Relevant** to domain-specific needs
- **Helpful** to navigate the breadth of a complex visualization design space that arises from heterogeneous data landscapes

Many different systems could conform to these design requirements. However, a system that is able to help domain experts rapidly view and meaningfully assess their data will be of greatest value. Such a system could be used in conjunction with traditional design study methodologies to speed up the process. Finally, the three stages (*fog and friction*, *informed ideation*, and demarcation of *final data*) may each require different solutions and defining transition boundaries between these stages would be an interesting area of future work.

6 Conclusion

Data reconnaissance and task wrangling are co-ordinated processes that domain experts undertake to familiarize themselves with an unfamiliar data landscape. These processes are essential to identify and triage viable datasets that can then be explored or analyzed in depth. However, these processes have been largely overlooked by the visualization research community. Here, we have provided definitions for data reconnaissance and task wrangling and have developed a conceptual framework that reifies how these processes are used to navigate an unfamiliar data landscape where some data is immediately accessible and other data is not. While our definitions and conceptual framework require further validation, we hope that providing these concrete definitions will enable visualization researchers and practitioners characterize and explicitly support domain experts in their attempts to understand unfamiliar data landscapes.

Acknowledgments

AC is supported by Vanier Canada Scholarship, UBC Four Year Fellowship, UBC Affiliated Fellowship, and UBC Public Scholar award. TM is supported by the NSERC Discovery Program. The authors also wish to thank the members of the UBC InfoVis group for their inputs and feedback: Madison Elliott, Shannah Fisher, Stephen Kasica, Zipeng Liu, and Michael Oppermann.

REFERENCES

[1] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 2019. doi: 10.1109/TVCG.2018.2865040

[2] L. Battle and J. Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Eurographics Conference on Visualization (EuroVis) 2019*, 38(3):1–15, 2019. doi: 10.1109/TVCG.2018.2865040

[3] M. Brehmer, S. Carpendale, B. Lee, and M. Tory. Pre-design empiricism for information visualization. In *Proceedings of the Fifth Workshop on Beyond Time and Errors Novel Evaluation Methods for Visualization - BELIV '14*, pp. 147–151. ACM Press, New York, New York, USA, 2014. doi: 10.1145/2669557.2669564

[4] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, Dec. 2013. doi: 10.1109/TVCG.2013.124

[5] A. Crisan, J. L. Gardy, and T. Munzner. On regulatory and organizational constraints in visualization design and evaluation. In *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, BELIV '16, pp. 1–9. ACM, New York, NY, USA, 2016. doi: 10.1145/2993901.2993911

[6] A. Crisan, G. McKee, T. Munzner, and J. L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ*, 6:e4218, Jan. 2018. doi: 10.7717/peerj.4218

[7] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist. Visual analytics for multimodal social network analysis: A design study with social scientists. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2032–2041, Dec 2013. doi: 10.1109/TVCG.2013.223

[8] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 99. Microsoft Research, 2009. doi: 10.1017/CBO9781107415324.004

[9] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Mller, and J. Stasko. Vispubdata.org: A metadata collection about ieee visualization (vis) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, Sep. 2017. doi: 10.1109/TVCG.2016.2615308

[10] H. Lam, M. Tory, and T. Munzner. Bridging from Goals to Tasks with Design Study Analysis Reports. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):435–445, 2018. doi: 10.1109/TVCG.2017.2744319

[11] D. Lloyd and J. Dykes. Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. *IEEE Trans. Visualization and Computer Graphics*, 17(12):2498–2507, Dec 2011. doi: 10.1109/TVCG.2011.209

[12] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans. Visualization and Computer Graphics*, 18(12):2431–2440, Dec 2012. doi: 10.1109/TVCG.2012.213

[13] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Trans. Vis. Comput. Graph.*, 8(1):52–65, 2002.

[14] J. Wood, R. Beecham, and J. Dykes. Moving beyond sequential design: Reflections on a rich multi-channel approach to data visualization. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2171–2180, Dec 2014. doi: 10.1109/TVCG.2014.2346323

[15] M. A. Yalçin, N. Elmqvist, and B. B. Bederson. Keshif: Rapid and Expressive Tabular Data Exploration for Novices. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2339–2352, 2018. doi: 10.1109/TVCG.2017.2723393