Survey on the Analysis of User Interactions and Visualization Provenance

Kai Xu[1], Alvitta Ottley[2], Conny Walchshofer[3], Marc Streit[3], Remco Chang[4] and John Wenskovitch[5]

[1]Middlesex University, UK; [2]Washington University in St. Louis, USA; [3]Johannes Kepler University Linz, Austria;
[4]Tufts University, USA.; [5]Virginia Tech, USA.

Abstract

There is fast-growing literature on provenance-related research, covering aspects such as its theoretical framework, use cases, and techniques for capturing, visualizing, and analyzing provenance data. As a result, there is an increasing need to identify and taxonomize the existing scholarship. Such an organization of the research landscape will provide a complete picture of the current state of inquiry and identify knowledge gaps or possible avenues for further investigation. In this STAR, we aim to produce a comprehensive survey of work in the data visualization and visual analytics field that focus on the analysis of user interaction and provenance data. We structure our survey around three primary questions: (1) WHY analyze provenance data, (2) WHAT provenance data to encode and how to encode it, and (3) HOW to analyze provenance data. A concluding discussion provides evidence-based guidelines and highlights concrete opportunities for future development in this emerging area. The survey and papers discussed can be explored online interactively at `https://provenance-survey.caleydo.org`.

1. Introduction

The definition of *provenance* is "The place of origin or earliest known history of something" [oxf89]. The term is often used in the context of "the history of ownership of a valued object or work of art or literature" [mer19]. The notion of provenance has been adopted and extended in the field of Computer Science and applied to concepts such as data, computation, user interaction, and reasoning. In this context, provenance is no longer limited to origin or history, but also includes the process and other contextual information. Provenance is a growing topic in the visualization and visual analytics subfields, and includes the development of systems to visualize provenance data, analyzing such data to understand user behavior, and personalizing systems in response to user interactions.

One of the key goals of visualization and visual analytics is to support data analysis and *sensemaking* – "how we structure the unknown so as to be able to act in it" [Anc12]. In the context of data analysis, sensemaking involves understanding the data, generating hypotheses, selecting analysis methods, creating novel solutions, and critical thinking and learning wherever needed. Due to its exploratory and creative nature, the research and development of visualization approaches and techniques to support sensemaking lags behind the quickly-growing user needs. As a result, sensemaking is often performed manually, and the limitations of human cognition can become a bottleneck [LS10].

Provenance supports a variety of sensemaking tasks, such as recall of the analysis process by visualizing the provenance information, including the sequence of the investigations performed with contextual information (such as parameters and motivation). Provenance consists of the results of each analysis stage (including the final results) as well as the process that leads from data to conclusion. Such information can also be used to communicate analysis

outcomes. Examples include providing an overview of what has been examined, revealing gaps such as unexplored data or solution possibilities, and supporting collaborative sensemaking and communication by sharing the rich context of the analysis process.

The literature on provenance analysis research is growing rapidly, covering aspects such as its conceptual framework, use cases and user requirements, and techniques that are designed to capture, visualize, and analyze provenance data. As a result, there is an increasing need to better organize the provenance-related research landscape, categorizing and connecting current work, and identifying knowledge gaps. In this state-of-the-art report, we structure our survey of provenance-related research around three primary questions: *WHY* analyze provenance data, *WHAT* provenance data to encode and ways to encode it, and *HOW* to analyze provenance data. Those three aspects can be embedded along the overall process of analytical provenance outlined in Figure 1.

Through our survey, we identified a broad variety of purposes that underlie the analysis of provenance data, ranging from user-centric goals such as storytelling and modeling to system-centric goals like creating adaptive systems and evaluating algorithms. To perform such analysis, we note four overarching methods for encoding provenance data: sequences, grammars, models, and graphs. Given such data, researchers then analyze user provenance through a variety of classification and probabilistic models, pattern analysis, and program synthesis. We note that fuzzy boundaries exist in our categorization schema, as these methods of provenance analysis often overlap and blur.

Following our survey of provenance-related research, we discuss opportunities for future research in provenance analysis, including both fundamental problems and long-standing challenges. These include active areas of research such as inferring high-level prove-
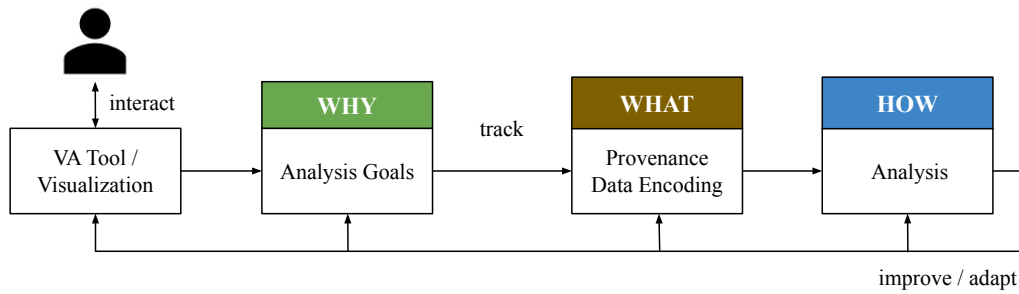
Figure 1: *A summary of the flow and structure of provenance-related research activities. We organize the survey around three primary questions: WHY analyze provenance data, WHAT provenance data to encode and how to encode it, and HOW to analyze provenance data.*

nance from low-level data, identifying groups within interaction sequences, and the use of provenance data to create truly adaptive systems. We discuss the need within the community for provenance standards, cross-tool integration, and reproducibility.

2. Related Work

In Computer Science, provenance has been studied in many fields, often under different names. The Human-Computer Interaction community relies on the analysis of *protocols* to understand user behaviors and intentions [DFAB03, PSR15]. Such protocols include audio/video recording, computer logging, and user notebooks. Their analysis goals are similar to those of provenance. The Database, Semantic Web, and e-Science communities have been studying provenance for almost two decades [BTC19]. *Data lineage* [BF05] and *data provenance* [LPG05] are used interchangeably in the discussion of provenance-related work [HDBL17] targeting issues such as process debugging, data quality, and accountability. This is closely related to work in the reproducible science community that aims to make complex scientific experiments reusable [FKSS08, ODOB18] and repeatable [CF17, IT18].

There is active ongoing research within the visualization community from both the scientific visualization [SFC07] and information visualization/visual analytics perspectives [NCE*11, XAJK*15, RESC15]. Many of the existing works focus on capturing [NXW*16] and visualizing [WSD*13, SLSG16, LAN19] provenance. There are few surveys or overviews on visualization-related provenance work. Xu et al. [XAJK*15] discussed the common techniques and open questions during the process of provenance analysis, namely modeling, capture, visualization, and its application in collaboration and trust. The work by Ragan et al. [RESC15] categorizes existing work based on the types of provenance information (data, visualization, interaction, insight, and rationale) and the purposes of the provenance (*recall*, *replication*, *action recovery*, *collaborative communication*, *presentation*, and *meta-analysis*). At a recent Dagstuhl workshop [FJKTX19b], leading researchers from the various provenance-related disciplines discussed the open challenges and outlined directions for possible solutions.

3. Definition and Scope

In this survey we focus on the **analysis of user interactions and provenance data**, whose main purpose is similar to the "meta-analysis" as defined by Ragan et al. [RESC15]. However, instead of a comprehensive review on all aspects of analytic provenance and the visualization of user histories, this survey focuses on the **analysis** of interaction and provenance in the field of visualization. As such, we only include existing work that incorporates meta-analysis based on user-generated (interaction) provenance data with the high-level goal of improving, enhancing, or understanding a visual analysis system, visualization process, or visual artifact.

To be included in this survey, we require the provenance data to constitute a cohort of recorded information from multiple users, a series of information from the same user, or both. As a result, a paper is not included if it only involves the analysis of a single piece of information provided by a user during an interactive visual analysis session. The same is true for non-trivial machine learning approaches, such as active learning methods. Here, the criteria is not the level of sophistication of the machine learning approach, but the amount and complexity of user input required. For example, it is not included if a sophisticated active learning technique only requires simple yes/no decisions from a user and requires no meta-analysis of the interaction data. We exclude user studies that collect user-generated data and work on collaborative sensemaking, if there is no additional analysis of the provenance information performed beyond recording and sharing.

4. Survey Methodology

Before diving into the review of provenance analytics, we describe our methodology for collecting the research papers that are included in this survey. For our literature review, we followed a three-stage systematic process as applied by Beck et al. [BBDW14]. We used tagging as a main instrument, starting with a list of freely assigned reasonable tags that are then iteratively merged, extended, and grouped to categories while working through the literature. As a result, we developed a typology for areas of application (WHY), encoding techniques (WHAT), and analysis methodologies (HOW) of provenance data. Even though we are aware that, for instance, the database provenance community already makes use of a three W terminology, namely WHY, WHAT, and WHERE, their application

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
CG&A	3	1	-	-	-	1	2	-	-	-	3
CHI	-	-	1	2	-	-	2	-	-	-	4
EuroVis	-	-	-	-	1	-	2	2	1	3	5
IUI	1	-	-	-	1	1	-	1	3	2	3
TIIS	-	-	-	-	-	2	-	1	1	1	-
TIST	-	1	-	-	-	-	-	-	-	-	-
TVCG	-	-	1	2	3	1	1	1	-	2	1
UIST	-	-	-	-	-	1	1	1	1	-	1
VIS	5	1	-	3	4	3	3	7	3	6	2
Sum ∑	9	3	2	7	9	9	11	13	9	14	19

Figure 2: *Number of candidate papers for the four journals and five conferences/symposium we screened from 2009-2019.*

is not the same [CCT07]. Thus, we particularly want to highlight that our state-of-the-art review puts emphasis on the meta-analyses of provenance data.

4.1. Corpus

To form the corpus of papers we discuss in this survey, we started by collecting work that we were aware of from our previous research and that were discussed in provenance-related survey papers as well as the report from the recent Dagstuhl workshop on 'Provenance and Logging for Sense Making' [FJKTX19a] (https://www.dagstuhl.de/18462). We continued with a systematic approach by manually scanning all issues from four journals and all proceedings from five conferences/symposia over the last eleven years (2009-2019):

- **Journals**
  - IEEE Computer Graphics and Applications (CG&A)
  - ACM Transactions on Interactive Intelligent Systems (TiiS)
  - ACM Transactions on Intelligent Systems and Technology (TIST)
  - IEEE Transactions on Visualization and Computer Graphics (TVCG)
- **Conferences and Symposium**
  - ACM Conference on Human Factors in Computing Systems (CHI)
  - EuroVis
  - ACM Conference on Intelligent User Interfaces (IUI)
  - IEEE Visualization Conference (VIS)
  - ACM Symposium on User Interface Software and Technology (UIST)

Figure 2 shows the number of publications per year from these journals and conferences that we included in this survey. The two main visualization conferences, VIS and EuroVis, have the largest number of relevant papers, which is not surprising. The IUI conference is a close third. Also, the total number of papers per year from all the journals and conferences have been increasing steadily over the ten years, and this topic is likely to receive even more attention in the near future.

4.2. Coding Process

For the systematic approach by screening the papers from the last eleven years, we continued with the tagging following the three stage process inspired by Beck et al. [BBDW14]:

1. **Explorative Tagging:** Every author screened at least one conference/journal. In the first round, we manually surveyed the title, keywords (e.g., *provenance analytics* and *model steering*), and the abstract, and used an open-coding approach to identify potential papers that make use of provenance data. This manual tagging allowed us to get an overview of relevant literature that deals with provenance analytics with the high-level goal of improving, enhancing, or understanding a visual analysis system. In total, this process resulted in a collection of 266 papers. The first coding round allowed us to review the entire body of work and prepared us for the second round of categorizing the tags into the three main sections: WHY to analyze, WHAT provenance data to analyze, and HOW to analyze provenance data.

2. **Category Tagging:** The aim of the second round of coding was to unify the categories and narrow down the scope. To achieve this, we developed a closed set of keywords for the spectrum of possible reasons (WHY) for doing meta-analysis on provenance data as well as for the state of the art of user interaction analysis (HOW). The process of categorizing the papers was reminiscent of a 'peer-review' because two of the authors independently revised the paper collection and coded the papers respectively. They analyzed only those papers from conferences/journals that they were not initially assigned to. In the case of an ambiguous and uncertain classification, all authors were in constant exchange. During this stage, we also continuously refined our scope and excluded papers accordingly.

3. **Supplementary Tagging:** In the last stage, we decided to further split up the two main categories, WHY and HOW, into subcategories. Therefore, we came up with six subcategories for WHY and five subcategories for the HOW. Some papers showed multiple reasons for conducting meta-analysis, for which we added more than one distinct subcategory to one paper. Similarly, when multiple techniques were applied to analyze provenance data, we accepted both categories. In the course of the analysis, we came to the conclusion that exclusively tagging the HOW section for provenance analytics is not sufficient, To address this, we introduced an additional WHAT aspect that allowed us to characterize the different encodings of provenance data in more detail. Similar to the WHY and the HOW, the WHAT shows double tags as well as no tags at all if no tags were applicable. After going through all three phases, we ended up with 105 papers, as summarized in Table 1.

The companion website, available at https://provenance-survey.caleydo.org, provides an overview of the WHY, WHAT, and HOW categories and allows users to filter and order the full list of publications by the categories and sub-categories interactively.

	Adaptive Systems	Evaluation of Systems and Algorithms	Model Steering and Active Learning	Replication	Report Generation	Understanding the User	Sum Σ
Classification Models	14	4	11	4	5	19	57
Pattern Analysis	9	7	6	5	3	13	43
Probabilistic Models	17	1	11	3	4	20	56
Program Synthesis	3	1	7	2	3	5	21
Interactive Visual Analysis	6	3	7	4	7	10	37
Sum Σ	49	16	42	18	22	67	

Figure 3: *The number of papers for each category in the HOW and WHY.*

	Grammar	Graph	Model	Sequence	Sum Σ
Classification Models	3	0	8	15	26
Pattern Analysis	3	5	4	11	23
Probabilistic Models	5	1	10	9	25
Program Synthesis	10	2	1	3	16
Interactive Visual Analysis	1	8	2	7	18
Sum Σ	22	16	25	45	

Figure 4: *The number of papers in the HOW and WHAT category.*

5. Structure of Survey

The structure of our survey is based on a high-level provenance analysis model (Figure 1) that we created to describe the important factors and their internal relationships in provenance analytics. All of the included works are based on user-generated (interaction) provenance data, wherefore we assume that the user interacts with a system. Analysis goals (WHY) are the reasons for provenance analysis and give rise to requirements such as what data to capture and how the data is encoded (WHAT). The encoded provenance data is then further evaluated by analysis techniques (HOW) such as classification methods or probabilistic models. At the end of this process, users gain either user-specific or system-specific knowledge that can be used to improve or adapt any of the process model components to enhance the overall provenance analysis iteratively. Based on this model, our survey aims to addresses the three main questions that will be faced by any researcher who chooses to conduct provenance analysis:

- WHY analyze provenance data?
- WHAT types of provenance data and ways to encode it?
- HOW to analyze provenance data?

Figure 3 and Figure 4 summarize the number of papers within each sub-category under WHY, WHAT, and HOW. These will be discussed in more details in the following sections. Section 6 provides an overview of the spectrum for possible purposes for conducting meta-analyses and outline six essential drivers for provenance analysis (WHY), followed by the encoding, representation, and storing of provenance data (WHAT) in Section 7. Section 8 continues the discussions by categorizing current provenance analysis methods according to their various approaches (HOW). We summarize our observations on goals, encoding methods, and analysis approaches in Section 9 and examine opportunities for further research in Section 10.

6. Goals: WHY Analyze Provenance Data

The spectrum of possible reasons for conducting meta-analyses on provenance data is broad. Our goal is to provide a comprehensive overview of the existing body of literature that analyzes provenance data for specific purposes. At a high-level, we can categorize the goals of the existing work as:

- Understanding the User
- Evaluation of System and Algorithms
- Adaptive Systems
- Model Steering
- Replication, Verification, and Re-Application
- Report Generation and Storytelling

6.1. Understanding the User

The goal of visualization is to create visual representations to support the user's reasoning and decision-making with data. Consequently, one of the primary reasons for analyzing provenance data is to *understand the user* and their sensemaking process [PJ09]. The ultimate goal of this category of research is to create theoretical and computational models that can describe the human analytical reasoning process. Some of the earlier research in the area works to uncover analysis patterns from interaction log data. For example, Dou et al. [DJS*09] demonstrated that it is possible to recover analysts' findings and strategies from log data. More recent work uses computational methods to uncover analysis patterns and workflows (e.g., [FPH19], [MRB19], and [LWD*17]). A promising set of work has also started to learn individual user characteristics, such as expertise, personality traits, and cognitive abilities from provenance data [BOZ*14, KWRK12, OYC15, SCC13]. Also in this category is work on modeling attention [OGW19] and exploration biases [GSC16, LDH*19, WBFE17] during analysis.

6.2. Evaluation of System and Algorithms

A few of the prior works have leveraged provenance data to understand the visualization system itself and to evaluate its usefulness [BKO*17, GL12, SML*09].

Here, it is important to distinguish between conducting statistical analysis on coarse user study metrics (e.g., speed, accuracy, and preference) and the *non-trivial analysis of provenance data for the primary purpose of evaluating a visualization design or system*. For instance, Bylinskii et al. [BKO*17] trained a neural network on mouse click data to create an automated model that learns the relative importance of visual elements for a given design. Smuc et al. captured the provenance to identify when users have insights [SML*09]. Gomez and Laidlaw modeled task performance on crowd workers to evaluate system design and help guide encoding choices [GL12]. Blascheck et al. [BJK*16] created a visual analytics system for evaluating an interactive visualization system. Among other techniques, they used pattern matching methods to uncover similarities within the provenance data of multiple users.

6.3. Adaptive Systems

A better understanding of the system and the user's analytic process give rise to opportunities to create *adaptive systems*. Such approaches are prominent in the existing literature and seek to improve the usability and performance of a visualization system, or the collaborative potential of the visual analytics tool. The body of prior work includes a wide variety of systems that recommend visualizations based on inferred tasks [GW09], provide guidance for a given interface [CGM*17, CGM19, CAS*18, WSL*19], or prefetch data to improve system performance [BCS16, KN19]. For example, Gotz and Wen [GW09] proposed behavior-driven visualization recommendation that infers a user's task in real-time and suggests an alternative visualization that might support the task better. A similar approach was adopted by Mutlu et al. [MVT16] by adapting visualization recommendations to the users' preferences. Fan et al. [FH18] trained a convolutional neural network on interaction data to create a faster and more accurate scatter plot brushing tool. By analyzing real-time interactions, Battle et al. [BCS16] demonstrated that incorporating provenance data into the prefetching pipeline improved system latency by 430%. To explore event sequence predictions, Guo et al. [GDM*19] preserve and aggregate records by their top prediction. In order to achieve a higher acceptance rate of the predictions, they showed multiple predictions and let the user choose.

6.4. Model Steering

Modeling steering leverages provenance data to improve the underlying data representations, machine learning models, or projection calculations in the case of high-dimensional datasets. Much of the work in this area uses active and reinforcement learning methods to learn from real-time interaction data and interactively improve the visualization. One noteworthy approach to model steering is *Semantic Interaction*, which defines the process of inferring model parameters as users directly manipulate data visualization components [EFN12a, EFN12b, ECNZ15]. For example, the *IN-SPIRE* system allows the user to directly manipulate the spatial layout of text documents to express perceived document similarity. Similarly, with *Dis-Function* [BLBC12] an analyst can update the parameters of

a distance function in a two-dimensional data projection by adjusting the positions of visual points. A similar approach is used by Hu et al. [HBM*13] with a spatialization algorithm to preserve semantics by allowing the user to move objects or highlight unmoved ones. Other research has applied model steering to refine data simulations [RWF*13a, SWR*13] or to steer approximation models of real-time streaming data [RWF*13b].

6.5. Replication, Verification, and Re-Application

Another usage of provenance data is to verify, replicate or re-apply analysis sessions. Here, we consider the body of work that goes beyond action recovery such as undo/redo. This category of research uses interaction logs to perform real-time or post-hoc quantification to validate the analysis results or to replicate the process when a similar problem arises. For example, in VisTrails [CFS*06] an analyst can create, edit, and compare the results of provenance dataflows. The *Harvest* [SGL09] system tracks interactions with data elements and recommends both notes and relevant views based on previous analyses in a collaborative environment. It is also common to convert the user interactions into executable scripts using a process called *program synthesis* – generating a script or executable sets of operations. *Wrangler* [KPHH11], for example, creates data transformation scripts based on passive observations of the user interactions. The scripts can then be re-applied to similar datasets. *Knowledge-Pearls* [SGP*19] allows users to rank and retrieve previous visualization states by formulating a string-based query. The query operates on a provenance graph containing automatically recorded user interactions and visualizations.

6.6. Report Generation and Storytelling

Finally, research has analyzed provenance data to automatically generate summary reports of an analysis session. Since a user's interaction history can be long and varied, "chunking" [HMSA08] to reduce the complexity of the history log, and "authoring" to generate reports and stories to reflect the relevant of the analysis are two common challenges. For example, *Click2Annotate* [CBY10] uses low-level tasks results to create insight summaries with automated annotations. Similarly, *InsideInsights* [MHK*19] produces automated data-driven reports that allow the analyst to edit and structure insights into hierarchical views. *Chart Constellations* [XBL*18] generates summary insights from observations in a collaborative system. Lastly, *CLUE* [GLG*16] supports a user to directly interact with the history (provenance) graph to generate a *story* from the user's analysis history.

7. Encodings: WHAT Types of Provenance Data to Analyze

Now that we have considered the different reasons WHY researchers analyze provenance data (see Section 6), the next challenge is to determine how the user's interactions can be encoded, represented, and stored. The choice of the encoding has a direct impact on the downstream analysis of the provenance data as well as the expected outcome. For example, recording the user's interactions as low-level keystroke or mouse movement events is apt for

the goal of reproducibility, but isn't adequate for higher level analysis for the purpose of real-time analysis recommendation or guidance. Conversely, representing a user's interactions as a sequence of discrete events has been successful for the purpose of user modeling. However, this encoding is not as flexible as a grammar-based approach that allows for future modification and reuse of the analysis process.

From the perspective of analysis techniques (see Section 8), the decision of how to encode the provenance data can also dictate the types of analyses that can be applied. For example, string analysis such as sequential pattern mining are best applied to sequence-based encoding, whereas signal-based analysis such as the use of Fourier and wavelet transform would assume that the data is represented as a continuous stream. Given the importance of encoding, in this section we categorize how provenance data have been represented in existing literature. In particular, we find four common encoding schemes:

- Sequence
- Grammar
- Model
- Graph

7.1. Sequence

Perhaps the most common of the encoding schemes, a *sequence-based* encoding records a user's interactions with a visualization tool into a temporally ordered list. This list is often represented as a *string* that consists of a discrete number of *symbols*, in which each symbol represents a type of interaction event.

Depending on the goal, the choice of the symbols may differ. For example, the symbols may be used to represent the interactive elements in a specific visualization (e.g., range selection in scatter plot), the data elements that the user interacted with (e.g., a page from a clickstream data), captured information about a user (e.g., the user's eye gaze movement), etc. In this section, we identify publications that: (1) record the user's interactions as a linear sequence of events, and (2) perform analysis over such sequences. Based on the representation of sequence-based encoding, we further group these publications into six types: **Interaction Type**, **Application State**, **User State**, **Taxonomy-Based Abstraction**, **Image Space**, and **Temporal Signal**.

7.1.1. Interaction Type

Arguably the most direct approach for recording a user's interactions with a visualization interface, this encoding approach can be considered as a log of all user actions. Typically, this log is generated from recording the callback functions executed during a user's session interacting with a visualization.

In some cases, the logged information can be low-level, such as keystrokes and the (x, y) positions of the mouse [GL12]. While these types of interactions do not contain semantic information, researchers have found that the analysis of such information can be used to classify types of users [BOZ*14]. More typically, interactions are captured at a higher semantic level that reflect the specific

capabilities afforded by the visualization tool itself. For example, Battle and Heer [BH19] record user interactions with Tableau, including actions such as "shelf-add," "shelf-remove," "show-me," etc. which are interaction elements specific to Tableau.

Since the captured interactions are application-specific, the analysis of the provenance data is largely focused on the understanding of the user and their reasoning processes. For example, Dou et al. [DJS*09] log interactions of expert financial analysts to examine how much of their reasoning process can be recovered. Similarly, Brown et al. [BOZ*14], Cho et al. [CWK*17], and Feng et al. [FPH19] use the interaction logs to classify users based on their performance, whether they might be under the influence of anchoring bias, and exploration strategies, respectively.

7.1.2. Application State

Instead of logging the user's interactions with a visualization, a system can also log the resulting state of the visualization. The reason for choosing an *Application-State*-based encoding over an *Interaction*-based encoding is often because the visualization itself affords few interaction elements to differentiate a user's exploration or analysis intent.

For example, clickstream data from a user's web-browsing history has low granularity in terms of a user's interactions (i.e., there are few types of actions that a user can perform, such as click on a link, refresh, go-back, etc.), but can be very rich if the system logs the specific (types of) websites that the user examined. In the work by Wei et al. to analyze users' purchase patterns on eBay [WSSM12], the authors encode the clickstream data into categories such as Title, Description, Pricing, Shipping, Picture, etc. Related, works by Liu et al. [LKD*17, LWD*17] use a similar approach to analyze branching behaviors and detect uncommon patterns in clickstream data.

Beyond clickstream data, researchers have used the *Application State* encoding approach in a variety of other contexts. Cavallo and Demiralp [CD18] log changes to a machine learning model (and its corresponding changes in performance metrics) in a collaborative data analysis task. Stitz et al. [SGP*19] record past visualization states and allow a user to retrieve the state (and the visualization) by querying the system. Guo et al. [GDM*19] use the recorded application-state log to predict and recommend possible visualizations using a recurrent deep learning model. Moritz et al. [MHHH15] capture the query execution trace to help improve query performance.

7.1.3. User State

In addition to recording user interactions or the states of the application, there are often additional data and information generated from the use of a visualization such as insights, annotations, etc. In our survey, we identify two types of such information: *active* user annotations and labels and *passive* user information such as eye-tracking data.

Active Actions: Instead of analyzing interaction history, Smuc et al. [SML*09] develop a tool to analyze the insights of the user as sequential data. The tool takes into account three such sequences: insights about the tool, insights about the data, and interactions with

the tool. Similarly, work by Choe et al. [CLs15] correlate fitness data with the user's annotations of their health state.

Related, researchers can manually code users' analysis sessions to identify patterns, commonalities, etc. For example, Boukhelifa et al. [BBT*19] perform an exploratory study on how experts collaboratively perform sensemaking with machine learning models. The experts' interactions are encoded as one of six possible high-level operations: initial exploration, new exploration, refine, compare, alternative, storytelling.

Passive Actions: In contrast to user's annotations and self-reported insights – which are data *actively* generated by the user – researchers have also included the use of *passive* data such as eye-tracking information and brain signals into the analysis of interaction logs. Works by Blascheck et al. [BJK*16, BBB*16] combine a number of data sources, including eye-tracking, audio, video, and other provenance information into an analysis environment to better understand and evaluate how a user uses a visual analytics tool.

Eye-tracking data have also been used in the visualization for other inferencing tasks. Bylinskii et al. [BKO*17] use eye-tracking data to learn visually salient features in graphic and visualization design. Steichen et al. [SCC14] and Smith et al. [SLMK18] demonstrate that analysis of eye-tracking data can be used to infer task difficulty and the user's confidence, respectively. Ottley et al. [OKCP19] track user's eye movements when reading texts that are embedded with visualizations and find that users do not integrate information well across the two representation styles.

In addition to eye-tracking data, recently researchers began using brain-sensing technologies to monitor a user's mental state when using a visualization. For example, Anderson et al. [APM*11] analyze EEG signals to determine a user's cognitive load when using different designs of box plots. Similarly, Peck et al. [PYO*13] use functional near-infrared spectroscopy (fNIRS) to compare users' levels of cognitive effort when using bar charts and pie charts.

7.1.4. Taxonomy-Based Abstraction

One shortcoming of an *Interaction Type* encoding strategy is that the interaction logs are specific to the application. As a result, if the goal of analyzing the provenance data is to compare users using different visualization systems, the use of an application-specific encoding strategy would be ineffective.

To generalize the user interactions, researchers have made use of taxonomies in the visualization of interaction types [YaKS07, LS10], task types [BM13], and analysis models [PC05]. Instead of recording each of the user's interactions at the application level, each interaction is first converted to an element in the taxonomy, thus unifying the symbols used to encode users' interactions in multiple visualizations.

In particular, Pohl et al. [PWM*12] and Guo et al. [GGZL15] encode user's interactions using the taxonomy by Yi et al. [YaKS07] to compare analysis paths from the use of different visualizations and identify interaction trails that lead to user insights, respectively. Xu et al. [XBL*18] develop a tool that organizes a user's analysis history (and the corresponding visualizations) using the task taxonomy by Brehmer and Munzner [BM13]. Loorak et al. [LTC18] take a similar approach to examine changes between visualizations

in Tableau. However, instead of utilizing an existing taxonomy, the authors proposed their own categorization consisting of six task types: encoding, filtering, analytics, arrange, data, and formatting.

Also using a model, Perry et al. [PJ09] take a different approach from the previously described work. Instead of encoding a user's interactions using an existing taxonomy, the authors first associate a user's interaction with one of the analysis states in the Sensemaking Loop by Pirolli and Card [PC05]. Treating the Sensemaking Loop as a Markov model, over time the system learns the transition probability of the edges and can therefore predict or recommend future analysis actions.

7.1.5. Image Space

Since many of the operations in a visualization relate to the user's interactions with the visual representations, these interactions can be encoded directly in the image space. Most common use cases of image-based encoding are visualization systems that support sketch-based query construction. In these systems, a user draws a pattern in the visualization and the system searches through the data to find data items that exhibit similar patterns. This technique has been demonstrated to work well for querying temporal [Wat01, CG16, MVCJ16] and spatial data [WCW*14].

In addition to sketches, Fuches et al. [FWG09] present a system that uses a genetic algorithm to learn interesting visual features from user-highlighted regions in the generated visualizations. Battle et al. [BCS16] analyze images produced by a visualization and extract features to predict a user's interests and future actions.

7.1.6. Temporal Signal

Lastly, we find one example of a sequence-based encoding scheme that makes use of the temporal aspect of the interaction logs. Instead of converting the interaction log into a string of discrete symbols, in the work by Feng et al. [FPH19] the authors treat the sequence as a continuous temporal signal. As a result the authors are able to apply signal analysis techniques such as wavelet transforms to analyze the interaction data.

7.2. Grammar

While the *Sequence-based* encoding scheme is robust and faithful in recording a user's interactions with a visualization, it is a static representation that does not afford future modifications and therefore reuse. In cases where a user's provenance information needs to be examined and re-applied to automate future analyses, researchers have developed techniques for recording the user's actions using rules and grammars. An early example of this approach is the *HomeFinder* system by Williamson and Shneiderman [WS92]: a user's interactions with the *HomeFinder* visualization result in the generation of SQL queries that are then executed by a back-end database.

Outside of the visualization community, one popular example of a *Grammar-based* encoding scheme is *Excel's* AutoFill and Flash Fill techniques. In *Excel*, a user can provide a few example values in cells and "drag" those values to other cells that are then automatically populated. Under the hood, *Excel* uses the few examples to learn regular expression rules [Gul11, GHS12] that are

then applied to the empty cells – a technique in the programming languages community known as program synthesis.

In this section, we identify publications in visualization and analytic provenance that encodes user interactions using a *Grammar-based* approach. We group these techniques into three categories: **Logic Rules**, **Languages and Scripts**, and **Specifications**.

7.2.1. Logic Rules

A common approach, especially when encoding a user's interactions with a multiple coordinated visual analytics system, is to encode each user interaction as a simple rule using first-order logic. For example, a user brushing over a range of values in the x-axis of a scatter plot can result in the rule ($x < 5$). These rules can then be chained together using Boolean or first-order logic that can be stored, modified, and reused. A paper by Weaver describes this type of encoding in a visualization system as "Conjunctive Visual Forms" [Wea09].

In the paper by Xiao et al. [XGH06], the authors apply this method to perform network traffic analysis. A user's interactions with their system results in queries in first-order logic with domain-specific clauses relating to network information (e.g., IP source, IP destination, etc.). Garg et al. [GNRM08] adopt a similar first-order logic representation based on Prolog, but uses an Inductive Logic Programming method for learning the rules. Srinivasan et al. [SPEB18] present the *Graphiti* system that learns Boolean logic rules based on a user's interactions when constructing a graph from a tabular data. Lastly, using a more explicit approach (instead of the implicit learning of rules), Koch et al. present a system that allows a user to interactively construct (Boolean logic) queries in patent search [KBGE09].

In a slightly different vein, Mutlu et al. [MVT16] use rules extracted from past visualization examples to recommend new visualization. Their system, *VizRec*, learns visual-data mappings from previous visualizations generated by the user and stores them as rules. These rules are then used to automatically map data attributes in a new dataset to visual attributes.

7.2.2. Languages and Scripts

In addition to first-order logic, researchers have used a range of other grammars and domain-specific languages to represent the user's interactions. While a full treatment of formal languages and their power is beyond the scope of this paper, the encoding methods using these grammars and languages often have higher expressive power over the use of first-order logic for capturing the nuances in a user's interactions with a visualization. For example, in the paper by Dabek and Caban [DC16], the authors encode the user's interactions as a deterministic finite automaton and leverage existing algorithms to learn a compact grammar from the user's interactions. These learned grammars encode sequence information that cannot be easily captured using the *Logic Rules* approaches.

Beyond formal grammars, researchers have developed their own domain-specific languages to encode the user's interactions with their system. In the papers by Kadivar et al. [KCD*09] and Chen et al. [CQW*14], the authors present the *CzSaw* system that generates a reusable script based on the user's interactive analysis of graphs.

Kandel et al. [KPHH11] propose the *Wrangler* system that helps a user perform data cleaning. In Wrangler, the system generates multiple plausible scripts from a user's interaction. A user can choose one of those scripts and apply them to the rest of the data (similar to AutoFill) or make modifications to them before the application. Muthumanickam et al. [MVCJ16] and the *Zenvisage* system by Siddiqui et al. [SKL*16] apply a similar technique to querying temporal data. Using a sketch-by-example approach, a user's drawing of a desired temporal pattern is first converted into a shape grammar whose design is inspired by regular expression. A user can edit and modify the expression to further refine the degree of smoothing and approximation of the query.

Lastly, although not strictly a grammar or language, in the works by Hoque et al. [HSTD18] and Setlur et al. [SBT*16a], the authors make use of principles from linguistic theory to disambiguate natural language queries. These systems augment a user's query with annotation functions like Continue, Retain, Shift to maintain the context of a continuous analysis session and make potentially ambiguous user queries meaningful to the visualization system.

7.2.3. Specifications

In some cases, each user interaction with the visualization might not be meaningful or relevant to the user's goal. Instead, through iterative interactions with the visualization, the user aims to populate a specification that in turn can be used for generative purposes. Note that our distinction between a specification-based encoding and a language-based encoding is not strictly based on formal programming language theory. Instead, it reflects how the user's interactions are represented – either as a script that is open-ended, or as a means to generate a specification with prescribed properties.

For example, similar to the *HomeFinder* example [WS92], in the work in Ferreira et al. [FPV*13] for visual exploration of large-scale urban data, the user's interactions are translated into parts of the WHERE clause in a SQL query. In this case, each of the user's interactions is not of particular relevance and does not need to be recorded. Instead, it is the final constructed query that is of interest to the system and the analysis. Walker et al. [WSD*13] take a similar approach in developing their visual analytics tool to analyze "human terrain" information. Users' interactions and analysis states are stored in a specification format called *ProvML* that is XML-based extension of the Open Provenance Model. Similarly, Rubel et al. [RB18] store user analysis of mass spectrometry imaging (MSI) data in a format proposed by the authors to enable sharing and reproducibility of analysis. Lastly, *Voyager*, a system by Wongsuphasawat et al. [WMA*16], can recommend new visualizations based on the user's previous exploration pattern. In *Voyager*, visualizations are specified using the *Vega-Lite* specifications [SMWH16] and the system identifies patterns between these specifications to make future recommendations.

Recently, there has been a number of papers on the interactive specification and generation of bespoke visualizations. Although the goals of these research projects are not for the purpose of analytic provenance or tracking of the user's interaction histories, some of them use similar encoding techniques. For example, Lyra [SH14] uses the *Vega* specification [SRHH15], and the framework of *Charticulator* [RLB19] is a new specification proposed by its authors. In

both cases, a user's interaction with the systems result in populating fields in these specifications, which are then used to generate a bespoke visualization design.

7.3. Model

In visual analytics systems, the goal of provenance and interaction analysis is often expressed as the (machine learning) model that a user is constructing, steering, or exploring. In these cases, how the user interacts with the data or the visualization might not be the primary focus and are therefore not directly encoded. Instead, the visual analytics system performs *inferencing* over the user's interactions that results in updates to the underlying models.
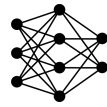
Sometimes known as interactive model steering, in this section we identify publications that: (1) use a sequence of interactions to derive the model, (2) make explicit, quantitative, and recordable representations of these models, or (3) present novel inferencing techniques for analyzing a user's interactions. We categorize papers in this section into two groups: **Machine Learning Models** and **User Models**.

7.3.1. Machine Learning Models

In interactive model steering, a common way to record the history of the analysis process is to encode the state of the model itself. An example of such model steering is the work in learning a distance function for a 2D projection of high dimensional data by Endert et al. [EHM*11] and Brown et al. [BLBC12]. In these systems, the user manipulates the positions of data points in a 2D projection, and the system learns the parameters of the underlying distance function that would make such a projection possible. For example, when using a weighted Euclidean distance function, the model can be represented as a vector where each value in the vector represents the weight of a dimension. Users' provenance data from interacting with such a system can then be visualized as trails for the purpose of cohort analysis and comparison [BYC*19].

Similar model steering techniques can be applied to other types of models, such as those to learn the relative importance of documents and keywords in texts [BNHL14, EFN12b], temporal sequences [KC14], ranking [WDC*18], projection planes [KCPE16, KKW*17], concept graphs [CCI*15, MSW10], visual features [FWG09], and visualization recommendation [BCBL13].

While most of these works do not directly record the user's interactions, in the work by Hossain et al. [HOG*12] the authors explicitly models a user's interactions with the underlying clustering model as a matrix of constraints. Each row and column of the matrix represents one cluster, and by toggling on or off each cell, the user can interactively "gather" or "scatter" the data points and steer the clustering model.

7.3.2. User Models

In some cases, the purpose of tracking a user's analysis behavior is to learn a model about the user. For example, Gotz et al. [GSC16] model a user's interactions with a visual analytics system to detect selection bias during a user's analysis of high-dimensional data.

Each of the user's interactions is modeled as a probability distribution over the data space, and bias is defined by measuring differences in these distributions. Wall et al. [WBFE17] take as similar approach to detect bias, but instead model the user's interactions as Markov chains.

Similarly, Healey and Dennis [HB12] and Ottley et al. [OGW19] both aim to model "user interests" in a visualization. Healey and Dennis develop a Bayesian user model using a boosted Bayesian network classifier that takes into account the user's explicit and implicit inputs on their interests in the shown visualization and the underlying data attributes. Ottley et al. use a Hidden Markov Model approach that models the user's attention, where low-level features in a visualization (e.g., color, size, positions, etc.) are modeled as hidden states and the user's interactions as the observable states in the model.

Lastly, Nguyen et al. [NHC*20] model a user's interactions as probability distributions over analysis sessions. Using Latent Dirichlet Allocation, each of the user's interactions is considered a "word" and each analysis session is a "document." With this model, the authors can generate hierarchical profiles of users based on their analysis behaviors.

7.4. Graph

Since the purpose of using a visualization is to explore data, discover patterns and relations, and eventually build knowledge, many visualization systems encode the user's interactions as knowledge graphs, concept graphs, or history graphs. In most cases, the nodes in these graphs represent a data item (e.g., a document, a location, an entity), an abstract concept (e.g., user annotations), or a visualization state (e.g., user histories). Edges then represent connections or relations between these entities. In this section, we identify publications that that use a *Graph-based* encoding approach and group them into two categories: **Entity and Concept Graphs** and **History Graphs**.

7.4.1. Entity and Concept Graphs

The use of entity or concept graphs is most commonly associated with sensemaking and collaboration. In these applications, the user's interactions with a visualization or visual analytics system result in some modification to the graph. Similar to the *Model-Based* encoding strategy, since the goal of the visual analysis process is defined (in this case, the construction of the graph itself), a user's interaction log can be modeled as changes to the graph over time.

In these systems, the choice of the representations of the nodes and edges reflect the purpose of the system. For example, the *VizCept* system [CYM*10] is designed for the purpose of collaborative sensemaking. In VizCept, a node represents a concept, which can be an entity such as a name, a location, an object (extracted from text documents) or a word that a user types in. Multiple users can simultaneously interact with these nodes by connecting them, forming relations. The *CLIP* system by Mahyar and Tory [MT14] and the *KTGraph* system by Zhao et al. [ZGI*18] use a similar encoding of nodes and edges. The *CLIP* system has an additional

emphasis on the temporal order of events over the *VizCept* system, whereas the *KTGraph* system is designed specifically for asynchronous collaboration.

We find one example where the authors make additional use of a entity or concept map beyond sensemaking. In the *Candid* system by Shadoan and Weaver [SW13], the user interactively constructs an entity (attribute) relations graph, which is then translated into a hypergraph querying language that can express complex n-ary multi-dimensional conjunctive relations.

7.4.2. History Graphs

Closely related to *Sequences* of *Application States*, the *History Graph* encoding strategy records a user's interactions as a graph structure instead of a linear list of temporal events in the *Sequence-based* approach. Although more complex, the graph-based approach can reveal patterns in the graph structure that cannot be easily found in a sequence-based representation, such as cycles (representing repeated analysis steps), high vertex connectivity (representing a commonly re-visited analysis state), cliques (potentially representing detailed analysis), etc. Further, with additional analysis to identify semantically meaningful labels for the nodes, these graph representations can be used for the purpose of reporting and storytelling.

Systems such as *VisTrails* [CFS*06], *Graphical Histories* [HMSA08], *GraphTrails* [DHRL*12] all use a graph-based encoding of a user's interactions. Each node in these graphs represents an action taken by the user. After the construction of a graph, the system can then perform additional operations over the graph, for example to reduce the graph's size and complexity [HMSA08]. Dabek and Caban [DC16] take a similar approach, but use a finite automaton (which is a directed acyclic graph) as its internal representation. Also using a directed acyclic graph (or more precisely a hierarchy), Dextras-Romagnino and Munzner [DM19] present the *Segmentifier* system that helps a user iteratively refine sequences of interaction data into meaningful segments.

For the purpose of reporting and storytelling, in the work by Gratzl et al. [GLG*16] a user directly interacts with the history (provenance) graph to generate a *story* from the user's analysis. Mathisen et al. [MHK*19] present the *InsideInsights* system that generates a report of a user's analysis by first annotating the visualization states and aggregating the states into narrative schemas.

History graphs can also be used to represent steps in data transformation and cleaning. In the work by Bors et al. [BGM19], data cleaning (wrangling) operations are stored as a directed acyclic graph. The user can explore the graph and evaluate the quality of the data cleaning along the process. Similarly, Schlinder et al. [SWR*13] present a dataflow model for data transformation. Although this model is not specific to recording user interactions, it uses a graphical encoding to represent the stages of data transform and analysis that is similar to provenance graphs.

Lastly, we find two examples of a hybrid approach that combines both a concept graph with a history graph. In the work by Shrinivasan et al. [SvW08, SvW09, SGL09], the authors present systems that track and maintain a history graph while allowing a user to manually construct a concept graph that represents the user's analysis process. The two graphs are coordinated such that a user clicking on a node in the concept graph will take the user back to the corresponding analysis step(s) in the history. In SenseMap [NXB*16], user online browsing history is shown as a graph ("*history map*") with webpages as nodes and visited links as edges. There is a additional "*knowledge map*" in which user can create concept graph with information collected during online exploration as node and the edges are created by user (not the visited link) to connect similar or relevant items.

### 8. Techniques: HOW to Analyze Provenance Data

User interactions collected from a visual data exploration or analysis session can be analyzed in a variety of ways. In the most simplistic cases, the user interaction data can be stored as part of the "undo/redo" mechanism with little data processing required. In this section, we focus on complex analysis methods that researchers apply to the interaction data to derive insight into the user's analysis intent, re-purpose past analyses, predict future user actions, or create analysis summaries. We organize the observed methods into five primary categories:

- Classification Models
- Pattern Analysis
- Probabilistic Models / Prediction
- Program Synthesis
- Interactive Visual Analysis

8.1. Classification Models

The most common technique for evaluating provenance data is the use of classification and statistical modeling techniques to differentiate sequences of user actions [BOZ*14, WBFE17, GGZL15, KPS*17, OYC15, BCN*19, DC16]. The overarching goal of such techniques is to map a user action to one or more categories. A number of surveys in the literature have demonstrated the application of these techniques to a variety of data types relevant to provenance analysis, including text [AZ12, SYD*14] and images [LW07, NMK*14]. Indeed, many of the common types of insights that users wish to identify in data necessitate a classification phase, including comparison, correlation, distribution, and trend insights [CLs15]. In this section, we identify publications that classify provenance data into groups of similar user actions via a variety of methods, ranging from straightforward clustering through complex machine learning processes.

Perhaps the most frequently applied method for classification in general research is *k*-means clustering. The goal of this unsupervised method is to partition a collection of observations into *k* clusters such that each observation is assigned to the cluster with the nearest mean. The standard algorithm for *k*-means is straightforward to implement, leading to the popularity of this technique. However, *k*-means has limitations, and is best able to identify clusters that are convex and with similar covariance [Llo82], a property that is not guaranteed in interaction logs. The algorithm is found in systems such as *Chart Constellations* [XBL*18], which permits
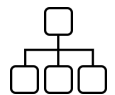
an analyst to interact with a collection of charts: projecting, clustering, filtering, and connecting results. Sherkat et al. [SNMM18] develop an interactive *k*-means approach to permit users to interact with eye-gaze patterns. A similar approach that is less-frequently used is the *k*-nearest neighbors classifier, in which an observation is assigned to the class most common among the *k* nearest observations. Pezzotti et al. [RWF*13b] use a similar Forest of Randomized KD-Trees approach to create a steerable t-SNE (t-Distributed Stochastic Neighbor Embedding) method for data exploration.

A common supervised method for classification is regression analysis, which attempts to determine the relationship between a dependent variable and a collection of independent variable inputs. Regression models appear in a variety of forms, including linear, logistic, and polynomial regression. The difference between these models lies in the type of function used to model the dependent variable. Toker et al. [TLC17] experiment with linear regression, among other techniques, to predict the phase of a user's skill acquisition when interacting with bar graphs, but separately use logistic regression in a previous experiment with similar eye tracking data and goals [TSG*14]. Hu et al. [HDG*19] also make use of linear regression in their *VizNet* experiments, working to understand the influence of user task and data distribution on visual encoding effectiveness.

Support-vector machines (SVMs) present another method for classification analysis, though the goal of the algorithm switches from approximating a relationship to identifying an optimal boundary between classes. SVMs can also efficiently perform non-linear classification by means of a kernel function that maps the input space into another that is more computationally tractable. SVMs are found classifying provenance data in a study processing eye-gaze data by Steichen et al. [SCC14], as well as in "Finding Waldo" [BOZ*14]. Toker et al. [TLC17] also test SVMs against linear regression in their prediction study.

Topic modeling reduces a broad collection of terms into a smaller collection of topics, simplifying the analysis and often enabling the outcome to be more easily visualized in a two-dimensional projection [ECNZ15]. Techniques such as Nonnegative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) are used by Sherkat et al. [SNMM18] in their adaptive clustering implementation. Latent Semantic Analysis (LSA) and Latent Semantic Modeling (LSM) work similarly, demonstrated by Wegba et al. [WLLW18] as their work identifies a relationship between users and movie recommendations. Boukhelifa et al. [BBT*19] use dimensionality reduction more generally to reduce a model exploration space, aiding analysts in exploring complex model results.

Artificial neural networks are learning methods inspired by biological neural networks. These come in a broad variety of forms, such as the Fully Convolutional Networks (FCNs) used by Bylinskii et al. [BKO*17] to predict important regions in an interface, enabling automatic design retargeting and thumbnailing by analyzing user interactions. Steichen et al. [SCC14] and Toker et al. [TLC17] also experiment with neural networks in their studies.

Hierarchical techniques such as decision trees model sequences of decisions and their consequences. The techniques are particularly suited for provenance analysis, as sequences of interactions in-

clude similar branching behavior, as seen in *CoreFlow* [LKD*17]. Similarly, hierarchies can be used to map low-level interactions to higher-level intents and reasoning processes [BWD*19, DJS*09].

8.2. Pattern Analysis

Pattern analysis refers to the detection of patterns in data or logs. Such analysis often comes in two forms. First, **Automated Pattern Analysis** often consists of the stages leading up to a prediction or classification, mapping the detected patterns in the provenance data to an outcome either as part of a continuing automated process or as a preprocessing step before an analyst begins exploration of the patterns. Second, **Manual Pattern Analysis** refers to user-driven exploration and analysis of patterns in provenance data. When considering the analysis of provenance data, detecting patterns in interaction logs by either the manual or automated approach can enable systems to predict future interactions, as well as providing users with insight into their own behaviors. In this section, we identify publications that examine large-scale patterns in provenance data, classifying these works into automated or manual groups by the initiator of the analysis.
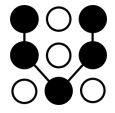
8.2.1. Automated Pattern Analysis

One common method for automated pattern analysis is to traverse a graph representation of the provenance data. For example, Dabek et al. [DC16] encode a collection of user interactions as a directed acyclic graph, and then extract common sequences from the graph for later analysis. Shrinivasan et al. [SGL09] also traverse a graph generated from provenance data to identify patterns and sequences for the purpose of automated annotation and recommendation. Shadoan et al. [SW13] take a similar approach, representing user queries as a hyper-graph that can be used in future analysis.

Systems and studies also use machine learning and similar intelligent approaches to identify patterns in provenance data. Gotz et al. [GSC16, GSC*17] use Adaptive Contextualization, monitoring and modeling a user's data selection activity and computing metrics over that model. Bylinksii et al. [BKO*17] use a neural network to extract and prune patterns that are later presented to users for interaction. Nguyen et al. [NTA*18] make use of the Generalized Sequential Patterns algorithm to identify frequent patterns from a set of user sessions.

Other approaches for automated pattern analysis include techniques such as the automatic extraction and visualization of branching patterns in event sequences as seen in *CoreFlow* [LKD*17]. In a similar visualization-centric approach, Liu et al. [LWD*17] use sequence clustering and sequential pattern matching on collections of websites visited by users, permitting patterns to be automatically extracted and pruned before presenting the patterns to users for interaction. *HARVEST* is used by Shrinivasan et al. [SGL09] to create a context-based retrieval algorithm that uses notes, views, and concepts from past analyses to identify patterns most relevant to a user, providing these patterns to a user within a web-based visual analytic system.

8.2.2. Manual Pattern Analysis

Sketches are a common means for manually presenting patterns to a system, representing a user-driven approach for pattern matching. The type of sketch varies across systems. Wang et al. [WCW*14] permit users to sketch 2D maps for querying trajectory information. Both Correll et al. [CG16] and Muthumanickam et al. [MVCJ16] support sketches of time series, which are then converted to queries for pattern retrieval.

Rule-based systems are a second approach for user-driven pattern matching. Cappers et al. [CvW18] present a system that allows users to construct rules to encode events, after which queries are automatically generated into a format similar to regular expressions. The *KnowledgePearls* system [SGP*19] supports a similar query-based approach for searching for past visualization states.

A variety of other techniques also permit analysts to query information about patterns in past interactions. *Segmentifier* [DM19] iteratively refines collections of interactions into meaningful sequences, which can subsequently be queried through a visual interface. *Perfopticon* [MHHH15] provides a similar interactive tool to explore query execution traces in distributed databases, while MacInnes et al. [MSW10] present an expert-guided clustering technique to identify patterns and link those patterns to their semantic meaning.

8.3. Probabilistic Models / Prediction

Precise classifications are not necessarily the only predictions that can result from the analysis of provenance data. Indeed, the nuance and randomness inherent in interpreting and inferring from such imprecise data often necessitates probabilistic interpretations [MA19]. In this section, we identify publications that make use of probabilistic techniques to identify trends and predict possible future interactions in provenance data.
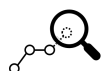
The most straightforward probabilistic approaches rely on traditional statistical models. For example, the approximated and steerable t-SNE created by Pezzotti et al. relies on the underlying t-distribution to model data via this dimensionality reduction strategy [RWF*13b]. Similarly, Feng et al. explore multiple metrics for visualization interaction behavior to cluster groups of users, including the chi-square distribution among others such as frequency counts and TF-IDF [FPH19].

The next stage beyond basic probablistic approaches is the use of Bayesian probability and inference. Naive Bayes classifiers are relatively simple probabilistic classifiers that incorporate substantial independence assumptions between the features under consideration, and are used as one of the models tested in the pupillometry and head distance analysis by Toker et al [TLC17]. Healey et al. [HB12] also make use of Bayesian classification to identify data items within large datasets that are of potential interest to the user.

The output of the neural networks discussed in Section 8.1 are often fuzzy classification results, not precise predictions. The Time-Aware Neural Networks (TRNNs) used by Guo et al. [GDM*19] and the Convolutional Neural Networks used by Smith et al. [SLMK18] demonstrate two separate prediction approaches, supporting the ability to predict tasks as distinct as next

actions and user confidence. Probabilistic prediction models can be simpler than neural networks as well, even including probabilistic views on regression approaches such as logistic [BCN*19] and linear [MSM*17] prediction models.

Markov models are used to model randomly changing systems, assuming that future states are only dependent on the current state rather than events that occurred previously. Predicting such future states often makes use of a state-based approach, in which edges represent transitions between the states [PWM*12,PSM12,BCS16, PJ09]. Ottley et al. use a hidden Markov model to learn from and anticipate mouse interactions during exploratory data analysis [OGW19], while Wei et al. [WSSM12] follow a similar approach for Web clickstream data.

The nuances in provenance data are also apparent in natural language interfaces. These systems must interpret ambiguous or complex instructions provided by users [SBT*16b], who could adopt a number of different means to express the same goal. Such approaches range from the use of standard libraries such as word2vec for keyword tagging [XBL*18] to the use of topic modeling, a classification method also addressed in Section 8.1 that is probabilistic in nature [CWK*17, NHC*20]. Linguistic theory can aide in the analysis of sentences to understand queries that are generated by users for these approaches [HSTD18].

8.4. Program Synthesis

Another provenance analysis approach is to perform program synthesis – generating a script or executable sets of operations based on past user interactions. In these systems, a common goal is to convert the user interactions into an executable script that can be applied to a new dataset. However, it is also possible to perform additional analyses over these higher-level scripts or grammar to extract additional information. In this section, we identify publications that present such provenance analysis as grammars or scripts, noting an emphasis on visual history and recall in this area of the literature.

There is substantial overlap between the first stage of this approach and the discussion of grammar-based encodings in Section 7.2, as the use of a grammar-based approach [DC16], or the design of a domain specific language [KPHH11,KCD*09,CQW*14] to encode the provenance data often underlies this approach. These approaches are varied, with analysis taking place on encodings ranging from text-based approaches such as SQL queries [FPV*13] and regular expressions [MVCJ16] to more complex forms like graphs [KCD*09, SPEB18] and Boolean logic [KBGE09, Wea09]. We point the reader to Section 7.2 for a more thorough discussion of these encodings.

Expanding the discussion in this section more broadly than only the analysis of these scripts and grammars, we find studies detailing ways in which reviewing provenance data permits a user to both better explore past actions and to understand the effects of potential future actions within a system. For example, Ragan et al. [RGT15] performed a controlled study to evaluate how the presence of a visual history aid affects the memory of a user with recalling the details of their exploration process, finding that providing such an aide for a user to explore their past interactions

was a substantial benefit. A similar conclusion was reached by Chen et al. [CQW*14] in their propagation-based parametric model approach, also noting that analysts were better able to recall the chronological states of the analysis process with their symbolic model. In terms of future effects, *Wrangler* [KPHH11] enables analysts to understand the effects of a variety of potential operations in the interaction space.

Understanding the cognitive behavior and strategies of users from interaction logs is a further benefit in this area. A think-aloud study by Dou et al. [DJS*09] was able to identify several strategies, methods, and findings of an analysis process through the examination of an analyst's interaction log. Similarly, Blascheck et al. [BVV*18] performed an analysis of interaction logs and eye movement data to identify analyst exploration strategies. An interview study of data analysts by Madanagopal et al. [MRB19] further demonstrated that the need for provenance information changes during the course of the analysis process and dependent upon the role of the analyst who is performing that analysis.

8.5. Interactive Visual Analysis

Beyond these common methods, researchers have also explored other types of data representations and analysis methods for a variety of purposes. There are a wide variety of internal representations and user interactions that can be supported for provenance analysis, as demonstrated by Liu et al [LS10]. This section summarizes publications that represent these additional analysis methods, with an overarching theme of interacting with visual representations of provenance data and generating insight from such interactions.

8.5.1. Semantic Interaction

Semantic interaction is a demonstrational technique for model manipulation introduced by Endert et al [EFN12a]. Analysts provide feedback to an underlying learning routine by demonstrating a relationship that they seek in the visualization, and the system in turn attempts to produce this relationship across all observations by altering the models that produce the visualization. As such, these models are indirectly the altered analysis object, leading to a great deal of overlap with the Model discussion in Section 7.3.1.

The interaction techniques supported by these semantic interaction systems are the direct means by which an analyst evaluates and updates the model state. The variety of supported interactions is quite broad, ranging from interactions on observations in the form of projected glyphs [BLBC12,BNHL14] and table rows [WDC*18] to interactions on the axes [KCPE16] and background [KKW*17] of the plot. These interactions often involve manipulating layout constraints or coordinates [HBR14, KC14], but can also include interactive clustering [HOG*12] or simply highlights regions of interest [FWG09].

8.5.2. Analysis Via Visual Analytics

Semantic interaction represents one portion of the overall space of using visual analytics techniques and systems to analyze provenance data. For example, Blascheck et al. [BBB*16, BJK*16] developed a visual analytics tool to analyze multiple streams

of user interaction data, including audio, video, eye gaze, content of analysis, and more. Similarly, SensePath [NXW*16], *SenseMap* [NXB*16], and *InsideInsights* [MHK*19] provide web-based interfaces for the analysis and exploration of provenance data. These techniques further allow for the exploration of large provenance data collections, which can benefit safety and security tasks [SMvdWvW15] as well as scientific research [GLG*16]. The analysis can also be collaborative, as seen in the study using RCloud by North et al [NSUW15].

A further set of systems made use of graph-based approaches, in which the graph is central as both an interaction target and analysis goal. Continuing the collaborative theme, Mahyar et al. create an evidence graph or collaborative sensemaking, making use of visual analytics techniques on this graph to gain insight from its structure [MT14]. Zhao et al. further discuss the construction of a knowledge graph for the purpose of collaborative analysis [ZGI*18], while *Vizcept* records user insights in a concept graph structure for further analysis [CYM*10].

Using visual analytics approaches for evaluating provenance data is not only limited to the creation of visualization and interaction techniques. Indeed, other systems make use of provenance analysis as an analysis component connecting the visualization output to the underlying models. Setlur et al. [STD19] demonstrate this in their technique using rule-based inferencing to obtain missing information necessary to process natural language queries, making use of concepts known by the system to refine user utterances. A similar approach is used by *Flux Capacitor* [KN19] to reduce queries, prefetch, and cache data to minimize intensive workloads on databases, using knowledge of user interactions to perform these operations.

8.5.3. Generation and Analysis of Visual Design

The analysis of provenance data can also be used to generate new visualization designs, as well as to analyze the quality of existing visualizations. Smuc et al. [SML*09] develop a methodology that makes use of participatory design processes and formative evaluation strategies to design novel analytical tools, while Choe et al. [CLs15] discuss an empirical study to inform the design of systems for personal data systems. In both cases, the proposed and actual interactions of users drove the design process. New visualizations can also be generated automatically from user interactions, as seen in both the *Visualization-by-Sketching* [SK16] and "Visualization by Demonstration" [SKBE17] techniques.

9. Discussion

Figure 3 and 4 provide an interesting overview of the WHY, WHAT, and HOW questions among the papers we surveyed. It was to be expected that "Understanding the User" is the most common analysis goal: besides being an interesting research topic itself (particularly in the field of HCI), it is also the foundation of other analysis goals such as "Adaptive Systems" and "Evaluation of Systems and Algorithms." These two rely on the user knowledge to provide better support and evaluate analysis performance respectively.

While understanding users is undoubtedly useful, there are potential ethical and privacy concerns. For example, while the data

collection and analysis in user studies undergo the scrutiny of ethical approval process, the resulting methods or tools can be potentially used by third parties to recover sensitive data from user provenance or inferring individual characteristics (such as the method described in "Finding Waldo" [BOZ*14]). There is limited work on this so far, and it is a topic that certainly deserves more attention from us as a community.

Adaptive systems is the second most common goal. Understandably, it is over-represented in the IUI community, and to a lesser extent in the CHI community. In addition to the examples discussed in this survey, they are now commonly deployed in smart phone systems with the ability to for example notify users about traffic conditions based on previous travel records and calendar events. However, it can be difficult to create effective adaptive systems and there are many examples of failed attempts, with Microsoft Office *Clippy* being the most well-known example. Therefore, the community needs carefully and thoroughly investigate when and how to adapt. Guidance and adaptive systems can also introduce exploration bias into the sensemaking process if not carefully checked.

Our organization of WHAT (*Sequence*, *Grammar*, *Model*, and *Graph*) is meant to provide a broad overview of existing techniques. We acknowledge that encoding techniques can be imprecisely categorized resulting in overlaps or ambiguities. For example, should a user model based on a Bayesian model (such as the work by Healey and Dennis [HB12]) be categorized under *User Model* or a probabilistic *Graph*? In this survey, we do not clearly define these boundaries and hope that the high-level abstraction of the categories serves the purpose of an overview.

In addition, we recognize that data science "notebooks" such as Jupyter Notebook and RStudio can be considered a form of analysis provenance. Systems like RCloud [NSUW15] that support collaborative analysis with R intrinsically encodes and visualizes user analysis history. While the encoding strategy can be considered as a *Sequence* of commands, we do not include these tools because they are not strictly visualization or visual analytics systems. A more nuanced example that is also not included in this survey is the *Literate visualization* work by Wood et al. [WKD18]. Using a "notebook-like" approach, the authors present a system for documenting the process of designing a visualization. Since the purpose of this system is to record the changes to a visualization design but not to perform analysis over these visualizations, it is out of the consideration of this survey.

Figure 3 also demonstrates the diverse application of provenance analysis. More recent topics, particularly the Model Steering and Active Learning, have already received considerable attention within the provenance research community. The reflects the trend of growing popularity of machine learning-related work in the larger Visualization field, and illustrates the effectiveness of provenance as a viable approach to this type of problems.

Reproducibility is one of the important use cases for provenance. However, we noticed that not many papers make their implementation, methods, or datasets available. This make these works less reproducible. As a field that is a major contributor to the reproducible science research, we as a community should make sure the wide adoption of practices for research transparency and reproducibility, sharing our "research provenance."

9.1. WHEN to Analyze Provenance Data

Along the line of WHY, WHAT, and HOW, a few patterns emerged with respect to WHEN to perform provenance analysis. However, we think the related contents are not as substantial as the other three, so we include it in the discussion here instead of as a individual section. There are three common approaches: **retrospective analyses** that occur after the task is completed, **real-time applications** that happens during analysis, or **hybrid approaches** that mix the two.

Retrospective Analyses. The literature contains numerous examples of techniques for analyzing provenance data from past user sessions. For example, a series of researchers have examined batches of interaction logs to uncover reasoning processes and insights [BH19, DJS*09, FPH19, GGZL15]. The *Inside Insights* [MHK*19] system leveraged observations to generate a report of the user session automatically. When facing the timestamp for the analysis, post-hoc analyses were applied to uncover correlations between user attributes and interaction patterns [BOZ*14, OYC15].

Real-Time Applications. An emerging topic in the provenance literature focuses on leveraging interaction data to perform real-time predictions, system adaptations, or offer guidance to the user. Common examples apply techniques from machine learning and artificial intelligence to make inferences based on behavioral patterns. *ForeCache*, for instance, used provenance data to improve prefetching and system performance [BCS16]. Wall et al. [WBFE17] proposed using real-time interaction to identify and mitigate exploration bias, and Ottley et al. [OGW19] inferred future clicks from real-time observations. Brown et al. [BLBC12] and Endert et al. [EFN12a] demonstrated how mouse interaction could steer data projection models.

Hybrid Approaches. Although less common, the growing availability of large interaction datasets has led to hybrid approaches that leverage data from past sessions to refine the analytic process. One example, *Scented Widgets* [WHA07], exploited historical usage data to improve interface controls. Other work by Fan et al. created machine learning models based on past user data to improve the accuracy and speed of a real-time brushing tool [FH18].

10. Opportunities for Future Research

As a rapidly-growing field, provenance analysis faces many open challenges and research opportunities. In this section, we organize these following the WHY, WHAT, and HOW structure of the paper, with some additional discussions on provenance standards. The list is not meant to be exhaustive, but we aim to cover the critical gaps and emerging topics that we believe deserve more attention. Some are similar to those discussed in the recent literature, such as the report from the Dagstuhl Seminar on Provenance [FJKTX19b].

10.1. WHY

There are still many opportunities for further development within the use cases discussed in Section 6. For example, user intent modeling (Section 6.1) is still an open problem, and a fully automated solution is unlikely in near future. Any progress in improving or

supporting this process can not only enhance the efficacy and efficiency of visualization tools, but also can help to advance applications such as evaluation (Section 6.2), adaptive systems (Section 6.3), and storytelling (Section 6.6).

Among the use cases, we see a growing interest in supporting machine learning, demonstrated by the works discussed in the Model Steering section (6.4). It is not far-fetched to envision that such techniques can also be beneficial to critical, emerging topics in the machine learning field, such as model transparency and explainable AI.

Another potentially impactful use case is *exploration optimization*: analyses, particularly those involved in data science, often require exploration in three different spaces: the *data space*, which is the input data, the *approach space*, which consists of all possible approaches to the problem, and the *parameter space*, which is part of the approach space and can have a large impact on the effectiveness of the approach. These three spaces share two properties: (1) often large in size and (2) of high dimensionality. Without a systematic approach, exploration within these spaces can be time-consuming and ineffective.

There are a few attempts that aim to guide a user during such exploration by analyzing the results of completed prior explorations. The *Evolutionary Visual Exploration (EVE)* system [BCBL13] uses an evolution algorithm that learns optimization criteria from user interactions to help user explore large multivariate datasets. The *GEMSe* [FMH16] system provides users with the provenance of parameter space exploration of multi-channel segmentation algorithms, which was shown to considerably improve the efficiency of finding suitable parameters. The open challenge is to have more general methods that can be adapted to various data, approach, and parameter spaces. A similar effort is the *Visual Parameter Space Analysis (VPSA)* [SHB*14], designed to find the optimal settings for simulation models. The main difference is that *VPSA* focuses on visualization parameters, and the quality of the results can often be quantified. In contrast, exploration optimization also considers the data space, and the optimization function has to be learned from analysis provenance. Given the prevalence of exploratory data analysis, such a solution will have significant impact in many fields beyond provenance and visualization.

10.2. WHAT

Many of the long-standing challenges in provenance analysis relate to the multi-layer nature of provenance: from the low-level system logs such as mouse movement, through more abstract and system-independent analysis actions such as searching and comparison, and to reasoning-related provenance such as insight and rationale [RESC15]. This adds an extra dimension to the four encoding schemes discussed in Section 7: each scheme, such as a sequence or a graph, can have multiple layers. For example, a provenance sequence can have several layers, such as *events* (system log), (user) *actions*, *sub-tasks*, and *tasks* [GZ08]. While most existing methods, such as those covered in Section 7.1, choose to target a single layer, there are efforts that investigate the connections between the layers. For an example, see the encoding schemes that uses taxonomy-based abstraction (Section 7.1.4).

It has been argued that it is increasingly difficult to capture provenance when moving up through the layers: it is often straightforward to record system logs, but capturing abstract actions is much harder [GZ08]. Further, there is no effective solution to automatically capture user insights or reasoning process. However, many of the use cases discussed in Section 6 heavily rely on high-level provenance. For example, adaptive systems require the knowledge of user analysis preference and strategy to provide customization or prediction, and user modeling represents a similar case. Because of this difficulty, there are few successful examples. One successful example is the *InsideInsights* system [MHK*19] that utilizes this hierarchy for report generation. Such analysis is often done manually, such as the study by Dou et al. [DJS*09], and is usually time consuming.

The existing efforts either attempt to infer higher-level provenance automatically from lower-level data or encourage externalization through annotation and note taking. Neither approach has been very successful, both because of the poor inference performance (even human experts can only achieve about 2/3 accuracy) and because of the distraction and cognitive load introduced by the externalization. While there are conceptual models, such as the one by Bors et al. [BWD*19], that can guide such inference, there are still no concrete implementations, and its efficacy is still an open question.

There is a new provenance encoding paradigm that may bring a new perspective to some existing challenges. This approach models provenance as a high-dimensional vector sequence: each vector in such a sequence is a step in the analysis interaction, and the dimensions of the vector are the information that constitutes provenance, including the visualization state, data displayed, user interaction, and any other provenance captured. This encoding approach provides new perspectives for examine provenance data, such as transforming provenance analysis into problems that may have been studied in related fields. For one example, provenance visualization can be performed via projecting a vector sequence from high-dimensional space to 2D (or 3D) space, providing a perspective that is different from treating the provenance as a sequence or graph. This has been shown to be effective for analyzing dynamic networks [vHBv16], in which each snapshot of the network at a time point is represented as a high-dimensional vector. This approach can be a potential fit for the exploration optimization problem described earlier, with the provenance vector capturing the information from the data, approach, and parameter space.

10.3. HOW

There are many opportunities to apply different types of analysis techniques, such as those discussed in Section 8, to provenance use cases, particularly the more difficult ones such as user modeling and exploration optimization. Many of the existing works rely on the relatively simple form of these techniques, such as *k*-means clustering, and have already demonstrated considerable improvements. The rapid development of new techniques from fields like machine learning provide an ever-increasing collection of new power tools to tackle provenance challenges. Machine learning methods based on the deep neural networks proved to be wildly successful in the breakthroughs of some of the long-standing ma-

chine learning challenges. However, these methods often require large amount of data for initial training, and it will be interesting to explore how such methods can be adapted to provenance problems, in which user data is usually much smaller than those available from an image or text collection.

There are still some fundamental provenance analysis problems that await better solutions. One such problem is provenance *chunking*, which groups the steps in an interaction sequence based on the analysis actions that they belong to. In stock analysis for example, all of the steps involved in trend analysis would be in one chunk, and steps related to comparing company performance would be in a different chunk. Chunking is fundamental to other provenance analyses. For example, once chunking is completed, it will be easier to infer the higher-level abstract task for each chunk, which is an important open question itself.

Chunking can be treated as either a classification problem, i.e., whether there should be a break between the current and next step, or as a clustering problem, i.e., all the consecutive steps belong to the same cluster will be in one chunk. The classification approach is suitable when real-time chunking is needed, as it only requires the knowledge of a number of prior steps, whereas the clustering may work better in post-hoc analysis as the global knowledge of all the steps may help produce higher quality clustering.

The boundary between chunks may not always be clear-cut. There can be a transition between two groups of actions, and the steps within such transitions are better served with probabilities of belonging to either of the chunks. Chunking can also be hierarchical due to the hierarchical nature of analysis task. For example in stock analysis, steps belonging to price trend analysis can be separated from company performance comparisons. Within the latter, steps related to company financial status can be separated from those pertaining to company executive track records. Such sub-divisions can be recursive and have multiple levels.

Existing works related to chunking, such as those discussed in Section 7.4.2, often link to the graphical and multi-layer nature of provenance data. It is possible that these properties can be exploited to improve chunking performance. The report from the recent Dagstuhl Seminar on Provenance [FJKTX19b] contains a section on this topic ("Machine Learning and Provenance in Visual Analytics").

10.4. Provenance Standard and Cross-Tool Integration

Most of the techniques and systems reviewed here contain some provenance format that is specifically designed for the tool and/or analysis need. However, real-world analysis is rarely completed within a single tool or with a single dataset. Instead, it often involves integrating provenance data from multiple tools, requiring a provenance data format that is compatible with a wide range of tools. This is currently not possible, and leads to the need for a common standard that supports provenance capturing from multiple applications, as well as inter-application provenance analysis.

Ideally, such a standard will accommodate the requirements of a variety of use cases, such as those discussed in Section 6. It should also support the diverse types of provenance information and methods for encoding them (as discussed in Section 7, including the multi-layer structure of provenance. Finally, such a format must be compatible with various downstream analyses, such as those discussed in Section 8.

There is likely to be a trade-off between compatibility and semantic richness. At one end of this spectrum, provenance is captured as a screenshot, which makes it compatible with all visualization systems. However, the amount of semantic information that can be used for down-stream analysis is very limited. At the other end of the spectrum, a provenance standard meets all of the requirements discussed earlier and can store detailed descriptions such as system state, data involved, and user actions and intentions. These are valuable to down-stream analysis, but designing such a rich standard that can accommodate different systems, data, and user requirements presents difficulties.

This will be a challenging task, but essential and beneficial to the entire field and the wider research community. Any progress on this front will be valuable, encouraging new collaborations and enabling new research. It can start with a simpler standard or a set of related standards, or alternatively can build upon existing standards, such as the provenance formats used by popular visual analytics tools such as W3C PROV. Additionally, it may be useful to learn from the experience of designing similar standards in fields beyond visual analysis. This was also discussed at the recent Dagstuhl Seminar on provenance, and the details are recorded in its report [FJKTX19b].

11. Conclusion

In this survey, we present a systematic review of provenance-related research in data visualization and visual analytics. Focusing on *the analysis of provenance data*, we explore three primary questions: (1) WHY analyze provenance data, (2) WHAT provenance data to encode, and how to encode it, and (3) HOW to analyze provenance data. However, the work in the paper provides only a narrow perspective of analytic provenance. As outlined by Ragan et al. [RESC15] the field of provenance extends beyond meta-analysis and encompasses other goals such as collaborative communication and presentation.

Beyond visualization, the analysis of user interaction is an active area of research in a variety of domains such as Artificial Intelligence, Machine Learning, Databases, and Human-Computer Interaction. The body of work in this manuscript reveals a clear trend of adapting and extending knowledge from other fields. For example, a large portion of the literature employed either machine learning, dimensionality reduction, or even signal processing techniques. These techniques are essential to the provenance analysis innovation, and the research development in these fields can bring considerable improvement to provenance analysis. Similarly, as detailed in Section 7, there are many types of provenance data that we can collect, ranging from sequence and scripts to graphs and images. Storing and managing such data can be an issue (e.g., *Glassbox*), and specialized methods are often required from other fields for their analysis.

Looking forward, capturing and examining provenance data holds the potential to realize many of the untapped capabilities of visualization systems, such as expanding its role in data science and

contributing to the current problems in other fields such as explainable AI. It will be beneficial for the provenance community to expand collaboration across disciplines, increase the awareness of the provenance literature, and jointly develop new techniques that can be applied to provenance analysis. There is much work ahead for the visualization community, and we detail some of these research opportunities in Section 10. We hope that this report will serve as a central resource for researchers and inspires new investigations that build on the surveyed literature.

12. Acknowledgment

Table 1: The 105 articles we reviewed in terms of WHY, WHAT, and HOW to analyze provenance data. The table is sorted first by WHY and then by WHAT. It shows single tags as well as more than one tag per category and paper as well as uncategorized ones, when no WHY, WHAT, or HOW was applicable.

Paper	WHY						WHAT				HOW				
	Adaptive Systems	Evaluation of Systems and Algorithms	Model Steering and Active Learning	Replication	Report Generation	Understanding the User	Grammar	Graph	Model	Sequence	Classification Models	Pattern Analysis	Probabilistic Models / Prediction	Program Synthesis	Interactive Visual Analysis
Srinivasan et al. (2018) [SPEB18]	X	X					X							X	
Setlur et al. (2016) [SBT*16b]	X		X				X				X		X		
Dabek and Caban (2016) [DC16]	X					X	X				X				
Mutlu et al. (2016) [MVT16]	X					X	X					X	X		
Chen et al. (2014) [CQW*14]	X						X							X	
Koch et al. (2014) [KBGE09]	X						X							X	
Scheepens et al. (2015) [SMvdWvW15]	X	X		X	X			X							X
Shrinivasan and van Wijk (2009) [SvW09]	X			X	X			X				X			X
Gotz and Wen (2009) [GW09]	X					X		X				X	X		X
Endert (2014) [End14]	X		X			X			X						X
Endert et al. (2012) [EFN12a]	X		X						X				X		
MacInnes et al. (2010) [MSW10]	X								X		X	X			
Boukhelifa et al. (2013) [BCBL13]	X				X				X				X		
Endert (2015) [ECNZ15]	X								X		X				
Ottley et al. (2019) [OGW19]	X								X				X		
Gotz et al. (2016) [GSC16]	X								X		X				
Healey and Bennis (2012) [HB12]	X					X			X		X				
Hu et al. (2019) [HBL*19]	X								X		X				
Correll and Gleicher (2016) [CG16]	X	X								X	X	X			
Setlur et al. (2019) [STD19]	X									X					X
Steichen et al. (2014) [SCC14]	X					X				X	X				
Guo et al. (2019) [GDM*19]	X									X			X		
Wang et al. (2014) [WCW*14]	X	X										X			
Lee et al. (2019) [LDH*19]	X		X		X						X	X			
Ceneda et al. (2019) [CGM19]	X		X			X			X		X	X			
Ragan et al. (2015) [RESC15]	X		X			X			X		X	X			
Micallef et al. (2017) [MSM*17]	X		X			X			X				X		
Wegba et al. (2018) [WLLW18]	X				X						X				X
Toker et al. (2014) [TSG*14]	X					X					X				
Steichen and Conati (2013) [SCC13]	X					X					X		X		
Khan and Nandi (2019) [KN19]	X					X									X
Sacha et al. (2018) [SKB*18]	X										X				
Walch et al. (2019) [WSL*19]	X							X							X
Dextras-Romagnino and Munzner (2019) [DM19]		X				X		X				X			
Ren et al. (2019) [RLB19]		X					X					X			
Bors et al. (2019) [BGM19]		X						X				X			
Shadoan and Weaver (2013) [SW13]		X						X							X
Gomez and Laidlaw (2012) [GL12]		X			X	X			X				X		
Liu et al. (2017) [LWD*17]		X		X						X	X	X			
Smuc et al. (2009) [SML*09]		X		X						X					X
Bylinksii et al. (2017) [BKO*17]		X				X				X	X	X			
Fröhler et al. (2016) [FMH16]		X								X	X				
Blascheck et al. (2016) [BJK*16]		X								X					X

Table 1: The 105 articles we reviewed in terms of WHY, WHAT, and HOW to analyze provenance data. The table is sorted first by WHY and then by WHAT. It shows single tags as well as more than one tag per category and paper as well as uncategorized ones, when no WHY, WHAT, or HOW was applicable.

Article	Adaptive Systems	Evaluation of Systems and Algorithms	Model Steering and Active Learning	Replication	Report Generation	Understanding the User	Grammar	Graph	Model	Sequence	Classification Models	Pattern Analysis	Probabilistic Models / Prediction	Program Synthesis	Interactive Visual Analysis
Moritz et al. (2015) [MHHH15]			●				●			●		●			
Wongsuphasawat et al. (2016) [WMA*16]			●				●				●			●	
Muthumanickam et al. (2016) [MVCJ16]			●				●					●		●	
Hoque et al. (2018) [HSTD18]			●				●						●		
Weaver (2009) [Wea09]			●				●							●	
Ferreira et al. (2013) [FPV*13]			●				●							●	
Kadivar et al. (2009) [KCD*09]			●				●							●	
Shrinivasan et al. (2009) [SGL09]			●					●						●	
Schlinder et al. (2013) [SWR*13]			●					●							●
Kwon et al. (2017) [KKW*17]			●		●				●			●			
Bradel et al. (2014) [BNHL14]			●			●			●					●	
Pezzotti et al. (2015) [RWF*13b]			●						●		●		●		
Brown et al. (2012) [BLBC12]			●						●		●				
Endert et al. (2012) [EFN12a]			●						●		●			●	
Cook et al. (2015) [CCI*15]			●						●			●			
Kunkel et al. (2017) [KLZ17]			●						●				●		
Kim et al. (2016) [KCPE16]			●						●						●
Fuchs et al. (2009) [FWG09]			●						●	●	●				
Brown et al. (2014) [BOZ*14]			●			●				●	●		●		
Cavallo and Demiralp (2018) [CD18]			●							●				●	
Ribicic et al. (2013) [RWF*13a]			●							●					●
Hossain et al. (2012) [HOG*12]			●							●	●				●
Sherkat et al. (2018) [SNMM18]			●								●				●
Gao et al. (2015) [GDA*15]			●												●
Hottelier et al. (2014) [HBR14]			●												●
Callahan et al. (2006) [CFS*06]				●				●						●	
Rübel and Bowen (2018) [RB18]				●			●				●			●	
Kandel et al. (2011) [KPHH11]				●			●							●	
Chung et al. (2010) [CYM*10]				●	●			●							●
Gotz et al. (2017) [GSC*17]				●		●			●			●			
Xu et al. (2018) [XBL*18]				●	●					●	●		●		
Loorak et al. (2018) [LTC18]				●		●				●	●				
Smith et al. (2018) [SLMK18]				●		●				●				●	
Stitz et al. (2019) [SGP*19]				●						●		●			
Andrienko et al. (2011) [AAM*11]				●											●
Cappers and van Wijk (2018) [CvW18]				●								●			
Zgraggen et al. (2015) [ZDFD15]				●										●	
Ragan et al. (2015) [RGT15]					●	●	●							●	●
Walker et al. (2013) [WSD*13]					●		●							●	
Porteous et al. (2010) [PCC10]					●	●		●						●	
Mathisen et al. (2019) [MHK*19]					●			●							●
Gratzl et al. (2016) [GLG*16]					●			●							●
Choe et al. (2015) [CLs15]					●	●				●	●				●

Table 1: The 105 articles we reviewed in terms of WHY, WHAT, and HOW to analyze provenance data. The table is sorted first by WHY and then by WHAT. It shows single tags as well as more than one tag per category and paper as well as uncategorized ones, when no WHY, WHAT, or HOW was applicable.

	WHY						WHAT				HOW				
	Adaptive Systems	Evaluation of Systems and Algorithms	Model Steering and Active Learning	Replication	Report Generation	Understanding the User	Grammar	Graph	Model	Sequence	Classification Models	Pattern Analysis	Probabilistic Models / Prediction	Program Synthesis	Interactive Visual Analysis
Willett et al. (2013) [WGS*13]					■						■				
Bors et al. (2019) [BWD*19]						■		■			■				■
Wall et al. (2018) [WDC*18]						■			■		■				
Liu et al. (2017) [LKD*17]						■				■	■	■			
Cho et al. (2017) [CWK*17]						■				■	■	■			
Blascheck (2018) [BVV*18]						■				■	■			■	
Dou et al. (2009) [DJS*09]						■				■	■			■	
Boukhelifa (2019) [BBT*19]						■				■	■				
Guo et al. (2015) [GGZL15]						■				■	■				
Feng et al. (2019) [FPH19]						■				■			■		
Battle and Heer (2019) [BH19]						■				■		■	■		
Nguyen et al. (2018) [NTA*18]						■				■		■	■		
Wei et al. (2012) [WSSM12]						■				■			■		
Blascheck et al. (2016) [BBB*16]						■				■					■
Toker et al. (2017) [TLC17]						■					■		■		
Yu and Silva (2020) [YS20]						■					■				
Mannino and Abouzied (2019) [MA19]						■						■	■		
Kodagoda et al. (2013) [KAW*13]						■						■			
Madanagopal et al. (2019) [MRB19]						■								■	
Kondo and Collins (2014) [KC14]									■						■

References

[AAM*11] ANDRIENKO G., ANDRIENKO N., MLADENOV M., MOCK M., POELITZ C.: Identifying Place Histories from Activity Traces with an Eye to Parameter Impact. *TVCG 18*, 5 (2011), 675–688. 19

[Anc12] ANCONA B.: Sensemaking: Framing and acting in the unknown in the handbook of teaching leadership: Knowing, doing and being. scott, s., nohria, n., khurana, r, 2012. 1

[APM*11] ANDERSON E. W., POTTER K. C., MATZEN L. E., SHEPHERD J. F., PRESTON G. A., SILVA C. T.: A user study of visualization effectiveness using eeg and cognitive load. *Computer graphics forum 30*, 3 (2011), 791–800. 7

[AZ12] AGGARWAL C. C., ZHAI C.: A Survey of Text Classification Algorithms. In *Mining Text Data*, Aggarwal C. C., Zhai C., (Eds.). Springer US, Boston, MA, 2012, pp. 163–222. URL: http://link.springer.com/10.1007/978-1-4614-3223-4_6, doi:10.1007/978-1-4614-3223-4_6. 10

[BBB*16] BLASCHECK T., BECK F., BALTES S., ERTL T., WEISKOPF D.: Visual analysis and coding of data-rich user behavior. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2016), pp. 141–150. ISSN: null. doi:10.1109/VAST.2016.7883520. 7, 13, 20

[BBDW14] BECK F., BURCH M., DIEHL S., WEISKOPF D.: The State of the Art in Visualizing Dynamic Graphs. In *Eurographics Conference on Visualization (EuroVis)* (2014), p. 21. 2, 3

[BBT*19] BOUKHELIFA N., BEZERIANOS A., TRELEA I. C., PERROT N. M., LUTTON E.: An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–14. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300874, doi:10.1145/3290605.3300874. 7, 11, 20

[BCBL13] BOUKHELIFA N., CANCINO W., BEZERIANOS A., LUTTON E.: Evolutionary Visual Exploration: Evaluation With Expert Users. *Computer Graphics Forum 32*, 3pt1 (June 2013), 31–40. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=88800174&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12090. 9, 15, 18

[BCN*19] BATTLE L., CROUSER R. J., NAKESHIMANA A., MONTOLY A., CHANG R., STONEBRAKER M.: The Role of Latency and Task Complexity in Predicting Visual Search Behavior. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. URL: https://ieeexplore.ieee.org/document/8809742/, doi:10.1109/TVCG.2019.2934556. 10, 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 - SIGMOD '16* (San Francisco, California, USA, 2016), ACM Press, pp. 1363–1375. URL: http://dl.acm.org/citation.cfm?doid=2882903.2882919, doi:10.1145/2882903.2882919. 5, 7, 12, 14

[BF05] BOSE RAJENDRA, FREW JAMES: Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR) 37*, 1 (Mar. 2005). URL: https://dl.acm.org/doi/abs/10.1145/1057977.1057978. 2

[BGM19] BORS C., GSCHWANDTNER T., MIKSCH S.: Capturing and Visualizing Provenance From Data Wrangling. *IEEE Computer Graphics and Applications 39*, 6 (Nov. 2019), 61–75. doi:10.1109/MCG.2019.2941856. 10, 18

[BH19] BATTLE L., HEER J.: Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum 38*, 3 (2019), 145–159. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13678, doi:10.1111/cgf.13678. 6, 14, 20

[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. 5, 7, 13, 18

[BKO*17] BYLINSKII Z., KIM N. W., O'DONOVAN P., ALSHEIKH S., MADAN S., PFISTER H., DURAND F., RUSSELL B., HERTZMANN A.: Learning Visual Importance for Graphic Designs and Data Visualizations. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST '17* (2017), 57–69. arXiv:1708.02660. URL: http://arxiv.org/abs/1708.02660, doi:10.1145/3126594.3126653. 4, 5, 7, 11, 18

[BLBC12] BROWN E. T., LIU J., BRODLEY C. E., CHANG R.: Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 83–92. 5, 9, 13, 14, 19

[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. URL: http://ieeexplore.ieee.org/document/6634168/, doi:10.1109/TVCG.2013.124. 7

[BNHL14] BRADEL L., NORTH C., HOUSE L., LEMAN S.: Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2014), pp. 163–172. ISSN: null. doi:10.1109/VAST.2014.7042492. 9, 13, 19

[BOZ*14] 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. 4, 6, 10, 11, 14, 19

[BTC19] BUNEMAN PETER, TAN WANG-CHIEW: Data Provenance: What next? *ACM SIGMOD Record 47*, 3 (Feb. 2019), 5–16. doi:https://doi.org/10.1145/3316416.3316418. 2

[BVV*18] BLASCHECK T., VERMEULEN L., VERMEULEN J., PERIN C., WILLETT W., ERTL T., CARPENDALE S.: Exploration Strategies for Discovery of Interactivity in Visualization. *TVCG 25*, 2 (2018), 1407–1420. 13, 20

[BWD*19] BORS C., WENSKOVITCH J., DOWLING M., ATTFIELD S., BATTLE L., ENDERT A., KULYK O., LARAMEE R. S.: A Provenance Task Abstraction Framework. *IEEE Computer Graphics and Applications 39*, 6 (Nov. 2019), 46–60. doi:10.1109/MCG.2019.2945720. 11, 15, 20

[BYC*19] BROWN E. T., YARLAGADDA S., COOK K. A., CHANG R., ENDERT A.: Modelspace: Visualizing the trails of data models in visual analytics systems. In *IEEE Visualization Workshop on Machine Learning from User Interactions for Visualization and Analytics* (2019). 9

[CAS*18] COLLINS C., ANDRIENKO N., SCHRECK T., YANG J., CHOO J., ENGELKE U., JENA A., DWYER T.: Guidance in the humanâĂŞ-machine analytics process. *Visual Informatics 2*, 3 (Sept. 2018), 166–180. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468502X1830041X, doi:10.1016/j.visinf.2018.09.003. 5

[CBY10] CHEN Y., BARLOWE S., YANG J.: Click2annotate: Automated insight externalization with rich semantics. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 155–162. 5

[CCI*15] COOK K., CRAMER N., ISRAEL D., WOLVERTON M., BRUCE J., BURTNER R., ENDERT A.: Mixed-initiative visual analytics using task-driven recommendations. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2015), pp. 9–16. ISSN: null. doi:10.1109/VAST.2015.7347625. 9, 19

[CCT07] CHENEY J., CHITICARIU L., TAN W.-C.: Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases 1*, 4 (2007), 145–159. URL: http://www.nowpublishers.com/article/Details/DBS-006, doi:10.1561/1900000006. 3

[CD18] CAVALLO M., DEMIRALP Ã.: Track Xplorer: A System for Visual Analysis of Sensor-based Motor Activity Predictions. *Computer Graphics Forum 37*, 3 (2018), 339–349. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13424, doi:10.1111/cgf.13424. 6, 19

[CF17] CHIRIGATI F., FREIRE J.: Provenance and Reproducibility. In *Encyclopedia of Database Systems*, Liu L., Ãŭzsu M. T., (Eds.). Springer, New York, NY, 2017, pp. 1–5. URL: https://doi.org/10.1007/978-1-4899-7993-3_80747-1, doi:10.1007/978-1-4899-7993-3_80747-1. 2

[CFS*06] CALLAHAN S. P., FREIRE J., SANTOS E., SCHEIDEGGER C. E., SILVA C. T., VO H. T.: Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (2006), pp. 745–747. 5, 10, 19

[CG16] CORRELL M., GLEICHER M.: The semantics of sketch: Flexibility in visual query systems for time series data. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2016), pp. 131–140. ISSN: null. doi:10.1109/VAST.2016.7883519. 7, 12, 18

[CGM*17] CENEDA D., GSCHWANDTNER T., MAY T., MIKSCH S., SCHULZ H.-J., STREIT M., TOMINSKI C.: Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 111–120. doi:10.1109/TVCG.2016.2598468. 5

[CGM19] CENEDA D., GSCHWANDTNER T., MIKSCH S.: A Review of Guidance Approaches in Visual Data Analysis: A Multifocal Perspective. *Computer Graphics Forum 38*, 3 (2019), 861–879. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13730, doi:10.1111/cgf.13730. 5, 18

[CLs15] CHOE E. K., LEE B., SCHRAEFEL M.: Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Computer Graphics and Applications 35*, 4 (July 2015), 28–37. doi:10.1109/MCG.2015.51. 7, 10, 13, 19

[CQW*14] CHEN Y. V., QIAN Z. C., WOODBURY R., DILL J., SHAW C. D.: Employing a Parametric Model for Analytic Provenance. *ACM Trans. Interact. Intell. Syst. 4*, 1 (Apr. 2014), 6:1–6:32. URL: http://doi.acm.org/10.1145/2591510, doi:10.1145/2591510. 8, 12, 13, 18

[CvW18] CAPPERS B. C., VAN WIJK J. J.: Exploring Multivariate Event Sequences Using Rules, Aggregations, and Selections. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 532–541. doi:10.1109/TVCG.2017.2745278. 12, 19

[CWK*17] CHO I., WESSLEN R., KARDUNI A., SANTHANAM S., SHAIKH S., DOU W.: The Anchoring Effect in Decision-Making with Visual Analytics. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology* (Oct. 2017). URL: https://ieeexplore.ieee.org/document/8585665, doi:10.1109/VAST.2017.8585665. 6, 12, 20

[CYM*10] CHUNG H., YANG S., MASSJOUNI N., ANDREWS C., KANNA R., NORTH C.: Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 107–114. 9, 13, 19

[DC16] 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* (Jan. 2016). URL: https://www.ncbi.nlm.nih.gov/pubmed/27514057, doi:10.1109/TVCG.2016.2598471. 8, 10, 11, 12, 18

[DFAB03] DIX A., FINLAY J., ABOWD G. D., BEALE R.: *Human-Computer Interaction*, 3 edition ed. Prentice Hall, Harlow, England ; New York, Sept. 2003. 2

[DHRL*12] DUNNE C., HENRY RICHE N., LEE B., METOYER R., ROBERTSON G.: Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 1663–1672. 10

[DJS*09] DOU W., JEONG D. H., STUKES F., RIBARSKY W., LIPFORD H. R., CHANG R.: Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications 29*, 3 (May 2009), 52–61. doi:10.1109/MCG.2009.49. 4, 6, 11, 13, 14, 15, 20

[DM19] DEXTRASÂĂŘROMAGNINO K., MUNZNER T.: Segmentifier: Interactive Refinement of Clickstream Data. *Computer Graphics Forum 38*, 3 (2019), 623–634. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13715, doi:10.1111/cgf.13715. 10, 12, 18

[ECNZ15] ENDERT A., CHANG R., NORTH C., ZHOU M.: Semantic Interaction: Coupling Cognition and Computation through Usable Interactive Analytics. *IEEE Computer Graphics and Applications 35*, 4 (July 2015), 94–99. doi:10.1109/MCG.2015.91. 5, 11, 18

[EFN12a] ENDERT A., FIAUX P., NORTH C.: Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *EEE Transactions on Visualization and Computer Graphics* (Dec. 2012). URL: https://www.cc.gatech.edu/~aendert3/resources/Endert_TVCG2012_.pdf, doi:10.1109/TVCG.2012.260. 5, 13, 14, 18, 19

[EFN12b] ENDERT A., FIAUX P., NORTH C.: Semantic interaction for visual text analytics. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (Austin, Texas, USA, 2012), ACM Press, p. 473. URL: http://dl.acm.org/citation.cfm?doid=2207676.2207741, doi:10.1145/2207676.2207741. 5, 9

[EHM*11] ENDERT A., HAN C., MAITI D., HOUSE L., NORTH C.: Observation-level interaction with statistical models for visual analytics. In *2011 IEEE conference on visual analytics science and technology (VAST)* (2011), IEEE, pp. 121–130. 9

[End14] ENDERT A.: Semantic Interaction for Visual Analytics: Toward Coupling Cognition and Computation. *IEEE Computer Graphics and Applications 34*, 4 (July 2014), 8–15. doi:10.1109/MCG.2014.73. 18

[FH18] FAN C., HAUSER H.: Fast and Accurate CNN-based Brushing in Scatterplots. *Computer Graphics Forum 37*, 3 (June 2018), 111–120. URL: http://doi.wiley.com/10.1111/cgf.13405, doi:10.1111/cgf.13405. 5, 14

[FJKTX19a] FEKETE J.-D., JANKUN-KELLY T. J., TORY M., XU K.: Provenance Analysis for Sensemaking. *IEEE Computer Graphics and Applications 39*, 6 (Nov. 2019), 27–29. doi:10.1109/MCG.2019.2945378. 3

[FJKTX19b] FEKETE J.-D., JANKUN-KELLY T. J., TORY M., XU K.: Provenance and logging for sense making (dagstuhl seminar 18462). *Dagstuhl Reports 8*, 11 (2019), 35–62. URL: http://drops.dagstuhl.de/opus/volltexte/2019/10355, doi:10.4230/DagRep.8.11.35. 2, 14, 16

[FKSS08] FREIRE J., KOOP D., SANTOS E., SILVA C. T.: Provenance for Computational Tasks: A Survey. *Computing in Science Engineering 10*, 3 (May 2008), 11–21. doi:10.1109/MCSE.2008.79. 2

[FMH16] FRÃŭHLER B., MÃŭLLER T., HEINZL C.: GEMSe: Visualization-Guided Exploration of Multi-channel Segmentation Algorithms. *Computer Graphics Forum 35*, 3 (June 2016), 191–200. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=116877247&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12895. 15, 18

[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. 4, 6, 7, 12, 14, 20

[FPV*13] FERREIRA N., POCO J., VO H. T., FREIRE J., SILVA C. T.: Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2149–2158. doi:10.1109/TVCG.2013.226. 8, 12, 19

[FWG09] FUCHS R., WASER J., GROLLER M. E.: Visual Human+Machine Learning. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (Nov. 2009), 1327–1334. doi:10.1109/TVCG.2009.199. 7, 9, 13, 19

[GDA*15] GAO T., DONTCHEVA M., ADAR E., LIU Z., KARAHALIOS K. G.: DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology - UIST '15* (Daegu, Kyungpook, Republic of Korea, 2015), ACM Press, pp. 489–500. URL: http://dl.acm.org/citation.cfm?doid=2807442.2807478, doi:10.1145/2807442.2807478. 19

[GDM*19] GUO S., DU F., MALIK S., KOH E., KIM S., LIU Z., KIM D., ZHA H., CAO N.: Visualizing Uncertainty and Alternatives in Event Sequence Predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–12. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300803, doi:10.1145/3290605.3300803. 5, 6, 12, 18

[GGZL15] 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* (Aug. 2015). URL: https://ieeexplore.ieee.org/abstract/document/7192662, doi:10.1109/TVCG.2015.2467613. 7, 10, 14, 20

[GHS12] GULWANI S., HARRIS W. R., SINGH R.: Spreadsheet data manipulation using examples. *Communications of the ACM 55*, 8 (2012), 97–105. 7

[GL12] GOMEZ S., LAIDLAW D.: Modeling task performance for a crowd of users from interaction histories. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (Austin, Texas, USA, 2012), ACM Press, p. 2465. URL: http://dl.acm.org/citation.cfm?doid=2207676.2208412, doi:10.1145/2207676.2208412. 4, 5, 6, 18

[GLG*16] GRATZL S., LEX A., GEHLENBORG N., COSGROVE N., STREIT M.: From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum 35*, 3 (June 2016), 491–500. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=116877233&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12925. 5, 10, 13, 19

[GNRM08] GARG S., NAM J. E., RAMAKRISHNAN I., MUELLER K.: Model-driven visual analytics. In *2008 IEEE Symposium on Visual Analytics Science and Technology* (2008), IEEE, pp. 19–26. 8

[GSC16] GOTZ D., SUN S., CAO N.: Adaptive Contextualization: Combating Bias During High-Dimensional Visualization and Data Selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16* (Sonoma, California, USA, 2016), ACM Press, pp. 85–95. URL: http://dl.acm.org/citation.cfm?doid=2856767.2856779, doi:10.1145/2856767.2856779. 4, 9, 11, 18

[GSC*17] GOTZ D., SUN S., CAO N., KUNDU R., MEYER A.-M.: Adaptive Contextualization Methods for Combating Selection Bias During High-Dimensional Visualization. *ACM Trans. Interact. Intell. Syst. 7*, 4 (Nov. 2017), 17:1–17:23. URL: http://doi.acm.org/10.1145/3009973, doi:10.1145/3009973. 11, 19

[Gul11] GULWANI S.: Automating string processing in spreadsheets using input-output examples. *ACM Sigplan Notices 46*, 1 (2011), 317–330. 7

[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. event-place: Sanibel Island, Florida, USA. URL: http://doi.acm.org/10.1145/1502650.1502695, doi:10.1145/1502650.1502695. 5, 18

[GZ08] GOTZ D., ZHOU M. X.: Characterizing usersâĂŹ visual analytic activity for insight provenance. In *2008 IEEE Symposium on Visual Analytics Science and Technology* (2008), pp. 123–130. 15

[HB12] HEALEY C., BENNIS B.: Interest Driven Navigation in Visualization. *TVCG 18*, 10 (2012), 1744–1756. 9, 12, 14, 18

[HBL*19] HU K., BAKKER M. A., LI S., KRASKA T., HIDALGO C.: VizML: A Machine Learning Approach to Visualization Recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–12. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300358, doi:10.1145/3290605.3300358. 18

[HBM*13] HU X., BRADEL L., MAITI D., HOUSE L., NORTH C.: Semantics of directly manipulating spatializations. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2052–2059. 5

[HBR14] HOTTELIER T., BODIK R., RYOKAI K.: Programming by manipulation for layout. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14* (Honolulu, Hawaii, USA, 2014), ACM Press, pp. 231–241. URL: http://dl.acm.org/citation.cfm?doid=2642918.2647378, doi:10.1145/2642918.2647378. 13, 19

[HDBL17] HERSCHEL M., DIESTELKĂĂMPER R., BEN LAHMAR H.: A survey on provenance: What for? What form? What from? *The VLDB Journal 26*, 6 (Dec. 2017), 881–906. URL: https://doi.org/10.1007/s00778-017-0486-1, doi:10.1007/s00778-017-0486-1. 2

[HDG*19] HU K., DEMIRALP Ã., GAIKWAD S. N. S., HULSEBOS M., BAKKER M. A., ZGRAGGEN E., HIDALGO C., KRASKA T., LI G., SATYANARAYAN A.: VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–12. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300892, doi:10.1145/3290605.3300892. 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 (2008), 1189–1196. 5, 10

[HOG*12] HOSSAIN M. S., OJILI P. K. R., GRIMM C., MÃIJLLER R., WATSON L. T., RAMAKRISHNAN N.: Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2829–2838. 9, 13, 19

[HSTD18] HOQUE E., SETLUR V., TORY M., DYKEMAN I.: Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 309–318. doi:10.1109/TVCG.2017.2744684. 8, 12, 19

[IT18] IVIE P., THAIN D.: Reproducibility in Scientific Computing. *ACM Computing Surveys (CSUR) 51*, 3 (July 2018). URL: https://dl.acm.org/doi/abs/10.1145/3186266. 2

[KAW*13] KODAGODA N., ATTFIELD S., WONG B. W., ROONEY C., CHOUDHURY S.: Using Interactive Visual Reasoning to Support Sense-Making: Implications for Design. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2217–2226. doi:10.1109/TVCG.2013.211. 20

[KBGE09] KOCH S., BOSCH H., GIERETH M., ERTL T.: Iterative integration of visual insights during patent search and analysis. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), IEEE, pp. 203–210. 8, 12, 18

[KC14] KONDO B., COLLINS C.: DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec. 2014), 2003–2012. doi:10.1109/TVCG.2014.2346250. 9, 13, 20

[KCD*09] KADIVAR N., CHEN V., DUNSMUIR D., LEE E., QIAN C., DILL J., SHAW C., WOODBURY R.: Capturing and supporting the analysis process. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), Ieee, pp. 131–138. 8, 12, 19

[KCPE16] KIM H., CHOO J., PARK H., ENDERT A.: InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. *IEEE Transactions on*

Visualization and Computer Graphics 22, 1 (Jan. 2016), 131–140. doi:10.1109/TVCG.2015.2467615. 9, 13, 19

[KKW*17] KWON B. C., KIM H., WALL E., CHOO J., PARK H., ENDERT A.: AxiSketcher: Interactive Nonlinear Axis Mapping of Visualizations through User Drawings. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 221–230. doi:10.1109/TVCG.2016.2598446. 9, 13, 19

[KLZ17] KUNKEL J., LOEPP B., ZIEGLER J.: A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17* (Limassol, Cyprus, 2017), ACM Press, pp. 3–15. URL: http://dl.acm.org/citation.cfm?doid=3025171.3025189, doi:10.1145/3025171.3025189. 19

[KN19] KHAN M. A., NANDI A.: Flux capacitors for JavaScript deloreans: approximate caching for physics-based data interaction. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19* (Marina del Ray, California, 2019), ACM Press, pp. 177–185. URL: http://dl.acm.org/citation.cfm?doid=3301275.3302291, doi:10.1145/3301275.3302291. 5, 13, 18

[KPHH11] KANDEL S., PAEPCKE A., HELLERSTEIN J., HEER J.: Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (Vancouver, BC, Canada, 2011), ACM Press, p. 3363. URL: http://dl.acm.org/citation.cfm?doid=1978942.1979444, doi:10.1145/1978942.1979444. 5, 8, 12, 13, 19

[KPS*17] KODAGODA N., PONTIS S., SIMMIE D., ATTFIELD S., WONG B. L. W., BLANDFORD A., HANKIN C.: Using Machine Learning to Infer Reasoning Provenance From User Interaction Log Data: Based on the Data/Frame Theory of Sensemaking. *Journal of Cognitive Engineering and Decision Making 11*, 1 (Mar. 2017), 23–41. URL: http://journals.sagepub.com/doi/10.1177/1555343416672782, doi:10.1177/1555343416672782. 10

[KWRK12] KODAGODA N., WONG B. W., ROONEY C., KHAN N.: Interactive visualization for low literacy users: From lessons learnt to design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI âĂŹ12, Association for Computing Machinery, p. 1159âĂŞ1168. URL: https://doi.org/10.1145/2207676.2208565, doi:10.1145/2207676.2208565. 4

[LAN19] LEE A., ARCHAMBAULT D., NACENTA M.: Dynamic Network Plaid: A Tool for the Analysis of Dynamic Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (Glasgow, Scotland Uk, 2019), ACM Press, pp. 1–14. URL: http://dl.acm.org/citation.cfm?doid=3290605.3300360, doi:10.1145/3290605.3300360. 2

[LDH*19] LEE D. J.-L., DEV H., HU H., ELMELEEGY H., PARAMESWARAN A.: Avoiding drill-down fallacies with *VisPilot*: assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19* (Marina del Ray, California, 2019), ACM Press, pp. 186–196. URL: http://dl.acm.org/citation.cfm?doid=3301275.3302307, doi:10.1145/3301275.3302307. 4, 18

[LKD*17] LIU Z., KERR B., DONTCHEVA M., GROVER J., HOFFMAN M., WILSON A.: CoreFlow: Extracting and Visualizing Branching Patterns from Event Sequences. *Computer Graphics Forum 36*, 3 (June 2017), 527–538. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=123910133&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.13208. 6, 11, 20

[Llo82] LLOYD S.: Least squares quantization in PCM. *IEEE Transactions on Information Theory 28*, 2 (Mar. 1982), 129–137. URL: http://ieeexplore.ieee.org/document/1056489/, doi:10.1109/TIT.1982.1056489. 10

[LPG05] L S., PLALEBETH, GANNONDENNIS: A survey of data provenance in e-science. *ACM SIGMOD Record 34*, 3 (Sept. 2005). URL: https://dl.acm.org/doi/abs/10.1145/1084805.1084812. 2

[LS10] LIU Z., STASKO J.: Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov. 2010), 999–1008. doi:10.1109/TVCG.2010.177. 1, 7, 13

[LTC18] LOORAK M., TORY M., CARPENDALE S.: Change-Catcher: Increasing InterâĂŘauthor Awareness for Visualization Development. *Computer Graphics Forum 37*, 3 (June 2018), 51–62. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=130628132&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.13400. 7, 19

[LW07] LU D., WENG Q.: A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing 28*, 5 (Mar. 2007), 823–870. URL: https://www.tandfonline.com/doi/full/10.1080/01431160600746456, doi:10.1080/01431160600746456. 10

[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. 4, 6, 11, 18

[MA19] MANNINO M., ABOUZIED A.: Is this Real?: Generating Synthetic Data that Looks Real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans LA USA, Oct. 2019), ACM, pp. 549–561. URL: http://dl.acm.org/doi/10.1145/3332165.3347866, doi:10.1145/3332165.3347866. 12, 20

[mer19] *Merriam-Webster Dictionary*, 12th ed. Merriam-Webster, Incorporated, 2019. URL: https://www.merriam-webster.com/dictionary/provenance. 1

[MHHH15] MORITZ D., HALPERIN D., HOWE B., HEER J.: Perfopticon: Visual Query Analysis for Distributed Databases. *Computer Graphics Forum 34*, 3 (June 2015), 71–80. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=108442264&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12619. 6, 12, 19

[MHK*19] MATHISEN A., HORAK T., KLOKMOSE C. N., GRÃŸNBÃĘK K., ELMQVIST N.: InsideInsights: Integrating Data-Driven Reporting in Collaborative Visual Analytics. *Computer Graphics Forum 38*, 3 (2019), 649–661. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13717, doi:10.1111/cgf.13717. 5, 10, 13, 14, 15, 19

[MRB19] MADANAGOPAL K., RAGAN E. D., BENJAMIN P.: Analytic Provenance in Practice: The Role of Provenance in Real-World Visualization and Data Analysis Environments. *IEEE Computer Graphics and Applications 39*, 6 (Nov. 2019), 30–45. doi:10.1109/MCG.2019.2933419. 4, 13, 20

[MSM*17] MICALLEF L., SUNDIN I., MARTTINEN P., AMMAD-UD DIN M., PELTOLA T., SOARE M., JACUCCI G., KASKI S.: Interactive Elicitation of Knowledge on Feature Relevance Improves Predictions in Small Data Sets. *arXiv:1612.02487 [cs, stat]* (Jan. 2017). arXiv: 1612.02487. URL: http://arxiv.org/abs/1612.02487. 12, 18

[MSW10] MACINNES J., SANTOSA S., WRIGHT W.: Visual Classification: Expert Knowledge Guides Machine Learning. *IEEE Computer Graphics and Applications 30*, 1 (Jan. 2010), 8–14. doi:10.1109/MCG.2010.18. 9, 12, 18

[MT14] MAHYAR N., TORY M.: Supporting Communication and Coordination in Collaborative Sensemaking. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec. 2014), 1633–1642. doi:10.1109/TVCG.2014.2346573. 9, 13

[MVCJ16] MUTHUMANICKAM P. K., VROTSOU K., COOPER M., JOHANSSON J.: Shape grammar extraction for efficient query-by-sketch pattern matching in long time series. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2016), pp. 121–130. ISSN: null. doi:10.1109/VAST.2016.7883518. 7, 8, 12, 19

[MVT16] MUTLU B., VEAS E., TRATTNER C.: VizRec: Recommending Personalized Visualizations. *ACM Trans. Interact. Intell. Syst. 6*, 4 (Nov. 2016), 31:1–31:39. URL: http://doi.acm.org/10.1145/2983923, doi:10.1145/2983923. 5, 8, 18

[NCE*11] NORTH C., CHANG R., ENDERT A., DOU W., MAY R., PIKE B., FINK G.: Analytic provenance: process+interaction+insight. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (2011), pp. 33–36. URL: https://dl.acm.org/doi/abs/10.1145/1979742.1979570. 2

[NHC*20] NGUYEN P. H., HENKIN R., CHEN S., ANDRIENKO N., ANDRIENKO G., THONNARD O., TURKAY C.: VASABI: Hierarchical User Profiles for Interactive Visual User Behaviour Analytics. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 77–86. doi:10.1109/TVCG.2019.2934609. 9, 12

[NMK*14] NATH S. S., MISHRA G., KAR J., CHAKRABORTY S., DEY N.: A survey of image classification methods and techniques. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)* (Kanyakumari District, India, July 2014), IEEE, pp. 554–557. URL: http://ieeexplore.ieee.org/document/6993023/, doi:10.1109/ICCICCT.2014.6993023. 10

[NSUW15] NORTH S., SCHEIDEGGER C., URBANEK S., WOODHULL G.: Collaborative visual analysis with RCloud. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2015), pp. 25–32. ISSN: null. doi:10.1109/VAST.2015.7347627. 13, 14

[NTA*18] NGUYEN P., TURKAY C., ANDRIENKO G., ANDRIENKO N., THONNARD O., ZOUAOUI J.: Understanding User Behaviour through Action Sequences: From the Usual to the Unusual. *TVCG 25*, 9 (2018), 2838–2852. 11, 20

[NXB*16] NGUYEN P. H., XU K., BARDILL A., SALMAN B., HERD K., WONG B. W.: SenseMap: Supporting browser-based online sensemaking through analytic provenance. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2016), pp. 91–100. ISSN: null. doi:10.1109/VAST.2016.7883515. 10, 13

[NXW*16] NGUYEN P. H., XU K., WHEAT A., WONG B. W., ATTFIELD S., FIELDS B.: SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 41–50. doi:10.1109/TVCG.2015.2467611. 2, 13

[ODOB18] OLIVEIRA W., DE OLIVEIRA D., BRAGANHOLO V.: Provenance Analytics for Workflow-Based Computational Experiments: A Survey. *ACM Computing Surveys (CSUR) 51*, 3 (May 2018). URL: https://dl.acm.org/doi/abs/10.1145/3184900. 2

[OGW19] OTTLEY A., GARNETT R., WAN R.: Follow The Clicks: Learning and Anticipating Mouse Interactions During Exploratory Data Analysis. *Computer Graphics Forum 38*, 3 (2019), 41–52. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13670, doi:10.1111/cgf.13670. 4, 9, 12, 14, 18

[OKCP19] OTTLEY A., KASZOWSKA A., CROUSER R. J., PECK E. M.: The Curious Case of Combining Text and Visualization. In *EuroVis 2019 - Short Papers* (2019), Johansson J., Sadlo F., Marai G. E., (Eds.), The Eurographics Association. doi:10.2312/evs.20191181. 7

[oxf89] *Oxford English Dictionary*, 2nd ed. Oxford University Press, 1989. URL: https://www.lexico.com/en/definition/provenance. 1

[OYC15] OTTLEY A., YANG H., CHANG R.: Personality as a Predictor of User Strategy: How Locus of Control Affects Search Strategies on Tree Visualizations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (Seoul, Republic of Korea, 2015), ACM Press, pp. 3251–3254. URL: http://dl.acm.org/citation.cfm?doid=2702123.2702590, doi:10.1145/2702123.2702590. 4, 10, 14

[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. 7

[PCC10] PORTEOUS J., CAVAZZA M., CHARLES F.: Applying Planning to Interactive Storytelling: Narrative Control Using State Constraints. *ACM Trans. Intell. Syst. Technol. 1*, 2 (Dec. 2010), 10:1–10:21. URL: http://doi.acm.org/10.1145/1869397.1869399, doi:10.1145/1869397.1869399. 19

[PJ09] PERRY J., JANNECK C. D.: *Supporting Cognitive Models of Sensemaking in Analytics Systems*. Tech. Rep. 2009-12, DIMACS, 2009. 4, 7, 12

[PSM12] POHL M., SMUC M., MAYR E.: The user puzzle—explaining the interaction with visual analytics systems. *IEEE transactions on visualization and computer graphics 18*, 12 (2012), 2908–2916. 12

[PSR15] PREECE J., SHARP H., ROGERS Y.: *Interaction Design: Beyond Human-Computer Interaction*, 4th edition ed. John Wiley, Chichester, Feb. 2015. 2

[PWM*12] POHL M., WILTNER S., MIKSCH S., AIGNER W., RIND A.: Analysing Interactivity in Information Visualisation. *KI - Künstliche Intelligenz 26*, 2 (May 2012), 151–159. URL: http://link.springer.com/10.1007/s13218-012-0167-6, doi:10.1007/s13218-012-0167-6. 7, 12

[PYO*13] PECK E. M. M., YUKSEL B. F., OTTLEY A., JACOB R. J., CHANG R.: Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), pp. 473–482. 7

[RB18] RÄIJBEL O., BOWEN B. P.: BASTet: Shareable and Reproducible Analysis and Visualization of Mass Spectrometry Imaging Data via OpenMSI. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 1025–1035. doi:10.1109/TVCG.2017.2744479. 8, 19

[RESC15] 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* (Aug. 2015). URL: https://ieeexplore.ieee.org/document/7192714, doi:10.1109/TVCG.2015.2467551. 2, 15, 16, 18

[RGT15] RAGAN E. D., GOODALL J. R., TUNG A.: Evaluating How Level of Detail of Visual History Affects Process Memory. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (Seoul, Republic of Korea, 2015), ACM Press, pp. 2711–2720. URL: http://dl.acm.org/citation.cfm?doid=2702123.2702376, doi:10.1145/2702123.2702376. 12, 19

[RLB19] REN D., LEE B., BREHMER M.: Charticulator: Interactive Construction of Bespoke Chart Layouts. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 789–799. doi:10.1109/TVCG.2018.2865158. 8, 18

[RWF*13a] RIBIÄ■IÄĞ H., WASER J., FUCHS R., BLÄŰSCHL G., GRÄŰLLER E.: Visual analysis and steering of flooding simulations. *IEEE Transactions on Visualization and Computer Graphics 19*, 6 (2013), 1062–1075. 5, 19

[RWF*13b] RIBIÄ■IÄĞ H., WASER J., FUCHS R., BLÄŰSCHL G., GRÄŰLLER E.: Visual analysis and steering of flooding simulations. *IEEE Transactions on Visualization and Computer Graphics 19*, 6 (2013), 1062–1075. 5, 11, 12, 19

[SBT*16a] SETLUR V., BATTERSBY S. E., TORY M., GOSSWEILER R., CHANG A. X.: Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (2016), pp. 365–377. 8

[SBT*16b] SETLUR V., BATTERSBY S. E., TORY M., GOSSWEILER R., CHANG A. X.: Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16* (Tokyo, Japan, 2016), ACM Press, pp. 365–377. URL: http://dl.acm.org/citation.cfm?doid=2984511.2984588, doi:10.1145/2984511.2984588. 12, 18

[SCC13] STEICHEN B., CARENINI G., CONATI C.: User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on*

Intelligent user interfaces - IUI '13 (Santa Monica, California, USA, 2013), ACM Press, p. 317. URL: http://dl.acm.org/citation.cfm?doid=2449396.2449439, doi:10.1145/2449396.2449439. 4, 18

[SCC14] STEICHEN B., CONATI C., CARENINI G.: Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Trans. Interact. Intell. Syst. 4*, 2 (July 2014), 11:1–11:29. URL: http://doi.acm.org/10.1145/2633043, doi:10.1145/2633043. 7, 11, 18

[SFC07] SILVA C. T., FREIRE J., CALLAHAN S. P.: Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science Engineering 9*, 5 (Sept. 2007), 82–89. doi:10.1109/MCSE.2007.106. 2

[SGL09] SHRINIVASAN Y. B., GOTZ D., LU J.: Connecting the dots in visual analysis. In *2009 IEEE symposium on visual analytics science and technology* (2009), IEEE, pp. 123–130. 5, 10, 11, 19

[SGP*19] STITZ H., GRATZL S., PIRINGER H., ZICHNER T., STREIT M.: KnowledgePearls: Provenance-Based Visualization Retrieval. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 120–130. doi:10.1109/TVCG.2018.2865024. 5, 6, 12, 19

[SH14] SATYANARAYAN A., HEER J.: Lyra: An interactive visualization design environment. *Computer Graphics Forum 33*, 3 (2014), 351–360. 8

[SHB*14] SEDLMAIR M., HEINZL C., BRUCKNER S., PIRINGER H., MÜLLER T.: Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2161–2170. 15

[SK16] SCHROEDER D., KEEFE D. F.: Visualization-by-Sketching: An Artist's Interface for Creating Multivariate Time-Varying Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (Jan. 2016), 877–885. doi:10.1109/TVCG.2015.2467153. 13

[SKB*18] SACHA D., KRAUS M., BERNARD J., BEHRISCH M., SCHRECK T., ASANO Y., KEIM D. A.: SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 120–130. doi:10.1109/TVCG.2017.2744805. 18

[SKBE17] SAKET B., KIM H., BROWN E. T., ENDERT A.: Visualization by Demonstration: An Interaction Paradigm for Visual Data Exploration. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 331–340. doi:10.1109/TVCG.2016.2598839. 13

[SKL*16] SIDDIQUI T., KIM A., LEE J., KARAHALIOS K., PARAMESWARAN A.: Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *arXiv preprint arXiv:1604.03583* (2016). 8

[SLMK18] SMITH J., LEGG P., MATOVIC M., KINSEY K.: Predicting User Confidence During Visual Decision Making. *ACM Trans. Interact. Intell. Syst. 8*, 2 (June 2018), 10:1–10:30. URL: http://doi.acm.org/10.1145/3185524, doi:10.1145/3185524. 7, 12, 19

[SLSG16] STITZ H., LUGER S., STREIT M., GEHLENBORG N.: AVOCADO: Visualization of Workflow-Derived Data Provenance for Reproducible Biomedical Research. *Computer Graphics Forum 35*, 3 (June 2016), 481–490. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=116877232&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12924. 2

[SML*09] SMUC M., MAYR E., LAMMARSCH T., AIGNER W., MIKSCH S., GÄRTNER J.: To Score or Not to Score? Tripling Insights for Participatory Design. *IEEE Computer Graphics and Applications 29*, 3 (May 2009), 29–38. doi:10.1109/MCG.2009.53. 4, 5, 6, 13, 18

[SMvdWvW15] SCHEEPENS R., MICHELS S., VAN DE WETERING H., VAN WIJK J. J.: Rationale Visualization for Safety and Security. *Computer Graphics Forum 34*, 3 (June 2015), 191–200. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=108442262&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.12631. 13, 18

[SMWH16] SATYANARAYAN A., MORITZ D., WONGSUPHASAWAT K., HEER J.: Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 341–350. 8

[SNMM18] SHERKAT E., NOURASHRAFEDDIN S., MILIOS E. E., MINGHIM R.: Interactive Document Clustering Revisited: A Visual Analytics Approach. In *23rd International Conference on Intelligent User Interfaces* (New York, NY, USA, 2018), IUI '18, ACM, pp. 281–292. event-place: Tokyo, Japan. URL: http://doi.acm.org/10.1145/3172944.3172964, doi:10.1145/3172944.3172964. 11, 19

[SPEB18] SRINIVASAN A., PARK H., ENDERT A., BASOLE R. C.: Graphiti: Interactive Specification of Attribute-Based Edges for Network Modeling and Visualization. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 226–235. doi:10.1109/TVCG.2017.2744843. 8, 12, 18

[SRHH15] SATYANARAYAN A., RUSSELL R., HOFFSWELL J., HEER J.: Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 659–668. 8

[STD19] SETLUR V., TORY M., DJALALI A.: Inferencing underspecified natural language utterances in visual analysis. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19* (Marina del Ray, California, 2019), ACM Press, pp. 40–51. URL: http://dl.acm.org/citation.cfm?doid=3301275.3302270, doi:10.1145/3301275.3302270. 13, 18

[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* (2008), pp. 1237–1246. 10

[SvW09] SHRINIVASAN Y. B., VAN WIJK J.: Supporting Exploration Awareness in Information Visualization. *IEEE Computer Graphics and Applications 29*, 5 (Sept. 2009), 34–43. doi:10.1109/MCG.2009.87. 10, 18

[SW13] SHADOAN R., WEAVER C.: Visual Analysis of Higher-Order Conjunctive Relationships in Multidimensional Data Using a Hypergraph Query System. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2070–2079. doi:10.1109/TVCG.2013.220. 10, 11, 18

[SWR*13] SCHLINDER B., WASER J., RIBICIC H., FUCHS R., PEIKERT R.: Multiverse Data-Flow Control. *TVCG 19*, 6 (2013), 1006–1018. 5, 10, 19

[SYD*14] SONG G., YE Y., DU X., HUANG X., BIE S.: Short Text Classification: A Survey. *Journal of Multimedia 9*, 5 (May 2014), 635–643. URL: http://ojs.academypublisher.com/index.php/jmm/article/view/12635, doi:10.4304/jmm.9.5.635-643. 10

[TLC17] TOKER D., LALLÉ S., CONATI C.: Pupillometry and Head Distance to the Screen to Predict Skill Acquisition During Information Visualization Tasks. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces* (New York, NY, USA, 2017), IUI '17, ACM, pp. 221–231. event-place: Limassol, Cyprus. URL: http://doi.acm.org/10.1145/3025171.3025187, doi:10.1145/3025171.3025187. 11, 12, 20

[TSG*14] TOKER D., STEICHEN B., GINGERICH M., CONATI C., CARENINI G.: Towards facilitating user skill acquisition: identifying untrained visualization users through eye tracking. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14* (Haifa, Israel, 2014), ACM Press, pp. 105–114. URL: http://dl.acm.org/citation.cfm?doid=2557500.2557524, doi:10.1145/2557500.2557524. 11, 18

[vHBv16] VAN DEN ELZEN S., HOLTEN D., BLAAS J., VAN WIJK J. J.: Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 1–10. 15

[Wat01] WATTENBERG M.: Sketching a graph to query a time-series database. In *CHI'01 Extended Abstracts on Human factors in Computing Systems* (2001), pp. 381–382. 7

[WBFE17] WALL E., BLAHA L. M., FRANKLIN L., ENDERT A.: Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Phoenix, AZ, Oct. 2017), IEEE, pp. 104–115. URL: https://ieeexplore.ieee.org/document/8585669/, doi:10.1109/VAST.2017.8585669. 4, 9, 10, 14

[WCW*14] WANG F., CHEN W., WU F., ZHAO Y., HONG H., GU T., WANG L., LIANG R., BAO H.: A visual reasoning approach for data-driven transport assessment on urban roads. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2014), pp. 103–112. ISSN: null. doi:10.1109/VAST.2014.7042486. 7, 12, 18

[WDC*18] WALL E., DAS S., CHAWLA R., KALIDINDI B., BROWN E. T., ENDERT A.: Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 288–297. doi:10.1109/TVCG.2017.2745078. 9, 13, 20

[Wea09] WEAVER C.: Conjunctive Visual Forms. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (Nov. 2009), 929–936. doi:10.1109/TVCG.2009.129. 8, 12, 19

[WGS*13] WILLETT W., GINOSAR S., STEINITZ A., HARTMANN B., AGRAWALA M.: Identifying redundancy and exposing provenance in crowdsourced data analysis. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2198–2206. 20

[WHA07] WILLETT W., HEER J., AGRAWALA M.: Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov. 2007), 1129–1136. URL: http://ieeexplore.ieee.org/document/4376132/, doi:10.1109/TVCG.2007.70589. 14

[WKD18] WOOD J., KACHKAEV A., DYKES J.: Design exposition with literate visualization. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 759–768. 14

[WLLW18] WEGBA K., LU A., LI Y., WANG W.: Interactive Storytelling for Movie Recommendation Through Latent Semantic Analysis. In *23rd International Conference on Intelligent User Interfaces* (New York, NY, USA, 2018), IUI '18, ACM, pp. 521–533. event-place: Tokyo, Japan. URL: http://doi.acm.org/10.1145/3172944.3172979, doi:10.1145/3172944.3172979. 11, 18

[WMA*16] 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. 8, 19

[WS92] WILLIAMSON C., SHNEIDERMAN B.: The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (1992), pp. 338–346. 7, 8

[WSD*13] WALKER R., SLINGSBY A., DYKES J., XU K., WOOD J.: An Extensible Framework for Provenance in Human Terrain Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* (2013). URL: AnExtensibleFrameworkforProvenanceinHumanTerrainVisualAnalytics, doi:10.1109/TVCG.2013.132. 2, 8, 19

[WSL*19] WALCH A., SCHWÄRZLER M., LUKSCH C., EISEMANN E., GSCHWANDTNER T.: Lightguider: Guiding interactive lighting design using suggestions, provenance, and quality visualization. *IEEE transactions on visualization and computer graphics 26*, 1 (2019), 569–578. 5, 18

[WSSM12] WEI J., SHEN Z., SUNDARESAN N., MA K.-L.: Visual cluster exploration of web clickstream data. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 3–12. 6, 12, 20

[XAJK*15] XU K., ATTFIELD S., JANKUN-KELLY T., WHEAT A., NGUYEN P. H., SELVARAJ N.: Analytic Provenance for Sensemaking: A Research Agenda. *IEEE Computer Graphics and Applications 35*, 3 (May 2015), 56–64. URL: http://ieeexplore.ieee.org/document/7111922/, doi:10.1109/MCG.2015.50. 2

[XBL*18] XU S., BRYAN C., LI J. K., ZHAO J., MA K.: Chart Constellations: Effective Chart Summarization for Collaborative and MultiâĂŘUser Analyses. *Computer Graphics Forum 37*, 3 (June 2018), 75–86. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=bth&AN=130628134&site=ehost-live&authtype=sso&custid=s5409946, doi:10.1111/cgf.13402. 5, 7, 10, 12, 19

[XGH06] XIAO L., GERTH J., HANRAHAN P.: Enhancing visual analysis of network traffic using a knowledge representation. In *2006 IEEE Symposium On Visual Analytics Science And Technology* (2006), IEEE, pp. 107–114. 8

[YaKS07] YI J. S., A. KANG Y., STASKO 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. doi:10.1109/TVCG.2007.70515. 7

[YS20] YU B., SILVA C. T.: FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 1–11. doi:10.1109/TVCG.2019.2934668. 20

[ZDFD15] ZGRAGGEN E., DRUCKER S. M., FISHER D., DELINE R.: (slqu)eries: Visual Regular Expressions for Querying and Exploring Event Sequences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (Seoul, Republic of Korea, 2015), ACM Press, pp. 2683–2692. URL: http://dl.acm.org/citation.cfm?doid=2702123.2702262, doi:10.1145/2702123.2702262. 19

[ZGI*18] ZHAO J., GLUECK M., ISENBERG P., CHEVALIER F., KHAN A.: Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 340–350. doi:10.1109/TVCG.2017.2745279. 9, 13