Data-First Visualization Design Studies

Michael Oppermann *   Tamara Munzner †

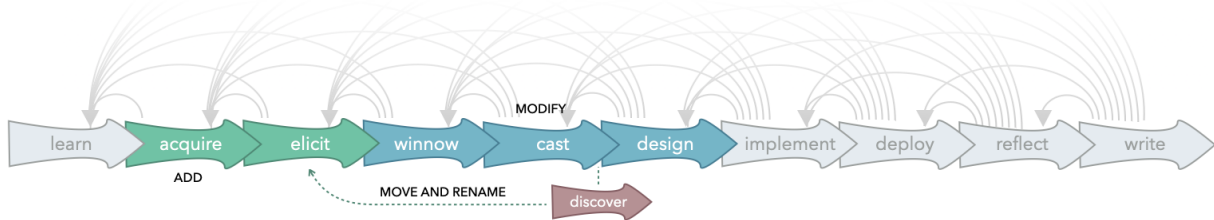University of British Columbia

Figure 1: Refined and extended framework for data-first design studies. A new *acquire* stage is added for both obtaining and abstracting data. *Discover* is moved and renamed to *elicit*, to emphasize the elicitation of tasks from potential stakeholders. *Winnow* focuses on analyzing the match between the abstract tasks of these stakeholders and the data abstraction. Nuances differ in the *cast* and *design* stages to incorporate the specific characteristics of the data-first approach.

ABSTRACT

We introduce the notion of a *data-first design study* which is triggered by the acquisition of real-world data instead of specific stakeholder analysis questions. We propose an adaptation of the design study methodology framework to provide practical guidance and to aid transferability to other data-first design processes. We discuss opportunities and risks by reflecting on two of our own data-first design studies. We review 64 previous design studies and identify 16 of them as edge cases with characteristics that may indicate a data-first design process in action.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods—Visualization theory, concepts and paradigms

1 INTRODUCTION

Design studies are frequently used to conduct and report problem-driven visualization research with domain experts. A design study is characterized by its highly iterative nature with a tight interplay between the analysis and abstraction of stakeholder needs, the transformation of domain-specific into domain-agnostic data, and the visualization design.

The design study methodology (DSM) by Sedlmair et al. [36] provides methodological guidance on how to conduct this type of applied visualization research, proposing a nine-stage sequence: *learn*, *winnow*, *cast*, *discover*, *design*, *implement*, *deploy*, *reflect*, and *write*.

The DSM focuses on connecting with collaborators in the form of domain-expert stakeholders at an early *winnow* stage with suggested criteria for narrowing down the set of potential collaborators, identifying their roles in the *cast* stage, and then undertaking a *discover* stage focused on problem characterization and task abstraction with the chosen stakeholder. The selection of an appropriate stakeholder dictates the relevant data, which is then abstracted from domain-specific to domain-agnostic form in the *design* stage, in parallel with

---

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

iteratively creating appropriate visual idioms. We now characterize this DSM approach as a **stakeholder-first** ordering.

We find that this framework falls short of capturing the nature of design studies that are primarily initiated by acquiring an interesting real-world dataset rather than selecting a specific stakeholder; we call these **data-first** design studies. In a data-first design study, the early selection of the data constrains appropriate choices for stakeholders. Data abstraction is carried out early, and followed by problem characterization and abstraction for multiple potential stakeholders. Stakeholders are chosen based on whether their tasks can be supported by the selected data source.

These non-traditional data-first design studies have not been explicitly reported or analyzed in the visualization literature. However, in retrospect we realize that they occur frequently in different variations, especially in class projects or visualization design competitions [10, 13, 19] but also in research contexts such as our own. We suspect that every visualization researcher has experienced the situation of stumbling across a dataset that aroused curiosity. In some cases, these serendipitous discoveries lead to the exploration of interesting design alternatives where the results and methods are shared on social media or documented in blog posts. In other cases, especially with more complex datasets or involving unique data characteristics, this process can lead to novel visualization contributions [34].

In this paper, we reflect on two of our own design studies that we now consider to be examples of a data-first approach, the *Bike Sharing Atlas* [30] and *Ocupado* [31], that guided our initial thinking about this alternative methodological approach for design studies. We follow up with a review of 64 previous design studies where we note 16 edge cases that may implicitly indicate a data-first ordering. The full list of 64 studies is provided in the supplemental materials, and the 16 identified examples are provided in Appendix A.

We contribute a first characterization of data-first design studies, namely those triggered by data instead of specific stakeholder questions. We propose an adapted version of the nine-stage design study methodology framework where we introduce a new *acquire* stage, move and rename *discover* to *elicit*, and adjust other stages accordingly. In addition, we present a review of previous design studies and discuss opportunities and risks of these non-traditional design studies. Our goal is to provide practical guidance and aid transferability of data-first design processes.

2 RELATED WORK

We discuss the extensive previous work on design studies and visualization process models to contextualize our proposed data-first approach. Literature on design studies emphasizes the need for real-world data and an early data understanding, but a stakeholder-first ordering is generally presumed.

In the original DSM [36] the order and timing of the data acquisition is not addressed head on; the closest guidance are the questions *"Does real data exist, is it enough, and can I have it?"* related to the pitfall of *"no real data available (yet)"* that is noted in the winnowing stage where potential stakeholders are being assessed.

The design study *lite* methodology [37] is an expedited version to fit within the time frame of undergraduate and graduate visualization courses. Recruiting community partners is a precondition in this framework. Although highly desirable, access to many stakeholders with course-relevant visualization needs is not always feasible. In larger class settings, students may not have access to stakeholders but could begin with data. The data-first methodology would provide a path for some or all such student groups to engage with domain experts at a later stage.

Munzner's nested model for visualization design [27] and different types of process models assume a user with a specific domain problem at the start, who may act as a data provider as well. The activity-centered framework [22] suggests to mandatorily request access to real data, or a data sample of sufficient size, early in the process. McKenna et al. [24] proposed a framework that links actions directly to the nested model with the four activities, *understand*, *ideate*, *make*, and *deploy*, whereby a project can start with any activity. The *understand* activity implies the data acquisition and task elicitation of the proposed data-first approach.

McCurdy et al. [23] suggest that a real-world domain problem is either expressed by domain experts or discovered by design researchers, whereas the latter is closely related to our framework adaptation. The human-centered design process by Lloyd and Dykes [21] highlights the effectiveness of real and interesting data to engage participants in requirement elicitation, which is a key objective in data-first design studies.

Crisan and Munzner [9] proposed a conceptual framework of data reconnaissance and task wrangling to explore unfamiliar data landscapes. This framework puts data first but focuses on domain experts and how they refine data to support a specific analysis goal.

Goodwin et al. [14] presented a creative design case study in the energy domain that explores possible visualization usage scenarios for a smart meter dataset, which is closely related to our suggested workflow but discusses primarily creativity techniques and assumes a set of users at the beginning.

In the context of evaluating visualization tools in large companies, Sedlmair et al. [35] distinguish between *employee-pull* and *researcher-push* solutions. Although the fundamental concept of pushing ideas from an outside perspective is similar, these authors regard it as advertising a specific tool. We instead propose to advertise data and the use of visualizations to solve a domain problem together.

3 CASE STUDIES

We introduce two of our own previously published design studies that we now consider to be examples of data-first design studies, before we generalize the underlying method in a refined DSM framework in Sect. 4 and discuss its implications in Sect. 5.

The Bike Sharing Atlas combines and visualizes distributed data from several hundred bike-sharing networks worldwide [12]. We show that the data produced by these networks reveal interesting insights, not only into patterns of bicycle usage, but also the underlying spatiotemporal dynamics of a city. By working with users from different domains including a bike sharing operator, a city planner, and urban sociologists, we illustrated how interactive visualization can help to open the data that is produced in our cities to a wider audience.

Ocupado is a set of visual decision-support tools centered around occupancy data for stakeholders in facilities management and planning [31]. We take WiFi device counts as a proxy for human presence, showing how to leverage data that had previously only been used for automation in building control systems in many new ways. We interviewed potential stakeholders from different domains and evaluated the conformity between tasks and data affordances. In a highly iterative process and with extensive feedback from a set of core stakeholders, we developed and deployed Ocupado, which was eventually adopted by our industry partner.

These two projects were different than regular design studies in the way that we engaged with stakeholders after acquiring the data.

We had worked with a bike sharing operator early on to analyze data from Vienna, but the discovery and collection of data from many station-based bike sharing networks worldwide [8] triggered additional use cases and aroused interest from new stakeholders working in different domains. The Ocupado project began through a collaboration with a startup [1] that estimates occupancy based on WiFi signals for building automation usage. We conjectured that with suitable visualizations, this data could be actively useful to many additional user groups for decision-making and resource management, and explicitly sought out these potential stakeholders.

In both design studies we had to consider whether to embed additional data sources into the mix to support potential stakeholder tasks. For example, for the bike sharing project, we included elevation profiles because the elevation differences between stations are an essential factor for the functioning of a bike sharing system, and the combined data can provide relevant insights. When we worked with custodial managers during the Ocupado project, we were asked to add work schedules that can be compared to occupancy statistics. We ultimately decided against an integration because these schedules were not relevant for other stakeholders and it was unclear if and how they facilitate core tasks related to occupancy analysis.

4 REFINED AND EXTENDED FRAMEWORK

We propose an alternative, extended version of the DSM framework, generalizing our experience as a guide for future data-first design studies. The data-first design study stages are shown in Fig. 1. One major change is to add an early *acquire* stage focused on both obtaining and abstracting data. The second change is to move the *discover* stage immediately afterwards and rename it *elicit* to reflect its task-based emphasis: obtaining information about and abstracting tasks. We keep the same ordering and general function of the *winnow* and *cast* stages, though *winnow* is heavily dependent on the initially acquired data, and nuances differ in the *cast* stage. The focus of the *design* stage is narrowed to cover only idioms, since the data abstraction occurs earlier in the new *acquire* phase. The subsequent stages *implement* and *deploy* are unchanged, as is the initial *learn* stage. *Reflect* and *write* are adjusted to reflect and report also on the peculiarities of the data-first process. As with the original DSM framework, jumping backwards to previous stages is common and necessary, in particular when adding new stakeholders.

We use the term *visualizer* to mean either a visualization researcher or practitioner. The adapted framework contains 10 stages:

Learn: This stage remains unchanged from the original because it concerns the visualization knowledge in general and independent of the specific domain problem. A solid understanding of the visualization literature is crucial for every design study.

Acquire data: A visualizer encounters, collects, generates, or obtains access to a dataset. The visualizer translates the data description into domain-independent language and begins to develop a data abstraction. Data sketches and descriptive statistics can provide first insights and guide the development of initial hypotheses about its
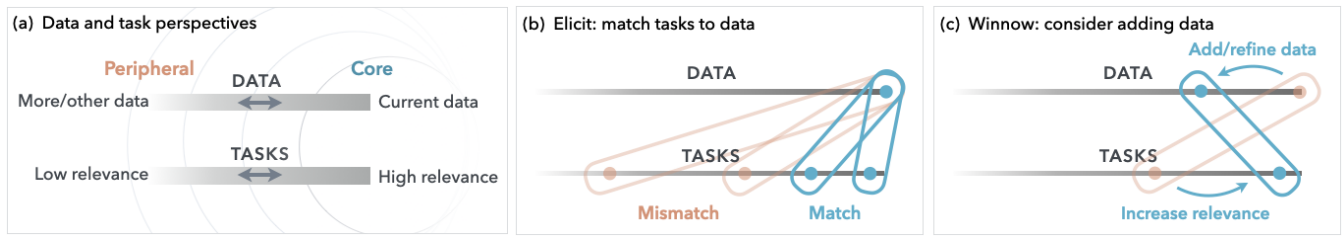
Figure 2: Simple conceptual model illustrating the challenges of finding an intersection of relevant tasks and data while keeping the problem space constrained. (a) The data and task axes both range from peripheral to core. (b) In the *elicit* stage, the visualizer examines the match between the initially acquired abstracted data and stakeholder tasks. (c) In the *winnow* stage, the visualizer can assess the benefits and risks of expanding the scope to additional data to expand the set of tasks and thus stakeholders.

underlying semantics.

The data-first DSM framework emphasizes iteration, as does the original DSM, and thus takes into account the idea that further data abstraction refinement will surely occur after task abstractions are identified. Relevant questions: *What type of data am I working with and what are the underlying semantics? Are there any data quality challenges and is long-term access guaranteed? What is special about this data and who would benefit of seeing and exploring it?*

Elicit tasks: The visualizer seeks out multiple potential stakeholders and elicits domain-specific tasks from them that might be relevant for the chosen data abstraction. The visualizer both explains the initial data abstraction and learns about unsolved stakeholder needs.

The visualizer translates the tasks into domain-independent language and considers what data is required to address them. *Is there a match between given data and stakeholder need? Is more or other data required to answer domain-specific questions, and if so, how difficult is it to acquire and integrate?*

Winnow: The goal of the data-first DSM winnowing stage is to assess and prioritize the set of potential stakeholders according to the match between their abstract tasks and the data abstraction. This process may involve multiple rounds of task and data abstraction, as the understanding of the stakeholder needs and of the dataset semantics evolves from partial to more complete. We discuss these data and task perspectives in Sect. 5.1. *How frequent are their data-relevant tasks? How central are these tasks to the stakeholder's primary mission? How many people in the organization deal with these tasks?*

Cast: The *cast* stage is similar to the original DSM with the goal to identify collaborator roles; here we note some previously proposed roles that have particular ramifications in the data-first case.

We named the *promoter* role in the Ocupado paper [31], and call it out here as essential in a data-first design study. The visualizer needs to serve in this role to reach out to potential stakeholders to present data and visualization opportunities. In addition, partners or stakeholders may act as promoters themselves when they undertake technical evangelism to other potential users about the benefits of a visualization system. When this evangelism includes demos of software prototypes, as it often does, the promoter may be involved as an intermediary in the task elicitation and winnowing stages. In this case, the promoters can serve as a conduit to relay information back to the visualizers about stakeholder needs and how they align with prototype capabilities. In retrospect, we identify an example of this role occurring within our own previous Overview work [7].

The *data producer* and *data consumer* roles were first proposed by Kerzner et al. [18] within stakeholder-first design studies. These two roles are most certainly held by different people in data-first studies, and therefore require particular consideration. A data provider may need to ensure long-term access to and the functioning of the data pipeline. The visualizer may sometimes take on the data provider

role in a later project stage, even if that role was played by somebody else at an early stage.

Design: The *design* stage focuses only on visual idioms, whereas it also includes data abstraction work in the original DSM; in our framework, the latter was carried out earlier, in the *acquire* stage.

Implement and **deploy:** Remain unchanged from the original DSM.

Reflect and **write:** Data-first design studies invite specific considerations during the reflection and writing stages. *How many potential stakeholders and domains have been considered, and ultimately chosen? What are the differences and commonalities among the stakeholder tasks? Did the original data suffice or was there a need to integrate data from additional sources?*

5 REFLECTIONS AND RECOMMENDATIONS

We reflect on the challenges of the winnowing stage, the opportunities of the data-first approach, and its risks.

5.1 Winnowing Tasks and Refining Data

The key challenge in this non-traditional design study approach is to identify which stakeholders are appropriate, and thus to identify whether there is a match between potential stakeholder tasks and the available data. To do so, the visualizer must assess the relevance of the identified task for the given data, and also consider whether to integrate secondary data sources. These considerations take place during the *winnow* stage.

Fig. 2 illustrates a simple conceptual model for reasoning about the conformity between data and tasks as aligned axes through a continuum with peripheral on the left, and core on the right. At the core are the current data and the high-relevance tasks, with other data and low-relevance tasks being relegated to the periphery. Fig. 2b shows several tasks assessed according to the initial data selection arising from the *elicit* stage, with some determined to be mismatches and some considered as matches.

Fig. 2c illustrates the tradeoffs of incorporating additional datasets into the scope of a project during the *winnow* stage. A possible benefit of adding secondary data sources could be to support more tasks, and thus to expand the set of eligible stakeholders. However, the visualizer also needs to ensure that the problem space under consideration maintains sufficient cohesion to have clear system goals and design targets. Cohesion can be reasoned about as constraints on the number and variety of stakeholders and of tasks.

5.2 Opportunities

We have identified several opportunities of a data-first approach.

Capitalizes on possibilities that do not fit traditional DSM process. We consider data-first design studies as an alternative way to approach visualization research collaborations, where interesting research questions can emerge through conversations that unfold

later in this process than in the traditional one. Analogous to market pull versus technology push dynamics in product innovation strategies [5], the originally acquired data is used to push ideas from an outside perspective instead of pulling analysis questions from a domain expert. For example, using existing WiFi access points to sense occupancy patterns in hundreds of rooms without installing additional hardware yielded a completely new opportunity that Ocupado stakeholders had not envisioned before.

Launching a data-first design study without a specific stakeholder collaboration commitment at the beginning of a project is a possible method to approach long-distance work and to engage with domain experts in imperfect circumstances, such as under COVID-19 restrictions. The *learn* and *acquire* stages are conducted independently by the visualizer. When reaching out to potential stakeholders, first ideas for solving a specific problem are presented and guide the discussion, instead of conducting a fully open-ended brainstorming session which can be challenging using only digital communication tools.

Allows very early data sketches or technology probes with real-world data. Launching a design study project without access to real data is a common pitfall [36]. Synthetic or toy data often results in additional project obstacles and delays [18, 39], and may even lead down the wrong path of data abstraction and visualization design. Access to real-world datasets at the beginning of the project allows the creation of data sketches [21] and technology probes [16], both to internally guide the development of hypotheses and possible usage scenarios, and to externally inquire needs and desires of target user groups.

Supports gradual expansion of stakeholder set. Multi-channel engagement with several stakeholders fosters serendipity and enables more impactful visualization designs, as demonstrated by Wood et al. [41] and confirmed in our own data-first design studies [30, 31], while providing breadth to a convincing evaluation [34]. However, addressing the needs of multiple stakeholders involves the risk of conflicting tasks and an ever-expanding project scope. In data-first design studies, stakeholders are selected by evaluating the conformity between tasks and data capabilities. Although additional data could be added, the primary data source serves as a guide to navigate the due diligence with stakeholders and to delimit the problem space.

Encourages progression of course or side projects into publishable results. Data-first design studies may start off during visualization courses or when exploring and visualizing data out of curiosity. Characteristics such as complex datasets, unique characteristics, or combinations of different data sources that address high-relevance stakeholder tasks may suggest a fruitful direction to expand initial efforts to result in interesting and novel knowledge contributions.

5.3 Risks

Hypothesized tasks. In the data-first approach, visualizers begin exploration immediately after acquiring the initial data, with conjectures about potential stakeholders and usage scenarios. We consider the delayed involvement of potential domain experts to be a major risk, because the initial implementation is based on hypothesized rather than verified tasks. To mitigate this risk, technology probes should be limited in functionality and open-ended with respect to use to avoid wasted effort [16].

Hammer looking for a nail.

The danger of having data at our fingertips is to converge on a specific idea prematurely and develop a full-fledged prototype without an in-depth engagement of stakeholders and the consideration of the broader design space. In a design study, stakeholder participation needs to be maximized and many design alternatives must

be considered, otherwise stakeholders are marginalized to design verifiers.

No actual users. In the worst case, the search for potential stakeholders may not succeed. In this case, other types of research contributions such as a novel visualization technique or algorithm may be possible, but a successful design study would not be achievable because it requires working with real users to solve their real-world problems. A major pitfall would be attempting to validate design decisions prematurely based on speculative tasks; care should be taken to avoid this mistake. A less extreme situation is when significant time is spent on promotion, leaving less energy available to direct towards research questions. The pitfall in this case would be a design study that requires longer than planned before the work is sufficiently complete to publish.

Mismatch between data and stakeholder tasks. A clear match between data and stakeholder tasks is not always immediately obvious and a short-cut to data and task abstractions can lead to a premature commitment to specific stakeholders. In some cases, proxy measures that stand in for a variable of interest entail a range of limitations. In other cases, the given data is out of date or needs to be augmented with other data sources. A misinterpretation can lead to project detours or to a failed design study.

Joy of discovery versus actual needs. The presentation of data through visualizations often sparks interest, particularly when potential stakeholders see their own data or data similar to theirs. We observed many meetings where domain experts enthusiastically began to conjecture about the many different ways they might be able to use the data. This process is encouraged to assess the conformity between the given data and relevant tasks. However, at these initial meetings ideas are often proposed based on limited information; the challenge for the researcher is to discriminate between initial enthusiasm and substantive user needs. A central concern is to understand whether the visualization addresses a core task of the target user group, to avoid the pitfall of a task that is only of peripheral relevance. Fellow tool builders and gatekeepers [36], who frequently act as intermediaries in *winnow*-stage meetings, might misunderstand specific needs of front-line analysts. The pitfall to avoid is premature selection of stakeholders whose actual tasks do not in fact align with the data, who would then abandon a visualization solution quickly.

Data promises. Decoupling the data producer from the consumer may increase the responsibility for the visualizer who needs to ensure the functioning of the data pipeline throughout all stages of the design study—as opposed to stakeholder-first design studies where data is typically provided by or collected together with the domain expert. A risk is therefore to have a strong collaboration with stakeholders but not deliver on the initial data promises. For example, generating interest by presenting data from one city, but not being able to follow through by collecting data from the city that is most relevant to the stakeholder, could be a major barrier for project success that is independent of the quality of the proposed visualization solution.

Although in many cases setting up a static database at the beginning is sufficient, for projects that rely on the availability of continuous data streams, external API or system level changes can jeopardize the success of a design study. To guarantee long-term availability and avoid continuous engineering effort, it is imperative to offload this responsibility.

Stakeholders likely depend on data providers even when the research project concludes. Visualization researchers may, but do not necessarily, act as data providers in a data-first design study.

Closed design approach. Data iteration [15], including the integration of secondary data sources in a visualization system, can be a crucial step to address a specific domain problem. One major risk in data-first design studies is to keep the focus too narrow on the

initial data without considering the wider problem space and ensuring extensibility and adaptability from an engineering and design perspective.

6 REVIEW OF PREVIOUS DESIGN STUDIES

In parallel with formalizing our proposal for a refined data-first design study methodology framework, we conducted a literature review of a diverse set of design studies. While we had our own direct experience of two projects that started with a dataset rather than a specific stakeholder as an existence proof, the ratio of these studies in the previous visualization literature was unclear. We were interested whether data-first approaches have been explicitly or implicitly described previously, what type of data they have used, and if there are any underlying common themes.

In total, we selected 64 design studies that were published between 1999 and 2019 (full list provided in supplemental material). We used the work by Lam et al. [20] who identified 39 design studies as a seed and extended this list with additional examples from Sedlmair [34] and by reviewing previous conference proceedings. In addition, we sourced design studies from vispubdata [17] by searching for the term "design study" in titles and abstracts. The vast majority of these design studies are published at InfoVis with a few examples from VAST and short papers.

In our analysis we identified 16 edge cases (25%) that indicate characteristics of a data-first design process in the motivation or process statements, listed in Appendix A. An accurate allocation to a stakeholder-first or data-first ordering is not possible based on this kind of retrospective interpretation. Nevertheless, these edge cases provide relevant insights and illustrate that data-first approaches are not at all unusual, although they are often underdocumented.

These potential instances of data-first design studies span a broad range of topics and data sources, such as music listening histories [4], code repositories [29], electricity consumption [14, 38], financial transaction flows [2], sports [32, 33], social media [25], hotel customer reviews [43], and genealogies [6]. The data is mostly acquired from publicly available databases but in some cases scraped or generated by researchers.

For instance, Miranda et al. [25] collected Flickr activity and tweets to capture spatio-temporal patterns in a city. The authors illustrated how interactive visualizations of this data can help experts specializing in different domains to better understand people's behavior in public places.

The BallotMaps [14] design study used electoral data from an open data repository to identify name and ordering biases in ballot papers. External domain experts are not apparent based on the publication but the study demonstrates with real-world examples how a visual approach can reveal significant biases that were not evident from statistical analysis alone.

Distinguishing between researchers being their own stakeholders and data-first design studies is most difficult, as there is sometimes no clear separation. Researchers with extensive expertise in a specific domain problem may be qualified to self-validate a visualization solution in an *autobiographical* design study, as we discuss in Sect. 7.3. In these 16 edge cases it is often unclear if the researchers stumbled upon data because of a broad interest in a topic or if they had a particular analysis question themselves. The level of domain expertise is sometimes indicated, for example, one of the authors of TenniVis [33] had 35 years of tennis playing experience, but in most design studies it is not evident.

While some authors [3, 32, 43] report on early interviews with domain experts to elicit requirements, many design study projects include external stakeholders rather late in the process, for example, to participate in a summative evaluation instead of influencing the design process. We consider this late involvement of target users a major risk (see Sect. 5.3), particularly for data-first approaches when real data is available and simple data sketches progress to more complex solutions.

While this review is an attempt to analyze many different types of design studies from the last two decades, our selection is certainly incomplete as many other design studies have been published. A further limitation is that our judgement is based on limited information on the design process and additional investigations are required to better understand implications of data-first approaches that were not explicitly reported. Interviewing visualization researchers would be a great opportunity for future work.

7 DISCUSSION

We discuss the role of the researcher's domain expertise, and design studies without engaging in collaboration with external domain experts.

7.1 Domain Familiarity and Data Understandability

The nature of data-first design studies requires a certain degree of domain familiarity and data understandability. Thus, not all datasets are appropriate for data-first design studies: the selection strongly depends on the personal background of the visualizer as it pertains to the research context. Initial data should be self-explanatory or well-documented, or the visualizer should have a strong personal interest.

Visualizers acquire and investigate data on their own before engaging into conversations with potential stakeholders. While specific use cases might not be immediately apparent and talks with different domain experts can lead to serendipitous findings, visualizers need to be aware of problems faced by domain experts and foreshadow potential tasks at the beginning of a project. We argue that visualizers need to be *task-curious* to advance this type of problem-driven research, which is methodologically very different to traditional design studies where only stakeholders come forward with tasks.

A thorough understanding of the selected data and its limitations is essential. Although this understanding will typically grow over time, a basic knowledge at the beginning is beneficial. Descriptive statistics and superficial explorations quickly reach their limits. In both our data-first design studies, we recorded data based on a live data stream. In the bike sharing project, although we recorded data from hundreds of cities, it followed a relatively simple and uniform scheme which allowed us to detect data limitations without having a deep domain knowledge. In the Ocupado project, we worked closely with our industry partner and data provider to better understand data semantics and quality issues.

7.2 Data Repurposing

The data-first approach entices visualizers to use available data and apply it to new problem contexts. For example, proxy variables may be used to answer questions when direct measurement is prohibitively expensive [26]. The repurposing of data [42] provides unique opportunities but entails a range of risks: data collectors may not have given consent for their data to be used in a particular way, the original data collection process may not be fully transparent and thus the data quality and underlying assumptions can not be judged reliably, and a lack of correlation with the variable of interest can ultimately entice analysts to draw wrong conclusions. Visualization researchers are responsible to determine if data is fit for the new use and need to understand all its ethical and analytical implications. As discussed in Sect. 5.1, one of the goals in the winnowing phase is to identify if the data is appropriate for a given task or if it would be better served with a different dataset.

7.3 Autobiographical Design Studies

Following the terminology of Neustaedter et al. [28], we define the process of designing and evaluating visualizations through self-usage as *autobiographical* design studies, that we distinguish from *data-first* design studies.

Although the DSM paper does not use this terminology, which was introduced at roughly the same time as its publication, the idea of it is clearly articulated within it: "While strong problem-driven work can result from situations where the same person holds both of these roles [visualization researchers and domain experts], we do not address this case further here" [36].

Autobiographical design draws on "extensive, genuine usage by those creating or building the system" [28]. Visualization experts act as primary stakeholders and design iterations are based on their own experiences. This type of first-person research [11] usually emerges from an unmet personal need and allows tinkering with an idea in a research context. The researcher should have or quickly gain extensive expertise related to the domain problem in order to validate the effectiveness of a visualization.

In contrast, data-first design studies follow the autobiographical design only at the very beginning. Domain familiarity is important but the process largely focuses on external domain experts [40] as target users and their feedback shapes the design decisions instead of insights from self-usage.

8 CONCLUSION

We contribute the first characterization of a data-first design study for applied visualization projects that begin with an interesting real-world dataset rather than stakeholder selection. The motivation behind this approach stems from our own experience in conducting two data-first design studies. Problem-driven research motivated and inspired by data enables opportunities that do not fit in the traditional design study process, although it does introduces risks and tensions. We propose an adaptation of the design study methodology framework to provide practical guidance for researchers interested in this alternative way to approach research collaborations. A preliminary review of published design studies revealed that roughly one quarter of them were edge cases which might have had an implicit data-first ordering. We conjecture that some of these researchers may have under-emphasized their true trajectory by bending their explanation closer to the previously articulated design study methodology framework. We hope this work encourages researchers to be more explicit in documenting and reflecting on various flavors of design studies.

REFERENCES

[1] Sensible Building Science, 2020. Accessed: 2020-07-07.

[2] A. Arleo, C. Tsigkanos, C. Jia, R. A. Leite, I. Murturi, M. Klaffenböck, S. Dustdar, M. Wimmer, S. Miksch, and J. Sorger. Sabrina: Modeling and visualization of financial data over time with incremental domain knowledge. In *Proc. IEEE Visualization Conference (VIS)*, pp. 51–55, 2019.

[3] R. C. Basole, T. Clear, M. Hu, H. Mehrotra, and J. Stasko. Understanding interfirm relationships in business ecosystems with interactive visualization. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2526–2535, 2013.

[4] D. Baur, F. Seiffert, M. Sedlmair, and S. Boring. The streams of our lives: Visualizing listening histories in context. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1119–1128, 2010.

[5] M. Baxter. *Product Design*. CRC Press, 1995.

[6] A. Bezerianos, P. Dragicevic, J.-D. Fekete, J. Bae, and B. Watson. Geneaquilts: A system for exploring large genealogies. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1073–1081, 2010.

[7] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2271–2280, 2014.

[8] CityBikes API. https://api.citybik.es, 2016. Accessed: 2020-07-07.

[9] A. Crisan and T. Munzner. Uncovering data landscapes through data reconnaissance and task wrangling. In *Proc. IEEE Visualization Conference (VIS)*, pp. 46–50, 2019.

[10] Data is Beautiful. DataViz Battle. https://reddit.com/r/dataisbeautiful. Accessed: 2020-07-07.

[11] A. Desjardins and A. Ball. Revealing tensions in autobiographical design in HCI. In *Proc. Designing Interactive Systems Conference (DIS)*, pp. 753–764, 2018.

[12] J. E. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Int. Joint Conf. on Artificial Intelligence*, 2009.

[13] GFDRR. VizRisk Challenge. https://understandrisk.org/vizrisk/, 2019. Accessed: 2020-07-07.

[14] 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 Trans. Visualization and Computer Graphics*, 19(12):2516–2525, 2013.

[15] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel. Understanding and visualizing data iteration in machine learning. In *ACM SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pp. 1–13, 2020.

[16] H. Hutchinson, W. Mackay, B. Westerlund, B. B. Bederson, A. Druin, C. Plaisant, M. Beaudouin-Lafon, S. Conversy, H. Evans, H. Hansen, et al. Technology probes: Inspiring design for and with families. In *ACM SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pp. 17–24. ACM, 2003.

[17] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A metadata collection about IEEE Visualization (VIS) publications. *IEEE Trans. Visualization and Computer Graphics*, 23(9):2199–2206, 2016.

[18] E. Kerzner, L. A. Butler, C. Hansen, and M. Meyer. A shot at visual vulnerability analysis. In *Computer Graphics Forum*, vol. 34, pp. 391–400, 2015.

[19] A. Kriebel and E. Murray. *#MakeoverMonday: Improving how We Visualize and Analyze Data, One Chart at a Time*. John Wiley & Sons, 2018.

[20] H. Lam, M. Tory, and T. Munzner. Bridging from goals to tasks with design study analysis reports. *IEEE Trans. Visualization and Computer Graphics*, 24(1):435–445, 2017.

[21] 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, 2011.

[22] G. E. Marai. Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Trans. Visualization and Computer Graphics*, 24(1):913–922, 2017.

[23] N. McCurdy, J. Dykes, and M. Meyer. Action design research and visualization design. In *BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, pp. 10–18, 2016.

[24] S. McKenna, D. Mazur, J. Agutter, and M. Meyer. Design activity framework for visualization design. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2191–2200, 2014.

[25] F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva. Urban pulse: Capturing the rhythm of cities. *IEEE Trans. Visualization and Computer Graphics*, 23(1):791–800, 2016.

[26] M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes. Measuring living standards with proxy variables. *Demography*, 37(2):155–174, 2000.

[27] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Trans. Visualization and Computer Graphics*, 15(6):921–928, 2009.

[28] C. Neustaedter and P. Sengers. Autobiographical design in HCI research: designing and learning through use-it-yourself. In *Proc. Designing Interactive Systems Conference (DIS)*, pp. 514–523, 2012.

[29] M. Ogawa and K.-L. Ma. code_swarm: A design study in organic software visualization. *IEEE Trans. Visualization and Computer Graphics*, 15(6):1097–1104, 2009.

[30] M. Oppermann, T. Möller, and M. Sedlmair. Bike Sharing Atlas: Visual

Analysis of Bike-Sharing Networks. *Int. Journal of Transportation (IJT)*, 6(1):1–14, 2018.

[31] M. Oppermann and T. Munzner. Ocupado: Visualizing Location-Based Counts Over Time Across Buildings. *Computer Graphics Forum*, 39(3), 2020.

[32] C. Perin, R. Vuillemot, and J.-D. Fekete. SoccerStories: A kick-off for visual soccer analysis. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2506–2515, 2013.

[33] T. Polk, J. Yang, Y. Hu, and Y. Zhao. Tennivis: Visualization for tennis match analysis. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2339–2348, 2014.

[34] M. Sedlmair. Design study contributions come in different guises: Seven guiding scenarios. In *BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, pp. 152–161, 2016.

[35] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, 10(3):248–266, 2011.

[36] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodolgy: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics*, 1(12):2431–2440, 2012.

[37] U. H. Syeda, P. Murali, L. Roe, B. Berkey, and M. A. Borkin. Design study "lite" methodology: Expediting design studies and enabling the synergy of visualization pedagogy and social good. In *ACM SIGCHI Conf. Human Factors in Computing Systems (CHI)*, pp. 1–13, 2020.

[38] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pp. 4–9, 1999.

[39] K. Williams, A. Bigelow, and K. Isaacs. Visualizing a moving target: A design study on task parallel programs in the presence of evolving data and concerns. *IEEE Trans. Visualization and Computer Graphics*, 26(1):1118–1128, 2019.

[40] Y. L. Wong, K. Madhavan, and N. Elmqvist. Towards characterizing domain experts as a user group. In *BELIV Workshop: Evaluation and Beyond-Methodological Approaches for Visualization*, pp. 1–10, 2018.

[41] 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, 2014.

[42] P. Woodall. The data repurposing challenge: new pressures from data analytics. *Journal of Data and Information Quality (JDIQ)*, 8(3-4):1–4, 2017.

[43] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1109–1118, 2010.

A APPENDIX: EDGE CASE DESIGN STUDIES

[1] A. Arleo, C. Tsigkanos, C. Jia, R. A. Leite, I. Murturi, M. Klaffenböck, S. Dustdar, M. Wimmer, S. Miksch, and J. Sorger. Sabrina: Modeling and visualization of financial data over time with incremental domain knowledge. In *Proc. IEEE Visualization Conference (VIS)*, pp. 51–55, 2019.

[2] R. C. Basole, T. Clear, M. Hu, H. Mehrotra, and J. Stasko. Understanding interfirm relationships in business ecosystems with interactive visualization. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2526–2535, 2013.

[3] D. Baur, F. Seiffert, M. Sedlmair, and S. Boring. The streams of our lives: Visualizing listening histories in context. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1119–1128, 2010.

[4] A. Bezerianos, P. Dragicevic, J.-D. Fekete, J. Bae, and B. Watson. Geneaquilts: A system for exploring large genealogies. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1073–1081, 2010.

[5] 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 Trans. Visualization and Computer Graphics*, 19(12):2516–2525, 2013.

[6] U. Hinrichs, S. Forlini, and B. Moynihan. Speculative practices: Utilizing infovis to explore untapped literary collections. *IEEE Trans. Visualization and Computer Graphics*, 22(1):429–438, 2015.

[7] B. Kim, B. Lee, S. Knoblach, E. Hoffman, and J. Seo. Geneshelf: A web-based visual interface for large gene expression time-series

[8] F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva. Urban pulse: Capturing the rhythm of cities. *IEEE Trans. Visualization and Computer Graphics*, 23(1):791–800, 2016.

[9] M. Ogawa and K.-L. Ma. code_swarm: A design study in organic software visualization. *IEEE Trans. Visualization and Computer Graphics*, 15(6):1097–1104, 2009.

[10] C. Perin, R. Vuillemot, and J.-D. Fekete. SoccerStories: A kick-off for visual soccer analysis. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2506–2515, 2013.

[11] T. Polk, J. Yang, Y. Hu, and Y. Zhao. Tennivis: Visualization for tennis match analysis. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2339–2348, 2014.

[12] A. Thudt, D. Baur, S. Huron, and S. Carpendale. Visual mementos: Reflecting memories with personal data. *IEEE Trans. Visualization and Computer Graphics*, 22(1):369–378, 2015.

[13] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Proc. IEEE Symp. Information Visualization (InfoVis)*, pp. 4–9, 1999.

[14] J. Wood, D. Badawood, J. Dykes, and A. Slingsby. Ballotmaps: Detecting name bias in alphabetically ordered ballot papers. *IEEE Trans. Visualization and Computer Graphics*, 17(12):2384–2391, 2011.

[15] 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, 2014.

[16] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1109–1118, 2010.

data repositories. *IEEE Trans. Visualization and Computer Graphics*, 15(6):905–912, 2009.