Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau

Leilani Battle[1] and Jeffrey Heer[2]

[1]Department of Computer Science, University of Maryland, College Park
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington

Abstract

Supporting exploratory visual analysis (EVA) is a central goal of visualization research, and yet our understanding of the process is arguably vague and piecemeal. We contribute a consistent definition of EVA through review of the relevant literature, and an empirical evaluation of existing assumptions regarding how analysts perform EVA using Tableau, a popular visual analysis tool. We present the results of a study where 27 Tableau users answered various analysis questions across 3 datasets. We measure task performance, identify recurring patterns across participants' analyses, and assess variance from task specificity and dataset. We find striking differences between existing assumptions and the collected data. Participants successfully completed a variety of tasks, with over 80% accuracy across focused tasks with measurably correct answers. The observed cadence of analyses is surprisingly slow compared to popular assumptions from the database community. We find significant overlap in analyses across participants, showing that EVA behaviors can be predictable. Furthermore, we find few structural differences between behavior graphs for open-ended and more focused exploration tasks.

1. Introduction

Exploratory visual analysis (or EVA) involves identifying questions of interest, inspecting visualized data, and iteratively refining one's questions and hypotheses. Visual analysis tools aim to facilitate this process by enabling rapid specification of both data transformations and visualizations, using a combination of direct manipulation and automated design. With a better understanding of users' analysis behavior, we might improve the design of these visualization tools to promote effective outcomes. However, as an open-ended process with varied inputs and goals, exploratory visual analysis is often difficult to characterize and thus appropriately design for.

Existing work provides many theories and assumptions regarding how EVA is conceptualized and performed, which prove invaluable in designing new EVA systems. However, we see in the literature that these contributions are generally defined in small snippets spread across many research articles. Existing surveys and frameworks touch on related topics, such as task analysis (e.g., [BM13, LTM18]), provenance (e.g., [ED16, HMSA08, RESC16]), and interaction (e.g., [HS12]), but not specifically for understanding EVA. Furthermore, we find several "schools of thought" on EVA, some of which may contradict one another. For example, it is argued that exploration can only be open-ended, without clear *a priori* goals or hypotheses (e.g., [AZL*19]). In contrast, we also see specific examples where analysts come to an EVA session with a clear goal or hypothesis in mind (e.g., [FPDs12, SKL*16]). This broad dispersion of the different definitions of EVA makes it difficult as a community to rigorously discuss, evaluate, and ultimately contribute to new research to advance EVA systems.

The goal of this work is to connect the many different ideas and concepts regarding EVA together and bring them into focus, enabling the community to more easily reflect on the way we motivate, analyze, and ultimately support visual exploration tasks. To this end, we make two major contributions:

- A review of relevant EVA literature, highlighting core ideas, themes, and discrepancies from across multiple research areas.
- An analysis of provenance data collected from an exploratory study using Tableau, to shed light on particularly contentious or seemingly undervalued EVA topics.

We reviewed 41 papers for insights into the EVA process. Three major themes emerged, centered around: 1) EVA goals and intent, 2) EVA behaviors and structure, and 3) EVA performance for both the analyst and the system. Within each theme, we identify one or more research questions that appear unanswered by the literature. Our initial aim is to provide additional context — not necessarily evidence — with respect to these questions and to encourage the community to take up these questions in future work.

To investigate further, we conduct a study with 27 Tableau users exploring three real-world datasets using Tableau Desktop. Our study design utilizes four task types of varying specificity, designed to match the common visual analysis tasks that occur during EVA, identified in our literature review. These tasks range from focused tasks with measurably correct answers to more open-ended, yet still goal-directed tasks. We summarize our main findings:

Evaluating Performance: We evaluate several performance metrics, such as pacing metrics (e.g., interaction rates, think time [BCS16]), as well as the variety and quality of task responses. Though not an exact match, we compare our accuracy results to prior reports of *false discovery rates* during EVA [ZZZK18]. Whereas prior work finds a 60% false discovery rate for "insights" reported during open-ended exploration, participants responding to our goal-directed prompts exhibit over 80% accuracy on focused tasks with measurably correct answers, and were generally cautious to avoid false discoveries. Furthermore, while *interaction latency* in EVA is frequently discussed [HS12,LH14,CXGH08,ZGC*16], the *pace* of exploration lacks realistic contextualization in some parts of the literature. In particular, a subset of the literature assumes that the time between interactions (or *think time*) is limited, constraining database optimization methods (e.g., [CGZ*16,BCS16,RAK*17]), at least for interactions of low cognitive complexity [LH14]. Our results show that high think times lead to slow analysis pace, and the pace of analysts is notably slower than assumed by this prior work, even with interaction taken into account. We also observe "bursty" behavior: participants spend some of their (think) time planning, then performing a relatively faster sequence of interactions. These results suggest that visual analysis evaluations could be improved via more realistic scenarios and accurate parameters.

Evaluating Goals and Structure: We next analyze participants' analysis *behavior graphs* — a structural model of "states" visited. We find that several assumptions of the structure of EVA are supported by our analysis. For example, participants' analysis sessions are consistent with a depth-first search process, confirming arguments made in prior work [WMA*16b]. However, our results also contradict other assumptions. The literature is inconsistent on whether EVA follows clear structure and patterns, and some argue that individual differences could make EVA behaviors unpredictable [ZOC*12]. We find that participants' analyses exhibit strong similarities and are somewhat predictable, but only at specific points in analysis sessions. The breadth and depth of analysis graphs are modulated by task, but the overall ratio of these measures is consistent across task types. Ultimately, we find that analysts' performance and strategies during open-ended tasks can be structurally similar to observations of more focused tasks, encouraging us to reconsider the distinctions made between open-ended exploration and more focused analysis. Though speculative, this similarity may be explained by a model in which analysts with open-ended aims formulate a series of more focused and goal-directed sub-tasks to pursue.

In sum, these results provide new perspectives on the content, structure, and efficacy of EVA sessions. We conclude with a discussion of how our findings might be applied to further the design of not only visualization tools, but also the way we evaluate them. All anonymized data artifacts generated by this work have been shared as a community resource for further study at `https://github.com/leibatt/characterizing-eva-tableau`.

2. Related Work

Our analysis builds on several areas of related work, such as logging interactions, modeling analysis states, and analyzing patterns in the resulting data structures. Visualization system state is often recorded via histories [HMSA08], interaction logs [ED16], and provenance tracking [LWPL11, CFS*06, BCC*05, SASF11]. We rely on built-in logging in Tableau [STH02] for analysis.

Visualization Theory & Process Models: Our work is informed by models of the visual analysis process developed through exploratory studies, such as those by Isenberg et al. [ITC08] and Grammel et al. [GTS10]. Isenberg et al. present a framework describing how teams collaborate during exploration [ITC08]. Grammel et al. present a model of how novice users construct visualizations in Tableau, and barriers to the design process [GTS10]. Brehmer and Munzner present a multi-level typology of visual analysis tasks [BM13], synthesized from a review of the literature on task analysis. Lam et al. review IEEE InfoVis design study papers to evaluate how high-level goals decompose into concrete tasks and visual analysis steps [LTM18]. However, like the many papers we evaluate in our literature review, existing theoretical work generally lacks a clear definition of exploratory visual analysis (also observed by Lam et al. [LTM18]), which we aim to contribute in this work. Furthermore, with our focus on log analysis, our metrics are primarily quantitative in nature, providing empirical context for a variety of EVA assumptions from the literature.

We focus on task or action-based models in our analysis [GZ09,BM13,YKSJ07,LTM18], particularly fine-grained models [HMSA08, YKSJ07]. We use the task model for Tableau proposed by Heer et al. [HMSA08], which assigns interactions in Tableau to five categories: "*shelf* (add, remove, replace), *data* (bin, derive field), *analysis* (filter, sort), *worksheet* (add, delete, duplicate), and *formatting* (resize, style) commands.".

Interaction Sequences: Interaction sequences reveal temporal relationships between observed interactions. Guo et al. [GGZL16] identify common sequences that lead to insights. Gotz and Wen [GW09] identify four common interaction sub-sequences, which they use to generate visualization recommendations. Others compute n-grams to identify common sub-sequences [BJK*16, BCS16]. Sequences are also used to build predictive (e.g., Markov) models, which can be used to compare analysts' performance [RJPL16], and predict users' future interactions [BCS16, DC17], task performance [GL12, BOZ*14a], or personality traits [BOZ*14a]. We use interaction sequences and more complex structures to track changes in analysis state over time. We contribute a new perspective on visual analysis patterns and structure using these quantitative methods.

Behavior Graphs: Problem-solving behavior can be modeled as a set of states along with a set of operations for moving between states [New72], including visual analysis [WMA*16b, STM17, WQM*17, SvW08, ST15, jJKMG07]. Graphs can capture more complex paths and patterns, such as back-tracking or state revisitation. Alternate analysis paths are depicted as branches from an analysis state. Branching can occur due to manipulation of analysis history (e.g., undo-redo interactions). Though many projects consider how to display history directly to users [HMSA08], past work lacks a characterization of the structure of exploratory analysis graphs across a range of conditions (e.g., tasks and datasets). Behavior graphs are also used in web browsing and click stream analysis [CPVDW*01, WHS*02, LWD*17, New72]. We leverage past work by similarly visualizing behavior graphs in Section 6, contributing a structural signature for analysis sessions, through which differences in analysis strategies can be measured during EVA.

Insight-Based Evaluations: Insight-based evaluations attempt to catalog "insights" generated through open-ended visualization use [SND05, SNLD06, LH14, GGZL16, ZZZK18]. These methods collect qualitative data about the EVA process, which researchers code and analyze. While useful for identifying meaningful cognitive events, the veracity of reported insights must be evaluated with care. However, having participants engage in open-ended exploration without clear goals and encouraging them to verbalize all insights that come to mind may decrease accuracy in visual analysis tasks. We use directed task prompts representative of visual analysis tasks that commonly occur during EVA, ranging from focused EVA tasks with verifiable answers to more open-ended, but still goal-directed, tasks. These tasks were identified through our literature review. We evaluate analysts' performance and analysis strategies, and compare with prior insight-based studies.

3. Themes in the EVA Literature

Through a review of the EVA literature we identify and discuss three major themes that appear frequently throughout: exploration *goals*, *structure*, and *performance*. We then present a summary definition of EVA based on our findings.

3.1. Review Methodology

We utilized the following paper selection method for our review:

1. Papers that analyze or design for EVA contexts were selected.
2. Papers described or referenced by papers from step 1 as also analyzing or designing for EVA were selected.
3. Tasks or topics from papers from step 1 that were described as relevant to EVA were identified. Task relevance suggested by paper authors *or* study subjects was considered (e.g., subjects' comments observed by Alspaugh et al. [AZL*19]). Then relevant, well-known papers that also discuss these tasks and topics were identified, such as the work by Kandel et al. [KPHH12]. These papers are only used to provide context, other irrelevant tasks or topics from these papers are excluded from our review.

Step 1 yielded 39 papers and step 2 yielded 2 papers [HS12, Shn96] for review. Step 3 Yielded 7 papers to provide additional context for specific EVA topics and tasks [Lam08, HMSA08, KS12, PC05, MWN*19, ZOC*12, KPHH12][†].

The selected papers were reviewed to identify major themes, and three themes emerged: EVA goals, structure, and performance. These themes occurred most frequently across the selected papers, and often as core priorities, for example Battle et al. prioritize system performance during EVA [BCS16], and Lam et al. prioritize understanding analysis goals in various contexts, including EVA [LTM18]. A subsequent review was made to capture similarities and differences between papers with respect to these themes.

3.2. Exploration Goals

Formulation and Evolution of Goals: An oft-stated goal of EVA is the production of new insights or observations from a given dataset (*insight generation*) [ED16, LTM18, jJKMG07, GGZL16, ZGC*16, ZZZK18, LH14, FPDs12]. Lam et al. [LTM18] describe the goal of exploration as "Discover[ing] Observation[s]"; however, this goal is vague in comparison to other visual analytic goals. Liu & Heer argue that EVA often "does not have a clear goal state" [LH14], which is a popular sentiment in both the visualization [Kei01, AZL*19, RJPL16] and database [IPC15] communities. For example, Idreos et al. [IPC15] describe EVA as a situation where analysts may not know exactly what they are looking for, but they will know something interesting when they see it. Keim makes a stronger argument: that EVA is *more* effective when the goals are *less* clear [Kei01]. Alspaugh et al. [AZL*19] take this idea even further by saying that exploration does *not* have well-formed goals; once clear goals are formed, the analysis is no longer exploration.

Others take a different view, saying that analysts' goals evolve throughout the course of an EVA session: the analyst starts with a vague goal, and refines and sharpens this goal as they explore [RJPL16, GW09, WMA*16b]. For example, Wongsuphasawat et al. [WMA*16b] describe the evolution of analysts' goals to motivate the Voyager system design: "Analysts' interests will evolve as they browse their data, and so the gallery [of suggested visualizations] must be adaptable to more focused explorations."

Bottom-Up Versus Top-Down Exploration: Exploration is often described as "open-ended", where many of the papers we reviewed equate exploration with at most vaguely defined tasks (e.g., [LH14, RJPL16, AZL*19, ZGC*16, ZZZK18]): visual analysis performed without an explicit objective, perusing a dataset for interesting observations. Open-endedness seems to be tightly coupled with the notion of *opportunistic* analysis [Tuk77, LH14, AZL*19, RJPL16]. For example, Alspaugh et al. [AZL*19] argue that during EVA, "... actions are driven in reaction to the data, in a bottom-up fashion...". Liu & Heer [LH14] suggest that "User interaction may be triggered by salient visual cues in the display...". There seems to be an argument in a subset of the literature that exploration must be unconstrained (e.g., by goals or tasks) to allow for an organic "bottom-up" process of uncovering new insights from a dataset.

In contrast, other projects describe scenarios where analysts come to an exploration session with a high-level goal or concrete hypothesis in mind. Liu & Heer [LH14] suggest that user interactions during EVA may be "... driven by *a priori* hypotheses...". Gotz & Zhou [GZ08] describe a specific example with a financial analyst exploring stock market data to identify and prioritize which stocks to invest in. Perer & Shneiderman [PS08] recount examples of domain analysts "trying to sift through gigabytes of genomic data to understand the causes of inherited disease, to filter legal cases in search of all relevant precedents, or to discover behavioral patterns in social networks with billions of people." Fisher et al. [FPDs12] study in-depth cases of EVA with three different analysts with specific goals; for example: "Sam is analyzing Twitter data to understand relationships between the use of vocabulary and sentiment." Kalinin et al. [KCZ14] describe two motivating scenarios, with users exploring stock data and astronomy data for records (i.e., stocks, celestial objects) with specific properties (e.g., stars with high brightness). Siddiqui et al. [SKL*16] describe three specific use cases, where scientists, advertisers and clinical researchers struggled to successfully explore their dataset for specific visual

[†] The full list of papers yielded from each step, along with our reasoning for the inclusion of each paper, is provided in the supplemental materials.

patterns. Zgraggen et al. [ZZZK18] motivate the multiple comparisons problem in EVA with the story of "Jean," an employee at a non-profit who is interested in exploring his organization's data to identify the best gift to send to their donors. In all of these examples, analysts are still performing EVA, but with concrete objectives to structure and focus exploration. These examples contradict the assumption of a purely bottom-up analysis strategy during EVA, indicating that, for realistic scenarios, top-down goals (including broader organizational objectives) need to be accounted for.

From our review, we observe that discussions of EVA include a spectrum of goal specifications, from no goals at all, to clear *a priori* goals and/or hypotheses. Analysts' positions within this spectrum may evolve as they learn more about their data. Furthermore, analysts may utilize *both* top-down (i.e., goal-directed) and bottom-up (i.e., opportunistic) strategies as they explore [RJPL16,LTM18]. Thus no one strategy completely represents how exploration unfolds, and *both top-down and bottom-up strategies should be considered* when analyzing and evaluating EVA use cases.

3.3. Exploration Structure

Phases of Exploration: EVA may involve iteration within and oscillation between *phases* of exploration, with analysts pursuing multiple branches of analysis over time [DR01, HMSA08]. However, the literature is inconsistent in defining exactly what the different phases of EVA are. Both Battle et al. [BCS16] and Keim [Kei01] assume that EVA follows Shneiderman's information-seeking mantra [Shn96]: "Overview first, zoom and filter, details on demand". Gotz & Zhou argue that users switch between two phases: browsing and querying of data to uncover insights, and recording their insights (e.g., writing notes) [GZ08]. Heer & Shneiderman [HS12] state that EVA "typically progresses in an iterative process of view creation, exploration, and refinement," where exploration happens at two levels: 1) as users interact with specific visualizations, and 2) in a larger cycle where users explore different visualizations. This concept is echoed by Grammel et al. [GTS10]. Perer & Sheiderman [PS08] say that analysts alternate between systematic exploration (searching with thorough coverage of the data space) and flexible exploration (or open-ended search). Wongsuphasawat et al. make a similar argument, inspired by earlier work [Tuk77]: "Exploratory visual analysis is highly iterative, involving both open-ended exploration and targeted question answering..." [WMA*16b]. The common theme is that EVA involves *alternating between open-ended and focused exploration.*

EVA and Search: Terms like "query" [GZ08, LKS13, KJTN14, DPD14, KCZ14, SKL*16], "browse" [LH14, GGZL16, BCS16], and "search" [KS12, WMA*16b, PS08] are frequently associated with visual exploration. In EVA, users are often searching for novel observations in a dataset, which could inform or validate future hypotheses [Kei01,LH14,GZ08,RJPL16,AZL*19,ZZZK18]. Jankun-Kelly et al. [jJKMG07] observe that earlier EVA systems "assume visualization exploration is equivalent to navigating a multidimensional parameter space," essentially a directed search of the parameter space of data transformations and visual encodings — a model subsequently adopted by visualization recommenders such as CompassQL [WMA*16a] and Draco [MWN*19]. Perer & Shneiderman [PS08] make a similarly strong connection between EVA and search by incorporating support for what they call "systematic exploration," an exploration strategy that "guarantees that all measures, dimensions and features of a data set are studied." Others [DPD14, KCZ14, VRM*15, SKL*16, DHPP17] propose techniques to automatically search the data space for interesting data regions or collections of visualizations for the user to review. The idea of searching for insights shares strong similarities with Pirolli & Card's Information Foraging loop [PC05], in particular the "Read and extract" action, where users extract observations or "evidence" that may "trigger new hypotheses and searches". Thus existing models of search behavior may play an important role in understanding behavioral patterns and analysis structure in EVA.

Analysis Tasks: Analysts seem to decompose their analyses into smaller tasks and subtasks [GZ08, RJPL16, AES05], where tasks may be re-used across datasets [PS08]. In the literature, we observe a consensus that EVA involves specific low-level visual analytics tasks and that specific classes of tasks occur frequently in EVA:

- understanding data correctness and semantics [PS08, AZL*19, KS12] (overlaps with "profiling" [KPHH12]),
- characterizing data distributions and relationships [Tuk77,PS08, IPC15, SKL*16, AZL*19, ZZZK18, CGZ*16, KS12, SKL*16, AES05] (overlaps with"profiling" and "modeling" [KPHH12]),
- analyzing causal relationships [PS08, HS12, STH02] (overlaps with "modeling" [KPHH12]),
- hypothesis formulation and verification [PS08, Kei01, LH14, RJPL16, SKL*16, AES05, AZL*19],
- and decision making [RJPL16, RAK*17, KJTN14].

For example, Stolte et al. [STH02] describe EVA as the process of "extract[ing] meaning from data: to discover structure, find patterns, and derive causal relationships." In similar spirit, Perer & Shneiderman [PS08] argue that during EVA, analysts seek to "...understand patterns, discern relationships, identify outliers, and discover gaps." Alspaugh et al. [AZL*19] find that analysts describe several of their own activities as exploration activities, which were re-classified by Alspaugh et al. as understanding data semantics and correctness or characterizing data distributions and relationships.

Interactions: EVA involves sequences of small, incremental steps (i.e., interactions) to formulate and answer questions about the data [HMSA08, GW09, WMA*16b]. Iteration could manifest as multiple interactions with the same data/visualization state, or a move to a new state. Interactions play an integral role in helping analysts explore their data [YKSJ07, HS12, jJKMG07, PSCO09]. For example, Jankun-Kelly et al. argue that "... the interaction with both the data and its depiction is as important as the data and depiction itself" [jJKMG07]. Intuitively this makes sense, as (inter)actions are the building blocks to complete low-level EVA tasks [GZ08].

Predictability: EVA is also described as "unpredictable" [STH02], where it may be unclear what the user will do throughout an EVA session. Many factors may influence predictability. A critical question is whether analysts will produce similar results when performing similar EVA tasks. If analysts approach an EVA task differently, then the outcomes will be hard to predict. If analysts arrive at similar answers, with notable overlap in strategies, then there may be opportunities to predict future outcomes [DC17, BCS16]. Ziemkiewicz et al. [ZOC*12] argue that differences in users' individual experiences drive differences in analysis outcomes with visual analysis tools. It is unclear whether analysts generally utilize similar analysis sequences during EVA, or arrive at similar answers to EVA tasks and subtasks, requiring further investigation.

3.4. Exploration Performance

An ambitious goal of visual analytics is to support "fluent and flexible use of visualizations at rates resonant with the pace of human thought" [HS12]. Liu & Heer divide [LH14] this goal into two specific research questions: "... understanding the rate of cognitive activities in the context of visualization, and supporting these cognitive processes through appropriately designed and performant systems." Here we discuss themes in the literature focused on measuring, supporting and improving: 1) the exploration pace and accuracy of end users and 2) the performance of EVA systems.

Pacing and Analyst Performance: A number of methods have been developed to measure the pace of exploration. Interaction rate, or the number of interactions performed per unit time, is a common measure of exploration pacing [LH14, ZZZK18, FPH19]. Insight generation rate is also a prominent pacing metric, particularly for open-ended exploration tasks [ED16,LH14,GGZL16,ZGC*16, ZZZK18]. Feng et al. [FPH19] propose new metrics, such as exploration uniqueness, to capture more nuanced information from casual, open-ended exploration sessions on the web.

Several observations regard how users' selection of interactions can affect exploration pacing. Guo et al. [GGZL16] find that more exploration-focused interactions lead to more insights being generated. More broadly, Lam [Lam08] observes that high cognitive load can impact visual analytic performance. Extrapolating from this observation, high cognitive load interactions, such as writing a SQL query, could lead to a slower exploration pace.

Zgraggen et al. [ZZZK18] argue that not only the number of insights, but also the quality of insights are critical to gauging the effectiveness of EVA. Their study finds a 60% rate of false discoveries (i.e., insights that do not hold for the population-level data) for unconstrained, open-ended exploration by novices. They ultimately argue that EVA systems should help users formulate a reliable mental model of the data, for example more accurate insights.

Wongsuphasawat et al. [WMA*16b] evaluate the number of unique data attribute combinations explored by users, to gauge whether exploration sessions increase in breadth when users are provided with useful visualization recommendations. Though not a direct pacing metric, exploration breadth can contribute to an overall understanding of analysts' performance.

System Performance: We note a general consensus within both the database and visualization communities that response time latency is a critical performance measure for EVA systems. For example, Liu & Heer [LH14] observe that high response time latencies (500ms or longer) can impede exploration performance and progress, where analysts may be more sensitive to high latencies for some interactions (e.g., brush filters) over others (e.g., zooming). Zgraggen et al. [ZGC*16] observe similar outcomes when evaluating progressive visualizations. Idreos et al. [IPC15] survey a range of database projects focused on optimization and performance for EVA contexts, and also observe that response time latency is the primary performance measure within these projects.

To study the effects of latency, both Liu & Heer [LH14] and Zgraggen et al. [ZZZK18] inject latency into EVA systems and measure the resulting interaction rates of analysts to gauge system performance. The idea is that latency will likely slow the user's exploration progress, resulting in fewer interactions over time. Crotty et al. [CGZ*16] propose optimizations to reduce system latency for big data EVA contexts, in an effort to improve interaction rates. Rather than measuring interaction rates, one can instead measure the average or worst case latencies observed per interaction, which several database research projects utilize to evaluate optimizations for EVA systems [CXGH08, KJTN14, BCS16, CGZ*16, RAK*17].

To measure effects over an entire EVA session, alternative metrics include total exploration time (i.e., the duration of a single EVA session) [DPD14, FPH19], and total user effort (i.e., total interactions performed) [DC17,DPD14,GW09,FPH19]. These metrics are often utilized to gauge whether recommendation-focused optimizations help users to spend less time and effort exploring the data to achieve their analysis goals [DPD14].

Pacing Optimization Constraints: Multiple projects further constrain EVA system optimization by not only positing latency constraints (e.g., system response time latencies under 500ms), but also assuming a *rapid pace* of exploration, where users quickly perform successive interactions. For example, Gotz & Zhou [GZ08] argue that "During a visual analytic task, users typically perform a very large number of activities at a very fast pace," implying that users perform interactions quickly during most visual analytic tasks (including EVA). Narrowing the scope to EVA, Fisher et al. [FPDs12] argue that "In exploratory data visualization, it is common to rapidly iterate through different views and queries about a data-set." In a more recent example, Battle et al. [BCS16] deploy new optimizations to reduce response time latency for pan-and-zoom applications by prefetching data regions (i.e., data tiles) that the user may pan or zoom to next. Battle et al. argue that due to the presumably fast pace of EVA, the system "... may only have time to fetch a small number of tiles before the user's next request," motivating a need for accurate and efficient prediction and prioritization of the set of tiles to prefetch before the user's next interaction. This work seems to argue that due to the fast pace of EVA, the time between interactions (or *think time*) is restricted, limiting how we deploy sophisticated (e.g., predictive) optimizations for EVA.

3.5. Synthesized Definition of EVA

Exploratory data analysis (or EDA, originally coined by John Tukey [Tuk77]) encompasses the tasks of learning about and making sense of a new dataset. We define *exploratory visual analysis* (or EVA) as a subset of exploratory data analysis, where visualizations are the primary output medium and input interface for exploration. EVA is often viewed as a high-level analysis goal, which can range from being precise (e.g., exploring an existing hypothesis or hunch) to quite vague (e.g., wanting to find something "interesting" in the data). During EVA, the analyst updates and refines their goals through subsequent interactions with and manipulation of the new dataset. Due to the inherent complexity in accomplishing high-level exploration goals, analysts often decompose their exploration into a series of more focused visual analysis subtasks, which in turn could be partitioned further into smaller subtasks, and so on. Several visual analysis subtasks are commonly associated with EVA:

assessing the quality and semantics of the data, discovering underlying relationships and statistical distributions within the data, formulating and verifying hypotheses, and evaluating causality or more complex models on the data. These subtasks are not unique to EVA, but occur often during EVA contexts nonetheless, and range from being more focused (i.e., data quality assessment) to more open-ended (i.e., causality analysis and modeling).

We note that this definition is quite broad, revealing a lack of precision and consistency in how EVA is discussed in the literature.

3.6. Summary

We now summarize our review and posit research questions, focusing on topics that lack consensus or corroborating evidence.

Goals Summary: Discrepancies across the literature suggest that we lack a shared definition that covers the various ways EVA goals are formulated. Analysts who perform EVA may come to it with a clear intent (i.e., explicit goals and tasks), with no clear intent (i.e., no pre-conceived goals or tasks), or somewhere in between (i.e., vague goals, a few initial tasks). As such, it seems beneficial to account for both top-down (i.e., focused) and bottom-up (i.e., opportunistic) exploration strategies when evaluating EVA behavior. We utilize these insights in the design of our exploratory study.

Structure Summary: Prior work suggests that actions and tasks during EVA are not purely opportunistic: exploration behavior is not only dependent on what is observed from the data, but also on the analyst's goals and experience [PS08, RJPL16]. Analysts may incorporate concrete analysis steps or tasks that they have performed in the past. However, analysts may select among these tasks opportunistically, based on what they observe in the data. Thus exploration sessions appear to have some structure to them, though the discrepancies in the literature make it difficult to reason about what structural properties to expect for a specific exploration use case (i.e., for real log data collected from a study on EVA). Here, we focus on two specific aspects of EVA structure, organization and overlap (or predictability):

- **(S1)**: How are focused and open-ended EVA sessions organized (e.g., are they breadth- or depth-oriented)? (Section 6.2)
- **(S2)**: How predictable are participants' EVA paths, given differences in task specificity/open-endedness? (Section 6.3)

Performance Summary: We find consensus around popular performance metrics for EVA (interaction rates, response time latency) and their outcomes (latency hinders exploration). However, more recent metrics and assumptions have only been measured in a limited number of experiments: the accuracy of insights, and exploration breadth, uniqueness, and pace. We focus on two metrics in our performance analysis, accuracy and pacing:

- **(P1)**: How does the accuracy of EVA compare for focused, goal-directed EVA? (Section 5.1)
- **(P2)**: How does the pace of EVA influence available time for deploying system optimizations? (Section 5.2)

4. Exploratory Study Design

Our literature review answers some questions, but also produces new ones. To further investigate our research questions in Section 3.6, we designed an exploratory study of analysis behavior. Our goal in the study design is to capture realistic visual analysis behavior during specific analytic subtasks relevant to EVA. To ensure that participants could use a familiar tool, we selected the commercial tool Tableau for analysis. Even one commercial tool can still provide useful insights into the strategies and needs of end users [GTS10, HMSA08, BDM*18]. Therefore, we focused our efforts on this one tool, allowing us ensure it was properly instrumented, designing realistic visual analysis subtasks for analysis, and recruiting local analysts with Tableau experience to participate. We discuss the limitations of our study design in Section 4.6.

4.1. Participants

We recruited participants via university mailing lists and local Tableau User Groups (e.g., message boards, meet-ups). 27 Tableau users participated in the study (10 male, 17 female, age 23-47 years). 22 participants were recruited from our university campus, 5 from our metropolitan area. Participants had no prior experience with the study datasets, and used Tableau either for work, or through classes. Participants varied widely in Tableau and data analysis experience, from just learning Tableau (including 13 students) to seasoned veteran analysts to Tableau power users.

4.2. Protocol

Participants completed an initial survey online. Qualifying participants completed a 90-120 minute in-person session (on campus, or at their workplace), consisting of: 1) study overview and consent form; 2) 5 minute warm-up with Tableau on a movies dataset; 3) 30 minute analysis block with one dataset; 4) 30 minute block with another dataset; and 5) exit survey. Analysis blocks included a 5 minute warm-up for the given dataset. Task sheets were provided, with 4 visual analysis subtasks to complete per block (8 total). Subtask prompts were printed for participants, as well as a dataset supplement document defining the data attributes, and a map of the USA. Participants were allowed to take notes, if desired, and were compensated with a $25 Amazon gift card. A 15-inch Macbook Pro with Tableau Desktop pre-installed was provided to participants. Datasets were loaded directly into Tableau Desktop 10.3 (the latest version that supported the logging features required for analysis).

4.3. Datasets

We evaluate three real-world datasets, selected for similar complexity, sufficiently large size to simulate large-scale analysis, relevance to real-world questions (irrelevant attributes were removed). Furthermore, we selected datasets used in previous studies of visual analysis and exploration behavior: flight performance data [LH14], wildlife strikes [WMA*16b], and weather data [KH18].

1. **FAA (31 columns, 34.5M rows, 5.36GB)**[‡]: recorded flights, with itinerary (destination, distance, etc.) and performance measures (arrival delays, cancellations, etc.).
2. **Weather (35 columns, 56.2M rows, 3.53GB)**[§]: daily weather station reports, containing measures (precipitation, temperature, etc.) and observed phenomena (e.g., tornados, ground fog, etc.)

‡ https://www.transtats.bts.gov/Tables.asp?DB_ID=120

§ https://www.ncdc.noaa.gov/ghcn-daily-description

Dataset	Subtask	Prompt
Birdstrikes	T1	Consider these four parts of the aircraft: engine 1 ([Dam Eng1]), engine 2 ([Dam Eng2]), the windshield ([Dam Windshield]), and wing/rotor ([Dam Wing Rot]). Which parts of the aircraft appear to get damaged the most?
Birdstrikes	T2	Which aircraft classes ([Ac Class]), if any, appear to be more susceptible to damage ([Damage]) from animal strikes? Note that [Damage] also records when no damage has occurred.
Birdstrikes	T3	What relationships (if any) do you observe involving weather conditions ([Precip], [Sky]) and strike frequency, or counts over time ([Incident Date])?
Birdstrikes	T4	What are the most common conditions for an animal strike? Note that this is not limited to weather conditions, any dataset columns that are interesting to you can be included.
FAA	T1	How do cancelled flights ([Cancelled]), diverted flights ([Diverted]), and delayed flights ([ArrDelay], [DepDelay]) compare in terms of counts or frequency?
FAA	T2	What patterns (if any) do you observe in the count of flights over time ([FlightDate])? If any patterns are observed, what deviations (if any) do you see for individual airlines ([UniqueCarrier])?
FAA	T3	What relationships (if any) do you find involving flight distance ([Distance]) and arrival delays ([ArrDelay])?
FAA	T4	Suppose Delta Airlines wants to expand 3 airports. Based on your analysis of the data, which 3 airports would you recommend to Delta Airlines (airport code DL)? Existing Delta Airlines airports, and/or airports that Delta doesn't cover, can be included in your analysis.
Weather	T1	Consider the following weather measurements: Heavy Fog ([Heavy Fog]), Mist ([Mist]), Drizzle ([Drizzle]), and Ground Fog ([Ground Fog]). Which measurements have more data? Which weather measures (if any) would you remove from the dataset?
Weather	T2	How have maximum temperatures ([T Max]) and minimum temperatures ([T Min]) changed over the duration of the dataset (i.e., over the [Date] column)?
Weather	T3	How do the wind ([High Winds]) measurements compare for the northeast and southwest regions of the US?
Weather	T4	What weather predictions would you make for February 14th 2018 in Seattle, and why?

Table 1: *Visual analysis subtask prompts given to participants for each dataset.*

3. **Birdstrikes (94 columns, 173K rows, 91MB)**¶: incidents of aircraft (e.g., airplanes) striking wildlife (e.g., deer, birds), with contextual details (e.g., weather conditions, total struck, etc.).

4.4. Visual Analysis Subtasks

The goal of this study is to better understand analysts' visual analytic behavior at fine granularity, with respect to EVA contexts. To support a "micro-analysis" of visual analysis behavior during EVA, we focus on specific subtasks that may occur during EVA sessions, according to our literature review in Section 3.3.

Selecting Visual Analysis Subtasks: We find in Section 3.3 that analysts utilize both a top-down and bottom-up exploration approach, which often includes performing multiple, focused subtasks. We find that a variety of subtasks are observed during EVA, including tasks that are not traditionally associated with EVA. Though the precise order of these subtasks may vary from analyst to analyst, there does appear to be a common progression through the different subtask types that maximizes the effectiveness of an EVA session. First, analysts explore to learn the data's structure and semantics, which lays the foundation for understanding more complex structures and phenomena in the data. Then analysts look for statistical patterns and relationships between different data variables. Given a basic understanding of the relationships between variables, analysts can then move to deeper exploration, tied to more complex subtasks such as causality analysis, forecasting, and decision making. We treat these visual analysis subtasks as separate categories, and design study tasks for each: 1) data quality assessment (T1); 2) evaluation of patterns and relationships between variables (T2, T3); and 3) causality and prediction analysis (T4, open-ended). All subtasks are listed in Table 1.

Ordering Subtasks: Given that learning effects are a natural part of the EVA process — and our aim to study realistic EVA behavior — we decided to have all participants complete subtasks in the same order (T1, T2, T3, T4), dictated by the natural order of their respective categories. For example, data quality assessment (T1) is often performed before causality analysis (T4) [KPHH12].

Managing Complexity: To ensure adequate breadth for analysis, multiple factors of complexity were balanced across subtasks: 1) total attributes analyzed per dataset, 2) diversity of attributes across subtasks (e.g., temporal versus spatial, dimensions versus measures), 3) relevance of subtasks to real-world equivalents. The complexity of all subtasks within a category (T1, T2, etc.) were balanced across the three datasets.

Subtask Durations: Zgraggen et al. asked participants to perform open-ended EVA tasks for 15 minutes [ZZZK18], where participants wrote down every insight they found. Given our more focused subtasks and less rigid recording, we chose shorter durations: 7 for the (simpler) subtasks T1-T3, and 9 minutes for the more open-ended subtask T4. We verified through pilots that these subtasks could easily be completed within these time frames before conducting the study.

¶ https://wildlife.faa.gov/

4.5. Data Collection & Processing

Data Sources: We collected raw logs from Tableau Desktop and Tableau Data Engine (or TDE, an internal database system). Screen capture was also recorded to contextualize our log analysis. Data from one participant was removed due to experiment error.

Interactions: All events were extracted from the logs and compared with the screen capture to identify interactions (e.g., add attributes to shelf, change mark type, etc.). These events were then labeled with established interaction types [HMSA08].

Shelf State and Visual Encodings: Tableau automatically records shelf state (Row, Column, Filter, and Encoding shelves). These entries were extracted and then stored with their corresponding interactions in a larger master log (matched via timestamps).

Data Transformations and Queries: Data transformations were extracted from Tableau shelf state (e.g., `sum:snow`, `count:damage`), as well as TDE logs. Tableau uses a separate query language to interact with the TDE. We developed a parser for the Tableau queries to perform the parameter extraction.

4.6. Study Design Limitations

Our study design differs significantly from previous studies of EVA behavior, which generally preserve the natural structure of entire exploration sessions, rather than capturing a series of focused subtasks. Our analysis does not capture a completely faithful representation of EVA, but instead simulates (some of) its individual parts. Understandably, this design will not allow us to make high-level inferences about EVA behavior. However, what we lose in contextual accuracy (i.e., capturing complete EVA sessions) we partially make up for in precision (i.e., knowing exactly what subtask is being performed and when). This study design helps us to better understand specific, low-level behavioral patterns, and the significance of these patterns within common visual analysis subtasks that can occur during EVA (as well as other analysis contexts).

We note here that some important aspects of EVA sessions are still preserved in our design, such as selecting subtasks that occur during EVA, and imposing a subtask ordering that aims to maximize the effectiveness of the selected subtasks for EVA contexts. For example, an analyst will not be able to make sense of relationships between data attributes (subtasks T2 and T3) without first verifying their understanding of the individual attributes themselves (subtask T1). Similarly, an analyst will not be able to effectively make predictions or assess causality amongst data attributes (subtask T4) without first identifying and understanding the underlying relationships between them (subtasks T2 and T3).

5. Task Performance

We report on the accuracy and pacing of study participants, and compare our results to assumptions in the literature.

5.1. Evaluating Task Completion Rates & Accuracy (P1)

Most participants successfully completed all tasks: Figure 1 shows the fraction of successful completions for tasks T1-T3 (tasks with measurable correctness). Incorrect answers stemmed from improper analysis, such as failing to consider the low number of long-distance flights for task FAA T3, or including strike records that do

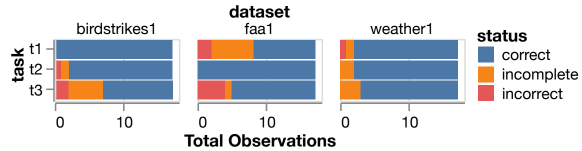


Figure 1: *Task performance by dataset, labeled as: correct answer, incorrect answer, or did not finish (incomplete).*

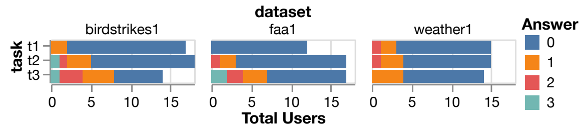


Figure 2: *Total answers observed for each dataset and task, and the total participants that arrived at each answer.*

not actually result in damage for Birdstrikes T2. A small number of participants had outwardly correct answers (i.e., similar in text as other participants'), but the observed analysis involved glaring errors; we conservatively labeled these cases as incorrect. Incomplete answers are cases where the participant ran out of time.

Most participants completed tasks on time. When a task was not completed, the cause often involved attempts to use complicated interactions, such as sophisticated filters (e.g., FAA task T1) or complicated grouping and binning operations (e.g., Weather task T3).

Participants provided correct answers: Overall, participants provided accurate responses (Figure 1): 80.6% (116/144) of task sessions had correct answers. Rates per task were 81.3% (39/48) for T1, 91.7% (44/48) for T2, 68.8% (33/48) for T3. For each dataset, over 50% of participants successfully completed all three tasks.

Participants were cautious analysts: We designed two tasks that asked participants to assess relationships between variables for which there was *no clear correlation* (Birdstrikes T3 and FAA T3). We observe the worst case error rate of 25% in FAA T3: 4 incorrect answers out of 16. In all other tasks, ~81% of participants who complete the task provide correct answers. Though not an exact comparison, Zgraggen et al. [ZZZK18] observed a 60% false discovery rate when asking participants to explore (i.e., search for) interesting relationships in the data. We believe the higher accuracies in our study stem from our focus on realistic task outcomes from experienced users, rather than all verbalized open-ended observations by novices.

Participants regularly compared visualizations with the raw data, to ensure that the visualizations matched their expectations. Regardless of correctness, participants qualified their answers with comments on their confidence in the rendered results. For task FAA T3, five participants made comments suggesting a distrust of looking only at their visualizations, including one participant who had given an incorrect answer. Two participants said that their hypotheses required further study, two participants commented on the disparity in available data between short and long distance flights, and one participant said the delays for long distance flights did not look "quite right." One participant described how "Just by looking, one might be tempted to say higher distance lower delay, but I wouldn't say that because there's more data for shorter delays," showing the cautious evaluation style adopted by many participants.

Most participants arrive at the same answers: Figure 2 shows

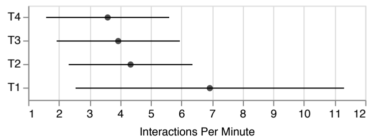


Figure 3: *Mean interaction rates and 95% CIs of the mean.*

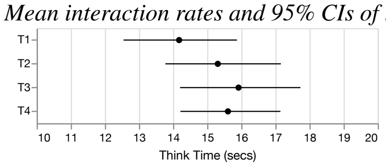


Figure 4: *Mean think times, with 95% CIs of the mean.*

the total unique answers observed for tasks T1-T3. We observed 1-4 answers per task. Most participants arrive at the same answer. For example, in the Birdstrikes T1 task, 15 participants conclude that the wing or rotor of aircraft are damaged most often during strikes. In the FAA T3 task, 10 participants conclude that there is no clear relationship between distance and arrival delays.

5.2. Evaluating Pacing: Interaction Rate & Think Time (P2)

We observe similar interaction rates across tasks: To assess the effects of task on interaction rate, we fit linear mixed effects models, with task as a fixed effect, and participant and dataset as random effects. We test significance by comparing the full model to a "null" model with the target fixed effect removed (task), using likelihood-ratio tests. Figure 3 plots expected mean interaction rates and 95% confidence intervals. We do not find a significant effect of task on interaction rates ($\chi^2(3, N = 192) = 2.194, p = 0.533$). Thus, interaction rates appear to be independent of task type.

Think times are long: An ambitious goal for EVA systems is to support interaction at the "speed of human thought" [HS12, CGZ*16]. We seek to better understand the pacing of visual analysis tasks that occur during EVA, where several projects seem to argue that EVA is generally fast-paced [GZ09, BCS16, FPDs12]. Many systems, particularly from the database community, propose specialized optimizations to improve performance under constrained pacing conditions (i.e., short time between interactions, or *think time* [BCS16]). However, the "speed of thought" might involve spans of viewing data, drawing comparisons, and planning next steps. If ample time is available between interactions for data pre-processing, complicated (e.g., predictive) optimizations may not be needed: we can use these think times to deploy simpler optimizations and achieve similar results.

Rendering times and query times in Tableau represent a small fraction of the time between interactions (1% and 14%, respectively). Thus the majority (85%) of this (think) time can be attributed to users (e.g., interpreting visualizations, selecting an interaction, etc.). We operationalize think times as the time from the end of the current interaction to the start of the next one, subtracting query and rendering times. Figure 4 shows mean think time across tasks. The means are notably high, ranging from 14 to 15 seconds, depending on the task. Median think times are 5-7 seconds, indicating skew. Depending on task, we find that 53-61% of think times are 5 seconds or longer, and 32%-41% are 10 seconds or longer.

Next we consider the mean think times preceding each interaction type (Figure 5a), which generally fall in the 10-20 second range. Median think times range from 4.7 (`undo`) to 29 seconds (`data-derive`). From our study observations, we find that `data-derive` is of high cognitive complexity, because it involves reasoning about and writing formula expressions. The gap between `data-derive` and other interactions shows that differences in cognitive complexity can directly impact the pacing of EVA. Lam also finds that high cognitive complexity leads to high interaction costs [Lam08], but refers specifically to selecting from a large space of possible interactions. We find that particular interactions with high complexity also lead to high interaction costs.

Consider the ForeCache system [BCS16] discussed in Section 3.4, which predicts the user's next interaction and fetches the corresponding data. Fetching data in ForeCache takes about 1 second, and users can perform 9 interactions (4 panning and 5 zooming directions). Suppose we observe similar interaction rates as Tableau (median navigation think time is 9 seconds). Without special optimizations, we could simply pre-fetch the data for all 9 potential interactions, one at a time, and with high probability fetch the required data before the user's next interaction. Though hypothetical, this example shows that without the full context of latency and pacing, we may devise unrealistic performance constraints for EVA.

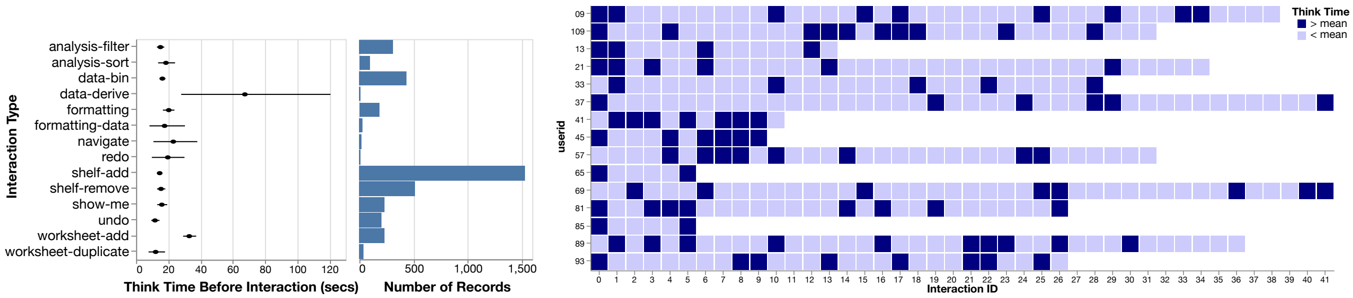
Interactions appear "bursty": We find that participants oscillate between spending relatively more time choosing an interaction, then less time on a subsequent sequence of interactions. Figure 5b shows the observed think times for the FAA T4 task, with relatively short think times in light blue, and longer think times in dark blue (i.e., below or above the mean think time for FAA T4, respectively). Shorter think times appear to be clustered together, which we observed across tasks, with participants performing on average 3-4 fast interactions per sequence. These results may suggest that analysts formulate a plan for the next few interactions (i.e., a subtask), and then execute their plans, before deciding what to analyze next.

6. The Structure of Analysis Sessions

We now analyze the structure of participants' analyses and compare with existing structural assumptions.

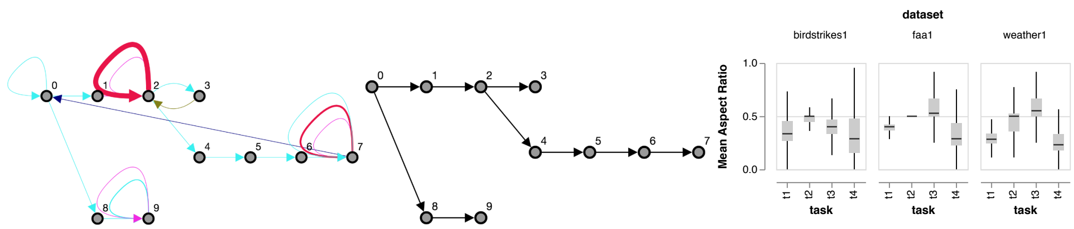
6.1. Defining Analysis States & Search Trees

As defined by Wongsuphasawat et al., an analysis state is the *set of attributes currently being analyzed* [WMA*16b, WQM*17], for which a user may specify visual encodings, apply filters, or group and aggregate the data. Interactions with Tableau can add to, remove from, or otherwise modify the current attribute set, producing new analysis states. We first construct a raw graph of all states and interactions Each time a participant adds a new attribute to the current visualization, a *forward edge* is included in the raw graph from the current attribute set to the new (larger) attribute set. When attributes are removed, a *backward edge* is included in the graph from the current attribute set to the new (smaller) set. We then remove the backward edges and self-loops to form *search trees*. Figure 6a shows the raw graph for one participant from task Weather T4. We create a new state for each observed attribute set, and a directed edge between corresponding states for each interaction. Edge color encodes interaction type, and width repeated interactions. Figure 6b shows an example search tree.



(a) (b)

Figure 5: *(a) Means and 95% confidence intervals for think time before each interaction (left), and total records per interaction type (right). (b) Observed think times for FAA T4, colored light blue when below mean think time, dark blue when above the mean.*



(a) *Graph Weather T4.* **(b)** *Tree Weather T4.* **(c)** *Box Plot for Tree Mean Aspect Ratios.*

Figure 6: *Raw graph (a) and matching search tree (b) for one participant, task Weather T4. (c) Distribution of tree aspect ratios.*

6.2. Breadth- Versus Depth-Oriented Analysis (S1)

Wongsuphasawat et al. define depth and breadth in terms of the total unique states analyzed [WMA*16b, WQM*17] (also considered by Sarvghad et al. [STM17]). However the interactions that produce these states, and thus the states themselves, are *not* independent. As we find in Section 3.3, interactions are part of a larger, hierarchical process, where they may be clustered together within a subtask (i.e., an analysis branch within a search tree). We extend these definitions of depth and breadth to consider the full EVA structure (i.e., search trees): branches in the tree correspond to breadth in subtasks or analysis trajectories, tree depth to effort or emphasis (i.e., total interactions) for a specific (sub)task. Greater breadth indicates that a participant more frequently backtracked to previous states and then branched off, greater depth indicates that a participant engaged in less backtracking.

Sessions are primarily depth-oriented: Branching to different sub-analyses (i.e., subtasks) was fairly common. Branches translate to leaves in the search trees (e.g., nodes 3, 7, and 9 in Figure 6b). 51.0% (98/192) of trees have multiple branches (i.e., >1 leaf), and 25/26 participants had multiple branches in at least one task. Figure 6b shows the two forms of branching we observed in Tableau: 1) within-sheet branching, and 2) between-sheet branching (e.g., creating sheets, returning to previous sheets). With the first form of branching, the user makes small pivots from the current analysis state (e.g., removing one of the current data columns being analyzed). To assess whether search trees tend to be depth- or breadth-oriented, we calculate an aspect ratio: we divide a tree's width (max breadth) by its height (max depth). Figure 6c shows the distribution of aspect ratios, which consistently have aspect ratios below one, showing *greater depth than breadth*. These findings are consistent with Wongsuphasawat et al.'s claims that analyses in Tableau are primarily depth-oriented [WMA*16b, WQM*17].

6.3. Predictability & Overlap (S2)

We assess the predictability of EVA by measuring structural similarity between sessions, which could indicate likely outcomes for a given task, or relevant interaction patterns, such as iteration. We compute similarity in two ways: 1) overlap in states visited across participants and 2) revisitation of states (i.e., self-loops, iteration).

Certain states show high overlap: Figure 7a shows a binned chart of overlaps, where the bins along the x-axis represent the total participants that visited each state in the given bin (i.e., blue circle). Circle area encodes the total unique states that had exactly x visitors, for each task (y-axis) and dataset (columns). For example, relatively large circles in the far left column (per datasets) means that most states were visited by only 1 participant. For task T4 we observe a steep drop-off, with only a few states being seen by most users. Tasks T1 and T3 exhibit more multi-modal distributions, with one cluster of states that only a few people visit, and another where many people see the respective states.

To further quantify overlap, we calculate a binary histogram of visited states per participant, where each bin represents a unique state, and the bin is set to 1 if this state was visited. We compare pairs of participants using a modified version of Jaccard similarity:

$$\frac{|A \cap B|}{min(|A|,|B|)}.$$

This measure avoids penalizing pairings where one participant visited many more states. The average similarity is as follows: T1 (0.56), T2 (0.61), T3 (0.57), T4 (0.22). We find high average overlap in T1-T3, suggesting somewhat predictable analysis steps or outcomes. Average overlap is lower for T4. As we discuss in greater detail later, we believe this is related to participants taking a high-level goal (T4) and decomposing it into focused subtasks.

Overlapping states align closely with task goals: We compute the top 3 overlapping nodes for each dataset and task, and compare them to participants' task goals (as indicated by task prompts). We observe that these states subsume the columns needed to complete the task for 8 of 12 tasks. Thus participants tend to overlap at the key states needed to complete a task. These particular nodes often appear later within the search trees (i.e., close to or as leaf nodes). For example, the top 3 overlapping states for each task appear as leaf nodes up to 30% of the time, depending on the task.

Self-loops signal key analysis states: Participants frequently iterate on the same analysis state, forming self-loops in the resulting graphs. 39.1% of all states have self-loops. We also find at least 1 self-loop in 95.8% (184/192) of analysis graphs. Most self-loops involve data transformations (42.2% of all self loops), followed by shelf (29.2%) and formatting (10.7%) interactions. Most shelf interactions involve moving the current attributes to different shelves. Self-loops highlight states where participants place more analysis effort, showing that this state may be a useful landmark within an EVA session. We computed the fraction of observed self-loops per state, participant, dataset, and task. We then summed the results per state, across all participants. In 11 of 12 tasks, at least one of the top 3 self loop-states also appears in the top 3 overlapping states. Thus self-loops serve as indicators of significant analysis states.

6.4. Open-Ended Versus Focused Analysis (S1, S2)

Specificity does not affect relative breadth/depth: We evaluate the effects of task on tree dimensions by fitting linear mixed-effects models on the *normalized* max depth and breadth of the trees (i.e., divided by tree size). We include task as a fixed effect, and participants and dataset as random effects. We compare with a "null" model (i.e., with task removed) using likelihood-ratio tests, and do not find significant effects of task on normalized max depth ($\chi^2(3, N=191) = 4.878, p = 0.181$) or breadth ($\chi^2(3, N=191) = 2.495, p = 0.476$). Figure 7 shows means and 95% confidence intervals. Our results indicate that open-ended and focused sessions are *structurally similar* in depth and breadth.

Open-ended tasks may decompose into focused tasks: On the surface, open-ended structures appear to diverge. For example, T4 has lower average modified Jaccard similarity (see Section 6.3). However, this analysis misses the hierarchical nature of more open-ended tasks, and of exploration in general, discussed in Section 3.2: participants likely decompose T4 into focused subtasks to make it more manageable. Individual subtasks could be characterized and compared to obtain a more accurate similarity measure, which we plan to investigate in the future.

7. Findings & Future Work

In this paper, we explore existing definitions of EVA, and identify three major themes in the literature: goal formulation, exploration structure, and performance. Within these themes, we highlight points of connection and contradiction, from which we formulate research questions for further study: organization (S1), predictability (S2), accuracy (P1), and pacing (P2) in EVA. We present a study in which 27 Tableau users completed a range of analysis tasks across 3 datasets, informed by our literature review. We use the resulting data to provide empirical context towards answering these research questions; our results provide useful insights that may help to inform future analyses and system evaluations.

We find that many *implicit assumptions* are made about EVA across the literature. EVA is often referenced, but rarely defined (with notable exceptions, e.g., [AZL*19, jJKMG07]). Lam et al. observe a similar dearth of explicit information on EVA [LTM18]. We hope that this work will help to motivate our community to rethink the way we perceive, design, and evaluate in EVA contexts.

7.1. Analysis Performance

Most participants successfully completed the tasks, with error rates of at most 25% per task. Our results differ from prior work [ZZZK18], which report error (false discovery) rates over 60% for participants assessing data properties and relationships. Though not a perfect comparison, these differences could be attributed to several factors. We provided specific tasks rather than encouraging more vague open-ended explorations; that structure may have helped to focus participants. Unlike earlier insight-based evaluations (e.g., [GGZL16, ZZZK18, LH14]), we did not require participants to vocalize all "insights" as they went along, regardless of overall relevance. Again, our participants were typically focused on a specific goal. We plan to evaluate how the articulation of specific questions and goals — structuring open-ended exploration into more explicit, albeit evolving, tasks — may affect analysis outcomes in other contexts. While both our study and that of Zgraggen et al. involved analyzing unfamiliar data, our study involved users with prior analysis experience in real-world environments, rather than a novice, student-only population. This difference in experience may have helped our participants exercise caution.

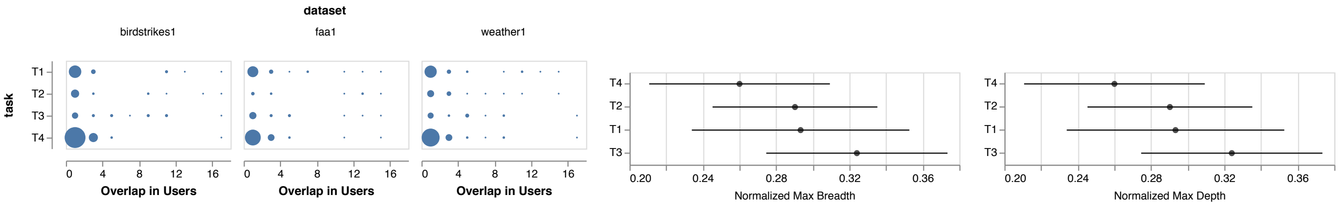
Insight (P1)*: Analysts tend to approach analysis tasks with care, as well as a priori goals and hypotheses, which may lead to better outcomes for analysis subtasks completed during EVA contexts.*

Latency is a key performance metric for evaluating EVA systems. However, the pace of EVA may be characterized unrealistically in parts of the literature [GZ09, BCS16, FPDs12], to further constrain (and sometimes complicate) system optimization contexts. We observe 14-15 seconds on average of idle time or *think time* between interactions, resulting in a slower pace for EVA, and providing ample time for simpler optimization methods to be deployed with presumably similar performance results. We measured think times for different interaction types and found similar results, with one exception: *data-derive* interactions take longer to reason about, resulting in longer think times. We also observe "bursty" interaction patterns, where analysts repeatedly have a longer think time (presumably planning next steps), followed by a sequence of shorter think times (i.e., plan execution). With a better understanding of pacing, we can construct more realistic evaluations, and test large-scale EVA systems with appropriate parameters.

Insight (P2)*: EVA system optimizations should consider not only latency, but also pacing and available resources, like think time.*

7.2. Analysis Goals and Structure

Participants' analyses were primarily depth-oriented, validating previous calls to prompt greater exploration breadth [WMA*16b, WQM*17]. We found strong similarities in interaction rates, task



(a) *Overlap in states visited.* **(b)** *Means and CIs for max tree breadth.* **(c)** *Means and CIs for max tree depth.*

Figure 7: *(a) Counts of users that overlap per state. Means and 95% confidence intervals for normalized max (b) breadth and (c) depth.*

answers, and relative breadth and depth across tasks, suggesting analysts use similar analysis strategies, regardless of task.

Insight (S1)*: Analysis search trees in Tableau are primarily depth-oriented, and consistent across subtasks of varying specificity.*

Individual differences between analysts could result in differences in EVA strategies and performance [ZOC*12], affecting the predictability of EVA behaviors. Our results hint towards a complex picture of predictability in analysis tasks in general. Strong overlap and predictable patterns do emerge, but in specific situations, often when the analysis becomes more focused, such as during more focused subtasks, towards the end of a subtask, or during iteration (i.e., self-loops in analysis graphs). Significant overlap in states visited across participants usually represented critical points in participants' analyses, where they were close to achieving task goals. Consecutive interactions within a state (i.e., self-loops) also correlate with critical analysis points. Less overlap was observed in more open-ended tasks, however our selected metrics may obscure participants' decomposition of open-ended subtask prompts into even smaller subtasks, requiring further study. These results could inform the design of EVA behavior models learned from log data, and help to generalize existing modeling techniques (e.g., [BCS16, BOZ*14b, DC17]) for EVA contexts.

Our results differ from past studies of broader EDA contexts [SU15]. However a variety of techniques beyond the scope of EVA are utilized, and teams, not individuals, perform the analysis.

Insight (S2)*: Predictable patterns do occur during analysis subtasks associated with EVA, but at specific points within analysis sessions in Tableau.*

7.3. Study Design and Analysis

Our study design enables evaluation of visual analysis behavior along multiple axes: task specificity, performance, and structure, as well as others like interaction and encoding types. We contribute our data as a community resource for further study at `https://github.com/leibatt/characterizing-eva-tableau`.

Significant manual effort was required to analyze native Tableau logs. A standardized process for curating system logs would greatly simplify the evaluation process, enable a variety of evaluations (e.g., meta-analyses, performance benchmarks), and improve reproducibility/comparability of results [BCHS17, BAB*18, E*16].

Though our study incorporates core (but decomposed) characteristics of EVA contexts and tasks, analysts still behave differently when in their own work environments (e.g., software tools, datasets, etc.), making it challenging to capture authentic EVA behavior in a controlled setting. It would be beneficial to repeat this study to evaluate how assumptions and insights change under different EVA contexts.

Our analysis does not examine differences between the analysis patterns of novice and expert analysts, a potentially interesting topic for future study. Furthermore, our focus on a single tool makes it difficult to distinguish between general and tool-specific analysis patterns. We hope to extend our study to other tools to better understand the influence of tool design on analysis behavior.

8. Conclusion

Exploratory visual analysis (EVA) is often considered a critical use case for visual analysis tools, however our understanding of EVA is arguably vague and incomplete. We sought to provide a more holistic view of how EVA is discussed across the literature, summarize and define EVA based on our observations, and to provide additional context for how analysts behave when performing EVA (sub-)tasks identified from the literature. We contribute a definition of EVA synthesized from a literature review of 42 research articles, as well as an empirical evaluation of several assumptions about how EVA is performed. We present the results of a user study with 27 Tableau users. Through a quantitative analysis of Tableau log data from the study, we evaluate multiple facets of task performance and analysis structure. We find that participants achieve over 80% accuracy across focused tasks with measurably correct answers. We find that the pacing of participants' analyses was surprisingly slow, compared to common performance assumptions observed in the literature. We find clear patterns and overlap across participants' analysis sessions, suggesting that some predictable behaviors do occur during tasks commonly associated with EVA. Furthermore, we find few differences between how more focused and more open-ended analysis tasks are structured. These findings suggest that analysts can be steady, cautious explorers, and that EVA may often contain familiar patterns and structures, helping us to build a more comprehensive view of visual analysis in the context of exploration. In the future, we aim to extend our analysis to better understand how differences in tool design (i.e., beyond Tableau) and analysts' experience may affect analysis performance, patterns, and structure.

Acknowledgements

We thank our participants for being part of our study. We thank the reviewers, the UW Interactive Data Lab, and the UMD Human-Computer Interaction Lab for their helpful comments on this paper. We thank Justin Talbot, Melanie Tory, and Richard Wesley for providing helpful feedback in parsing and interpreting Tableau log data. This research was funded by the Moore Foundation Data-Driven Discovery Investigator program.

References

[AES05] AMAR R., EAGAN J., STASKO J.: Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (Oct. 2005), pp. 111–117. doi:10.1109/INFVIS.2005.1532136. 4

[AZL*19] ALSPAUGH S., ZOKAEI N., LIU A., JIN C., HEARST M. A.: Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan 2019), 22–31. doi:10.1109/TVCG.2018.2865040. 1, 3, 4, 11

[BAB*18] BATTLE L., ANGELINI M., BINNIG C., CATARCI T., EICHMANN P., FEKETE J.-D., SANTUCCI G., SEDLMAIR M., WILLETT W.: Evaluating visual data analysis systems: A discussion report. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (New York, NY, USA, 2018), HILDA'18, ACM, pp. 4:1–4:6. URL: http://doi.acm.org/10.1145/3209900.3209901, doi:10.1145/3209900.3209901. 12

[BCC*05] BAVOIL L., CALLAHAN S. P., CROSSNO P. J., FREIRE J., SCHEIDEGGER C. E., SILVA C. T., VO H. T.: VisTrails: enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.* (Oct. 2005), pp. 135–142. doi:10.1109/VISUAL.2005.1532788. 2

[BCHS17] BATTLE L., CHANG R., HEER J., STONEBRAKER M.: Position statement: The case for a visualization performance benchmark. In *2017 IEEE Workshop on Data Systems for Interactive Analysis (DSIA)* (Oct 2017), pp. 1–5. doi:10.1109/DSIA.2017.8339089. 12

[BCS16] BATTLE L., CHANG R., STONEBRAKER M.: Dynamic Prefetching of Data Tiles for Interactive Visualization. In *Proceedings of the 2016 International Conference on Management of Data* (New York, NY, USA, 2016), SIGMOD '16, ACM, pp. 1363–1375. URL: http://doi.acm.org/10.1145/2882903.2882919, doi:10.1145/2882903.2882919. 2, 3, 4, 5, 9, 11, 12

[BDM*18] BATTLE L., DUAN P., MIRANDA Z., MUKUSHEVA D., CHANG R., STONEBRAKER M.: Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 594:1–594:8. URL: http://doi.acm.org/10.1145/3173574.3174168, doi:10.1145/3173574.3174168. 6

[BJK*16] BLASCHECK T., JOHN M., KURZHALS K., KOCH S., ERTL T.: VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 61–70. doi:10.1109/TVCG.2015.2467871. 2

[BM13] BREHMER M., MUNZNER T.: A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2376–2385. doi:10.1109/TVCG.2013.124. 1, 2

[BOZ*14a] BROWN E. T., OTTLEY A., ZHAO H., LIN Q., SOUVENIR R., ENDERT A., CHANG R.: Finding Waldo: Learning about Users from their Interactions. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec. 2014), 1663–1672. doi:10.1109/TVCG.2014.2346575. 2

[BOZ*14b] BROWN E. T., OTTLEY A., ZHAO H., LIN Q., SOUVENIR R., ENDERT A., CHANG R.: Finding Waldo: Learning about Users from their Interactions. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1663–1672. doi:10.1109/TVCG.2014.2346575. 12

[CFS*06] CALLAHAN S. P., FREIRE J., SANTOS E., SCHEIDEGGER C. E., SILVA C. T., VO H. T.: Managing the Evolution of Dataflows with VisTrails. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (Apr. 2006), pp. 71–71. doi:10.1109/ICDEW.2006.75. 2

[CGZ*16] CROTTY A., GALAKATOS A., ZGRAGGEN E., BINNIG C., KRASKA T.: The case for interactive data exploration accelerators (ideas). In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (New York, NY, USA, 2016), HILDA '16, ACM, pp. 11:1–11:6. URL: http://doi.acm.org/10.1145/2939502.2939513, doi:10.1145/2939502.2939513. 2, 4, 5, 9

[CPVDW*01] CARD S. K., PIROLLI P., VAN DER WEGE M., MORRISON J. B., REEDER R. W., SCHRAEDLEY P. K., BOSHART J.: Information Scent As a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2001), CHI '01, ACM, pp. 498–505. URL: http://doi.acm.org/10.1145/365024.365331, doi:10.1145/365024.365331. 2

[CXGH08] CHAN S.-M., XIAO L., GERTH J., HANRAHAN P.: Maintaining interactivity while exploring massive time series. In *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08* (Oct. 2008), pp. 59–66. doi:10.1109/VAST.2008.4677357. 2, 5

[DC17] DABEK F., CABAN J. J.: A Grammar-based Approach for Modeling User Interactions and Generating Suggestions During the Data Exploration Process. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 41–50. doi:10.1109/TVCG.2016.2598471. 2, 4, 5, 12

[DHPP17] DEMIRALP Ç., HAAS P. J., PARTHASARATHY S., PEDAPATI T.: Foresight: Rapid data exploration through guideposts. *arXiv preprint arXiv:1709.10513* (2017). 4

[DPD14] DIMITRIADOU K., PAPAEMMANOUIL O., DIAO Y.: Explore-by-example: An automatic query steering framework for interactive data exploration. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2014), SIGMOD '14, ACM, pp. 517–528. URL: http://doi.acm.org/10.1145/2588555.2610523, doi:10.1145/2588555.2610523. 4, 5

[DR01] DERTHICK M., ROTH S. F.: Enhancing data exploration with a branching history of user operations. *Knowledge-Based Systems 14*, 1 (Mar. 2001), 65–74. URL: http://www.sciencedirect.com/science/article/pii/S0950705100001015, doi:10.1016/S0950-7051(00)00101-5. 4

[E*16] EICHMANN P., ET AL.: Towards a benchmark for interactive data exploration. *IEEE Data Eng. Bull. 39*, 4 (2016), 50–61. 12

[ED16] ELTAYEBY O., DOU W.: A Survey on Interaction Log Analysis for Evaluating Exploratory Visualizations. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (New York, NY, USA, 2016), BELIV '16, ACM, pp. 62–69. URL: http://doi.acm.org/10.1145/2993901.2993912, doi:10.1145/2993901.2993912. 1, 2, 3, 5

[FPDs12] FISHER D., POPOV I., DRUCKER S., SCHRAEFEL M.: Trust Me, I'M Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 1673–1682. URL: http://doi.acm.org/10.1145/2207676.2208294, doi:10.1145/2207676.2208294. 1, 3, 5, 9, 11

[FPH19] FENG M., PECK E., HARRISON L.: Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan 2019), 501–511. doi:10.1109/TVCG.2018.2865117. 5

[GGZL16] GUO H., GOMEZ S. R., ZIEMKIEWICZ C., LAIDLAW D. H.: A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 51–60. doi:10.1109/TVCG.2015.2467613. 2, 3, 4, 5, 11

[GL12] GOMEZ S., LAIDLAW D.: Modeling Task Performance for a Crowd of Users from Interaction Histories. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2012), CHI '12, ACM, pp. 2465–2468. URL: http://doi.acm.org/10.1145/2207676.2208412, doi:10.1145/2207676.2208412. 2

[GTS10] GRAMMEL L., TORY M., STOREY M. A.: How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov. 2010), 943–952. doi:10.1109/TVCG.2010.164. 2, 4, 6

[GW09] GOTZ D., WEN Z.: Behavior-driven Visualization Recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2009), IUI '09, ACM, pp. 315–324. URL: http://doi.acm.org/10.1145/1502650.1502695, doi:10.1145/1502650.1502695. 2, 3, 4, 5

[GZ08] GOTZ D., ZHOU M. X.: Characterizing users' visual analytic activity for insight provenance. In *2008 IEEE Symposium on Visual Analytics Science and Technology* (Oct. 2008), pp. 123–130. doi:10.1109/VAST.2008.4677365. 3, 4, 5

[GZ09] GOTZ D., ZHOU M. X.: Characterizing Users' Visual Analytic Activity for Insight Provenance. *Information Visualization 8*, 1 (Jan. 2009), 42–55. URL: https://doi.org/10.1057/ivs.2008.31, doi:10.1057/ivs.2008.31. 2, 9, 11

[HMSA08] HEER J., MACKINLAY J., STOLTE C., AGRAWALA M.: Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov. 2008), 1189–1196. doi:10.1109/TVCG.2008.137. 1, 2, 3, 4, 6, 8

[HS12] HEER J., SHNEIDERMAN B.: Interactive dynamics for visual analysis. *Commun. ACM 55*, 4 (Apr. 2012), 45–54. URL: http://doi.acm.org/10.1145/2133806.2133821, doi:10.1145/2133806.2133821. 1, 2, 3, 4, 5, 9

[IPC15] IDREOS S., PAPAEMMANOUIL O., CHAUDHURI S.: Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2015), SIGMOD '15, ACM, pp. 277–281. URL: http://doi.acm.org/10.1145/2723372.2731084, doi:10.1145/2723372.2731084. 3, 4, 5

[ITC08] ISENBERG P., TANG A., CARPENDALE S.: An Exploratory Study of Visual Information Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1217–1226. URL: http://doi.acm.org/10.1145/1357054.1357245, doi:10.1145/1357054.1357245. 2

[jJKMG07] J. JANKUN-KELLY T., MA K., GERTZ M.: A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics 13*, 2 (March 2007), 357–369. doi:10.1109/TVCG.2007.28. 2, 3, 4, 11

[KCZ14] KALININ A., CETINTEMEL U., ZDONIK S.: Interactive data exploration using semantic windows. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2014), SIGMOD '14, ACM, pp. 505–516. URL: http://doi.acm.org/10.1145/2588555.2593666, doi:10.1145/2588555.2593666. 3, 4

[Kei01] KEIM D. A.: Visual exploration of large data sets. *Commun. ACM 44*, 8 (Aug. 2001), 38–44. URL: http://doi.acm.org/10.1145/381641.381656, doi:10.1145/381641.381656. 3, 4

[KH18] KIM Y., HEER J.: Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 157–167. 6

[KJTN14] KAMAT N., JAYACHANDRAN P., TUNGA K., NANDI A.: Distributed and interactive cube exploration. In *2014 IEEE 30th International Conference on Data Engineering (ICDE)* (Mar. 2014), pp. 472–483. doi:10.1109/ICDE.2014.6816674. 4, 5

[KPHH12] KANDEL S., PAEPCKE A., HELLERSTEIN J. M., HEER J.: Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics 18* (12 2012), 2917–2926. URL: doi.ieeecomputersociety.org/10.1109/TVCG.2012.219, doi:10.1109/TVCG.2012.219. 3, 4, 7

[KS12] KANG Y., STASKO J.: Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts. *IEEE Transactions on Visualization and Computer Graphics 18* (12 2012), 2869–2878. URL: doi.ieeecomputersociety.org/10.1109/TVCG.2012.224, doi:10.1109/TVCG.2012.224. 3, 4

[Lam08] LAM H.: A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov 2008), 1149–1156. doi:10.1109/TVCG.2008.109. 3, 5, 9

[LH14] LIU Z., HEER J.: The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 2122–2131. 2, 3, 4, 5, 6, 11

[LKS13] LINS L., KLOSOWSKI J. T., SCHEIDEGGER C.: Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2456–2465. doi:10.1109/TVCG.2013.179. 4

[LTM18] LAM H., TORY M., MUNZNER T.: Bridging from Goals to Tasks with Design Study Analysis Reports. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 435–445. doi:10.1109/TVCG.2017.2744319. 1, 2, 3, 4, 11

[LWD*17] LIU Z., WANG Y., DONTCHEVA M., HOFFMAN M., WALKER S., WILSON A.: Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 321–330. doi:10.1109/TVCG.2016.2598797. 2

[LWPL11] LU J., WEN Z., PAN S., LAI J.: Analytic Trails: Supporting Provenance, Collaboration, and Reuse for Visual Data Analysis by Business Users. In *Human-Computer Interaction âĂŞ INTERACT 2011* (Sept. 2011), Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 256–273. URL: https://link.springer.com/chapter/10.1007/978-3-642-23768-3_22, doi:10.1007/978-3-642-23768-3_22. 2

[MWN*19] MORITZ D., WANG C., NELSON G. L., LIN H., SMITH A. M., HOWE B., HEER J.: Formalizing visualization design knowledge as constraints: actionable and extensible models in draco. *To appear, IEEE Transactions on Visualization and Computer Graphics* (2019). 3, 4

[New72] NEWELL A.: *Human problem solving*. Prentice-Hall Englewood Cliffs, NJ, 1972. 2

[PC05] PIROLLI P., CARD S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (2005), vol. 5, McLean, VA, USA, pp. 2–4. 3, 4

[PS08] PERER A., SHNEIDERMAN B.: Systematic yet flexible discovery: Guiding domain experts through exploratory data analysis. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2008), IUI '08, ACM, pp. 109–118. URL: http://doi.acm.org/10.1145/1378773.1378788, doi:10.1145/1378773.1378788. 3, 4, 6

[PSCO09] PIKE W. A., STASKO J., CHANG R., O'CONNELL T. A.: The science of interaction. *Information Visualization 8*, 4 (2009), 263–274. 4

[RAK*17] RAHMAN S., ALIAKBARPOUR M., KONG H. K., BLAIS E., KARAHALIOS K., PARAMESWARAN A., RUBINFIELD R.: I've seen "enough": Incrementally improving visualizations to support rapid decision making. *Proc. VLDB Endow. 10*, 11 (Aug. 2017), 1262–1273. URL: https://doi.org/10.14778/3137628.3137637, doi:10.14778/3137628.3137637. 2, 4, 5

[RESC16] RAGAN E. D., ENDERT A., SANYAL J., CHEN J.: Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 31–40. doi:10.1109/TVCG.2015.2467551. 1

[RJPL16] REDA K., JOHNSON A. E., PAPKA M. E., LEIGH J.: Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization 15*, 4 (Oct. 2016), 325–339. URL: https://doi.org/10.1177/1473871616638546, doi:10.1177/1473871616638546. 2, 3, 4, 6

[SASF11] SILVA C. T., ANDERSON E., SANTOS E., FREIRE J.: Using VisTrails and Provenance for Teaching Scientific Visualization. *Computer Graphics Forum 30*, 1 (Mar. 2011), 75–84. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8659.2010.01830.x/abstract, doi:10.1111/j.1467-8659.2010.01830.x. 2

[Shn96] SHNEIDERMAN B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages* (Washington, DC, USA, 1996), VL '96, IEEE Computer Society, pp. 336–. URL: http://dl.acm.org/citation.cfm?id=832277.834354. 3, 4

[SKL*16] SIDDIQUI T., KIM A., LEE J., KARAHALIOS K., PARAMESWARAN A.: Effortless data exploration with zenvisage: An expressive and interactive visual analytics system. *Proc. VLDB Endow. 10*, 4 (Nov. 2016), 457–468. URL: https://doi.org/10.14778/3025111.3025126, doi:10.14778/3025111.3025126. 1, 3, 4

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (July 2005), 443–456. 3

[SNLD06] SARAIYA P., NORTH C., LAM V., DUCA K. A.: An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (Nov 2006), 1511–1522. 3

[ST15] SARVGHAD A., TORY M.: Exploiting Analysis History to Support Collaborative Data Analysis. In *Proceedings of the 41st Graphics Interface Conference* (Toronto, Ont., Canada, Canada, 2015), GI '15, Canadian Information Processing Society, pp. 123–130. URL: http://dl.acm.org/citation.cfm?id=2788890.2788913. 2

[STH02] STOLTE C., TANG D., HANRAHAN P.: Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (Jan 2002), 52–65. doi:10.1109/2945.981851. 2, 4

[STM17] SARVGHAD A., TORY M., MAHYAR N.: Visualizing dimension coverage to support exploratory analysis. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 21–30. 2, 10

[SU15] SILBERZAHN R., UHLMANN E. L.: Crowdsourced research: Many hands make tight work. *Nature News 526*, 7572 (2015), 189. 12

[SvW08] SHRINIVASAN Y. B., VAN WIJK J. J.: Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1237–1246. URL: http://doi.acm.org/10.1145/1357054.1357247, doi:10.1145/1357054.1357247. 2

[Tuk77] TUKEY J. W.: *Exploratory data analysis*, vol. 2. Reading, Mass., 1977. 3, 4, 5

[VRM*15] VARTAK M., RAHMAN S., MADDEN S., PARAMESWARAN A., POLYZOTIS N.: SeeDB: Efficient Data-driven Visualization Recommendations to Support Visual Analytics. *Proc. VLDB Endow. 8*, 13 (Sept. 2015), 2182–2193. URL: http://dx.doi.org/10.14778/2831360.2831371, doi:10.14778/2831360.2831371. 4

[WHS*02] WATERSON S. J., HONG J. I., SOHN T., LANDAY J. A., HEER J., MATTHEWS T.: What Did They Do? Understanding Clickstreams with the WebQuilt Visualization System. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2002), AVI '02, ACM, pp. 94–102. URL: http://doi.acm.org/10.1145/1556262.1556276, doi:10.1145/1556262.1556276. 2

[WMA*16a] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Towards a general-purpose query language for visualization recommendation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (2016), ACM, p. 4. 4

[WMA*16b] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 649–658. doi:10.1109/TVCG.2015.2467191. 2, 3, 4, 5, 6, 9, 10, 11

[WQM*17] WONGSUPHASAWAT K., QU Z., MORITZ D., CHANG R., OUK F., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 2648–2659. URL: http://doi.acm.org/10.1145/3025453.3025768, doi:10.1145/3025453.3025768. 2, 9, 10, 11

[YKSJ07] YI J. S., KANG Y. A., STASKO J., JACKO J.: Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov. 2007), 1224–1231. URL: http://dx.doi.org/10.1109/TVCG.2007.70515, doi:10.1109/TVCG.2007.70515. 2, 4

[ZGC*16] ZGRAGGEN E., GALAKATOS A., CROTTY A., FEKETE J.-D., KRASKA T.: How progressive visualizations affect exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics* (2016). 2, 3, 5

[ZOC*12] ZIEMKIEWICZ C., OTTLEY A., CROUSER R. J., CHAUNCEY K., SU S. L., CHANG R.: Understanding visualization by understanding individual users. *IEEE Computer Graphics and Applications 32*, 6 (Nov 2012), 88–94. doi:10.1109/MCG.2012.120. 2, 3, 4, 12

[ZZZK18] ZGRAGGEN E., ZHAO Z., ZELEZNIK R., KRASKA T.: Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada, 2018), CHI '18, ACM. doi:10.1145/3173574.3174053. 2, 3, 4, 5, 7, 8, 11