Doctoral Dissertations (All Dissertations, All Years)   Electronic Theses and Dissertations

2019-04-16

Quantifying, Modeling and Managing How People Interact with Visualizations on the Web

Mi Feng
Worcester Polytechnic Institute

Follow this and additional works at: https://digitalcommons.wpi.edu/etd-dissertations

Quantifying, Modeling, and Managing
How People Interact with Visualizations on the Web

by

Mi Feng

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Computer Science

April 16, 2019

APPROVED:
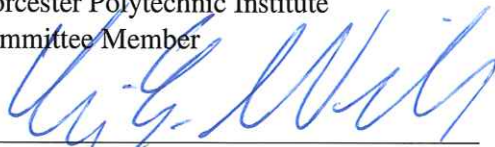
Professor Lane T. Harrison
Worcester Polytechnic Institute
Advisor

Professor Emmanuel O. Agu
Worcester Polytechnic Institute
Committee Member

Professor Craig E. Wills
Worcester Polytechnic Institute
Head of Department

Professor Elke A. Rundensteiner
Worcester Polytechnic Institute
Committee Member

Professor Alex Endert
Georgia Tech
External Committee Member

1

Abstract

The growing number of interactive visualizations on the web has made it possible for the general public to access data and insights that were once only available to domain experts. At the same time, this rise has yielded new challenges for visualization creators, who must now understand and engage a growing and diverse audience. To bridge this gap between creators and audiences, we explore and evaluate components of a design-feedback loop that would enable visualization creators to better accommodate their audiences as they explore the visualizations.

In this dissertation, we approach this goal by quantifying, modeling and creating tools that manage peoples open-ended explorations of visualizations on the web. In particular, we:

1. Quantify the effects of design alternatives on peoples interaction patterns in visualizations. We define and evaluate two techniques: HindSight (encoding a users interaction history) and text-based search, where controlled experiments suggest that design details can significantly modulate the interaction patterns we observe from participants using a given visualization.

2. Develop new metrics that characterize facets of peoples exploration processes. Specifically, we derive expressive metrics describing interaction patterns such as exploration uniqueness, and use Bayesian inference to model distributional effects on interaction behavior. Our results show that these metrics capture novel patterns in peoples interactions with visualizations.

3. Create tools that manage and analyze an audiences interaction data for a given visualization. We develop a prototype tool, ReVisIt, that visualizes an audiences interactions with a given visualization. Through an interview study with visualization creators, we found that ReVisIt make creators aware of individual and overall trends in their audiences interaction patterns.

By establishing some of the core elements of a design-feedback loop for visualization creators, the results in this research may have a tangible impact on the future of publishing interactive visualizations on the web. Equipped with techniques, metrics, and tools that realize an initial feedback loop, creators are better able to understand the behavior and user needs, and thus create visualizations that make data and insights more accessible to the diverse audiences on the web.

Contents

List of Figures

x

List of Tables

Acknowledgements

This dissertation would not have been possible without the tremendous support from a group of amazing people.

I would like to express my sincere gratitude to my advisor Professor Lane Harrison. It has been my great honor to be his first PhD student. I learned from him how to conduct solid visualization research, so that visualization to me changed from a simple hobby into a field of expertise. He helped me practice writing and presentation over and over again, so that I, a non-native English speaker, can effectively communicate research to people. He encouraged and trusted me when I was facing various kinds of challenges, so that I have become more confident in solving problems and diving into new areas.

I would like to thank my PhD committee members, Professors Elke Rundensteiner, Emmanuel Agu, and Alex Endert. They provided valuable feedback and suggestions that inspired me on the techniques used in our research, and asked many insightful questions that helped me better shape the structure and arguments of the dissertation.

I would like to thank the visualization researchers who acted as my mentors or collaborators: Evan Peck, Hua Guo, Meg Pirrung, Fumeng Yang, and Cheng Deng; and my labmates: Russ Davis and Hamid Mansoor. It is my honor and pleasure to work with these talented people with excellent research and technical expertise. They provided careful guidance, thoughtful questions and dedicated collaborations throughout my PhD study.

I would like to thank the MQP team of three undergraduate students at WPI: Madeline

Burke, John Dyer and Luke Gardner, who built the ReVisIt system from the ground up. I am so glad that we worked together, and I am impressed with their phenomenal progress as well as their excellent teamwork.

I would like to thank my friends: Wenzhao Xu, Zetian Zhang, Ruixiang Du, Zhongfang Zhuang, and many others, who provided valuable comments and suggestions to my work, and shared with me some of their own experiences and lessons on research.

Finally, I would like to thank my dear parents. They pushed me to look beyond the surface, which motivated me to pursue advanced studies. They trained me on engineering and problem-solving skills since I was a child (as both of them were engineers), which helped me along my PhD journey. I am indebted to them for their extensive support and encouragement from home throughout the years.

Chapter 1

Introduction

As *interactive visualizations* migrate from desktop applications to the web, numerous complex datasets have been made available for general public to freely explore [1, 14, 15, 16, 17]. However, unlike trained data analysts, general audiences may not be as adept at manipulating the vast quantity of information in front of them. Quite often, people do not explore as much as expected by the creators of these visualizations. *Only 10-15% of people click on buttons*, reports Archie Tse, the graphics editor at the New York Times [18].

One of the expectations creators may have when creating interactive visualizations is to empower the users to discover their own insights from complex data [4]. Such benefit for users is hardly achievable through web pages or static infographics. Consider the example from the New York Times in 2012 *"At the National Conventions, the Words They Used"* (Figure 1.1), where each circle represents a word. Without any interaction, one can learn from the annotations that, e.g., the word "Women" is used more frequently by Democrats. However, clicking a circle reveals the dialogues containing the word. One can thus go beyond the surface by clicking other words, such as "immigration", and finds that Democrats and Republicans used the word an equal amount of times, but in very different contexts.

(A) the default view   (B) selecting *"Women"*   (C) selecting *"Immigration"*

An annotation on *"Women"*
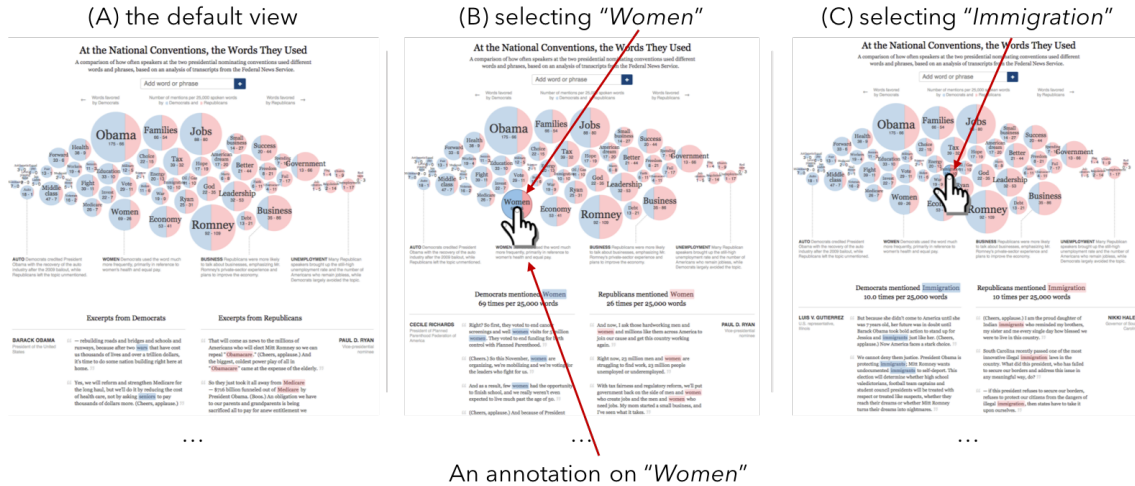
Figure 1.1: An interactive visualization from the New York Times – "At the National Conventions, the Words They Used" [1]. (A) is the default view. By clicking a circle at the top, a person sees the dialogs containing the word displayed at the bottom. (B) selecting an annotated circle "Women". (C) selecting "Immigration" which is in the middle of the visualization.

Nevertheless, there appears to be a gap between visualization creators' expectations and the reality, *e.g.*, people do not explore through interaction as much as expected. People's lack of interaction may be due to various reasons [4, 10, 18]: Maybe a user was less interested in the topic, did not have enough time to explore, did not understand the visualization, was not aware of the interactivity, could not find what she wanted, or got lost when navigating through the data points, etc. However, apart from these guesses, creators have little evidence to support why the gap between expectation and reality exists, or to inform their future design improvements.

1.1 Problem Statement

Despite the rapid growth of interactive visualizations on the web, there is a gap between the reality and the expectations from visualization creators. For example, creators may

expect their audiences to discover personal insights through actively interacting with a given visualization. Results from our initial studies [12, 13] suggest that visualization creators do not have sufficient evidence-based methods for understanding how their audiences engage with the visualizations and for how to improve the design.

One possible solution is to establish a *design-feedback loop* for web visualization creators to understand and support their audiences' explorations. As shown in Figure 1.2, such feedback loop enables creators to collect and make sense of their audiences' interaction data after deploying the visualization. The promise of this solution is informed by the positive results from two threads of research: *evidence-based website design workflow* [3, 19, 20] and *visualization interaction analysis* [6, 21, 22]. However, challenges still hinder the successful establishment of a feedback loop. For example, it remains unclear how to make sense of people's diverse interactions with visualizations on the web.

Evidence-based website design workflow. After deploying a website, designers often utilize people's usage data to evaluate their design and learn about their audiences. For example, through A/B testing [19], designers can choose between the design alternatives of a button based on the click-through rates. More advanced techniques have also been used to make sense of people's website clickstreams through statistical and visual analyses [3, 20, 23]. The evidence-based workflow of website design, including the strategies to *quantify, model and analyze users' interactions*, has inspired our research aiming to assist the creators of visualizations on the web. However, people's interactions with visualizations are fairly different from their interactions with websites, which makes it difficult to directly adopt the specific techniques used in the website design workflow.

Visualization interaction analysis. In the visualization field, researchers collect and analyze people's interactions with visualizations. This collection serves for various purposes [21], *e.g.*, to evaluate a visualization system or technique [4, 24], to recover and reuse people's analytic provenance [6, 22], or to identify the relationship between be-

havioral patterns and user characteristics [5, 25], etc. This research thread equips us with useful techniques to analyze people's interactions with visualizations, *e.g.*, providing metrics to quantify a user's visualization exploration: total exploration time and the number of interacted visual elements. However, this thread mostly focused on data or domain experts, often with specific analytic tasks. In contrast, the web environment is more diverse – there are users from different backgrounds, visualizations on different topics, and creators with different goals in mind. People's behavior patterns may be different in every situation. Thus, challenges still exist on how to make sense of people's interactions with visualizations in *diverse web contexts*.
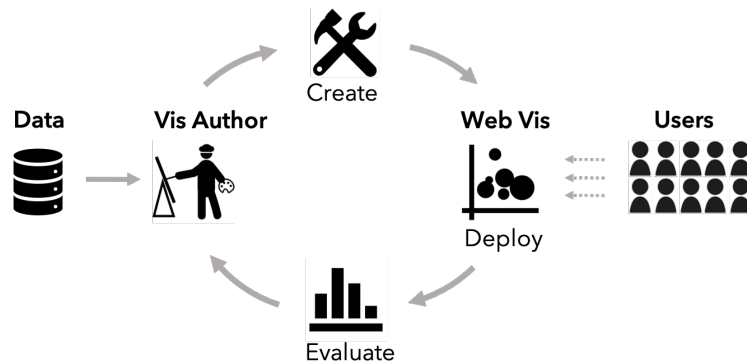
Figure 1.2: We explore and evaluate components of a design-feedback loop that would enable visualization creators to better accommodate their audiences as they explore the visualizations.

1.2 Research Scope, Questions and Tasks

In this dissertation, we explore and evaluate components of a design-feedback loop that would enable visualization creators to better accommodate their audiences as they explore the visualizations. We approach this goal by quantifying, modeling and creating tools that manage people's explorations of visualizations on the web. We focus on the types of exploration behavior that are relatively unique in web contexts. These are open-ended explorations that may be motivated by people's diverse goals, instead of traditionally studied visual analytical tasks, such as fraud detection [6].

4

Given the diverse context of the web, it remains unclear whether small changes to visualization designs lead to measurable differences in user interaction behavior.

RQ1: Can we detect behavioral differences of design alternatives from people's interactions with visualizations on the web?

Current metrics quantifying interaction behavior, *e.g.*, exploration time and the number of interacted elements, are insufficient to fully capture people's diverse explorations.

RQ2: How can we better express facets of people's explorations of visualizations on the web via their interaction traces?

Statistical analysis alone may not be adequate for visualization creators to achieve their goals in a flexible way. Visual explorations may help them understand diverse behavior.

RQ3: How can we help visualization creators to achieve their goals by feeding back people's usage data and enabling visual exploration?

We answer each of the research questions through three step-by-step research tasks:

Can we detect any behavioral differences of design alternatives at all from people's diverse interactions with visualizations?

T1: Quantifying the behavioral effects of design alternatives to initially demonstrate the efficacy of analyzing diverse explorations. Namely, we evaluate two interaction techniques, and evidence from controlled studies indicates people's increasing engagement.

T1.A: As reported by creators, general audiences seldom interact with web visualizations. We develop the technique *direct encoding of personal interaction history (HindSight)* aiming to encourage user exploration. We apply the technique to three existing visualizations, examining its impact through online user experiments. The results show that HindSight impacts users' breadth of exploration, and nudges them to different parts

5

of a visualization.

T1.B: Creators occasionally integrate search functionality into visualizations. Yet it remains unclear how search might impact users' exploration behavior. We evaluate the effects of adding *text-based search functionality* to visualization through a user study across five visualizations. Quantitative results show that search encourages users' diverse engagement with data, and that the impact of search is affected by users' familiarity with the underlying dataset.

How can we fully capture the quality of people's diverse explorations of visualizations on the web from their interaction traces?

T2: Characterizing and modeling peoples exploration behavior to fully utilize the information in their interaction traces. Specifically, we derive new metrics and adopt Bayesian statistics to analyze interactions, and the results show that these techniques are able to capture novel aspects of the visualization exploration process.

T2.A: Current user-centered metrics quantifying interaction behavior, *e.g.*, exploration time and the count of interacted elements, suffer from over-aggregating characteristics of users' exploration processes. We capture several facets of people's exploration behavior by deriving new metrics, such as exploration diversity and pacing. These metrics are shown to uncover additional characteristics of exploration.

T2.B: Interaction data from the web may be noisy, which sometimes makes it difficult to draw useful conclusions. We move beyond the standard practice of making statistical inferences, and adopt Bayesian analysis process to model people's interactions. We obtain more precise and conclusive results from Bayesian analysis compared with Frequentist analysis.

How can we help visualization creators to achieve their goals by feeding back people's usage data and enabling visual exploration?

T3: Re-visualizing peoples explorations to provide scalable and low-effort feedback to visualization creators. In the diverse web context, which may include users from different backgrounds, visualizations on different topics, and visualization creators with different goals, statistical analysis only may not be adequate to reveal people's behavioral patterns in a flexible way. Visual explorations instead, may enable visualization creators to ask and answer more open questions. We thus develop a visual analytic prototype *ReVisIt*, re-visualizing raw interaction data as well as automatically-computed high-level statistics. We evaluate the effectiveness of the visualization system through interviews with creators of visualizations on the web, and the creators express the effectiveness of making sense of their audiences interactions through visual exploration.

1.3   Organization

Research Question	Task	Chapter
RQ1: *Can we detect behavioral differences of design alternatives from people's interactions with visualizations on the web?*	T1.A	§3 Quantifying the Effects of Hind-Sight
	T1.B	§4 Quantifying the Effects of Text-based Search in Visualization
RQ2: *How can we better express facets of people's explorations of visualizations on the web via their interaction traces?*	T2.A	§5 Characterizing Diverse Exploration Behavior with Visualizations on the web
	T2.B	§6 Modeling Diverse Interactions with Visualizations using Bayesian statistics
RQ3: *How can we help visualization creators to achieve their goals by feeding back people's usage data and enabling visual exploration?*	T3	§7 Re-visualizing People's Interactive Explorations of Visualizations on the Web

Table 1.1: The organization of this dissertation.

Chapter 2

Background

We provide in-line discussion of the most related work as we describe each research task. Herein we provide a brief introduction of the related works as background.

2.1 Website Design and Clickstream Analysis

2.1.1 Behavior-driven Website Design

Researchers and practitioners base the design of web interfaces on people's interaction behavior. For instance, as introduced by Agapie *et al.* [2], changing the edge color of the search bar as people type leads them to input longer searching queries (Figure 2.1. Willett *et al.* create scented widgets [26], using visualizations to represent other users' aggregated navigation history. They find users exploring unfamiliar data make up to twice as many unique discoveries using these widgets imbued with social navigation data.

2.1.2 Website Clickstream Analysis

Clickstream research includes the data processing, analysis and visualization methods to analyze users' website visit logs [3, 20, 23, 27, 28]. For example, Liu *et al.* [23]

An empty query box has a pink halo.

As the person starts to type, red hue starts to fade.

As the query gets longer, the halo becomes progressively bluer.

Long queries are displayed with a blue halo.
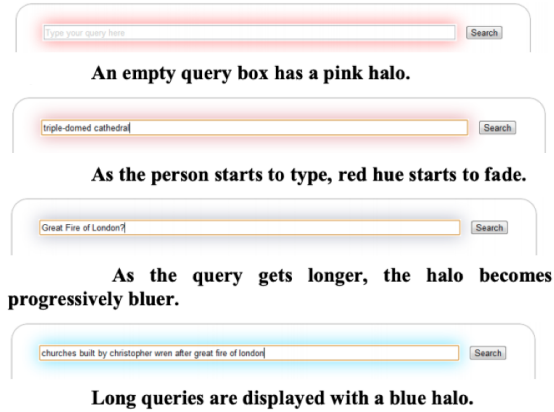
Figure 2.1: Agapie *et al.* found that changing the edge color of the search bar as people type nudges people to typing longer queries [2].

developed algorithms to extract sequence patterns from clickstreams. Zhao *et al.* [3] created a visualization called MatrixWave to compare two clickstream datasets, and found it to scale better than commonly used Sankey diagrams. Chi *et al.* [29] quantified the saliency of a user's visit to a website when modeling users information needs and actions on the web. Heer *et al.* [30] further used this measure to cluster web users. These efforts influence our work of visualization interaction analysis, in that a user's open-ended exploration of a visualization containing visual elements can be considered analogous to the exploration of a website.

(a) event sequences   (b) Sankey   (c) MatrixWave
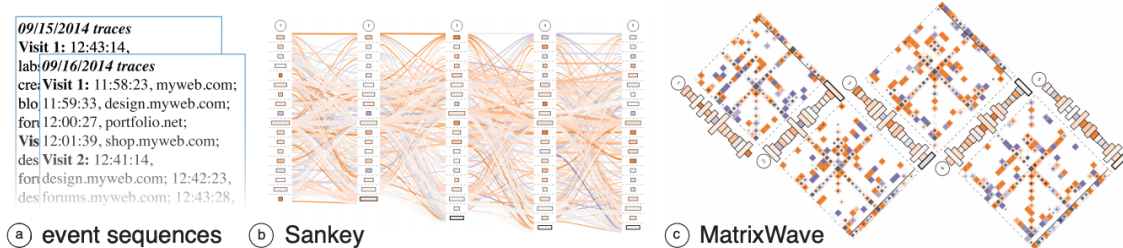
Figure 2.2: Zhao *et al.* [3] re-visualized users' website interaction logs (a) using MatrixWave (c), and found it to scale better than commonly used Sankey diagrams (b).

9

2.2 Visualization Interaction Analysis

Researchers and practitioners analyze people's interactions with visualizations for various purposes [21], e.g., to evaluate visualizations, to understand user performance and characteristics, and to recover and reuse people's analytical provenances.

2.2.1 Evaluating Visualizations

One goal of analyzing people's visualization interactions is to examine the comparative impact of competing design techniques on user behavior [4, 9, 22, 24, 31]. Boy *et al.* [4] evaluated the effectiveness of storytelling by comparing users' exploration time and raw interaction counts (hovers and clicks) between the experimental and control groups (Figure 2.3).

(a)   (b)
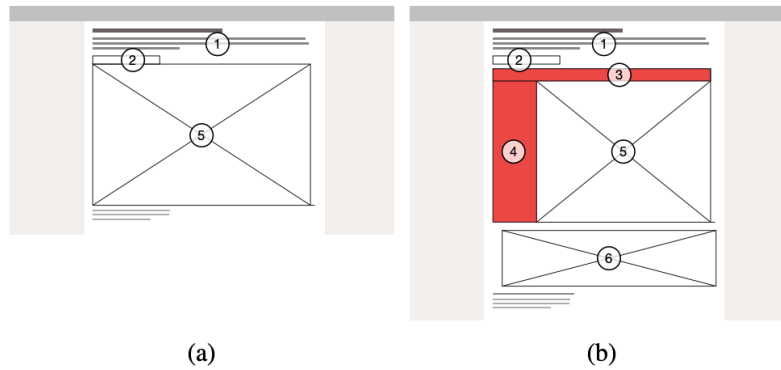
Figure 2.3: Boy *et al.* [4] evaluated the impact of "storytelling" in visualization on users' exploration behavior.

Liu *et al.* [24] measured the effects of latency on users' exploration behavior of visual analytics by using raw interaction counts (drag, brushing and linking, etc). Guo *et al.* [22] evaluated visualization design through sequence analysis – extracting the sub-sequences containing specific individual interactions, and then counted the sub-sequences for each user. Heer *et al.* [9] did an exploratory study examining the social support in a visualiza-

tion on users' interaction behavior.

2.2.2 Understanding User Performance and Characteristics

Ottley *et al.* [5] found that people's searching strategies of a tree visualization reveal their personality, i.e., locus of control. They used aggregated maps to show different exploration patterns of two types of tree visualizations. Brown *et al.* [25] also found that locus of control affects people's graph searching patterns. They identified the pattern through classifying the interaction logs using SVM.

(a) Externals with indented tree (b) Internals with indented tree (c) Externals with dendrogram (d) Internals with dendrogram
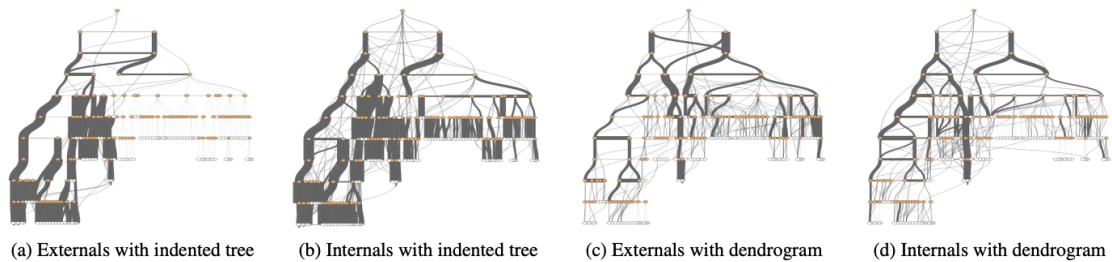
Figure 2.4: Ottley *et al.* [5] observed that people's searching strategies of a tree visualization reveal their locus of control.

2.2.3 Recovering and Reusing Analytical Provenances

Visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces" [32]. The *process* of the reasoning, shown to be as important as the reasoning results [33], can partially be informed by a user's interactions with a visualization. Thus there exists the research thread of *analytic provenance* that focuses on utilizing a user's interactions to understand her reasoning process [6, 7, 34, 35, 36].

Dou *et al.* [6] showed that interactions with visual analytic systems can reveal some of the reasoning steps taken by users (Figure 2.5 . Blascheck *et al.* [37] introduced a visual analytic approach to study users' interactions with visual analytics. Ragan *et al.*

11

[38] studied how users' interactions can infer their intentions during analysis. Endert *et al.*'s[7, 39] semantic interaction reused users' analytical provenances to enable users to steer the underlying model of visual analytics. Wall *et al.* [36] proposed six metrics to measure cognitive bias during visual analysis process, including data coverage, data distribution and attribute coverage/distribution, etc.
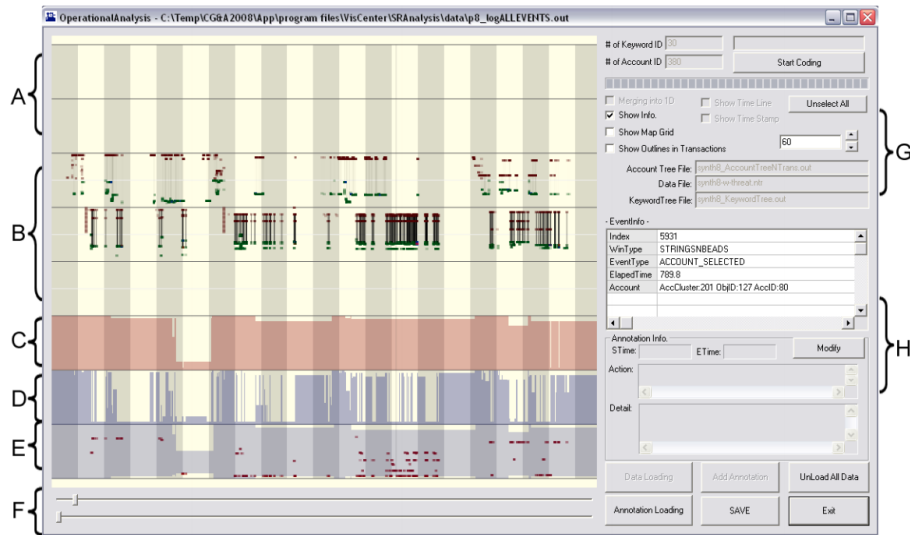
Figure 2.5: Dou *et al.* [6] showed that interactions with visual analytic systems can reveal some of the reasoning steps taken by users.
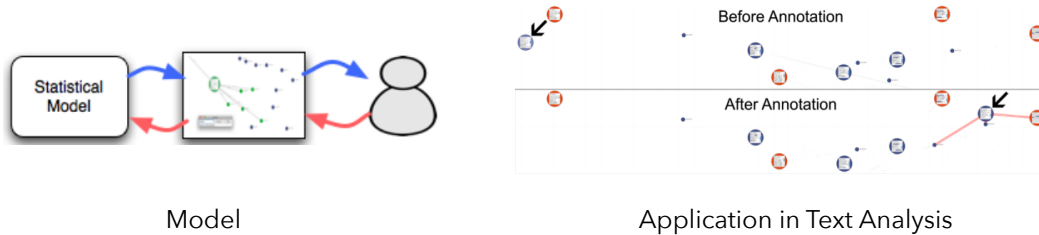
Figure 2.6: Endert *et al.* [7] reused users' analytical provenances to enable users to steer the underlying model of visual analytics.

This research thread equips us with useful techniques to analyze people's interactions

with visualizations. However, this thread mostly focused on data or domain experts, often with specific analytic tasks. In contrast, the web environment is more diverse – there are users from different backgrounds, visualizations on different topics, and creators with different goals in mind. People's behavior patterns may be different in every situation. Thus, challenges still exist in how to make sense of people's interactions with visualizations in the diverse web context.

2.3 Interactive Visualizations on the Web

Visualization researchers have been seeking ways to engage general audiences with data through certain techniques, including narattive visualizations, social support and initial exploration support.

2.3.1 Narrative Visualizations

One of the techniques used to engage users is letting creators to add pre-defined stories to the visualization to guide their audiences. Segel and Heer used a spectrum spanning from author-driven to user-driven structure, to categorize the visualizations on the web (mostly journalism) [8]. There are three types of visualizations placed in the spectrum, Martini Glass Structure, Interactive Slideshow, and Drill-Down Story. The first two are closer to the author-driven structure, and the last is closer to the user-driven structure.
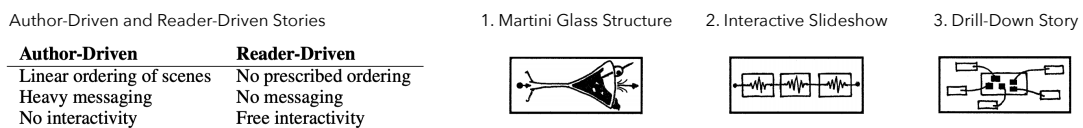
Figure 2.7: Segel and Heer used a spectrum spanning from author-driven to user-driven structure, to categorize the visualizations on the web [8]

13

2.3.2 Social Support

Heer investigated the role of social data analysis in supporting user explorations. Several techniques have been developed and evaluated through users' interaction activities and insight generation [9, 26, 40, 41, 42] (see Figure 2.8 for example). Social data analysis can effectively support user explorations through inspirations from each other. In this work, we are investigating the techniques that support *personal explorations*. A user should be able to use them independently from social groups. These techniques should also be simple enough to be quickly understood by general audiences, and of low development costs.
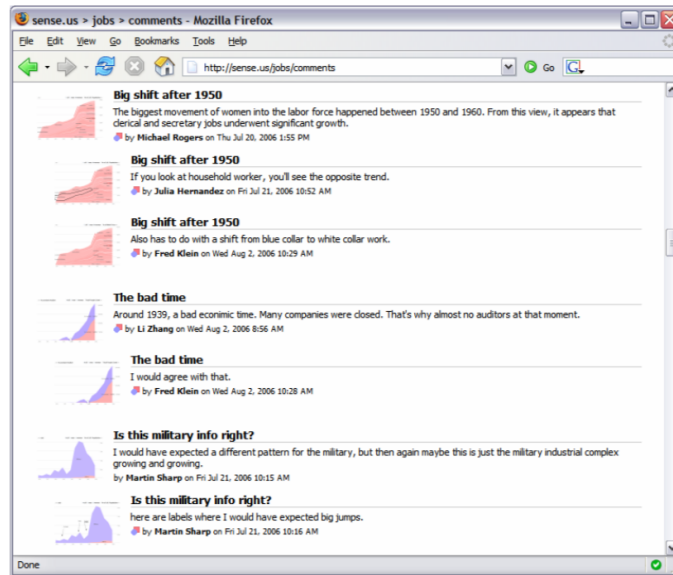
Figure 2.8: Heer *et al.* [9] demonstrated through sense.us, that social data analysis can effectively support user explorations through inspirations from each other.

2.3.3 Initial Exploration Support

Users can be supported either before or during their open-ended explorations. Boy *et al.* [4] examined whether an initial data story provided ahead of an exploratory visualization

could improve the engagement of users' open-ended explorations. Boy *et al.* [10] also designed and evaluated the perceived affordance techniques to invite users to explore visualizations (Figure 2.9). There are also hints from web visualization developments that certain techniques applied to the visualizations appear to engage users by initially situating them in the data [43, 44]. While in this work, we focus on the strategies to support users to navigate in a pool of visualization elements *during exploration*. We believe it is essential to provide people with both a sense of orientation and freedom when they face a large quantity of information in front of them.
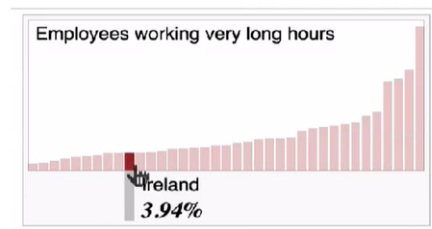
Figure 2.9: Boy *et al.* [10] designed and evaluated the perceived affordance techniques to invite users to explore visualizations. The figure is the design condition SI-4 that was shown to be effective.

Chapter 3

Quantifying the Effects of HindSight

HindSight encodes a user's interaction history directly in the visualization. In this case, visited charts are darker.

Compared to the original, history encodings make it easier to find a chart previously explored, even when the visualization order changes.

Original

HindSight Enabled

a

Original

It's possible to leverage a chart's existing visual encodings to show history. In this case, visited lines are made slightly darker and larger.

HindSight Enabled

b

Where are the Big Polluters since 1971?

Current position

Prior positions

HindSight Enabled

Directly encoding history in visualizations make it possible to segment what has been explored versus what remains.

HindSight is applicable to many existing visualizations. i.e. Boy *et al.* 2015

Original

c

Figure 3.1: Visually encoding a user's interaction history – a technique we call "Hind-Sight" – can be easily implemented in many existing visualizations and is shown to significantly impact both exploration and insights. Here we show the three visualizations from our experiment, encoding interaction history through: a) chart opacity, b) line width and opacity, c) color (red highlighting), and "shadows" of previous marker positions

Physical and digital objects often leave markers of our use. Website links turn purple after we visit them, for example, showing us information we have yet to explore. These "footprints" of interaction offer substantial benefits in information-saturated environments – they enable us to easily revisit old information, systematically explore new

information, and quickly resume tasks after interruption. While applying these design principles have been successful in HCI contexts, direct encodings of personal interaction history have received scarce attention in data visualization. One reason is that there is little guidance for integrating history into visualizations where many visual channels are already occupied by data. More importantly, there is no firm evidence that making users aware of their interaction history results in benefits with regards to exploration or insights.

Following these observations, we propose HindSight – an umbrella term for the design space of representing interaction history directly in existing data visualizations. We examine the value of HindSight principles by augmenting existing visualizations with visual indicators of user interaction history (e.g. How the Recession Shaped the Economy in 255 Charts, NYTimes). In controlled experiments of over 400 participants, we found that HindSight designs generally encouraged people to visit more data and recall different insights after interaction. The results of our experiments suggest that simple additions to visualizations can make users aware of their interaction history, and that these additions significantly impact users' exploration and insights.

3.1 Introduction

During exploratory data analysis (EDA), people navigate through unseen data for an indeterminate amount of time until an unknown insight is discovered. As a result, EDA aligns with some of the fundamental goals of information visualization. Data Exploration is generally defined in the context of scientific workflows, yet it is quickly becoming a part of peoples day-to-day lives through news organizations and broadly accessible analysis tools.

Exploration takes time, creating a tension with our biological capacity for memory, a tension that is not supported by the visualization itself. Our memory's capacity to re-

member recent interactions is severely limited in both amount and decay [24, 45]. As a result, even when a visual design is aligned with our perceptual abilities, we struggle to remember and track parts of the data we have encounted, creating a barrier to exploration and engagement. These limitations suggest that a refinement of visualization techniques to support memory in interactive contexts may have broad impact in supporting user explorations.

The call to support *history* operations in data visualization is not new. Many systems leverage formal representations of visualization state to capture and analyze scientific provenance [41, 46, 47]. Shneiderman identified *history* as an important visualizaton task to "allow users to retrace their steps" [48]. Gutwin realized Shneiderman's hypothesis, showing that indicators of exploration history helped users identify which parts of the data they have seen [49, 50, 51]. Collaborative analysis has also been a focus, where users are shown a history of operations from their collaborators to support situational understanding [52, 53]. Despite these advances, interaction history is not common in visualization systems today. One reason for this scarcity is that there is currently little guidance on how interaction history can be incorporated into the visualization itself. More importantly, however, there is little evidence for the possible benefits making users aware of their history, beyond supporting a user's ability to retrace their steps.

To uncover new opportunities in this space, we applied Wexelblat and Maes' interaction history framework [54] to the current state-of-the-art in visualization. Wexelblat and Maes identified six design properties – *proxemic vs. distemic, active vs. passive, rate/form of change, degree of permeation, personal vs. social, kind of information* – that can be used to characterize interaction history systems, or in this case, shed light on unexplored regions of the design space. We focus on two dimensions that expose a hole in the current design space – how history is directly tied to an object (*degree of permeation*) and whether history represents personal or group activity (*personal vs. social*).

18

As a direct result of this analysis, we propose **HindSight – a representation of personal interaction history that directly encodes interaction history as a visual variable on the data.** At its most basic level, HindSight modifies the saliency of data after a user engages with it, leaving visual markers of interaction history. Given an indication of what they have visited, users can quickly segment what parts of the data they have explored as well as what remains unexplored– using their perceptual system rather than their memory. The technical barrier of integrating HindSight into visualizations is low, requiring only simple modification to existing visualization infrastructure.

Direct encoding of interaction history on data has potential benefits that align with aspects of Shneiderman's arguments for direct manipulation [55]: increased visibility of object and actions, for example, or rapid and incremental actions with immediate feedback. Direct encoding puts interaction history right in front of the user, supporting visual recognition of previous interactions rather than relying on recall, short-cutting the mental translation of history information. Compared to indirect history encoding techniques common in visualization research [41, 46], direct encoding doesn't require users to process spatially separate regions to relate history information back to the data.

Given these observations, we hypothesized that the combination of direct encoding and personalized histories in HindSight would positively impact user behavior during exploratory analysis. To test our hypotheses, we applied HindSight to three visualizations, analyzing exploration behavior during interaction, as well as user-reported insights after exploring the visualization. Our cases include:

- *"The Rise and Decline of Ask MetaFilter" by Jim Vallandingham* ($N = 92$): 16 line charts of topic trends over time at MetaFilter that can be reordered by Count or Name.

- *"How the Recession Reshaped the Economy, in 255 Charts" by the NYTimes* ($N = 116$): a scatterplot of 255 line charts showing how jobs have changed across indus-

tries over the past 10 years.

- *"Where are the Big Polluters since 1971" by Jeremy Boy* ($N = 206$): a coordinated view map and line graph showing CO2 emissions that can be filtered by year or country [4].

In controlled experiments of over 400 participants, we found that HindSight designs encouraged people to visit more data and recall different insights after interaction. These results illustrate that the longstanding design principles developed by visualization research– principles that allow us to effectively map data to visual variables– can also be used to encode interaction, allowing us to leverage our perceptual system in interactive exploration and sensemaking.

3.2   Related Works

Interaction becomes a key mechanism in exploratory data analysis when the size or complexity of the data eclipse what the visual display can handle [48]. To this end, research has historically focused on interaction techniques that empower users to effectively reveal and re-configure data in visualization systems. More recent work addresses the challenges of supporting user exploration and their awareness in the information foraging process. We describe several seminal results and research threads in this area, focusing on how they shape our contributions.

3.2.1   Wexelblat and Maes' Interaction History Framework

Objects are *history-rich* if they contain "historical traces that can be used by people in the current time" (p. 270, [54]). In the physical world, we note the wear on a tool to help us understand how it has been gripped in the past, or observe footprints in the snow to help us see areas that have previously been already explored. Embedding history rich objects
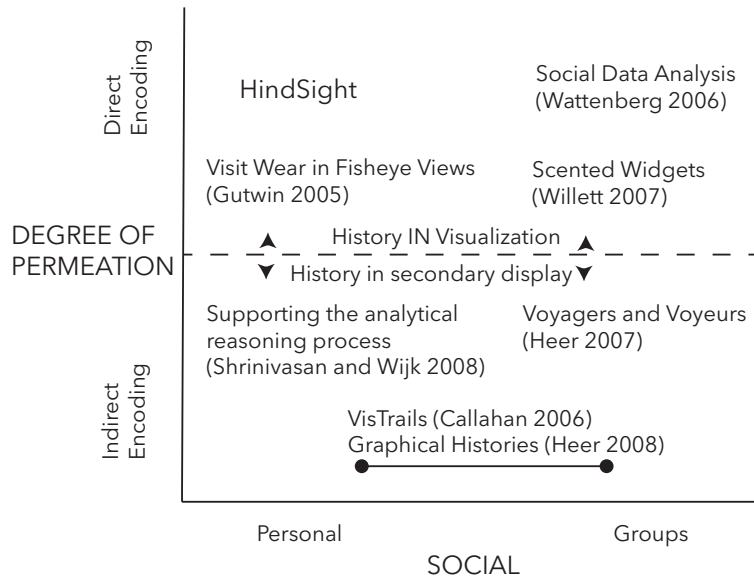
Figure 3.2: With the exception of Gutwin's implementation of visit wear in fisheye views, research in data visualization has typically focused on three quadrants defined by Wexlblet and Maes. HindSight lies in the fourth– a direct encoding of personal interaction history.

into the digital realm enables people to either leverage their own experience that they have accumulated over time, or leverage the combined experience of people who have interacted in the same space. Citing results from Pirolli and Card, Wexelblat and Maes argue that without interaction history we are "forced to become information foragers over and over again" [54, 56].

Wexelblat and Maes describe six properties to articulate a design framework for interaction history: the extent to which people find a space to be transparent and easily understood vs. needing background or training to engage with it (*proxemic vs. distemic*), the degree of effort needed to record history (*active vs. passive*), the degree to which an object is changed by history (*rate/form of change*), the extent to which history is directly tied to an object or recorded separately (*degree of permeation*), whether history is tied to an individual or a group (*personal vs. social*), and finally the information we choose to represent history (*kind of information*). Each dimension of these six properties will nudge

21

user behavior as they engage or use their own histories.

Consider the interaction when we click a link on a webpage – an example of an information-rich environment. The link that I click (*high degree of permeation*) automatically (*passive*) turns purple (*form: color as history*), and indicates whether I (*personal*) visited the site or not (*kind of information, binary rate of change*). Contrast this interaction with how our browser represents visit history. Our browsing history is also automatically collected (*passive*), but contains more detailed information than the purple links (*kind of information: time, url, etc.*). However, seeing our visit history requires us to navigate to a history page that is spatially separated from the original data (*low degree of permeation*). This shift from a high to low degree of permeation enables focused views of our browsing history, but sacrifices the availability of that information by relegating it to a secondary display.

These design tradeoffs are critical to weigh when designing history-rich tools and have implications for guiding exploration or engagement in any information foraging task. In particular, the change in permeation from the previous example shifts the notion of history from "How did I get here?" to "Where have I been before?" and "What is left to explore?". In the next sections, we highlight the benefits of reframing history in this manner, and explore whether these same benefits can be translated to data visualization contexts.

3.2.2   Interaction History from HCI to Visualization

The direct encoding of interaction history has been studied in HCI since the early 90s, when Hill *et al.* proposed the notion of computational wear ('read wear' and 'edit wear') to display authorship history [57]. Alexander *et al.* later analyzed principles of wear mechanisms – in this case marks on the scrollbar – to return to previously edited regions of a document. They found that marking the scrollbar with interaction history decreased

visitation time, was highly preferred by participants, and was scalable to a large number of marks [58].

Following these foundational papers, researchers in HCI have applied interaction history to support users in novel ways. Gutwin, for example, visualized the traces of multiple mouse-pointers in a collaborative system to make users aware of where other people were focusing [49]. They found that a direct representation (or *high degree of permeation*) of interaction history (the pointer trail) was easy to understand, and helped users understand the context of their collaborators current actions. Bridging the gap from HCI into data visualization, Skopik and Gutwin, introduced the notion of "visit wear" in the context of fish-eye pointers [59]. Using visual indications of history, they show that users were more readily able to trace their previous steps. Building on this work, Gutwin and Anton examined the extent to which users could remember their path after information history was removed [50]. Gutwin also carried some of these findings back to HCI, by integrating a "recency cache" in a list-interface to improve revisitation [51].

Beyond this, however, we also hypothesize that directly encoding interaction history is useful beyond revisitation. As we will demonstrate, even the most simple indications of history not only benefit revisitation, even more so, they impact the exploration patterns and insights of users.

3.2.3   Interaction History in Visualization

Broadly, several threads of visualization research have focused on interaction history. In formal terms, Jankun-Kelly *et al.* propose a model for capturing the exploration process [47]. This work enabled several extensions, including VisTrails from Bavoil *et al.* , which used formal models of exploration to support scientific provenance in visualization systems [46], and Shrinivasan and van Wijk, who propose methods of transferring these provenance techniques to visual analytics [60]. However, the combination of direct, per-

23

sonal representations of history in the HCI community has not be suitably transferred and explored in the context of data visualization (see Figure 3.2).

Direct vs. Indirect encoding

In visualization, interaction history widgets typically use **indirect encoding** to represent history in secondary displays. This spatial separation from the data allows history to be expressed using a diverse palette of design characteristics that will not interfere with existing visual encodings. For example, textual or graphic representations of history may be spatially organized as a linear sequence of items, on continuous timelines, using branching metaphors, or in network diagrams [41, 61]. In addition, these views support a broad set of operations on historical information such as navigation, editing, annotation, searching and filtering, and exporting [60]. For a more thorough examination of these displays, see [41].

Outside of Gutwin's "visit wear" study, examples of visually encoding interaction history directly onto the data are more difficult to come by. Since interaction is represented in the same space as the data, the design space is constrained to visual features that are separable from the visual encoding. However, direct encoding of interaction history on data has clear usability benefits because it situates history signifiers directly onto the data. For example, Willet's *Scented Widgets*, which places small data visualizations next to interface widgets to guide exploration, found that users exploring unfamilar data make up to twice as many unique discoveries [62]. Instead of relegating interaction history to a secondary display that requires a mental translation, direct encoding leverages preattentive processes to spatially put interaction history next to or on top of the data itself.

Personal History vs. Social History

A second distinction we make is the use of history to communicate personal interactions with the data or group-driven interactions with the data. While most work in this space has focused on facilitating collaboration, we believe that directly encoding interaction history can improve *personal data exploration* with a fraction of the overhead.

History-focused interface widgets in data visualization typically appear in the context of asynchronous collaboration [52], or are shown indirectly through secondary displays [41]. A relevant example similar to our proposed work is Wattenberg and Kriss [53] who, when describing the visual encodings used in NameVoyager, briefly mention directly encoding personal interaction histories (p. 556):

> *color by history ...causes any visited series to appear in gray... We refer to this as road-less-traveled navigation: Instead of using previous visits as a cue to importance, as in traditional social navigation interfaces, we treat it as a cue to staleness and hope to draw a users eye to new territory, thus suggesting a unique perspective to each user.*

We propose that this concept can be broadened into a general design principle for interactive data visualizations: directly encoding personal interaction histories, or HindSight. In the context of exploratory data visualizations and in contrast to indirect displays of history which capture a "moment in time", encoding history directly on the data frees users to explore new spatial organizations without losing context. We hypothesize that HindSight-inspired techniques will encourage personal exploration of data and yield benefits such as higher levels of engagement, more systematic exploration, and as a result, more diverse insights about a particular dataset. While we have included an experiment that targets these measures, we first discuss the design process of building interaction history directly into existing visualizations.

25

3.3 HindSight Design Process

The core idea of HindSight is that designers can architect visualizations not only by visually encoding data, but also by encoding their users' interactions in the visualization itself. In this section, we pose questions for designers when they are considering to apply HindSight – *how do we define history, how do we represent history, and is it worth it?* – and share the principles we have developed while applying HindSight to a range of existing visualizations.

3.3.1 What type of history is important to this visualization?

As we mentioned in the previous section, HindSight shifts our perspective of history from "How did I get here?" to "Where have I been before?" and "What is left to explore?". As a result, HindSight may be most beneficial for visualizations in which exploration is a design goal. For example, when interactive news visualizations reveal important context only after users hover over data, encouraging exploration may lead to more nuanced insights that complement the story.

On the other hand, HindSight is less suitable when it is important for users to retrace their steps. Since spatial encodings are likely already in use by a visualization, it is not able to represent sequence data without interfering with the existing design. While we see this as the primary limitation of direct encoding, designers must generally make informed decisions about framing the user's mental model of history.

What data entities best represent a 'unit' of history?: Since we can refer to data at various levels of abstraction in a graph (*e.g.* chart-level vs. data-level), it is important to carefully weigh the entities we choose when applying HindSight. For example, in the small multiples visualization in Figure 3.1.a, we could consider interaction with each chart as meaningful (encoding history at the chart level) or we could consider interaction

within each chart to be meaningful (for example, highlighting explored regions of the area graph). In this case, because chart reordering was a core interaction mechanism in the visualization, we encoded HindSight at the chart level, enabling visited charts to remain salient even as the data is reorganized. Additionally, encoding HindSight at the chart level encourages exploration of different topics in the MetaFilter visualization rather than secondary trends within a single topic. Choosing an appropriate level of coding for HindSight has the potential to unify exploratory goals with the capabilities of our perceptual system, making user history immediately available for further exploration and discovery.

What duration of user interaction represents *meaningful* interaction?: Interaction history is dynamic. Users may visit charts multiple times, or accidentally visit a chart when en route to another. In our initial pilots, we found that triggering a "visit" immediately was not ideal, whereas a short delay (*i.e.* 500ms) led to more predictable results. While definitive guidance on timing is beyond the scope of this work, a general principle is to delay for long enough that the visit is considered "intentional".

3.3.2 Which visual channels should be used?

One broadly applicable way of encoding interaction history is changing the opacity of the element after interaction. Opacity is just one of many visual channels that may be used, however. Designers should be aware of the relative efficacy of visual channels such as position and color, as well as concepts such as integral and separable channels [63]. A poor choice of encoding– significantly increasing line size, for example– may severely interfere with the other data in the visualization, especially as the user spends more time interacting. Here we give high-level guidelines for selecting visual channels based on the current design of the visualization and the goals of the designer. We categorize three use-cases for applying HindSight encodings:

- *augmentation:* when unused visual channels are available, augment existing data with additional visual encodings to the target visualization to show interaction history. For example, we identified opacity as an unused visual channel that could be used to encode interaction history in the area charts shown in Figure 3.1.a.

- *addition:* There is often empty space available in a visualization that can be repurposed for interaction history. When history can be represented in unused regions of a chart, modify unoccupied visual layers with interaction data. Transforming the background of a scatterplot into a heat map, for example, could clearly communicate regions of the plot that were already explored.

- *adaptation:* when no visual channels are available but displaying history is deemed important, adapt the target visualization to show interaction history by modifying visual channels that are already occupied by data. If there are no available visual channels, existing encodings can be manipulated to represent interaction history. Note that this approach runs the risk of undermining the perceptual benefits of some visual encodings.

How important is interaction history to the goals of the visualization?: One helpful way of assessing design tradeoffs is to consider interaction history as an additional data attribute. Weighing interaction history's impact on understanding in relation to other data attributes enables designers to use the *principle of importance ordering* to map both data *and* interaction history onto visual variables. For example, encouraging exploration in a complex news visualization may be critical enough to the success of a graph that representing interaction using color will yield stronger results than using that same channel to encode an additional data dimension.

Similarly, in The New York Times "255 Charts" visualization, there are many visual variables which could be used to encode history (see Figure 3.1.b). Line charts are the primary encoding in this visualization, representing the most important information – the

financial growth of the particular industry. Color is also used on each line chart to show whether a particular industry has grown (green) or fallen (red). Since color is a redundant encoding, we may decide that the benefits of representing interaction history outweigh the benefits of aligning multiple visual channels with a single dimension of data.

However, assessing the importance of encoding interaction raises the inevitable question: what are the benefits? While prior work such as Gutwin *et al.* suggest that showing users where they've been can help when revisiting previously visited elements [49, 51], it is not clear from existing research whether making users aware of their interaction history impacts any other aspects of the exploration process. The duration of this work, in particular our three experiments, are dedicated to examining this question.

3.4 Quantifying the Effects of HindSight

The goal of our study was to determine the effect of directly encoding personal interaction history on the following factors:

- **exploration behavior:** how does HindSight impact exploration behavior such as number of charts visited, total time spent exploring the data, and patterns of exploration?
- **post-interaction insight:** how does HindSight impact the insights that people recall immediately *after* interacting with a visualization?

To this end, we used a between-subjects design to test HindSight principles in three different interactive data visualizations. Two were selected to vary in complexity and design, and the third was chosen to draw comparisons with recent work by Boy *et al.* [4] that evaluates exploration and engagement in visualization. In each visualization, we tested conditions with and without HindSight:

- **control:** we present an interactive visualization in its original form, removing only

29

extraneous information

- **hindsight:** we apply a straightforward encoding of user's interaction history.

3.4.1 Procedure and Tasks

Participants were recruited through Amazon's Mechanical Turk (AMT) to participate in a maximum of one of our three studies. AMT is a crowdwork platform where "Workers" select from a range of available tasks, including research experiments [64, 65]. Each participant was randomly assigned to either the *control* (original-visualization) or *hindsight* (original with HindSight techniques) condition. Based on time data in pilot experiments, participants were paid $1.00 in order to exceed US Minimum Wage. All participants were shown a standard consent form before continuing.
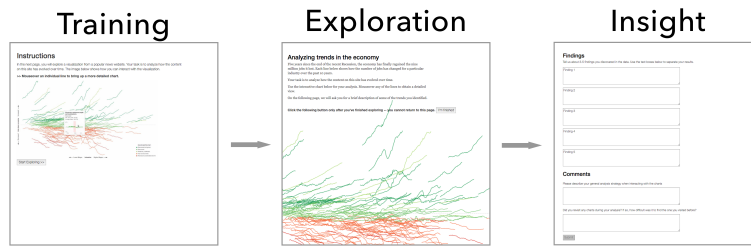
Figure 3.3: The experiment procedure to evaluate HindSight.

Our procedure consisted of three phases: *Training*, *Exploration*, and *Insight*. In the *Training* phase, we provided participants with an instruction page that briefly described their task and the interaction mechanisms in the visualization. For example, for the *meta filter* experiment participants were told:

> *In the next page, you will explore a visualization from a popular social media site. Your task is to analyze how the content on this site has evolved over time. The image below shows how you can interact with the visualization.*

In the *hindsight* condition, an extra sentence explained that visited charts would be made visually distinct, and an image showed HindSight being triggered.

Following training, the *Exploration* phase began with a paragraph that introduces participants to the visualization and their task. Participants were instructed that they may interact with the visualization without any minimum or maximum time limit. They were also reminded that after they finish, they would be asked to describe several of their findings. When participants finished exploring the visualization, they advanced to the *Insight* phase through a button press.

As a final step, participants entered the *Insight* phase. After the visualization was hidden, participants were instructed to describe 3-5 of their findings in individual text boxes. Additional text boxes were included to allow for more freeform comments about their experience.

3.4.2 Measures

Given the *Exploration* and *Insight* phases of the experiment, we draw on both quantitative and qualitative measures for evaluation. For quantitative exploration metrics, we build on work from Boy *et al.* [4], recording *visited* items and *exploration-time*. We also include the *revisit* metric from Gutwin *et al.* [49].

- *visited* : the number of unique charts that a person directly interacts with during exploration.
- *revisited* : the number of instances when a user interacts with a previously visited chart.
- *exploration time* : the total amount of time spent interacting with charts. We use this metric to try and capture *active* use of the visualization, mitigating when external distractions artificially inflate the time spent in the exploration phase.

For qualitative metrics, we referred to work by Saraiya *et al.* on analyzing insights

from interactive data visualization [35]. We used faceted coding, where independent coders mark what elements of the visualization (*e.g.* a particular topic or year) appear in the comments.

- *mentions* : the number of times a chart is directly referenced in findings during the *Insight* phase of our experiment.

Finally, we asked participants to describe their general analysis strategy and to reflect on the difficulty of revisiting charts.

3.4.3 Pilots, Analyses, and Experiment Planning

We conducted several pilot experiments using the *metafilter* visualization to help establish our measures and procedure. In response to concerns about the limitations of null hypothesis significance testing [66, 67], we model our analyses on recent visualization research that seeks to move beyond these limitations [68], primarily focusing on confidence intervals and effect sizes. Following Cumming [66], we compute 95% confidence intervals using the bootstrap method, and effect sizes using Cohen's d– which is the difference in means of the conditions divided by the pooled (*i.e.* both conditions') standard deviation. While we include significance testing and related statistics, it is with the intention of supplementing these analyses.

The results of our pilots showed some measures from the *Exploration* phase were non-normally distributed, according to a Shapiro-Wilk test. These measures include *exploration − time*, *visits*, and *revisits*, all of which were right-skewed with long tails. Because common transforms (*i.e.* log, square-root) did not cause a significant change in the Shapiro-Wilk result, we use the non-parametric Mann-Whitney test to compare the *control* and *hindsight* conditions.

Analyzing the findings left during the *Insight* phase, we turned to three independent coders and inter-coder reliability metrics. The coders were undergraduate students who

	metafilter	255charts	storytelling
control	44	57	99
hindsight	48	59	107
Total	*92*	*116*	*206*

Table 3.1: We tested HindSight using a between-subjects design on three visualizations. The table above shows participant numbers for each visualization, which were determined by running effect size and power analyses on pilot studies.

had little-to-no visualization experience, and were not involved in this project. The coders annotated each comment by assigning tags to indicate the entities mentioned (*e.g.*, the social media topic mentioned). Fleiss' Kappa was calculated to measure the agreement among the three coders [69]. We took the majority agreement when 2 out of 3 coders agreed on all entities mentioned in a given comment. If all coders disagreed, the comment was discarded from analysis.

In order to ensure our experiments included enough participants to reliably detect meaningful differences between the *hindsight* and *control* conditions, we conducted effect size and statistical power analyses. Specifically, we used pilot experiments to estimate the variance in our quantitative measures, and combined these with the observed means to approximate how many participants were needed. This procedure was repeated for each of our three experiments (see Table 3.1).

3.5   Visualization 1: Metafilter

We first chose to apply HindSight to a relatively simple interactive visualization. Many interactive visualizations people encounter on a day-to-day basis consist of a few views and simple interactions such as clicks and hovers to uncover more information. From an experiment control perspective: a simple visualization should lead to less variance between participants, making it more likely to detect reliable effects.

After evaluating several alternatives, we selected an interactive small-multiples area

HindSight: Experiment Results

MetaFilter   255 Charts   Storytelling

Exploration For each visualization, we plot the visit frequency of the participants in the control condition, hindsight condition, and a direct comparison of the two. The smaller 'thumbnail' visualizations show visit patterns for hindsight (orange) and the control condition (purple). The larger visualization maps differences in visitation patterns between hindsight and the control conditions.

(layout: hindsight control   difference)

HindSight 100 50 25 10 5 1   0 1   5 10 25 50 100 Original

Average number of visits given 100 hypothetical users.

a

b

c

HindSight appears to nudge users to visit different parts of the data.

Participants with HindSight visited more individual charts.

HindSight appears to change behavior more when used with complex visualizations.

Visited
0 5 10 15
p=.004, W=1426

Visited
0 10 20 30 40
p=.110, W=1971

Visited-Years
0 5 10 15
p=.153, W=5957

Revisited
0 5 10 15
p=.965, W=1062

Revisited
0 10 20 30 40
p=.028, W=2077

Visited-Countries
0 5 10 15
p=.496, W=5053

Exploration Time
0 20 40 60
p=.467, W=1150
d

Exploration Time
0 50 100 150 200
p=.016, W=2177
e

Exploration Time
0 50 100 150 200
p=.931, W=5308
f

Insight Similar to above, we plot the frequency of data mentioned in participants' findings (or insights) across each condition. The smaller 'thumbnail' visualizations show insight patterns for the hindsight and control conditions. The larger visualization shows differences in findings frequencies between the conditions.

(layout: hindsight control   difference)

HindSight 100 50 10 5 1   0.5 0 0.5   1 5 10 25 100 Original

Average number of mentions given 100 hypothetical users.

g

h

i

HindSight also impacted the insights users generated after exploring datasets.

In 255charts, users in the HindSight condition mentioned more charts of the center, while those in the control focused onthe perimeter.

Again, in low information datasets HindSight appears to have some effect, but not much.

Figure 3.4: Experimental results comparing basic HindSight encodings with three visualizations. Exploration metrics suggest that HindSight generally encourages more exploration and nudges users towards investigating different parts of the data.

chart - *The Rise and Decline of Ask MetaFilter*. Obtained from a popular data visualization blog [70], it depicts posting trends across topic categories in a community weblog. There were twenty area charts in total. Mousing-over any chart brought up a cursor at the corresponding x-axis (time) location on all other charts, and a toggle button allowed the charts to be reorded either by alphabetical order or post count. We used a between-subjects design with the following conditions:

- **control:** the original design of the visualization
- **hindsight:** interaction history was encoded through a small change in opacity. If a chart was visited for more than 500 milliseconds, it received a slight increase in opacity and became more salient in the visualization.

The original visualization and HindSight encoding can be seen in Figure 3.1.a. Pilot experiments with *metafilter* coupled with a power analysis indicated that at least 76 participants would be needed to detect a large effect (*e.g.* a difference of 3+ charts visited).

3.5.1 Results

We recruited 92 participants through AMT for this experiment. Through random assignment, we gathered 48 responses for the *hindsight* condition and 44 responses for the *control* condition.

Behavior/Interaction Analysis

Shown in Figure 3.4.d, the average participant in the *hindsight* condition visited more area charts ($M = 9.4$ visits 95% CI $[7.5, 11.3]$) than those in the *control* condition ($M = 5.4$ $[4.4, 6.5]$). Given the upper and lower limits of the confidence intervals, the average participant visits at least 1 additional chart with *hindsight*, and up to 7 more ($d = 0.75$ $[0.34, 1.11]$). There was little difference in participants' time spent interacting with charts in the *hindsight* condition ($M = 43.4$ seconds $[32.6, 65.6]$) compared to the *control*

condition ($M = 36.1$ $[25.7, 51.6]$, $d = 0.15$ $[-0.27, 0.53]$).

Qualitative analysis of visits indicate that participants in the *control* condition tend to focus on the top region of the chart (*i.e.* the top two rows). While this trend held for HindSight, additional visits were more evenly spread across the entire chart (see Figure 3.4.a).

Insight Analysis

92 participants left a total of 363 findings in the Metafilter experiment. Following the methodology, three people independently coded each finding to determine whether a specific posting topic was referenced. A statistical analyses of the 363 comments indicate strong agreement for the posting topics mentioned ($\kappa = 0.89$). For posting topic, there were 362 comments with majority agreement (*i.e.* at least two out of three coders agreed), and one comment with complete disagreement (this was discarded).

Qualitative analysis of posting topics mentioned in findings for the *metafilter* experiment indicate that participants in the *hindsight* condition overwhelmingly referenced the bottom region of the chart-grid more often than in the *control* condition (see Figure 3.4.g). In contrast, findings from participants in the *control* condition appear to tend are more evenly distributed across the chart-grid.

We also analyzed the number of unique charts referenced by a minimum number of participants (analyzed by 1, 2, ..., up to 5). From this data, we see very little difference between the *hindsight* and *control* conditions (see Figure 3.4.j). This suggests that in the *metafilter* visualization, HindSight did not encourage more findings overall, but different findings. Turning to the open-ended comments, this change in behavior may be the result of HindSight enabling more systematic exploration strategies. When asked about their approach, participants in the *hindsight* condition often responded with a clearly defined strategy similar to the following: *"I looked at every chart one by one, sorted by 'Count'"*.

3.6   Visualization 2: 255 Charts

The datasets people encounter on a daily basis are often larger and more complex than the *metafilter* visualization discussed in the last section. For this reason, we examined whether HindSight impacted behavior with more advanced interactive visualizations.

Towards this goal, we adapted a popular interactive visualization from The New York Times titled, "How the Recession Shaped the Economy, in 255 Charts" [71]. Shown in Figure 3.1.b, the 255*charts* condition includes 255 line charts distributed across the viewport in a scatterplot-like fashion. Each line in 255*charts* represents how a particular industry of the US Economy – Home Health Care Services or Air Transportation, for instance – grew or declined from 2004 to 2014. Mousing-over an industry's chart brought up a detailed view showing specific values, years, and industry information.

The original article included multiple stages with animations, transformations, and annotations, which the user controlled through scrolling. To better control our experiment, we isolated the part of the visualization where users are given the opportunity to freely explore the charts. We also repositioned the introductory explanation to avoid obscuring any part of the data.

The open-ended nature of 255*charts* coupled with its large data size makes it an ideal candidate for examining how HindSight impacts exploration with more complex data. Again, we used a between-subjects design with the following conditions:

- **control:** the design of the visualization as described above.
- **hindsight:** if a line chart was visited for more than 0.5 seconds, it received a slight increase in width and opacity to represent interaction history.

Running a power analysis on pilot experiments of 255*charts* indicated that at least 102 participants would be needed to reliably detect a large effect (*e.g.* a difference of 5 or more charts visited).

3.6.1 Results

We recruited 116 participants through AMT for this experiment. Through random assignment, we gathered 59 responses for the *hindsight* condition and 57 responses for the *control* condition.

Behavior/Interaction Analysis

Shown in Figure 3.4.d, the average participant in the *hindsight* condition visited more charts ($M = 28.4$ visits $[23.3, 34.2]$) than those in the *control* condition ($M = 20.7$ $[17.9, 24.2]$). Given the limits of the confidence intervals, the average participant will at least visit the same number of charts with HindSight, and up to 16 more ($d = 0.44$ $[0.12, 0.75]$). In addition, the average participant in the *hindsight* condition appears to revisit more charts ($M = 7.9$ *visits* $[5.6, 13]$) than the *control* condition ($M = 4.6$ $[3.2, 6.7]$, $d = 0.32$ $[-0.06, 0.57]$). Similarly, we see that the average participant in the *hindsight* condition may spend more time interacting with charts ($M = 119.7$ *seconds* $[93.1, 166.9]$) compared to the *control* condition ($M = 79.1$ $[59.7, 110.6]$, $d = 0.36$ $[-0.02, 0.68]$).

Qualitative analysis of visits in the 255*charts* experiment indicate that participants in the *hindsight* condition tend to focus more attention than the *control* condition on industries in the center of the visualization, where the data density is at its highest (Figure 3.4.b). In contrast, participants in the *control* condition appear to focus on charts in periphery, particularly the top left and bottom right.

Insight Analysis

116 participants recorded a total of 492 findings in the 255*charts* experiment. Three people independently coded each finding to determine whether a specific industry was referenced. The statistical analyses of the 492 comments indicate moderate agreement for the industry mentioned ($\kappa = 0.59$). For the industry mentioned, there were 444 comments

with majority agreement (*i.e.* at least two out of three coders agreed), and 48 comments with complete disagreement (these were discarded).

Qualitatively, the maps showing referenced findings (Figure 3.4.h) indicate trends that mirror behavioral patterns. When compared to the *control* condition, participants in the *hindsight* condition were more likely to reference industries in their findings that were spatially in regions of high data density. This is also reflected in the map of industries participants visited (Figure 3.4.b). We also analyzed the number of unique charts mentioned by at least 1 participant, 2 participants, etc (see Figure 3.4.j). While most findings still gravitated towards a handful of charts, in contrast to *metafilter*, the trends in Figure 3.4.h and Figure 3.4.j suggest that HindSight not only encouraged a different set of findings, but more diverse set of findings. These benefits were reflected in open-ended comments: *"... it was relatively easy to find the chart that I wanted to see again because it had been changed to a bolder and darker line which is a great feature seeing as how there are a whole bunch of lines mixed up together."*

It's possible that the increased data in 255*charts* amplified the effect of HindSight in comparison to *metafilter*, however, more experiments would need to confirm this hypothesis.

3.7   Visualization 3: StoryTelling

We turn to existing research in exploratory data analysis to choose our third visualization. In a recent study, Boy *et al.* examined the impact of storytelling techniques across several quantitative measures of user engagement [4]. While we adapt several of the measures they use throughout our experiments, we also replicate one of the conditions of their experiment, thanks to their releasing the study's experiment materials online.

The *CO2 Pollution Explorer* was one of the primary interactive visualizations in Boy

et al. [4]. Consisting of a world map, a year selector, and a line chart showing a country's pollution over time, this interactive visualization allows users to compare pollution from a particular country across several decades (see Figure 3.1.c). User interactions included the ability to hover on a country to highlight the corresponding trend on the line chart, and click on a year to update the map and year marker in the line chart. As in previous visualizations, we used a between-subjects design with the following conditions:

- **control:** the design of the visualization as described above.
- **hindsight:** if a country shown in either the map, list, or line chart view was visited for more than 0.5 seconds, its opacity increased slightly in each view. Similarly, the color of a visited year button changed from gray to light red, and a light red border also appears in the line chart indicating the year's range.

Pilot experiments with *storytelling* coupled with a power analysis indicated that at least 177 participants would be needed to reliably detect a medium effect (*e.g.* a difference of 5 or more years or countries visited). In contrast to the *metafilter* and 255*charts* conditions, the variance in behavioral metrics in the *storytelling* pilots was higher, leading to a larger number of participants needed.

3.7.1 Results

We recruited 206 participants through AMT for this experiment. It took approximately one day to gather all responses. Through random assignment, we gathered 107 responses for the *hindsight* condition and 99 responses for the *control* condition.

Behavior/Interaction Analysis

The behavioral metrics for the *storytelling* visualization differ slightly from the previous graphics. Specifically, instead of reporting "visited" items, the original work from Boy *et al.* distinguishes between years visited and countries visited. We adopt their approach

here.

The visit quantities for years and countries were largely the same. The average participant in the *hindsight* condition visited a similar number of countries ($M = 7.2$ $[5.9, 8.7]$) as those in the *control* condition ($M = 7$ $[6, 8.1]$, $d = 0.04$ $[-0.24, 0.3]$). Participants in the *hindsight* condition also appeared to visit a similar number of years ($M = 6.7$ $[5.6, 8.5]$) as the *control* condition ($M = 5.4$ $[4.4, 6.9]$, $d = 0.19$ $[-0.09, 0.47]$). Qualitatively, the maps showing which years and countries participants visited were largely similar (see Figure 3.4.c), particularly when compared to the differences in the *metafilter* and 255 maps (Figure 3.4, a and b). In terms of timing, participants in the *hindsight* condition spent roughly the same amount of time in the exploration phase of the experiment ($M = 140.3$ $[117.4, 180.7]$) compared to the *control* condition ($M = 148.2$ $[123.3, 188.6]$, $d = -0.05$ $[-0.32, 0.22]$).

Meta Analysis. In their study, Boy *et al.* analyzed additional metrics such as hover and click interactions. We also tracked these metrics in our experiment to facilitate a meta-analysis with the results of [4]. While the raw data from Boy *et al.* was not available, we carefully inferred means from the confidence interval plots in [4]. Their experiment hypothesized that the addition of storytelling prompts would increase several of these measures. However, they found the opposite occurred – users in the experiment condition generally interacted less with the visualization. In contrast, we found that HindSight produced small gains across the board in identical behavioral metrics when compared to our control (Table 3.2).

Insight Analysis

206 participants left a total of 831 findings in the *storytelling* visualization. Coders labeled two dimensions – whether a specific country or year was referenced in the comment. A statistical analyses of the 831 comments indicate strong agreement for the country men-

| | HindSight, 2016 | | Boy *et al.* , 2015 | |
Metrics	**Exp**	**Control**	**Control**	**Exp**
meaningful interaction	54.8	48.6	44	33
meaningful hover	22.8	19.8	35	26
meaningful click	32.2	28.8	8	6
semantic - inspect	14.6	13.6	26	17
semantic - connect	8.1	6.2	10	8
semantic - select	21.1	19.8	5	3
semantic - explore	8.7	6.6	3	2
semantic - filter	2.44	2.39	0.2	0.1
exploration time	140.3	148.2	108.8	54

Table 3.2: Meta-analysis of HindSight applied to one of the primary visualizations from Boy *et al.* , 2015. While the control condition in the present experiment led to generally higher results, HindSight appears to reliably outperform the other conditions– past and present.

tioned ($\kappa = 0.87$), and substantial agreement for the year mentioned ($\kappa = 0.76$). For the country and year mentioned respectively, there were 821 and 826 comments with majority agreement (*i.e.* at least two out of three coders agreed), and 10 and 5 comments with complete disagreement. The latter were discarded.

Qualitatively, behavioral visitation trends did not transfer to year or country references in the findings (Figure 3.4.i). While countries of increased interest in *hindsight* appear to reflect the most significant stories in the data, the effect is not strong enough to make more generalizable claims. We also found that participants in the *control* condition referenced a more diverse set of years from the visualization while participants in the *hindsight* condition focused their findings on major trends in the data. We will contextualize these findings in the discussion section.

3.8   Discussion

HindSights simple encoding of interaction history generally changed users' behavior as well as the details that they remembered. In both *metafilter* and 255*charts*, we saw

significantly increased interaction with data. As indicated in the insight maps (Figure 3.4), users also reflected on a more diverse set of findings with HindSight, although they identified dominant outliers and trends less often.

In the *storytelling* condition, we noticed slightly different results. There were few differences in the amount of data explored (*e.g.* visited countries, years, and exploration time). We did see, however, a small improvement in most behavioral exploration metrics recorded in the original study (Table 3.2). This change raises the question: when should we expect techniques like HindSight to cause a noticable change in user performance?

The results of these experiments generally confirm our hypothesis that subtle indications of interaction history impact user behavior in data visualizations, while the degree of impact may vary across different visualizations, e.g., 255*charts* versus *storytelling*. Our goal now is to discuss the implications of these findings more broadly and make recommendations for the use and development of HindSight.

3.8.1 Benefits on Exploration, Engagement and Insights

We found that HindSight generally encourages people to interact with more data. We also observed that HindSight impacts the findings that users report after viewing a visualization – nudging users towards areas that are typically unexplored in a visualization (for example, areas of high data density). While it is difficult to make value judgements about exploration patterns, our findings suggest that at the very least, HindSight redirects attention to *different* data. Whether more interaction is a good thing – for instance leading to a deeper understanding of the dataset as a whole – remains an open question for future research.

The quantitative results suggest that the effects of HindSight may be amplified by larger, more complex data visualizations. This observation is supported by the comparison of results between *metafilter* and 255*charts* visualizations. As the amount of data

between *metafilter* and 255*charts* increased (20 to 255), the effect of HindSight on exploration time also increased (see Figure 3.4.e).

We also believe that HindSight improves levels of the sustained attention on a visualization, which is one marker of engagement. This raises the question: Why does HindSight nudge exploration behavior?

One plausible explanation is that HindSight helps negate attentional biases related to the spatial placement of data on a page by making people more aware of their own navigational patterns. As an example, the visit spatial pattern of the control condition in *metafilter* appears to mirror the typical F-shaped gaze patterns observed in eye-tracking studies of product websites [72]. In these website studies, users typically explored the top rows and down the left side of a webpage, avoiding the center. While some form of top-to-bottom bias still holds for HindSight in the *metafilter* visualization, visit patterns and findings suggest users with HindSight engaged with the bottom row of charts much more frequently than in the control condition. Another possible explanation is that HindSight gamifies interaction by providing immediate visual feedback and anchor points from which users can systematically navigate complex data.

The *storytelling* condition is of note because we did not observe the same changes in behavior and insight. There were several factors that made the *storytelling* visualization unique, however– countries were not available to interact with due to limitations of the underlying dataset, and several participants commented that the animated pollution clouds interefered with their ability to select European countries. In the insight maps, country references were largely focused on just a handful of nations, suggesting that *storytelling* contained fewer significant insights that could be gleaned from the data. The regions of the map in which HindSight provoked the most findings tended to align with the major pollution contributors (Figure 3.4.i). These factors suggest that HindSight may help users more systematically navigate datasets where fewer insights are to be found. In other

words, when considering techniques like HindSight, designers should ensure that their data contains many possible stories that may benefit from exploration (*i.e.* not just a few outliers).

Overall these results confirm that HindSight impacts user engagement and exploration patterns. As visualization research continues to add language and metrics that capture user interaction strategies (*e.g.* Ottley *et al.* [5]), techniques such as HindSight should be developed in parallel to help support the cognitive task of exploration in interactive visualizations.

3.8.2 Low Technical Barrier

The cost of implementation effort versus the added value to users is a tradeoff rarely discussed in visualization design. We see this dynamic as one of the core advantages of HindSight. HindSight can be applied to existing visualizations by adding just a few lines of code and without changing any technical infrastructure. For example, modifying the visual encoding of data in response to mouse behavior is a trivial change in dominant visualization libraries such as d3js. This enables designers to leverage the benefits of interaction history we have established without having to dramatically alter existing code bases (necessary for indirect coding approaches) or by adding server-side storage mechanisms (necessary for social applications). We envision future research targeting the long-term support of visualization navigation (*i.e.* beyond a single-session), similar to the topic of analytic provenance from the visual analytics community [32].

3.8.3 Design Tradeoffs

HindSight's direct encoding of interaction history, much like Gutwin's "visit wear", can be compared to the concept of direct manipulation as defined by Shneiderman [55] and

following research. While changes in visual encoding occur passively, they are triggered by explicit actions. This encoding creates a continuous and dynamic indication of data of interest, allowing users to rapidly and incrementally tweak their interaction strategy.

As a result, some of the same advantages of direct manipulation outlined by Shneiderman and Plaisant can also be considered within the context of HindSight [73]. Immediate visibility of user actions a) results in reduced error rates, b) promotes usage by novices with minimal knowledge or instruction [55], and c) encourages exploration [73]. While we did not investigate error rates or visualization expertise, exploration benefits are reflected in our results. Looking forward, the concepts explored in direct manipulation (*e.g.* reversible actions) may serve as inspiration for future research related to HindSight.

We must also consider the constrained design space of directly encoding interaction history onto visualizations. HindSight's definition of history to this point has shifted from the traditional notion of "How did I get here?" to instead focus on "Where have I been before?" and "What is left to explore?". In designs that already map several data variables to visual variables, identifying additional separable channels is difficult [63]. Over-representing history information, for example, may interfere with existing spatial encodings of data. While there is no silver bullet for design, the examples and principles we lay out in the design space are intended to help architects of interactive visualizations maximize benefit and minimize tradeoffs.

3.9 Conclusion

As visualization becomes more widely used by everyday people, research should focus on low-barrier interaction support techniques that can benefit people without expertise or training. We believe that HindSight offers an opportunity to do exactly that.

Building on preliminary evidence from Gutwin *et al.* , we used Maes and Wexelblat's

interaction history framework to identify gaps in existing interaction history encoding approaches used in visualization. A direct encoding of personal interaction history not only is trivial to apply to many web-based visualizations, but as we discovered, can yield high benefits for the low cost. In three experiments, we found simple applications of HindSight techniques changed exploration behavior – increasing the amount of data covered and the range of insights articulated after encountering a visualization.

HindSight provides cognitive support for interaction through visual encodings, and yields benefits beyond enabling users to "retrace" previous steps. Our results suggest that HindSight may hold immediate benefits for practitioners. News organizations who are building expository visualizations similar to the designs we tested in our experiment may use HindSight to help encourage their users engage more deeply with the data presented. As visualization research continues to define and understand the interaction process, techniques like HindSight should be further developed and evaluated to ensure users have as much cognitive support for exploratory data analysis as possible [1].

---

[1] **To facilitate future work, all experiment materials, participant data, and analyses scripts are available online:** `https://github.com/wpivis/hindsight`.

Chapter 4

Quantifying the Effects of Text-Based Search in Visualization

Search mechanisms in interactive visualizations are thought to support users' personal exploration and discovery.

a) In this example, searching "Lasseter" reveals movies written by John Lasseter.

b) The partial search "Steph" reveals writers including Stephen Mazur and Stephanie Meyer.

c) Searching "Toy" reveals the Toy Story series, which uses the movie name attribute rather than the names of writers.

Many existing interactive visualizations can be augmented with search to enable diverse information seeking goals.

Board of Directors (The Wall Street Journal)

Search provides a direct route to identifying companies of interest.

How the Recession Reshaped the Economy (The New York Times)

Search enables users to find a particular industry of industries sharing keywords.

College Admissions (experimental)

Search allows users to find a college of interest.

Exoplanets (experimental)

Search allows users to find a planet of interest.

Results from five crowdsourced experiments indicate that search shapes users' experience and performance with interactive visualizations.

Figure 4.1: Search mechanisms in interactive data visualizations have been used sporadically throughout research and in practice. Little is known, however, about how search impacts how people interact with visualizations. We contribute an analysis of search mechanisms in visualization. Our experiment results indicate that most users will use search when available, and that search leads to positive increases in measures related to engagement. (The example on the left is from an interactive visualization *Women in Films* on the web [11].)

The widespread use of text-based search in user interfaces has led designers in visualization to occasionally add search functionality to their creations. Yet it remains unclear how search may impact a person's behavior. Given the unstructured context of the web, users may not have explicit information-seeking goals and designers cannot make assumptions about user attention.

To bridge this gap, we observed the impact of integrating search with five visualizations across 830 online participants. In an unguided task, we find that (1) the presence of text-based search influences people's information-seeking goals, (2) search can alter the data that people explore and how they engage with it, and (3) the effects of search are amplified in visualizations where people are familiar with the underlying dataset. These results suggest that text-search in web visualizations drives users towards more diverse information seeking goals, and may be valuable in a range of existing visualization designs.

4.1   Introduction

Text-based search is widely used on the web in order to enable users to meet a variety of goals. Whether it is to browse between webpages, locate a keyword of interest on a particular page, or facilitate quick actions that shortcut tedious manual navigation on mobile devices, search has largely become an interface expectation and necessity. Thus, it comes as no surprise that data visualization designers have begun to add search to the visualizations they create for the web.

To demonstrate search within the context of data visualization, consider the *Women in Films* visualization in Figure 4.1 that explores gender diversity in high grossing films. A text-based search box at the bottom left allows users to enter either a writer or a film name. After three characters, any film whose writer or name matches the substring is

highlighted, while others fade out. This functionality empowers users to rapidly search for specific films without resorting to an exhaustive, guess-and-check strategy.

Despite these clear benefits, it remains unclear how the mere *presence* of text-based search impacts how website visitors explore data. When people explore a visualization, they may have explicit information-seeking goals (*e.g.* looking for a particular point of data), implicit information-seeking goals (*e.g.* opting to meander through a dataset until a goal is formed), or they may arrive at a website with no data-centric goals at all. From a design perspective, content creators may need to weigh several questions when considering to add search functionality to their visualizations:

- If search is built into a visualization, do people notice and use it?

- How does search impact a user's experience of the visualization? Does it change their goals or interaction patterns?

Given the resources of time and effort that it takes to create a compelling data visualization, designers cannot simply assume people will use search, or that search will benefit exploration. From a research perspective, it is unclear whether making relatively small additions to a visualization, like adding text-based search, results in a significant difference in how the user will engage with data. Motivated by the intuition of adding search to visualizations and the unanswered questions of its benefits and trade-offs, we isolate and quantitatively study its effect on users' goals and behavior in the context of open-ended web exploration.

Defining and bounding search. "Search" has many definitions in human-computer interaction and data visualization. For the scope of this work, we refer to search as **text-based search functionality integrated with interactive visualizations**. To clarify, below are a set of juxtapositions with altering definitions and scopes of "search".

Search as a task vs. search as a functionality. A search task in HCI can refer to a user task, such as seeking information in a system. Search functionality, then, is defined

broadly as the features the system provides to support users to complete their tasks. This might mean, for example, menu bars and button layouts in addition to text-based search.

Faceted search vs. text-based search. Faceted search includes user-interface functionalities for accessing information organized according to a faceted classification system, which can allow users to retrieve certain parts of information by applying multiple filters. Text-based search, in contrast, generally refers to functionality which accepts text input and displays results matching the input.

Herein we contribute a study on the effects of text-based search in interactive data visualizations. In an experiment with five stimuli and 830 crowdsourced participants, we quantify how search can shape user behavior and goals with visualizations on the web-even when they are not explicitly given a task to complete.

The results of these experiments suggest that (1) people generally use search when it is present, (2) the presence of search encourages people to actively look for individual data items, (3) search encourages users to spend more time examining detail in the data, and (4) search nudges users towards more diverse exploration patterns. Finally, we find that these effects are modulated when search accompanies a dataset in which people have no familiarity.

4.2   Related Works

The widespread use of search in general computing systems has led to taxonomies and in-depth studies on the design space of search within the HCI community. While a full review of this space is beyond the focus of this work, we find that recent work from Wilson *et al.* is particularly relevant [74, 75, 76]. In *Search User Interfaces* [74], Wilson described a design space of search user interfaces (SUIs), including issues such as faceted search and auto-complete. These results directly inform the dimensions we consider of

the design space of search in visualization. Additionally, the metrics from Wilson *et al.* 's evaluation of search interfaces, such as how search can lead to engagement with individual pieces of information, inform the metrics we use in our experiment [76].

Most research on search in visualization has focused on visualizing the results of search queries rather than search as an interaction mechanism. Nevertheless, there is some overlap in this thread of prior work and the goals of this study. To that end, we turn to SUIs in visualization.

4.2.1   Search User Interfaces and Visualization

Visualization has been used extensively to support users' search processes. With the growth of SUIs, structured 2D visualizations were introduced to display search results to support or substitute standard results lists [74]. Several forms of visualizations have been explored in these systems. *Treemaps* were used to show search results in ResultMaps [77], an interface to a digital library. *Faceted search* was used in the systems including Dotfire [78], Envision [79], and List and Matrix Browser [80], grouping specific facets of metadata using both the horizontal and vertical axes. *Timelines* were used in Perspective Wall [81] and Continuum [82] to display the search results in the form of time series. More recently, more complex visualizations have been created to support search systems. In PivotPaths [83], after typing search keywords, the user can explore the search results of faceted information resources displayed in an interactive visualization.

What is common between "search mechanisms for visualization" and "visualizing search results" is that they both have visualization and search components. This raises considerations for the present work, such as the impact of search on the visual display. These works also differ from the present focus in several ways. First, many prior systems do not support textual search, rather relying on graphical methods to construct queries [78, 82]. Second, many systems use search as the starting point for analysis, meaning

that subsequent searches change the dataset display in the visualization [79, 80, 81, 83]. Of these systems, ResultMaps most closely resembles the use of search as an interaction mechanism. In ResultMaps, an initial visualization of the data is given as a treemap, and search is used as a means to highlight sub-sections of the treemap.

4.2.2 Query-Based Interfaces

Query-based interfaces are part of a long thread of research in data visualization. Queries are core components of well-known systems such as Polaris [84] and HomeFinder [85]. Evaluating query interfaces consisting of sliders, Ahlberg *et al.* found that queries enabled people to quickly hone in on data of interest. Keim and Kriegel emphasize the notion of using boolean logic to join queries and ask more complex questions of data [86]. Text-based search could potentially be used as a mechanism for more complex queries, using schemes such as the ones described here.

4.2.3 Natural Language Interfaces

Setlur *et al.* 's Eviza system [87], a *natural language interface* for visual analysis, is closely related to the focus of this work. Eviza uses a text-based search bar (or voice) to allow users to ask questions of the data. In a user study, Setlur *et al.* found that users produced queries aligning with several visualization tasks: navigation, calculation, comparison, and more. Our goal is complementary- acknowledging that search mechanisms have been included in prior systems and visualizations on the web, and that they will become more powerful thanks to work similar to Setlur *et al.* - how do these mechanisms shape users' experience and understanding of a visualization?

4.2.4 Design of Search in Visualization

Even after narrowing our focus to text-based search on the web, there are a variety of potential design choices- some of which are unique to data visualization. As opposed to the typical results page of a search engine, designers must bear in mind the perceptual interactions between visual encodings in a visualization, such as integral and separable features [63]. Motivated in part by these challenges, as well as the search design space articulated by Wilson in *Search User Interface* [74], we use the following characteristics to describe how visualizations in the past have defined search:

- *search scope:* Do searches access just the primary labels (often names) of the data or do they access the full dataset, including metadata?

- *trigger:* How should search be triggered? Search can be triggered, for example, by clicking a "search" button, pressing an "enter" key, or updating continually as the user types.

- *autocomplete:* As the user types, does the search box suggest queries based on the dataset?

- *transition:* How will the user be notified that the results have been updated? In most search interfaces, only the search results are shown, and the others are hidden from the user. While this may be desirable in some cases for data visualization, it's also possible to increase the saliency of selected data elements, decrease the saliency of remaining data, or lend focus to search results through automated zooming.

- *encoding change:* What visual encoding changes will accomplish the aforementioned increases and/or decreases in saliency (*e.g.* color, opacity, width, size)?

The results of categorizing several prior research systems and visualizations on the web are shown in Table 4.1. Besides these examples, text-based search has also been supported in some visualization development tools, such as Prefuse [104] and Tableau Software [105], where visualization designers can choose from different design options

Source	Year	Title	Search Scope	Trigger	Autocomplete	Transition	Encoding Change
Paper	2002	SpaceTree [88]	tree node name	on click	(unclear)	highlight	color
Web	2003	WordCount/QueryCount [89]	words or queries	on enter	no	filter out other data	position
Web	2004	Zipdecode [90]	zip codes	while typing	no	highlight and zoom in	color
Paper	2006	NameVoyager [31]	baby names	while typing	no	filter out other data	position
Paper	2006	TimeTree [91]	person or position names	on click	no	highlight	color
Paper	2007	NewsLab [92]	news content	(unclear)	(unclear)	(unclear)	(unclear)
Paper	2007	VisLink [93]	words	(unclear)	(unclear)	highlight	color
Paper	2009	ResultMaps [77]	metadata	on click	(unclear)	highlight	color
Paper	2010	VizCept [94]	node names	on click	(unclear)	highlight	color
Paper	2010	GeneaQuilts [95]	any entry or attribute	(unclear)	(unclear)	highlight	color
Paper	2013	GPLOM [96]	car properties	while typing	yes	highlight	color
Paper	2014	Footprints [97]	document text content	on enter	(unclear)	reposition	position
Paper	2014	Overview [98]	document text content	on click	(unclear)	highlight	color
Paper	2015	VAiRoma [99]	location or article names	on click	(unclear)	highlight	addition
Web	2015	Clustergram [100]	gene names	on click	yes	highlight and zoom in	color and size
Paper	2016	ResViz [101]	staff names	(unclear)	(unclear)	(unclear)	(unclear)
Web	2016	Who Marries Whom [102]	job names	on enter	yes	highlight	opacity and size
Web	2016	Women in Films [11]	film names	while typing	no	fadeout other data	opacity
Web	2016	NBA 3-Point Record [103]	player names	on enter	yes	fadeout other data	opacity

Table 4.1: Text-based search has appeared in multiple visualizations throughout research and the web. The above are a sample. We categorize each across several dimensions, including the scope of the search, how the encoding changes, and others. Notably, some prior research systems do not contain sufficient detail to determine how text-based search is used in the visualization.

related to search. Researchers have also expressed intuitions on the potential benefits of search. For example, in NameVoyager [31], where users can search baby names by prefix, the authors mention: *"A user might not think that searching the data set by prefix would be interesting, but seeing the striking patterns for single letters like O or K could encourage further exploration."* In a study on the social impact of NameVoyager [9] by Heer *et al.* , search functionality was also specifically mentioned: *"Many participants searched for their own occupations and those of friends and family."*

These works including the search examples and the social impact studies motivate the need to isolate and quantitatively study the broader effect of search, and inform our experiment design.

4.3   Exploring the Impact of Search

Our study on the effect of text-based search on visualizations aims to investigate open-ended user exploration on the web, where users may not have explicit analytical goals. We aim to examine (1) how users' exploration strategy is influenced by the **presence of**

search, and (2) how users' exploration behavior is influenced by the **use of search**.

We used a **between subjects design** in which each participant was randomly assigned to either the **no search** or **search present** condition. In the *search present* condition, a search box was always present in the visualization, enabling text-based search. Functionally, users had to click the text-box and type queries to activate the search-based highlighting. In order to maintain ecology validity in the study, *i.e.* recognizing that users may pursue open-ended exploration rather than specific data-seeking tasks, we did not force a user to use search when it was present. In the resulting analyses, therefore, we focus in part on the group of participants who **used search**.

By drawing on analytic approaches from several recent studies examining user behavior and performance with interactive visualizations on the web [4, 12, 106, 107, 108], we frame our research questions as follows:

- **self-reported exploration strategy:** does the presence of text-based search impact peoples' reported exploration strategies? When search is present, what proportion of users make use of it? Does dataset familiarity matter?

- **exploration behavior:** does the use of text-based search impact measures of behavior, such as total exploration time, the location of data investigated, or the proportion of time spent viewing detailed information about chart elements?

4.3.1 Procedure and Tasks

Participants were recruited through Amazon's Mechanical Turk (AMT) to participate in a maximum of one of the five visualization stimuli. Each participant was randomly assigned to either the *no search* or *search present* condition. Based on completion times in pilot experiments, each participant was paid $2.00 in order to exceed US Minimum Wage. All participants viewed an IRB-approved consent form.

Our procedure consisted of four phases: *Training*, *Exploration*, *Insight/Strategy*, and
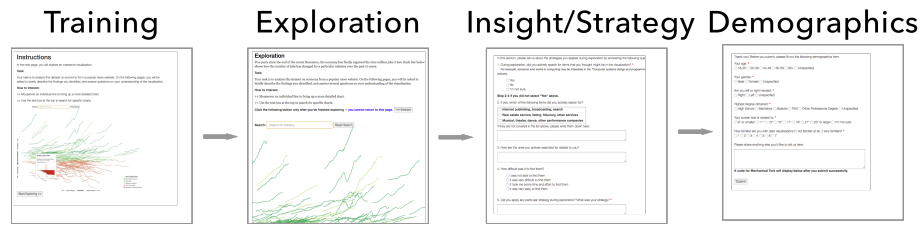
Figure 4.2: In our experiments with five visualizations, participants completed a training phase before heading to the exploration section. When they were finished exploring the interactive (no time limit), they moved to the next section where they describe their insights and strategies of exploration. In the final section, they provided demographic information.

Demographics.

Training: we provided participants with an instruction page that briefly described their task and the interaction mechanisms in the visualization. For example, for the 255*Charts* experiment participants were told:

> *In the next page, you will explore an interactive visualization. Your task is to analyze data on the economy from a popular news website. On the following pages, you will be asked to briefly describe the findings you identified, and answer questions on your understanding of the visualization.*

Participants were shown an animation of the interactive features available. In the *search present* condition, an extra sentence explained that the text box could be used to search for specific charts. No other indication of search functionality was provided.

Exploration: The *Exploration* phase began with a paragraph that introduces participants to the visualization and their task. Participants were instructed that they may interact with the visualization without any time limit. When participants indicated they were finished exploring the visualization, they advanced to the next phase.

Insight/Strategy: Participants were asked about findings they made in the visualization and the strategies they used during exploration. Specifically, participants were asked

57

During exploration, did you actively search for items that you thought might be in the visualization? They were then asked to list any such data items they specifically sought out during their exploration.

This protocol included additional steps to help ensure reliability in participants' self-reported answers. First, an example case was provided tailored to the visualization stimuli to help understand the question, *e.g.*, in 255*Charts*, the example was "someone who works in computing may be interested in the 'Computer systems design and programming' industry". Second, we included options for uncertainty in the single-choice response, *i.e.*, the participants chose among yes, no, and not sure. Third, we provided participants with a list of items they interacted with as a memory trigger. Specifically, participants who indicated they had actively sought specific data items, were asked to select which data items they sought, choosing from a dynamically generated list of the items they interacted with for more than 500ms. We refer to these engagements with data items as "visits"; the 500ms threshold mitigates accidental visits from stray mouse movement.

Demographics: Participants provided basic demographic information.

4.3.2 Experiment Stimuli

Each of the following visualizations were equipped with two conditions: *no search* and *search present*.

VIS 1: 255 Charts (The New York Times)

The first visualization we augmented with text-based search is from The New York Times, titled "How the Recession Shaped the Economy, in 255 Charts" [71] (see Figure 4.3). We refer to this as 255*Charts* through the remainder of this chapter.

Representation and Data: 255*Charts* includes 255 line charts distributed across the viewport in a scatterplot-like fashion. Each line in 255*Charts* represents how a particular
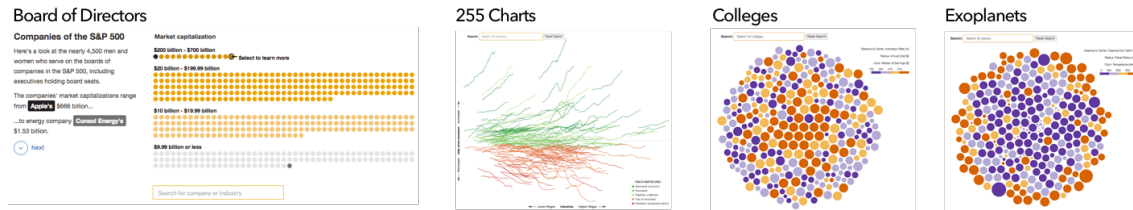
Figure 4.3: Experimental stimuli used to evaluate the effects of text-based search on visualization use and exploration. Each stimuli has been augmented to include search. From left to right: "Inside America's Boardrooms" from the Wall Street Journal- a multi-section visualization exploring company leaders. "How the Recession Reshaped the Economy, in 255 Charts" from The New York Times- showing how industries recovered or fell after the recent US recession. The final two visualizations are used to test specific hypotheses about the value of visualization, *e.g.*, whether the general familiarity of the dataset impacts the likelihood of users making use of search. (Not shown) An identical version of the third chart was also tested, with anonymized college names.

industry of the US Economy – Home Health Care Services or Air Transportation, for instance – grew or declined from 2004 to 2014. Mousing-over an industry's chart reveals a detailed line-chart view showing specific values, years, and industry information.

Search Design: For participants in our *search present* condition, the search box appeared at the top-left, allowing users to search "Industry Names" with auto-complete available. Search is triggered by an update of each character, allowing partial searches (*e.g.* "comp" for "computer sales" or "computer engineering").

VIS 2: Board of Directors (The Wall Street Journal)

We augmented an interactive visualization from The Wall Street Journal titled, "Inside America's Boardrooms" [16]. This scrolling visualization includes multiple stages with the same basic view (a grid of dots, see Figure 4.3, far left).

Representation and Data: The *Board of Directors* visualization includes companies from the S&P 500. The companies are represented by colored dots, and grouped into views, where they are sorted according to the market capital, the percentage of directors

who are women, and other related fields. The user can navigate through the views in a storytelling form by clicking the "Next" button on the lower left, or jump to a certain view by clicking buttons at the top of the visualization. Mousing-over a company brings up a detailed view, including the company's name, the industry it belongs to, and a list of other data attributes.

Search Design: Search was added on the bottom-right of the visualization, supporting queries on "Company Names" and "Industry Names" with auto-complete enabled. Search was triggered as each character was typed, and partial queries were possible. To display the search results, the selected data was highlighted by dark gray outlines, while unselected charts decreased slightly in opacity. The search box remained visible across all views.

VIS 3-5: Familiar and Unfamiliar Bubble Charts

One factor worth considering in text-based search is that its effectiveness may be limited by whether a person *knows what to search for*. In other words, does the familiarity of the dataset impact search behavior? It is with this in mind that we designed three additional visualizations of similar form (bubble charts), but with varying familiarity (Figure 4.3).

Representation and Data: The data sources and mapping for the datasets are as follows:

- *Colleges*: we selected 300 colleges from the College Scoreboard dataset [109]. Each college was represented by a circle, of which the radius, color, distance to center mapped to the college's annual cost, median earning of the students, and admission rate.

- *AnonColleges*: we used exactly the same data source and mapping as *Colleges*, except we anonymized the names of colleges. College names were anonymized via a script that combined fictitious town names and a typical college/university prefix

or suffix (*i.e.*, *X* university, university of *X*, *X* community college, etc.)

- *Exoplanets*: We selected 300 data points from the extrasolar planets dataset, to control for data size relative to the college datasets. Each planet was represented by a circle, of which the radius, color, distance to center mapped to the planet's radius, temperature and distance to the solar system.

In each bubble chart, circles represented data elements that contain three data attributes, represented by color, size, and distance to the center of the chart. We selected 300 data points from each of the three datasets, to control for data size. Mousing-over a circle brings up a detailed view, showing text values for the underlying data element.

Search Design: Search appeared on the top-left of the bubble chart, with auto-complete enabled. Searches and highlighting were triggered on character press. To display search results, the selected data items maintained opacity, while unselected items were deemphasized through a slight decrease of opacity.

4.3.3 Measures

We include both quantitative and qualitative measures derived across the phases of the experiment.

In the *Strategy* phase, quantitative measures include:

- *intent*: the proportion of participants who indicated that they intentionally sought specific data items in the visualization.

- *active search count* the number of data items participants selected as items they intentionally sought in the visualization (as opposed to incidental findings).

Self-reported quantitative measures were collected via steps described in Procedure and Tasks. Through free-response questions, we also collect participant comments on their strategies and experience of the experiment.

In the *Exploration* phase, we collect which data elements each participant visited

(*i.e.* interacted with for longer than 500ms), as well as any search queries. Quantitative measures include:

- *exploration time*: the total time a participant spent on the *Exploration* phase.
- *average visit time during exploration*: the average time a participant spent viewing the details of a data item during exploration.
- *average visit time during search*: the average time a participant spent viewing the details of a data item while an active search query was highlighting items in the visualization (*search present* condition only).
- *average visit time outside search*: the average time a participant spent viewing the details of a data item while outside of a search query (*search present* condition only).

4.3.4 Pilots, Analyses, and Experiment Planning

We conducted several pilot studies to help establish our measures and procedure. In response to concerns about the limitations of null hypothesis significance testing [66, 67], we model our analyses on HCI research that seeks to move beyond these limitations (*e.g.* Dragicevic [110]), primarily focusing on confidence intervals and effect sizes. Following Cumming [66], we compute 95% confidence intervals using the bootstrap method, and effect sizes using Cohen's d- which is the difference in means of the conditions divided by the pooled standard deviation. While we include significance tests and related statistics, it is with the intention of supplementing these analyses.

The results of our pilots showed some measures from the *Exploration* phase were non-normally distributed, according to a Shapiro-Wilk test. These measures, such as *exploration − time*, were right-skewed with long tails. Because common transforms (*i.e.* log, square-root) did not lead to changes in the Shapiro-Wilk result, we use the non-parametric Mann-Whitney test to compare these conditions.

	no search	search present	used search	total
255Charts	57	102	(72, 70.6%)	*159*
BoardofDirectors	47	151	(49, 32.5%)	*198*
Colleges	68	93	(75, 80.6%)	*161*
AnonColleges	53	103	(68, 66.7%)	*156*
Exoplanets	61	95	(65, 68.4%)	*156*

Table 4.2: We evaluate the impact of text-based search using a between-subjects design across multiple visualizations. The table shows participant numbers for each experiment, determined by running effect size and power analyses on pilot studies. More participants were added to the *search present* condition based on proportions of use derived from pilot studies.

To ensure our experiments included enough participants to reliably detect meaningful differences between the conditions, we conducted effect size and statistical power analyses. Specifically, we used pilot studies to estimate the variance in our quantitative measures, and combined these with the observed means to approximate how many participants were needed. Additionally, from pilot studies we estimate the percentage of users who are likely to use search, adding more participants to the *search present* condition to ensure roughly equal numbers of participants in the "used search" and "no search" groups (see Table 4.2 for specific proportions and outcomes).

4.4   Results

In total, we recruited 830 participants through Amazon's Mechanical Turk for the study. For each visualization, participants were assigned into one of the two conditions, *search present* and *no search*.

4.4.1   Proportion of People who Use Search When Present

Exploration behavior: when search is present, what proportion of users make use of it?

In general, a majority of people used text-based search when present in a visualization.

Figure 4.4: Experimental results comparing original visualizations with versions that integrate search. The results suggest that adding search enables a subset of users to identify specific data of interest in visualizations, and that in many cases this leads to more time spent with individual data items, an indicator of greater engagement with data. Maps showing items visited during search (orange) versus items visited when users did not have search (purple) suggest that search leads users to different parts of the data.

The proportion of participants that used search were similar in the most of the visualizations (70.6% for 255*Charts*, 80.6% for *Colleges*, 66.7% for *AnonColleges* and 68.4% for *Exoplanets*). However, the proportion was lower for *BoardofDirectors* (32.5%). We visit possible reasons for this outlier and design implications that follow this finding in the discussion.

4.4.2 Search's Effect on Information Seeking Goals

Self-reported exploration strategy: does the presence of text-search impact peoples' reported exploration strategies?
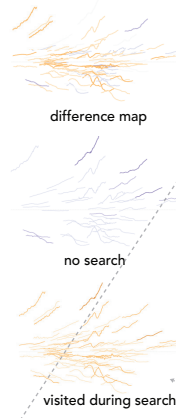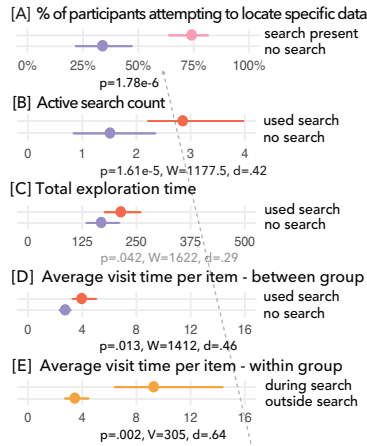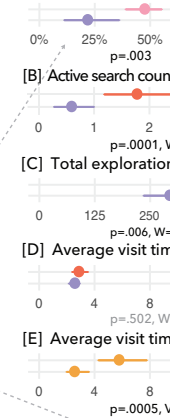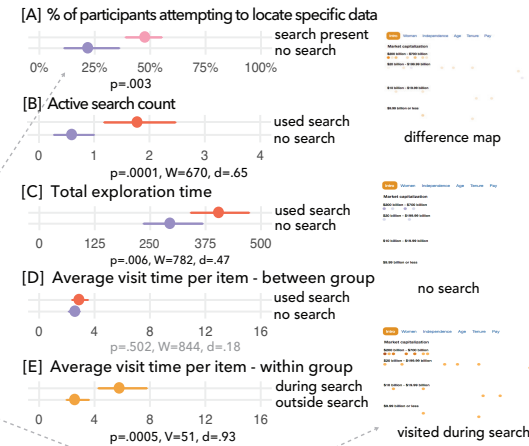
For all except one condition, *AnonColleges*, significantly more participants indicated that they actively sought specific data items when search was present (significance determined by a two-proportion z-test, see also the top chart in each column of Figure 4.4).

Of note is that *AnonColleges* and *Colleges* are identical in visual form and data attributes, except for the plaintext college-name field. Specifically, in *Colleges*, the proportion difference was 25.7% (95% CI [9.5%, 42.1%]), with 69.9% affirmative in the *search present* condition and 44.1% in the *no search* conditon. *AnonColleges*, on the other hand, showed a difference of 12.2% (95% CI [-3.0%, 27.4%]), with 31.1% affirmative in the *search present* condition and 18.9% in the *no search* conditon. We consider differences in these findings related to exploration strategies and their implications further in the discussion.

4.4.3 Search's Effect on Information Seeking Patterns

Exploration behavior: does the use of text-based search impact measures of behavior, such as the location of investigated data?

If a participant indicated that they intentionally sought specific data items in the visu-

alization, they were shown a list of every item they interacted with (defined as viewing for more than 500ms, to account for accidental interaction). Given these sets, we generate maps that show patterns of where participants visited. Specifically, we normalize the number of visits to each data item, and add a diverging gradient that indicates how often data items were selected by participants who used search versus those who did not use search.

Figure 4.4 shows three of these normalized maps for each visualization, including proportions for participants who used search, those who did not, and a diverging map showing the difference. Notably, across all conditions, with the possible exception of *Exoplanets*, participants select a **more diverse** set of data items. For example, in the *Board of Directors* visualization, participants who did not use search generally indicated their intent for items in the upper left of the view. In contrast, participants who used search indicate a wider range of values, spanning more of the range of the data.

4.4.4 Search's Effect on Exploration Time

Exploration behavior: does the use of text-based search impact measures of behavior, such as the total time spent on exploration?

We measure exploration time, *i.e.* the total time spent interacting with the visualization, at three levels of granularity. First, we collect the overall time, meaning the time from which the participant begins exploring, to the time they click to indicate they are finished and ready for the next section. Second, we collect the amount of time a person spends looking at the details of a data item. Finally, for participants who use search, we distinguish between "visit" times when a search is active (*i.e.* data items are highlighted) and inactive. In the latter case, the user is examining item detail without the aid of search.

At the overall exploration time level, significant differences are only found for the "in the wild" visualizations. For example, in *Board of Directors* we found that the average

66

participant who used the search functionality spent more time ($M = 117$ seconds 95% CI [93.1, 147.5]) than those in *no search* condition ($M = 76.9$ seconds 95% CI [55.5, 109.8]). Following Cumming's methodology for interpreting confidence intervals [66]. Given the upper and lower limits of the confidence intervals, the average participant in the group spends at least the same time on exploration, and up to 92 seconds more.

We note that longer exploration time, while reported in prior studies (*e.g.* Boy *et al.* [4]), may indicate greater engagement of participants, it could also indicate difficulty in using aspects of the visualization, like search. For this reason, it is necessary to further differentiate aspects of time, such as time spent examining individual data items.

4.4.5   Time Examining Individual Data Elements

Exploration behavior: does the use of text-based search impact measures of behavior, such as the proportion of time spent viewing detailed information about chart elements?

At the second level of time-granularity, we analyze the average time participants spent viewing the details of each data element, which we term a "visit". This difference was significant only in 255*Charts*, where the average participant who *used search* spent more time visiting a data item ($M = 3.9$ seconds 95% CI [3.2, 4.9]) than those in the *no search* condition ($M = 2.7$ seconds 95% CI [2.3, 3.1]). Given the upper and lower limits of the confidence intervals, the average participant who used the search functionality spends at least same time visiting a data item, and up to 2.6 more seconds ($d = 0.46$ [0.17, 0.7]). We revisit this finding in the discussion, as 255*Charts* is also different from all other visualizations in that a "visit" brings up a secondary chart.

At the third level of granularity, we compare visits within the search condition, specifically visits that occur while search is active, against those that occur when search features are not in use. As shown in the bottom confidence interval charts in each column of Figure 4.4, participants spent significantly more time with data items when search was active, for

all conditions except *Exoplanets* ($p = .10$).

These results suggest that, in most cases, data items that are visited during search are examined for longer. This effect is particularly strongest in the news visualizations, where visits during search are higher than all population-level visit times (see Figure 4.4). For example in 255*Charts*, the average participant spent more time visiting a data item found by using text-based search ($M = 9.2$ seconds 95% CI [6.5, 13.8]) than through browsing ($M = 3.4$ seconds 95% CI [2.7, 4.3]). Given the upper and lower limits of the confidence intervals, the average participant in the group spends at least 2.2 more seconds visiting a data item found through text-based search, and up to 11.1 more ($d = 0.64$ [0.34, 0.87]). In *Board of Directors*, we found that the average participant spent more time visiting a data item found by using text-based search ($M = 5.7$ seconds 95% CI [4.3, 7.8]) than through browsing ($M = 2.5$ seconds 95% CI [2, 3.4]). Given the upper and lower limits of the confidence intervals, the average participant in the group spends at least 0.9 more seconds visiting a data item found through text-based search, and up to 5.8 more ($d = 0.93$ [0.43, 1.41]).

4.5   Discussion

As shown in Figure 4.4, the results of these experiments suggest that the *mere presence* of text-based search in visualization can impact users' self-reported exploration strategy, the data they explore, and how long they explore specific items of data. Results from the controlled variation of dataset familiarity suggest that the effects of text-based search change depending on the topic of a visualization. We turn our attention to possible causes for these findings, notable uses of search by participants, and the implications these findings carry for the design of visualizations.

4.5.1 Search Encourages Personalized Information-Seeking

Our results indicate that most people use text-based search when it appears alongside a visualization. Furthermore, people who utilized search were more likely to indicate that they actively sought specific data items in the visualizations. While these results may not be surprising in their own right, our observations suggest that the data people looked for while using search was often deeply personal. As one participant who searched for 'Duke' stated:

> *Duke University is very expensive at \$61,000 a year... when I was a kid I wanted to go to Duke.*

Similarly, a participant that used a partial query 'Tech', provided the finding:

> *Tech colleges promise the most consistent ROI... [I'm] interested in science. Also, my brother applied to these schools*

Quotes like those above suggest that simple interaction mechanisms such as text-search have the capability of changing user's relationship with the visualization. While it is possible that these participants could have arrived at their insights without the use of search, doing so may have been more haphazard or time-consuming given the initial interaction schemes and visual forms.

From a design perspective, it may be important to emphasize that some people did *not* use search, even when it was present. Use of search ranged from a high of 81%, *Colleges*, to a low of 33% *Board of Directors*. This low value is an outlier, but remains an interesting case worthy of further investigation. One possibility for the low use of search is that people simply didn't notice it. Due to constraints in the form of the visualization, search appeared in the bottom-right (Other positions were possible, but it was unknown apriori that position may have an effect).

Another possibility is that the interaction scheme of *BoardofDirectors* is what practitioners and researchers sometimes refer to as "scrollytelling", where the main narrative of the visualization is controlled by user scrolling or clicking to advance the "slides". Effects like these raise questions of whether there is an upper limit on the number of available interactions that a given person will make use of during exploration.

4.5.2   Search Encourages Diverse Engagement with Data

When participants used search queries, they engaged with individual data items for significantly longer than when search queries were inactive. A likely explanation for this trend is that search queries serve as an implicit indicator of interest. However, it's worth noting that this deeper engagement was facilitated by the presence of search.

However, the strength of the effect differed across conditions. In 255*Charts*, for example, data items that were visited during searches outpaced non-search visits (according to 95% CIs) by at least 2.2 seconds, and up to 11.1 on average. One possible reason for these differences corresponds to the depth of detail available to users on-demand. A unique feature of 255*Charts* is that, on mouseover, a secondary line chart appears, showing additional data for the given industry (see Figure 4.4). In contrast, the details shown in the bubble charts consist of a few simple data items: college cost, planet temperature, etc. The effect was similar in *BoardofDirectors*, where multiple data elements about companies were shown on mouseover.

The results show a longer exploration time found in the *BoardofDirectors* visualization. Unlike the other visualizations, *BoardofDirectors* has multiple tabs, which may have led users to compare highlighted search results in different views by switching between tabs.

A possible consideration for design, then, is that the value of search increases alongside the amount and quality of data revealed in detail views in an interactive visualisation.

Future research might investigate the role of detail quantity and quality of exploration patterns, as well.

4.5.3 Text-based Search in Visualization Task Taxonomies

Beyond measures of behavior, the value of search mechanisms in visualization may be more clearly articulated through existing task taxonomies. One that is particularly relevant is Brehmer and Munzner's typology of abstract visualization tasks [111]. In their taxonomy, *search* is used as a general term referring to multiple user goals, including lookup, browse, locate, and explore. We contextualize each of these within the *Colleges* condition of our visualization:

- *lookup* (location known, target known): the user knows exactly which college they are looking for and exactly where it is in the visualization
- *browse* (location known, target unknown): the user has characteristics of a college that they are interested in (ex: high tuition) guiding them to a region of the visualization, but does not have any specific college in mind.
- *locate* (location unknown, target known): the user is looking for information about a specific college, but does not know where in the visualization that college might be represented.
- *explore* (location unknown, target unknown): the user is not looking for any particular college or characteristic.

While the visualizations we tested largely support *browse* and *explore*, they fall short in *locate* goals without search. Because of the density of the data, labels are supported through interaction mechanisms rather than natively appearing on the page. As a result, finding a specific college, industry, planet, or company can be challenging.

4.5.4 Search Enables Creative Exploration of Unfamiliar Data

The flexible nature of linking text-based queries to visual encodings (such as highlighting) in visualizations enables some users to investigate data in surprising ways. For example, in the *Exoplanets* condition, one participant used partial queries to investigate relationships in the naming schemes of the planets. As shown in Figure 4.5, the participant appears to have arrived at a query of 'hat' - a naming prefix of the exoplanets discovered by the Hungarian Automated Telescope (HAT) network. Analysis of interaction logs shows that this participant began exploration by mousing over planets at random, until noticing that some had this common prefix. In the free-response section, the participant described their strategy:

> *I compared different properties of the different groups of planets with similar names to those with different names*

Queries like this demonstrate a possible ancillary benefit of text-based search: partial queries across data fields allow people to segment unfamiliar data in novel ways, even if the data is unfamiliar to them.

4.5.5 Keyboard-based Features for Accessibility

In addition to exploration behaviors and strategies, our experimentation with text-based search raised questions of accessibility in visualization. Visualizations can be difficult to interact with for people with motor deficiencies, *i.e.* people who cannot use a mouse to generate precise movements, as interactive elements may be only a few pixels wide. However, the W3 Standards organization lists extensive accessibility principles for web designers [112]. Text-based search mechanisms in visualization, applied at the appropriate scope, increase accessibility by supporting keyboard based interaction, which is a key recommendation of W3. While accessibility has not been addressed broadly in the
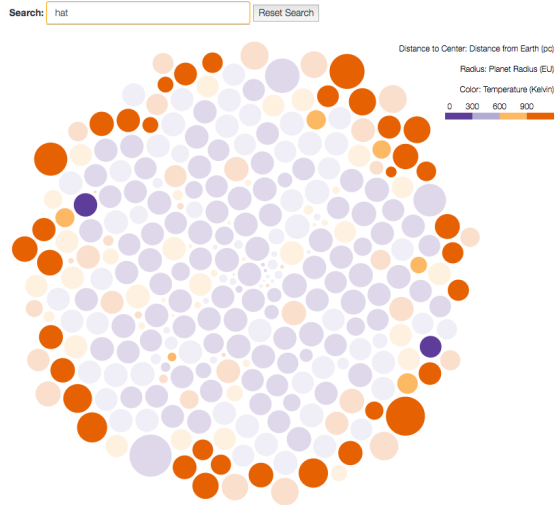
Figure 4.5: Some participants used text-based search to explore the data in creative ways. In one case, a participant noticed that some planets had common substringsin their names. They arrived at the query "hat", and produced a finding about common data features among "hat" planets. ("HAT" happens to be the organization that discovered these planets.)

visualization community, the results of this study, along with other findings that multimodal interaction mechanisms are generally beneficial [113, 114], add a perspective to this ongoing thread.

4.6 Limitations

Our study of text-based search in visualizations was within a limited scope in three aspects: (1) data characteristics, (2) visualization types and (3) user background. First, all the visualizations used in our study consist of 200-500 data items. Each data item has at least one key (*e.g.*, industry name in 255*Charts*), which is used for text-based search. Second, the visual representation of the visualizations was single view including all data items, with details revealed by mouseover. More complex representations such as coordinated multiple views were not used in this first study. Third, participants of our study were

closer to a general population with diverse backgrounds, not domain experts. In addition, there are alternative mechanisms supporting text-based search, such as drop-down boxes and sliders, which may yield different behavioral results and raise new design trade-offs. The generalization of our results beyond these constraints is open to investigation.

The effect of search on comprehension is a likely a delicate dance in which design, data, target audience, and encoding interact to nudge its effect on the user. While we investigated the impact of search in different visualizations, we do not know the effect of varying choices in the visualization design space as it relates to search. Future research can build upon these experiments to investigate increasingly diverse combinations of search and interaction mechanisms to generate clearer design guidelines (for example, when is search *not* useful or harmful?)[1].

Finally, the measures we have for understanding the overall impact of any interaction mechanism still leave a lot to be desired. In this study, we used a combination of behavior, open-response, and survey questions to try and understand the overall impact of search in visualization. However, in a realistic environment in which goals are not prescribed to the user, they do not always translate cleanly to clear success/failure outcomes - is the person who found their home institution in the *Colleges* condition but visited nothing else less successful than the person who broadly explores the entire visualization? More research is needed to understand exactly when a visualization succeeds or fails in the open web environment. Future work in this area will likely require close collaboration with practitioners who create visualizations for the masses.

---

[1] **To facilitate future work, all experiment materials, participant data, and analyses scripts are available online: `https://wpivis.github.io/search-in-vis`.**

4.7 Conclusion

Across the web, designers build thousands of data-dense visualizations for the public to explore and comprehend. Surprisingly, only a very small subset of these visualizations are accompanied by text-based search mechanisms. While text-based search has often been used in conjunction with large datasets for analysts, our results suggest that its inclusion in everyday visualizations, even those with relatively small amounts of data, may encourage engagement and support user information seeking goals that are difficult with other forms of interaction. Through experiments with five visualizations, we find that in most visualizations, a majority of users will use text-based search features if present, and that search can shape people's experience and behavior with visualizations. Results of the experiments also indicate the average participant who used text-based search engaged with individual data items for longer, and explored different parts of the data. The results of these experiments have practical implications for design, and more broadly serve as a case study in how interactive data visualizations can be augmented to support diverse information seeking goals.

Chapter 5

Characterizing Diverse Exploration Behavior with Visualizations on the Web

The diverse and vibrant ecosystem of interactive visualizations on the web presents an opportunity for researchers and practitioners to observe and analyze how everyday people interact with data visualizations. However, existing metrics of visualization interaction behavior used in research (e.g., *exploration-time* and *number-of-interacted-elements*) do not fully reveal the breadth of peoples open-ended explorations with visualizations. One possible way to address this challenge is to determine high-level goals for visualization interaction metrics, and infer corresponding features from user interaction data that characterize different aspects of peoples explorations of visualizations.

In this chapter, we identify needs for visualization behavior measurement, and develop corresponding candidate features that can be inferred from users interaction data. We then propose metrics that capture novel aspects of peoples open-ended explorations, including exploration uniqueness and exploration pacing. We evaluate these metrics along with four other metrics recently proposed in visualization literature by applying them to interaction data from prior visualization studies. The results of these evaluations suggest that these new metrics 1) reveal new characteristics of peoples use of visualizations, 2) can be used

to evaluate statistical differences between visualization designs, and 3) are statistically independent of prior metrics used in visualization research.

We discuss implications of these results for future studies, including the potential for applying these metrics in visualization interaction analysis, as well as emerging challenges in developing and selecting metrics depicting visualization explorations.

5.1 Introduction

As interactive visualizations migrate from standalone applications to the web, visualization users have expanded from domain experts to the general population. Alongside this expansion of both visualization creators *and* consumers comes an expansion in the goals of both - from casual exploration to focused analysis. But do the metrics we use to assess visualizations capture this diversity in objectives? In this work, we explore how the rapid development of expressive and interactive forms on the web has demanded an extension of the metric toolbox in which we equip content creators, and how we can better align assessment with the goals of the designers.

Consider an example where someone explores an interactive scatterplot visualization showing a company's profit and income. Each point represents a company, and upon mousing over a point the user will uncover the company's income over several years, the employees' age distribution, etc. A person's goals can be diverse here, ranging from specific (gathering information on a possible stock purchase) to broad (getting to know more companies). Two likely metrics to describe their behavior include *time spent on exploration* and *points interacted with*. These metrics could be used to answer basic questions about how an audience uses a published visualization, for example "how many points did the average person interact with?" or "how long did the average person explore the visualization?". Yet despite their diversity in goals, it's possible that users interact

with a similar number of points and engage with the visualization for a similar amount of time. While simple metrics might not reveal differences between users, in reality, their behavior may not align with what the creator of the visualization had in mind for their audience.

Although research has made strides in designing and evaluating interaction in visualization, we lack *low-barrier, expressive* metrics that capture the breadth of user interaction [4, 10, 22, 24]. Various analysis strategies have been used to answer these questions, including *statistical* and *visual* approaches (*e.g.*, [4, 5, 37]). However, these existing approaches have limitations with characterizing user explorations precisely. Many of the metrics used to summarize activity tend to over-aggregate behavior, failing to identify differences between users, or by failing to capture detailed information such as *how long* has been spent on *which visual elements*. On the other hand, the visual approaches usually keep the details of users' interaction logs, but visual inspections can hardly lead to reliable inferences.

One possible way to bridge this gap is to develop metrics, *i.e.*, statistical measures, which take into account more information in peoples' interaction logs, and to better reveal facets of peoples' explorations. Related efforts can be found in the field of HCI. Chi *et al.* [29] quantified the saliency of a user's visit to a website when modeling users information needs and actions on the web. Heer *et al.* [30] further used this measure to cluster web users. These efforts influence our work of visualization interaction analysis, in that a user's open-ended exploration of a visualization containing visual elements can be considered analogous to the exploration of a website. However, it is impractical to directly adapt these methods developed to analyze website explorations, due to the differences between the website clickstream analysis and visualization interaction analysis, such as different scales (*i.e.*, usually millions of users versus tens to thousands of users) and different complexity of interaction types.

The aims of this work are three-fold:

1. Derive a requirements space to categorize existing and new metrics that quantify facets of users' exploration of data visualization and in doing so identify emerging analysis needs.

2. Derive two new metrics centered around user exploration diversity and pacing that provide new perspectives into users' open-ended exploration.

3. Evaluate both these new metrics and metrics recently proposed in visualization literature across hundreds of interaction traces from previously published visualization experiments [1].

We further discuss how these metrics can help both statistical and visual approaches to analyze interaction logs, such as 1) quantifying the impact of visualization designs on user behavior; 2) organizing the visual representation of interaction logs; and 3) serving as features to machine learning models.

5.2 Background

5.2.1 Characterizing Website Exploration

One closely related thread of research is clickstream analysis and visualization for websites or applications [3, 20, 23, 27, 28], under a broader research topic of event sequence analysis [115, 116, 117, 118, 119]. Clickstream research includes the data processing, analysis and visualization methods to analyze users' website visit logs. For example, Liu *et al.* [23] developed algorithms to extract sequence patterns from clickstreams. Zhao *et al.* [3] created a visualization called MatrixWave to compare two clickstream datasets, and found it to scale better than commonly used Sankey diagrams.

---

[1] The experiment data and analysis scripts are available on the Open Science Framework: `https://osf.io/dx43q`

5.2.2 Characterizing Visualization Explorations

Characterizing user behavior through interaction logs has been used for various purposes, such as learning user characteristics [5, 25], understanding system usage [120] and the reasoning process [6], and evaluating visualization design [4, 12, 22]. Various approaches have also been used for these interaction analyses, including visual and statistical approaches.

Visual Approaches

Visual approaches refer to strategies of showing users' interaction logs with visualizations [6, 34, 35, 117]. The interaction logs can be shown in an aggregated way in order to reveal the behavioral differences of user groups in experiment analyses. For example, Ottley *et al.* [5] used aggregated maps to show different exploration patterns of tree visualizations. Users' interaction logs can also be shown individually. Blascheck *et al.* [37] introduced a visual analytic approach to study users' interactions with visual analytics. These visual approaches have the advantage of preserving the details of the interaction logs. However, visual examination alone cannot provide robust analyses of user behavior, as they are often better paired statistical approaches. [118]

Statistical Approaches

Commonly used metrics to depict a user's exploration include *total exploration time* spent by a user, and *number of raw interactions* performed by a user during exploration, such as hovering and clicking. Boy *et al.* [4] evaluated the effectiveness of storytelling by comparing users' exploration time and raw interaction counts (hovers and clicks) between the experimental and control groups. Liu *et al.* [24] measured the effects of latency on users' exploration behavior of visual analytics by using raw interaction counts (drag, brushing and linking, etc). There are many other works using the basic metrics to characterize

users' interaction with visualizations [22, 25, 41].

However, these raw counts have limitations with delivering semantic meanings of user explorations. Interaction coding was thus used to describe interaction behavior. Boy *et al.* [10] and Guo *et al.* [22] coded the raw interactions into semantic interactions, such as selecting, filtering and inspecting, according to Yi *et al.* 's [121] visualization interaction framework. They counted the coded interactions afterwards.

Some work went beyond counting individual interactions (including raw and semantic ones), in order to reveal more characteristics of user explorations. Guo *et al.* [22] further extracted the sub-sequences containing specific individual interactions, and then counted the sub-sequences for each user. Wall *et al.* [36] proposed six metrics to measure cognitive bias, including data coverage, data distribution and attribute coverage/distribution, etc.

In this work we create a feature space to categorize the existing metrics, and develop new metrics by filling the gaps in the framework, by fully utilizing the information in interaction sequences, in order to reveal more characteristics of users' visualization explorations.

5.3   A Requirements Space for Metric Development

The aim of this work is to explore and evaluate metrics that characterize the diversity of peoples' explorations with interactive visualizations on the web. We therefore situate our metric development activities by deriving a set of requirements (top-down) and examining the possible common data sources (bottom-up) from which new metrics can be derived, which translate into two questions:

1. What do we *need* to measure for behavior analysis?

2. What *can* we measure given users' interaction logs?

These questions drive two dimensions in this requirements space: 1) identifying unfilled measuring needs for visualization behavior analysis, and 2) deriving low-level measurable features from visualization interaction logs.

We form the structure of this requirements space based on O'Connell *et al.*'s[122] framework for deriving metrics measuring human interaction with interactive visualizations. In their framework, high-level needs include human-interaction heuristics, *i.e.*, measures that assess how well visualizations empower analysis, collaboration, ease of use, etc. O'Connell *et al.* then derive corresponding metrics for each of the heuristics by utilizing features from users' interaction data such as number of interactions performed by a user.

One aim of this work is to move beyond system-specific visualization interaction metrics towards metrics that can be applied across a range of visualizations and for a range of creator goals, whether they be visualization practitioners or researchers. Our requirements space therefore expands and generalizes O'Connell *et al.*'smetric framework in two ways. First, centered on needs of visualization creators, we identify desirable avenues for visualization interaction metric development. Second, by examining commonly available interaction data in web-based visualizations we derive novel features which are intended to enable new means for comparison and reasoning about how audiences interact with visualizations.

5.3.1 Visualization Interaction Analysis: Identifying Needs

What are unfilled needs in visualization interaction analysis?

Researchers and practitioners often develop metrics to capture various dimensions of user experience. While there is currently no widely used metrics framework targeting visualization activity, we might consider frameworks developed in human-computer interaction as a starting point. Consider the HEART metrics framework for web applications

proposed by Rodden *et al.* [123]. In the context of evaluating web applications, Rodden *et al.* define five categories: *happiness*, *engagement*, *adoption*, *retention*, and *task success*. This framework, while not originally intended for visualization, can be used to some extent to categorize existing efforts in developing metrics for interactive visualizations. For example, some existing metrics aim to reveal users' *task success*. Time spent on exploration has been used as an indicator of how efficiently a task is performed by the user [106?]. Targeting bias, Wall *et al.*'smetrics aim to capture the quality of exploration in visual analytics contexts. Other metrics target proxies of *engagement* with visualizations. For example, interaction coverage metrics have been explored in several recent works, typically targeting numbers of specific types of interaction events (*e.g.*, [4, 12, 24, 124]).

The complexity of interactive visualization makes it difficult to develop metrics that fulfill all dimensions of measurement needs, especially by only adapting metrics from web applications. Challenges surrounding metrics development are especially salient given the question of measuring user "engagement" in visualizations, which indicates *users willingness to invest effort to explore further and gain more information from the visualization* [4, 108, 125]. For example, longer exploration may indicate engagement, but may also be reflective of confusion and difficulties faced by a user learning a new interaction scheme. Similarly, an increase in interaction counts could either reflect users' interest or possibly their random clicks to orient themselves in a new environment. Given these ambiguities, one aim of this work is to explore new formulations and perspectives on metrics that may serve as useful proxies for capturing part of the user experiences of an interactive visualization.

In this work, we seek new metrics to reveal more aspects of user engagement with visualizations. The next section deals with how to enumerate these metrics given users' interaction data.

Notation	Description
$E = \{e_1, ..., e_N\}$	The set of N interactive elements in the visualization.
$U = \{u_1, ..., u_M\}$	The set of M users who explore the visualization.
$|E(u)|$	The number of visualization element a user u interacts with.
$A = \{click, hover...\}$	The set of available action types of interaction.
t	The moment when an event occurs.
$I = (t, a, E_i, d)$	An interaction event, including the moment t when it occurs, the type of action $a \in A$, the set of interacted visualization elements $E_i \subseteq E$, and the duration of the interaction d.
$Ex(u) = (t_{start}, I_1, ..., I_k, t_{end})$	The interaction log of a user u exploring the visualization, including a time-series ordered sequence of events, the moment when the exploration starts t_{start}, followed by k ordered interaction events $I_1, ..., I_k$ and the end moment t_{end}.
$C(u_m, e_n)$	The count of interactions with the visualization element $e_n \in E$ by the user u_m.
$T(u_m, e_n)$	Time spent by user u_m interacting with the vis object $e_n \in E$.

Table 5.1: Notations used to describe user interactions with visualizations, and to describe the metrics in this chapter.

5.3.2 Deriving Features from Visualization Interaction Data

What can we measure from peoples' visualization interaction data?

Working from the bottom-up, we observe that multiple candidate low-level features can be extracted from peoples' visualization interaction data. For example, Wall *et al.* [36] listed two measurable features, *types of interaction* (*e.g.* clicking and hovering) and *objects of interaction* (*i.e.*, the visual elements interacted with), that can be extracted from users' interaction data and used in metric development. Similarly, Blascheck *et al.* [126] point out that *the time spent for inspecting particular data items* is a widely available and useful feature to be considered in evaluating user behavior.

Based on the existing literature of visualization interaction analysis [36, 126], visual-

ization interaction frameworks [121], and the related topic of website clickstream analy-sis [27, 127], we list several low-level interaction features that can commonly be obtained from a user's exploration session of visualizations encountered on the web (see Table 5.1 for precise descriptions using mathematical notations):

- **type**: What type of interaction is it (*e.g.* selection, hovering, or the types from existing frameworks such as Yi *et al.* [121])?

- **element(s)**: Which element(s) in the visualization are users interacting with?

- **duration**: How long does the user interact with the visualization element(s)?

- **order**: In what order do the interactions take place?

- **moment**: At what moment in exploration does each interaction take place?

- **exploration time**: How long does the user spend exploring facets of the visualization?

We list several metrics characterizing users' visualization explorations and identify the underused features. There are several basic metrics that are commonly used by re-searchers and practitioners. We simplify the description of the metrics by assuming that there is only one type of interaction in the visualization:

- ***number-of-actions*** [4, 22, 24]: equals to k, the number of certain type of interactions performed by the user during exploration.

- ***number-of-visited-elements*** [12]: equals to $|E(u)|$, the number of unique visualization elements visited by the user. A *visit* is defined as a meaningful (*i.e.* non-accidental) interaction with an element, which lasts for at least a short amount of time (*e.g.*, 500ms) [4, 12, 13].

- ***exploration-time*** [4, 12, 13, 24]: equals to $t_{end} - t_{start}$, the time spent by the user on the exploration.

A recent study in the visual analytics area from Wall *et al.* [36] propose metrics to measure bias in visualization exploration, by using multiple features from users' interac-

tion data.

bias-data-point-coverage measures bias based on the user's coverage of the data points in the visualization. Since one data point is often mapped to one element in the visualization, in this work, a data point is considered equivalent to a visual element.

$$b_{DC} = 1 - min(\frac{|E(u)|}{\hat{\kappa}(E(u))}, 1)$$

where $|E(u)|$ denotes the number of unique visualization elements interacted by the user, and $\hat{\kappa}(E(u))$ denotes the expected value of the number of unique elements visited in k interactions.

$$\hat{\kappa}(E(u)) = \frac{N^k - (N-1)^k}{N^{k-1}}$$

Another metric Wall *et al.* propose is *bias-data-point-distribution*, measuring bias toward repeated interactions with individual data points or subsets of the data.

$$b_{DD} = 1 - p$$

where p is the p-value obtained from the χ^2 distribution with N - 1 degrees of freedom

$$\chi^2 = \sum_{n=1}^{N} \frac{(C(u,e_n) - \hat{C}(u,e_n))^2}{\hat{C}(u,e_n)}$$

where $\hat{C}(u, e_n) = [k/N]$.

Among the existing metrics, we observe that the basic metrics (*e.g.*, *number-of-visited-elements*) reveal the *end results* of the exploration process, meaning that the *details in the process* (*e.g.*, elements and duration of each interactions) are not preserved in the metrics.

Wall *et al.*'sbias metrics go beyond traditional methods by distinguishing the *element* of each interaction in the exploration process. However, we still observe that some other features revealing details are underused by existing metrics, *e.g.*, *duration* and *moment* of interaction. Examining the existing metrics such as *visits* and *exploration time* in light of these recent studies raise new questions about user exploration in visualizations. For example, instead of *how many* elements are visited in the visualization, what about the *diversity* of elements visited? And instead of *how much* time is spent in exploration, what about the *pacing* of peoples' exploration inside visualizations?

Given both the unfilled needs of measurement and underused features in users' interaction data, we propose two additional metrics that take into amount more features, and thus offer new perspectives user engagement, *i.e.*, the diversity and pacing of peoples' open-ended explorations with interactive visualizations.

5.4   Proposed Metrics

5.4.1   Exploration Uniqueness

The *exploration-uniqueness* (*EU*) metric aims to capture the diverse engagement of peoples' open-ended explorations, *i.e.*, to quantify how unique a user's exploration pattern is compared to patterns from others. The *EU* metric is defined as *aggregated visit duration over visual elements, weighted by the uniqueness in comparison to the crowd*. A low *EU* value suggests that a user's time distributed visiting the visual elements align with common patterns of exploration in the visualization. A high *EU* value suggests that a user's exploration strategy differentiates itself from most other users.

Measurable features: There are many ways to define a unique exploration pattern, *e.g.*, visiting a set of unique elements, or visiting elements in an unique order. This metric characterizes the uniqueness of a user's visited element sets, instead of visit orders. In

open-ended explorations, visit orders may carry too much variance, obscuring the trends of unique explorations. We thus compute uniqueness based on the distribution of time spent visiting the data items in the visualization. The features from users' interaction logs taken into account are the *elements* and *duration of interaction*, *i.e.*, the *order of interaction* and *exploration time* are discarded.

Modeling approach: As we develop the metric, we model the interaction behavior of a group of users as a *matrix* (Equation (5.2)), where each row represents a user, and each column represents an element in the visualization. An alternative approach is to model each user's interaction sequence as a Markov chain, which is used in Wall *et al.*'sbias metrics. Each interaction with a visual element is a state in a state space. A user performing the {element, interaction} combination has transitioned to the associated state in the Markov chain. We adopt *matrix* to model interactions for the *EU*metric mainly for two reasons. First, a user's behavior is represented as a vector, instead of a sequence. This representation aligns with the selection of features, *i.e.*, focus on the data distribution, rather than the visit order. Second, matrix enables comparisons across users, *i.e.*, each user's behavior can be compared to the crowd.

Adapting the Concept of Term Frequency-Inverse Document Frequency (TF-IDF)

In order to depict the uniqueness of a users' exploration process, we need to know how unique each of her visit is, compared to other users' visits. From the field of information retrieval, we find one adaptable concept, *Term Frequency-Inverse Document Frequency (TF-IDF)*, describing how unique a word is in a document collection.

The uniqueness of the appearance of a word in a document is the product of TF and IDF. *Term frequency (TF)* is the frequency of the word in a particular document. *Inverse document frequency (IDF)* is the inverse proportion of documents the word appears in. IDF acts as a weight to TF, rewarding the words appear in less documents, and penalizing

those appear in more documents. For example, *a* and *the* tend to have lower IDF than *house*.

Given a document collection D, a word w, and an individual document $d \in D$, we calculate a TF-IDF value for each word in a document:

$$TFIDF(w,d) = TF \times IDF = \frac{f(w,d)}{|d|} \times log(\frac{|D|}{f(w,D)}) \tag{5.1}$$

where $f(w,d)$, equals the number of times w appears in d, $|d|$ is the number of words in d, $|D|$ is the size of the corpus, and $f(w,D)$, equals the number of documents in which w appears in D [128].

In several research initiatives, TF-IDF has been extended from characterizing words to modeling user behavior. For example, In the field of HCI, TF-IDF has been used to describe the uniqueness of peoples' visits to a website [29, 30, 129], and peoples' geospatial movement [130]. Herein we adapt TF-IDF to calculate the uniqueness of a user's visit to an element in a visualization. Specifically, we map a user's interaction log to a *document*, each visit of a visual element to a *word* in the document, and a collection of exploration sessions of multiple users to a *corpus*. A visual element visited by more users has a higher IDF value than an element visited by fewer users.

Metric Calculation Steps

The *exploration-uniqueness* metric is computed in three steps.

Step 1: Form a matrix $V_{N \times M}$ representing the distribution of visits from the M users to the N visual elements in a collection of interaction logs.

$$V_{M \times N} = \begin{pmatrix} T(u_1, e_1) & \dots & T(u_1, e_N) \\ \vdots & \ddots & \vdots \\ T(u_M, e_1) & \dots & T(u_M, e_N) \end{pmatrix} \qquad (5.2)$$

where each row represents a user u_m, each column represents a visual element e_n, and each element $T(u_m, e_n)$ is the aggregated time (measured in ms) the user u_m spent visiting the e_n.

Step 2: For each element in the matrix $V_{M \times N}$, calculate a TFIDF value. We adapt Equation (5.1) to calculate the TF-IDF values, which represent how unique the visit is from each user u_m to each visualization element e_n.

$$TFIDF(u_m, e_n) = \frac{T(u_m, e_n)}{\sum_{i=1}^{N} T(u_m, e_i)} \times log(\frac{M}{f(e_n, Ex)}) \qquad (5.3)$$

where $T(u_m, e_n)$ is the aggregated time the user u_m spent visiting the element e_n, M is the total number of the users, and $f(e_n, Ex)$ denotes the number of users in the exploration collection Ex who spent time on the visual element e_m.

To calculate the *Term Frequency* of TF-IDF, we choose to use a user's aggregated time spent on a visual element $T(u_m, e_n)$, divided by the total time the user spent visiting all the visual elements. There are two alternative options, 1) to use the count of visits from a user to an element $C(u_m, e_n)$, divided by the total number of visits from the user to all the elements, and 2) to use the binary value $\{1, 0\}$ to mark a user's visit to an element (1 if visited, 0 if not visited), and then divide it by the total number of elements visited by the user. We choose the time option over the other two to minimize noise, *i.e.*, during open-ended explorations, a user might accidentally interact with an element, and aggregating the time spent by the user on the element can better indicate the user's intentional visit to

the element.

Step 3: Aggregate the uniqueness scores *Uniq* **for each user** u_m**.**

$$Uniq(u_m) = \sum_{n=1}^{N} TFIDF(u_m, e_n) \tag{5.4}$$

where $TFIDF(u_m, e_n)$ is the TF-IDF value calculated for each visit from a user u_m to a visual element e_n, using Equation (5.3).

By aggregating the TF-IDF values of the visits from one user to all the visual elements, we get a metric depicting the overall uniqueness of the user's exploration. This aggregation process can omit the variation of the TF-IDF distribution, *i.e.*, a user having only one visit with extremely high TF-IDF value may have the same uniqueness metric value as another user having many visits with low TF-IDF values. But the aggregation has at least two advantages. First, it enables the comparison among any users in a group. Before aggregation, each user's interaction behavior is modeled as a vector. A vector supports pair-wise comparison between two users, which can be very useful under some circumstances (see Section 5.6.1). However, it does not support comparison among more than two users. The aggregated metric, instead, supports comparison between subgroups containing any number of users, which is useful for evaluating alternative visualization designs through user interaction. Second, a sum-based aggregation preserves all the unique visits during a user's exploration, *i.e.*, if a user has visited *any* rarely-visited element, it will be preserved in the final metric. Considering an alternative aggregation approach – averaging the uniqueness of a user's visits by the visited elements, if a user has visited lots of frequently-visited elements and one rarely-visited element, the latter will be averaged out after aggregation.

5.4.2 Exploration Pacing

The pacing metric aims to differentiate temporal strategies by users. Given that the temporal features, *e.g.*, *duration* and *moment* of interactions, are underused by existing metrics, we seek metrics that can utilize them and reveal peoples' diverse exploration patterns. The temporal information from a user's exploration process can be viewed as time-series signal, with the moments "visiting any element" marked as non-zero values, and the other moments marked as zeros.

In fields related to signal processing, *e.g.*, image and audio processing, people extract features not only from the temporal aspect, but also the frequency aspect of the signal, *e.g.*, the high-frequency parts of an audio piece. Similarly, the frequency-related information may also carry users' exploration characteristics. While users may visit the same set of visualization elements and spend same amount of time on exploration, the duration and frequency of those visits may be different, and thus may reflect different exploration strategies. By characterizing these differences with *exploration-pacing*, we may begin to quantify another aspect of user engagement with a visualization.

The *exploration-pacing* (*EP*) derived in this section is defined as *the density of a user's high-frequency visits to the visual elements during exploration.* A higher *EP* value suggests a user that rapidly moves from item to item. A lower value might reflect a user that explore individual elements for more time.

Measurable features: We compute the pacing metric based on the distribution of a users' interaction frequency. The features related to time from users' interaction logs (Section **??**) to calculate the metric, *i.e.*, the *moments* and *duration of interaction*, and *exploration time*. We discard the other less relevant features, *e.g.*, *elements*.

Transform approach: One essential step to develop the metric is to select a function to transform a user's visit sequence over time to a visit-frequency sequence. One key intuition here is that merely averaging or binning time durations of a user's interaction with

visual elements is insufficient for developing a single metric, as in-depth interactions will be dominated by multiple shorter-duration interactions. Instead, we observe that common mathematical techniques, *e.g.*, the wavelet transform, can readily transform duration data from the *time* domain to the *frequency* domain.

Adapting the Concept of Continuous Wavelet Transform

The *Continuous Wavelet Transform (CWT)* is often used for extracting the frequency information from a time-series signal by conducting a convolution of the signal with a wavelet function [**?**]. The CWT of a function $x(t)$ at a scale s ($s > 0$, $s \in \mathbb{R}^{+*}$) and translational value $\tau \in \mathbb{R}$ is expressed by Equation (5.5):

$$Wave_w(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t)\, \overline{\psi}(\frac{t - \tau}{s}) dt \qquad (5.5)$$

where $\overline{\psi}(t)$ is the *mother wavelet*, a continuous function in both the time domain and the frequency domain, and the over-line represents operation of complex conjugate.

The *power* of CWT has an interpretation as time-frequency wavelet energy density called the wavelet power spectrum (Equation 5.6). This coefficient can be used to indicate the energy distribution of every moment in a user's exploration, *e.g.*, a moment with higher power values on high-frequency ranges indicate rapid visits to visual elements.

$$Power(s, \tau) = \frac{1}{s}|Wave_w(s, \tau)|^2 \qquad (5.6)$$

Compared to the traditional *short-time Fourier transform* (STFT), CWT can also construct a time-frequency representation of a signal that offers reliable time and frequency localization. This property makes it better at extracting frequency features from the non-periodic time-series user interaction sequences. Thus we adapt CWT to automatically
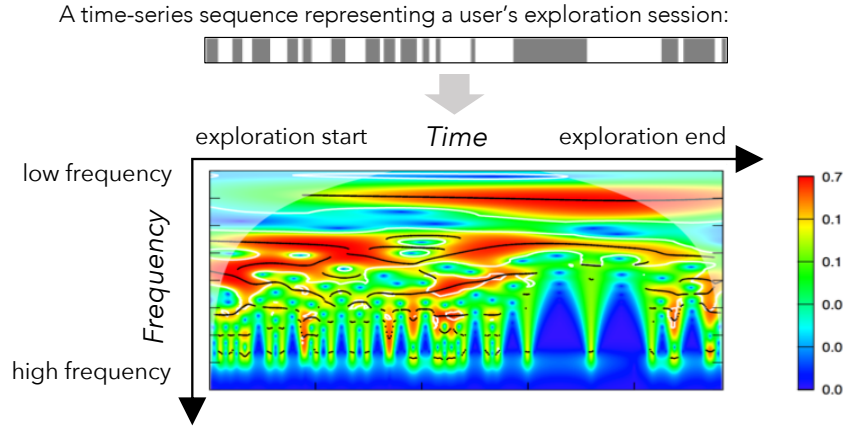
Figure 5.1: A user's exploration interactions can be transformed into a time-series signal with $\{0,1\}$ representing her visiting status. (Time used to visit an element is marked as gray.) The signal sequence can further be transformed to a 2D wavelet power spectrum through continuous wavelet transform.

detect the frequency distribution of a user's visualization exploration over time.

Metric Calculation Steps

The *exploration-pacing* metric is computed in three steps.

Step 1: For each user, form a time-series sequence $S(t)$, representing the user's visiting status over time, from a user's exploration interaction sequence $Ex(u)$. $S(t)$ contains a sequence of values of $\{0,1\}$ over time, sampled from a user's exploration process. If at moment t, the user is visiting an element, then it is marked as 1; otherwise if the user is not visiting any element, it is marked as 0. We sample the data every 0.1 second. If a user visited an element for one second at the beginning of her exploration, then the sequence $S = (1,1,1,1,1,1,1,1,1,1,0,0,0,...)$.

Step 2: Apply *continuous wavelet transform* to the sequence $S(t)$ to obtain a 2D time-frequency wavelet power spectrum. We use the R package "WaveletComp" [131], which computes CWT and obtain a wavelet power spectrum (Figure 5.1) according to Equation (5.5) and (5.6). The Morlet wavelet is used as the mother wavelet for the convolution.

94

Step 3: Obtain the metric value $Pacing_{HF}$ by computing the average power over time and a high-frequency range $[f_{min}, f_{max}]$.

$$Pacing_{HF} = \frac{1}{(f_{max} - f_{min})(t_{end} - t_{start})} \sum_{f=f_{min}}^{f_{max}} \sum_{t=t_{start}}^{t_{end}} Power(t, f) \qquad (5.7)$$

where t_{start} and t_{end} denote the start and end moments of a user's exploration session.

We use 1/32 Hz and 1/8 Hz as the minimum and maximum bounds for the high-frequency range to compute the metric. Given that the sampling rate we use is once per 0.1 second, the high-frequency range corresponds to a period range of 0.8 and 3.2 seconds. This range aims to generally align with high-frequency (*i.e.*, rapid) visit behavior, and to mitigate possible accidental interactions, but can vary depending on visualization design. Precise modeling and parameters given visualization type and user behavior may be a valuable route for future work. For example, the range parameter can be changed to extract users' power density for other frequencies (*e.g.*, low frequency).

5.5   Metric Evaluation
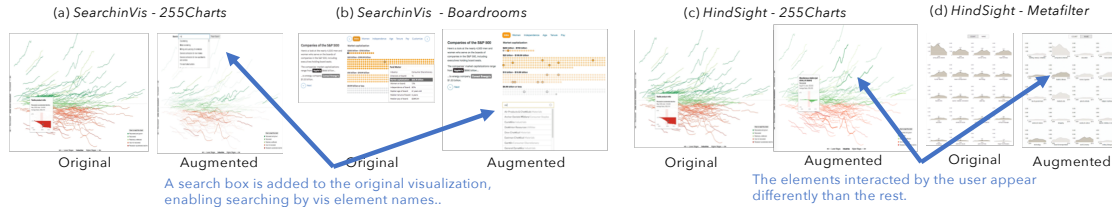
Figure 5.2: Four experiment datasets from two previous studies [12, 13] were used for the metric evaluation: *SearchinVis-255Charts*, *SearchinVis-Boardrooms*, *HindSight-255Charts* and *HindSifght-Metafilter*. Each dataset includes the interaction data of two groups of participants. Each group interacted with either the original or the augmented visualization.

Metrics for Individual Cases

SearchinVis - 255Charts (Dataset Naming: STUDY - EXPERIMENT)

Baseline Map showing % of visits
(each circle is an element in the vis):
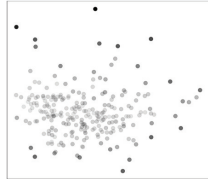
Participants who explored the same number of visited elements (25 elements),
with lower, medium, and higher *exploration uniqueness (EU)*:

EU = 0.9   EU = 1.4   EU = 2.0

user A   user B   user C
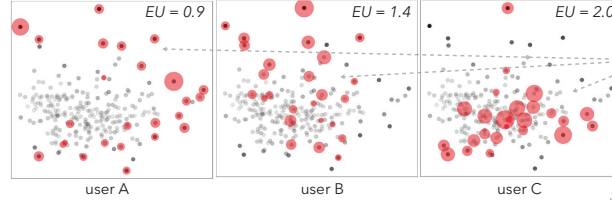
Participants with lower EU tend to focus on the elements at the periphery, which are also frequently visited by others, while those with higher EU tend to explore the middle (rarely-visited) parts of the vis.

Participants who explored for similar amount of time, with lower, medium, and higher *exploration pacing (EP)*:

user D   EP = 0.06
user E   EP = 0.12
user F   EP = 0.17

(The timelines represent the participants' interaction logs. The moments visiting an element are marked as gray.)

HindSight - 255Charts

Baseline map showing % of visits
(each circle is an element in the vis):
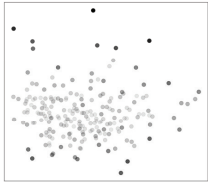
Participants who have the same *number of visited elements* (22 elements),
with lower (left), medium (middle), and higher (right) *exploration uniqueness (EU)*:

EU = 0.7   EU = 1.2   EU = 1.7

user G   user H   user I
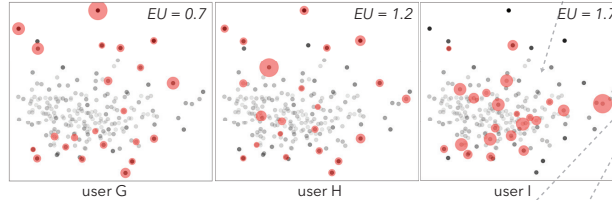
Participants with lower EP tend to explore with lower paces, and focus on individual elements for longer time, while those with higher EP tend to explore the vis in rapid paces.

Participants who explored for similar amount of time, with lower, medium, and higher *exploration pacing (EP)*:

user J   EP = 0.04
user K   EP = 0.10
user L   EP = 0.15

Metrics for Experiment Analyses

(The error bars represent 95% Confidence Intervals. $p < 0.5$ *, $p < 0.1$ **, $p < 0.001$ ***)

■ Experimental Group
■ Control Group

SearchinVis - 255Charts	SearchinVis - Boardrooms	HindSight - 255Charts	HindSight - Metafilter
number of visited elements	*number of visited elements*	*number of visited elements**	*number of visited elements***
0 10 20 30 40	0.0 12.5 25.0 37.5 50.0	0 10 20 30 40	0 4 8 12 16
exploration time	*exploration time***	*exploration time*	*exploration time*
0 100 200 300 400	0 125 250 375 500	0 100 200 300 400	0 50 100 150 200
bias (data point coverage)	*bias (data point coverage)*	*bias (data point coverage)*	*bias (data point coverage)***
0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00
bias (data point distribution)	*bias (data point distribution)***	*bias (data point distribution)*	*bias (data point distribution)****
0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00	0.00 0.25 0.50 0.75 1.00
*exploration uniqueness****	*exploration uniqueness*	*exploration uniqueness**	*exploration uniqueness**
0.8 1.1 1.4 1.7 2.0	0.8 1.1 1.4 1.7 2.0	0.8 1.1 1.4 1.7 2.0	0.4 0.5 0.6 0.7 0.8
*exploration pacing****	*exploration pacing*	*exploration pacing*	*exploration pacing*
0.00 0.03 0.06 0.09 0.12	0.000 0.025 0.050 0.075 0.100	0.00 0.03 0.06 0.09 0.12	0.00 0.03 0.06 0.09 0.12

Metric Correlation and Independence

SearchinVis - 255Charts   (positive / negative correlations)

exploration time
number of visited elements
bias (data point coverage)
bias (data point distribution)
exploration uniqueness
exploration pacing

The metrics exploration time and number of visited elements have a moderate correlation.

The metric bias (data point coverage) and bias (data point distribution) have a strong correlation.

Both exploration uniqueness and exploration pacing metrics can capture different aspects of user explorations.

The exploration pacing metric reveals that those participants in the experimental group tend to explore the vis in lower paces.

The exploration uniqueness metric reveals that those participants in the experimental group tend to have a more unique exploration compared to others.

Figure 5.3: We applied the proposed *pacing* and *uniqueness* to two previous studies, with results suggesting that they capture different facets of user explorations of visualizations on the web.

5.5.1 Interaction Data from Two Studies

We evaluate the two proposed metrics by applying them, together with other four metrics used in visualization literature, to interaction data collected from two previous studies from Feng *et al.* [12, 13]. ***SearchinVis*** [13] studied the effects of adding text-based search functionality to interactive visualizations on the web. Search enables a user to highlight elements in the visualization by typing keywords in a search box. ***HindSight*** [12] studied the effects of directly encoding people's personal interaction history in visualizations. During a user's exploration, visual elements the user interacts with are visually augmented to appear distinct from unexplored elements. Figure **??** shows visualizations from these studies whose data was adapted for the present work.

Participants and Conditions: Both studies include multiple experiments, one for each study-visualization pair, *e.g.*, *SearchinVis-255Charts*. All experiments had a between-subjects design, conducted on Amazon Mechanical Turk. Every participant was randomly assigned to interact with either the original or augmented version of a visualization. Table 5.2 shows the participant numbers of the control and experimental groups in each experiment.

Task and Procedure: Each participant was asked to analyze the visualization for as long as they liked before answering questions. The authors used an open-ended exploration task in order to simulate people's real-world explorations of web visualizations. Each participant went through several phases, including *introduction*, *exploration*, and *insight/strategy*. In the *introduction* phase, the participant was given instructions to interact with the visualization in any way they saw fit. Afterwards, the participant entered the *exploration* phase, where they could interact with the visualization without time limit. When the participant finished exploring, they entered the *insight/strategy* phase, where they answer questions about their findings aimed at highlighting possible differences in the control and experiment conditions.

Studies	**Searchin Vis**		**HindSight**	
Experiments	*Colleges*	*255Charts*	*255Charts*	*Metafilter*
control	67	57	57	44
experimental	72	72	59	48
Total	*139*	*129*	*116*	*92*

Table 5.2: Number of participants in 4 experiments in the previous studies [12, 13]. The control user group in each experiment includes the users randomly assigned to explore original visualizations. The experimental group includes those exploring augmented visualizations.

During the *exploration* phase, each participant's interactions with visual elements were recorded as $Ex(u) = (t_{start}, I_1, ..., I_k, t_{end})$, where t_{start} and t_{end} are the start and end time of the *exploration* phase, and $I_i(0 \leq i \leq k)$ denotes every interaction (Table 5.1). Each interaction was recorded as $I = (t, a, E_i, d)$, where t is the start moment of the interaction, a is the interaction type (hovering), E_i is the set of interacted elements (there is one and only one element affected by each interaction), and d is the duration of the interaction.

Specifically, we select from these studies visualizations that were adapted from published visualizations on the web, including 255*Charts* and *Boardrooms* from the *SearchinVis* study, and 255*Charts* and *Metafilter* from *HindSight*. We exclude the remaining visualizations in these studies in the present analysis, such as a bubble chart of exoplanet data (*i.e.*, condition 5 in [13]), because they were designed to test specific hypotheses in the prior studies, such as the effects of dataset familiarity and search behavior.

5.5.2   Applying Interaction Metrics: Case Studies

Can the proposed metrics reveal new characteristics of people's interaction with visualizations?

To evaluate the extent to which the proposed metric *exploration-uniqueness*(*EU*) shows different exploration patterns, we examine several individual exploration sessions in which

users visited the same number of visual elements (*NVE*), but varied on the *EU* values. Then we visually examine whether an exploration with a higher metric value includes more elements rarely-visited by others.

In order to distinguish frequently- and rarely-visited elements, we calculate the percentage of visits for each element, and plot it as a baseline map (Figure 3.4), where each circle represents a visual element, and its opacity mapped to the percentage of users who visited it. From the baseline maps of both *SearchinVis-255Charts* and *HindSight-255Charts*, visual inspection suggests that the elements at the periphery of the visualization are frequently visited, while the elements in the middle are rarely visited.

As shown in Figure 3.4, user A, B and C from ***SearchinVis-255Charts*** visited the same number of elements (25) during their explorations. However, these explorations vary at *EU*, *i.e.*, A's is the lower (0.9), B's is the medium (1.4), and C's is the higher (2.0). For each of them, we plot all visited elements on top of the baseline map, where each circle represents a visit, with its size corresponding to how long the user spent visiting this element. Visually comparing the visit maps from the user A, B and C, we find that the elements visited by A (lower *EU*) are mostly at the periphery of the visualization, which are frequently visited by other users, according to the baseline map. Instead, user C (higher *EU*) visited more elements located at the lower-middle part of the visualization, which are rarely visited by other users in the study. The elements visited by user B (middle *EU*) include some frequently-visited and some rarely-visited elements.

Similarly, user G, H and I from ***HindSight-255Charts*** visited the same number of elements (22) with varying *EU*, *i.e.*, G's is the lower (0.7), H's is the medium (1.2), and I's is the higher (1.7). By visually comparing the visit maps from G, H and I, we see that the elements visited by G (lower *EU*) are mostly at the periphery of the visualization (*i.e.*, frequently-visited elements). User I (higher *EU*) visited more elements at the middle part of the visualization, which are rarely visited by other users in the study. The elements

visited by user H (middle *EU*) include both frequently-visited and rarely-visited elements.

We also find that *exploration-pacing* (*EP*) can reveal differences in the pacing of user explorations. Specifically, we examine the individual cases from *SearchinVis* (user D, E, F) and *HindSight* (user J, K, L). Each of these users explored the visualization for similar amount of time, but with lower, medium or higher paces. User D, for example, appears to intersperse rapidly-paced interactions with longer interactions. User F, in contrast, spends nearly all of their time performing rapid exploration. These cases illustrate that the pacing metric can aid in distinguishing between the temporal behavior of users, essentially by transforming the temporal observations to the frequency space.

5.5.3 Metrics for Experiment Analyses

Can the metrics provide additional insight in experiment analyses?

After examining individual cases to check the validity of the proposed metrics, we explore their effectiveness on one of the potential application scenarios, *i.e.*, to show the impact of visualization designs on user interaction behavior. Specifically, we compare the metric values between two user groups (experimental and control) in each study by applying the same statistical tests as in the original studies. We also evaluate four other metrics proposed or used in visualization literature, *number-of-visited-elements*, *exploration-time*, *bias-data-point-coverage*, and *bias-data-point-distribution*.

Following the statistical methods used in the previous studies, we compute 95% confidence intervals using the bootstrap method, and effect sizes using Cohens d - which is the difference in means of the conditions divided by the pooled standard deviation. We also use the non-parametric Mann-Whitney test to compare different user groups.

Study 1: SearchinVis

We apply the existing and proposed metrics to the interaction logs from the two exper-iments of the study *SearchinVis*, *i.e.*, *255Charts Boardrooms* visualization stimuli, to examine the behavioral impact of the text-based search functionality.

255Charts: We filtered out 5 (from 129) users who did not interact with any elements of the visualization. We found that, compared to the users in the control group, the users from the experimental group show significantly more unique explorations (***exploration-uniqueness***), and had fewer rapid-pace visits (***exploration-pacing***). The experimental group had a higher *exploration-uniqueness* on average (M=1.7 95% CI [1.6, 1.8]) than the control group (M=1.4 95% CI [1.3, 1.5]). The Mann-Whitney test shows that $W = 1165, p = 0.0002$, and the effect size is $d = 0.7\ [0.3, 1]$. The experimental group has a lower *exploration-pacing* value on average (M=0.07 95% CI [0.06, 0.07]) than the control group (M=0.11 95% CI [0.1, 0.12]). The Mann-Whitney test shows that $W = 3024, p = 2.3^{\times}10^{-8}$, and the effect size is $d = -1.2\ [-1.57, -0.8]$.

Importantly, these metrics align with and quantify the findings and intuitions of that study. The addition of search to *255Charts* encouraged more diverse spatial patterns of exploration, while also nudging users to look more in-depth at specific visual elements.

Boardrooms: We filtered out 11 (from 96) users who did not interact with any visual element. We found significant differences on the ***exploration-time*** and ***bias-data-point-distribution*** between the two user groups. Specifically, the experimental group spent longer time on average (M=405 95% CI [337, 480]) in seconds than the control group (M=290 95% CI [234, 368]). The Mann-Whitney test shows that $W = 598, p = 0.007$, and the effect size is $d = 0.51\ [0.05, 0.96]$. This difference has been reported in the previous study [13]. The experimental group has a lower *bias-data-point-distribution* value on average (M=0.03 95% CI [0.01, 0.09]) than the control group (M=0.16 95% CI [0.09, 0.25]). The Mann-Whitney test shows that $W = 1109.5, p = 0.01$, and the effect size is

$d = -0.6\,[-0.95, -0.14]$.

In summary, by applying the existing and proposed metrics to two experiments in the study, we found that our proposed metrics appear to provide additional insight in experiment analyses that could have appeared in previous published work, *i.e.*, by quantifying the impact of text-based search functionality on the uniqueness or "diversity" of user explorations. At the same time, we find that in Feng *et al.*'stext-based search study, the presence of the search functionality appears to have a different impact on user behavior when added to the 255*Charts* and *Boardrooms* visualizations.

Study 2: HindSight

We apply existing and proposed metrics to the interaction data from the two experiments of the study *HindSight* [12], which includes *255Charts*, a visualization from The New York Times, and a comparatively simpler *Metafilter* visualization stimuli. The aim here is to examine the behavioral impact of the direct encoding of personal interaction history, by comparing the two user groups in each experiment.

255Charts: We filtered out one (from 116) user who did not interact with any visual elements. We found significant differences on the metric **number-of-visited-elements** and **exploration-uniqueness** between the two user groups, while the other metrics are similar between groups. Specifically, the experimental group visited more visual elements on average (M=28 95% CI [24, 35]) than the control group (M=21 95% CI [18, 24]). The Mann-Whitney test shows that $W = 1392.5, p = 0.11$, and the effect size is $d = 0.41\,[0.07, 0.73]$. This difference has been reported in the previous study [12]. The experimental group has a higher *exploration-uniqueness* on average (M=1.44 95% CI [1.34, 1.54]) than the control group (M=1.26 95% CI [1.17, 1.35]). The Mann-Whitney test shows that $W = 1165, p = 0.0002$, and the effect size is $d = 0.48\,[0.11, 0.84]$.

Metafilter: We found significant differences on the metric **number-of-visited-elements**,

bias-data-point-coverage, *bias-data-point-distribution*, *exploration-uniqueness* between the two user groups, while the other metrics are similar between groups. Specifically, the experimental group has a higher *number-of-visited-elements* on average (M=9.4 95% CI [7.7, 11.3]) than the control group (M=5.4 95% CI [4.3, 6.5]). ($W = 686, p = 0.004, d = 0.75$ $[0.37, 1.12]$) This difference has been reported in the previous study [12]. The experimental group also has a higher *exploration-uniqueness* on average (M=0.7 95% CI [0.6, 0.8]) than the control group (M=0.6 95% CI [0.6, 0.7]). The Mann-Whitney test shows that $W = 801.5, p = 0.047$, and the effect size is $d = 0.44$ $[0.02, 0.86]$. In addition, we found that the experimental group has a higher *bias-data-point-coverage* value on average (M=0.9 95% CI [0.8, 0.9]) than the control group (M=0.8 95% CI [0.7, 0.8]). The Mann-Whitney test shows that $W = 646.5, p = 0.001$, and the effect size is $d = 0.71$ $[0.26, 1.13]$. The experimental group also has a higher *bias-data-point-distribution* value on average (M=0.5 95% CI [0.4, 0.6]) than the control group (M=0.2 95% CI [0.1, 0.3]). The Mann-Whitney test shows that $W = 616.5, p = 0.0006$, and the effect size is $d = 0.82$ $[0.41, 1.24]$.

In summary, by applying the existing and proposed metrics to two experiments in the study *HindSight*, we found that our proposed metrics can provide additional insight in experiment analyses, *i.e.*, uncover the impact of direct encoding of personal interaction history on the uniqueness of user explorations. We also find that the HindSight technique has a different impact on user behavior, specifically, users' bias levels, when added to the real-world and less complex visualizations.

5.5.4   Metric Correlation and Independence

Are the metrics correlated or independent when applied to real data?

We compute correlations between each pair of the metrics across all experimental datasets (Figure 3.4), *SearchinVis-255Charts*, *SearchinVis-Boardrooms*, *HindSight-255Charts*

and *HindSight-Metafilter*. We expect that the metrics measuring different high-level aspects of user explorations are independent from each other.

We found strong and moderate correlations, $r = [0.5, 1)$, between the metric *bias-data-point-coverage* and *bias-data-point-distribution*. Specifically, they are strongly correlated in *SearchinVis-255Charts* ($r = 0.72, p < 0.001$) and *HindSight-Metafilter* ($r = 0.76, p < 0.001$), and moderately correlated in *HindSight-255Charts* ($r = 0.67, p < 0.001$) and *SearchinVis-Boardrooms* ($r = 0.61, p < 0.001$). This indicates that for these cases, the two metrics play similar roles characterizing exploration behavior.

We also found a moderate correlation, $r = [0.5, 0.7)$, between *number-of-visited-elements* and *exploration-time*, in *SearchinVis-255Charts* ($r = 0.57, p < 0.001$).

Both *exploration-uniqueness* and *exploration-uniqueness* have weak correlations or less with the other metrics across all the datasets. These results suggest that the proposed metrics carry different information than the others when applied to user exploration data. However, we note that linear correlation is just one of many possible measures of dependence, and further analyses with larger datasets may be necessary to make definitive claims.

5.6 Discussion

The proposed uniqueness and pacing metrics aim to reveal new facets of how people interact with visualizations. To examine this claim, we applied the proposed metrics, together with metrics from prior work, to interaction data from prior information visualization studies. The results suggest that, first, the proposed metrics do reveal new characteristics of peoples exploration behavior in visualizations. Second, the proposed metrics can be used as target metrics in comparative experiments, *i.e.*, quantitative analysis comparing control and treatment groups. Third, the proposed metrics are also generally independent

of prior metrics used in visualization research, as indicated by the correlation analysis, implying that they may be a source of new information for exploratory visualization design and research.

In the analysis of the study *SearchinVis*, we found that the results differ for 255*Charts* and *Boardrooms*. The metric values of *exploration-uniqueness* and *exploration-pacing* are significantly different between groups in 255*Charts* while in *Boardrooms* they are similar. One possible explanation is that the *Boardroom* visualization is in a storytelling form with multiple types of interaction. Larger interaction sets, then, may pose new challenges and opportunities for measures of interaction behavior.

We also found that in the experiment *HindSight-Metafilter*, the experimental user group has higher values in the bias metrics on average than the control group. By further examining the interaction logs, we found that a higher value in the bias metric is contributed by the revisits to previously-visited visual elements. On the other hand, the users in the experimental group visited more charts on average than those in the control group, in the whole exploration process. The proposed metrics thus lead to new questions, such as how much bias measures should weight revisits against the breadth of peoples' exploration?

5.6.1   Potential Applications of Interaction Metrics

New and newly evaluated metrics for visualization interaction analysis may open several opportunities in visualization research.

As *metrics* to quantify the impact of visualization designs. One goal of researchers and practitioners is to examine the comparative impact of competing techniques on user behavior. However, users' open-ended explorations of visualizations can be complex, and cannot be adequately summarized into *number of actions taken* or *total time spent on exploration*, which are the basic metrics commonly used in previous evaluations, *e.g.*, [4,

105

12, 24]. Instead, we have shown that from these low-level user interaction components we can develop metrics that provide new perspectives into users' open-ended explorations, which may allow us to better assess the impact of a given visualization or interaction technique on user behavior.

As *proxies* **to infer user characteristics and reasoning processes.** Behavioral patterns of exploration can be used to infer users' characteristics (*e.g.*, locus of control [5, 122]), reasoning processes [6], and insight generation [22]. By providing new ways to characterize users' exploration behavior, we could possibly explore new avenues of individual differences in visualization use and preference.

As *attributes* **to visualize users' interaction logs.** One advantage of visualization approaches for analyzing interaction data is that detailed information can be logged during user interaction, as shown in previous works that center on visualizing interaction logs *e.g.*, Blascheck *et al.* [37, 126]. One limitation of this approach is that it relies primarily on expert analysts to visually identify trends across user interaction traces. Fortunately, recent work has begun to explore automatic approaches to assist in navigating interaction traces, such as sequence search and extraction [37]. We contend that new metrics can also aid in this direction of research, by serving as relatively low-barrier features that could be encoded in interaction log visualization, for example ordering or coloring by uniqueness, bias, or pacing [132].

As *features* **to support machine learning algorithms.** Machine learning techniques have also been used to analyze users' interactions with visualizations, *e.g.*, to classify user characteristics [5], to extract interaction sequences [25], and to cluster users by behavioral patterns [37]. The proposed metrics in this work, as well as intermediate variables generated from the computation process of the metrics, may be useful as features for these machine learning approaches. For example, users' explorations can be clustered using the *feature vectors* containing the time spent on each visual element (Equation 5.2), the vec-

tors of TF-IDF values (Equation 5.3), or the highest level exploration uniqueness metric values. Similarly, both the two-dimensional wavelet power spectrum and the corresponding pacing metric can be used as features for machine learning algorithms.

5.6.2 Benefits and Tradeoffs of Interaction Metrics

All of the metrics in our evaluations may prove beneficial to user behavior analysis in visualizations, due to their ability to uncover different facets of peoples' explorations. However, potential adopters need to be aware of certain properties of these metrics in order to apply them correctly. We now compare the metrics in our evaluations according to a list of criteria focusing on barriers such as interpretability, and derive initial guidelines on when and how to use these metrics.

The criteria used for metric comparison are adapted from the works in relevant fields evaluating metrics [133, 134, 135], and are listed from lower to higher perspectives (*i.e.*, from metric computation to human perception and cognition):

Computational cost (computational level): *How much does it cost for the metric computation?* The computation of some metrics is trivial, such as *number-of-visited-elements*. However, there is a certain level of complexity required to compute other metrics *e.g.*, *exploration-uniqueness*. The computation of *exploration-pacing* requires more resources and its complexity depends on the choice of convergence parameters. The computation of *bias-data-point-coverage* includes power operations on the number of interactions performed by the user, *i.e.*, N^k where N is the number of all the interactive elements in the visualization, and k is the number of interactions performed by the user. This suggests that extra steps may be needed to avoid the overflow caused by large numbers when dealing with an exploration session where a user interacts with the visualization a lot, *e.g.*, $k > 100$.

Computational context (computational level): *Does the metric computation require*

extra context, i.e*., out of the single user scope?* Among all the metrics in our evaluation, the computation of *exploration-uniqueness* depends on the interaction logs not only of the current user being considered, but also those of the other users within the same group, while the computations of other metrics, *e.g.*, *exploration-pacing* are only based on the current user, *i.e.*, no extra context needed. This property influences the practical usage of a metric, *e.g.*, a reasonable number of users should be selected when computing the *exploration-uniqueness* metric.

Comparability (application level): *(How) can the metric values be compared?* All of the evaluated metrics are comparative because they are quantitative measures of scale. The comparability of *exploration-uniqueness* is constrained because the metric values are only comparative within a TF-IDF computation group, *i.e.*, it is not feasible to compare the metric values of two users in different computation groups. The values of the *exploration-pacing* metric can be compared across user groups if the same set of parameters are used for the computation.

Interpretability (cognition level): *How easily can the metric be understood or interpreted by human?* The proposed metrics have different levels of interpretability. Metrics such as *exploration-time* and *number-of-visited-elements* could be considered readily interpretable, since people can easily understand the meaning of the values (*e.g.*, 10 elements, 15 seconds). The values of some other metrics may require more cognitive effort to interpret, *e.g.*, *exploration-uniqueness*, *exploration-pacing* and *bias-data-point-coverage*.

Knowledge coverage (cognition level): *How much additional knowledge does the metric cover given other metrics?* This dimension evaluates whether proposed metrics can uncover characteristics of users' exploration that other metrics do not capture. By examining correlations between metrics, we found that both proposed metrics, *exploration-uniqueness* and *exploration-pacing* may reveal different perspectives from other metrics.

5.7 Future Work & Conclusion

Despite their advantages, there are also limitations with the proposed metrics. For example, the uniqueness metric focuses on an interaction set, ignoring ordering effects, which may be another source of exploration diversity. The visualizations tested thus far are also constrained in interaction scope, implying that more complex interaction schemas may require more sophisticated approaches for developing useful metrics. Further, there are currently no established guidelines in the visualization community for evaluating other characteristics of proposed metrics, such as consistency, discriminability, and reliability. Given the changing landscape of visualization in the world, addressing challenges such as these may be fruitful areas for future work.

Each day, thousands of people interact with thousands of interactive visualizations across the web's vibrant and growing visualization ecosystem. However, our metrics for quantifying facets of peoples' open-ended explorations with these visualizations are lacking, as they are primarily based on low-level metrics such as *elements visited* or *time spent exploring*. The aim of this work is to characterize, develop, and evaluate metrics for visualization interaction that can be used in a variety of settings. We introduce two new metrics, *uniqueness* and *pacing*, and evaluate these metrics alongside those proposed in earlier and more recent research in visualization. The results of these evaluations suggest that, indeed, new metrics may provide new perspectives on how people interact with the visualizations they come across. We discuss the broad potential applications of new metrics for visualization interaction analysis, and enumerate some of the challenges future work in interaction metrics may face in the future.

Chapter 6

Modeling Diverse Explorations with Visualizations using Bayesian statistics

Studies targeting how people interact with visualizations have yielded advances in techniques for evaluating competing designs, methods for guiding peoples exploration process in-situ, and guidance for human-in-the-loop analytics approaches, among other successes. At the same time, these studies typically make use of dichotomous statistical analysis practices, which, given the diverse ways in which people explore visualizations, may lead to overly general conclusions about the nature of interaction in visualizations. One possible means for addressing this gap is to employ emerging statistical paradigms involving Bayesian statistics, which offer more expressive ways of analyzing and reporting differences in observed behavior.

In this chapter, we aim to demonstrate some of the possible benefits of Bayesian statistical modeling for examining visualization behavior by re-analyzing interaction data from previous studies open datasets regarding visualizations on the web. Specifically, we use Bayesian statistics to re-examine the effects of interaction design choices on quantitative aspects of participants exploration sessions, such as the number of elements visited in an interactive visualization. These analyses go beyond previously reported dichotomous

results, and yield a more precise distribution of the overall effects of design choices in the studied interactive visualizations. Beyond re-analysis, we demonstrate how these new statistical paradigms can address a new range of questions about interactive behavior. For example, we are able to identify both design-level and element-level effects on peoples interaction count, interaction duration, and the probability of a participant mentioning an element after exploration, given characteristics of their exploration behavior.

We discuss the implications of the study, including potential applications of models for capturing and predicting interactive behavior with visualizations, as well as the benefits and tradeoffs of incorporating Bayesian techniques into studies on how people interact with visualizations.

6.1 Introduction

Researchers and practitioners study people's interactions with visualizations for various purposes, e.g., to evaluate the effectiveness of alternative designs, to infer people's characteristics from their behavior, and to recover and reuse their visual analytic provenance, etc. However, these studies typically make use of dichotomous statistical analysis practices that may lead to overly general conclusions that are less practical [136].

For example, a typical research question is to examine the effectiveness of a novel design technique on people's interaction behavior. Null hypothesis significance testing (NHST) provides a "yes-no" answer to the question, i.e., determining whether or not there exists an effect. However, this conclusion cannot lead to a practical design guideline. If there *is* an effect, one might wonder whether the effect *always* exists – if not – the probability of its occurrence. That probability would help determine, when there are competing design options having the same effect, which design to choose.

The emerging statistical paradigms involving Bayesian statistics may help build prac-
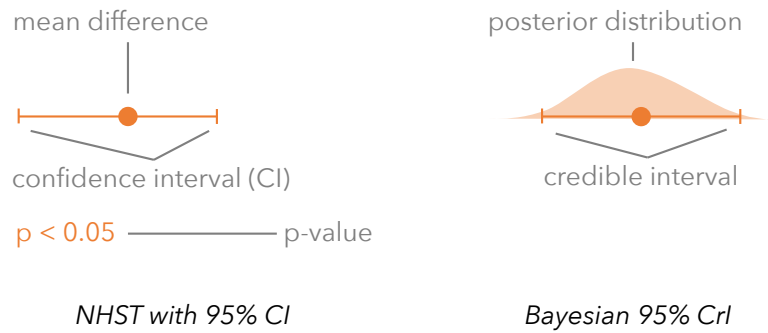
Figure 6.1: Typical result outputs of null hypothesis significance testing (NHST) with 95% confidence interval and Bayesian statistics, measuring the difference of means between two groups.

tical behavior-driven guidelines by offering more expressive ways of analyzing and reporting differences in observed behavior (Figure 6.1). This alternative statistical practice has been advocated and demonstrated by some researchers through the studies of visualization perception and human-computer interaction [137, 138, 139]. However, Bayesian statistics has not yet been incorporated into the studies of human interaction with visualization. We see lots of potential benefits it can bring to the visualization behavior research, such as deriving more practical behavior-driven design guidelines, being able to precisely model people's diverse behavior patterns, etc. Nevertheless, these potential benefits remain under-explored in the field of visualization interaction analysis, so are the trade-offs of alternating the standard practice of statistical inference.

To bridge this gap, in this work, we aim to demonstrate some of the possible benefits of Bayesian statistics for examining visualization behavior. In the meantime, we aim to point out the potential trade-offs through our demonstration, that some parts of the analysis workflow require extra statistical and domain knowledge compared to the current standard practice, and may lead to erroneous results due to incorrect practice. To do so, we re-analyze interaction data from previous studies' open datasets regarding people's

interactions with visualizations on the web, *HindSight* and *SearchinVis* [12, 13]. First, we re-examined some of the results by using Bayesian statistics to answer the same analysis questions as in the previous study. Then we explored new questions facilitated by Bayesian techniques. The **contributions** of this work are three-fold:

1. We use Bayesian statistics to re-examine the effects of interaction design choices on quantitative aspects of participants exploration sessions, and found that these analyses go beyond previously reported dichotomous results, and yield a more precise distribution of the overall effects of design choices in the studied interactive visualizations.

2. We demonstrate how these new statistical paradigms can address a new range of questions about interactive behavior. Specifically, Bayesian statistics facilitate multiple comparisons, enabling us to investigate the effects on multiple levels (e.g., participant-level and element-level).

3. We derive new findings on people's interactive explorations of visualizations, including both design-level and element-level effects on peoples interaction count, interaction duration, and the probability of a participant mentioning an element after exploration, given characteristics of their exploration behavior.

We further discuss the implications of the study, including potential applications of models for capturing and predicting interactive behavior with visualizations.

6.2 Background

In this section, we first introduce the current and emerging alternative statistical paradigms, Null Hypothesis Significance Testing (NHST) and Bayesian statistics, followed by the concerns regarding the current practice and the advocacy of Bayesian statistics. Then we introduce the two previous studies from which we reanalyzed the open datasets.

113

6.2.1 NHST and Bayesian Statistics

Both *Null Hypothesis Significance Testing (NHST)* and *Bayesian Statistics* are the methodologies to make *statistical inference*, which is the process to go from an *effect* back to a *cause* [140]. For example, given the *data* (*i.e.*, *effect*) of the heights of some high school students and some elementary school students, estimate the true difference of heights between the two populations (*i.e.*, *cause*).

In NHST, the parameters of the probability model is *fixed* (*i.e.*, the true difference of the heights between the two populations is fixed), and we can approach the parameter through *varying* data generated from an infinite sequence of repetitions.

In Bayesian Statistics, the parameters of the system are *varying* (*i.e.*, the true difference of the heights between the two populations is varying, *e.g.*, there is a 60% chance that the difference is around 20 cm, and a 10% chance that the difference is around 10 cm). Bayesian statistics approaches the statistical inferences via the Bayesian rule (Equation 6.1).

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{6.1}$$

where $p(\theta|data)$ denotes the probability of a hypothesis θ given the actual *data* obtained.

6.2.2 Reflections on the Current Statistical Paradigm

The concerns with *replication crisis* in some scientific fields have led to reflections of the *validity* of prior studies in some other scientific fields, especially those where the studies involve human subjects, *e.g.*, human-computer interaction (HCI) and data visualization (dataVis). [137, 141].

In the field of dataVis and HCI, the validity of a study could be compromised due to several reasons. Kosara *et al.* [141] listed six possible reasons, including statistical fluke,

114

questionable research practices, analysis problems, study design flaws, overgeneralized conclusions, and misinterpreted results. Of those six reasons, half are to related to the current standard of *statistical methodologies* used in research, directly or indirectly:

- *Statistical Fluke*: A study can lead to a statistically significant finding by accident. The common cutoff of alpha=0.05 still allows for a 5% false-positive rate (or 1 in 20). With the current standard practice of statistical analysis, 1 of 20 studies could get a significant result by chance. Kay *et al.* [137] also argue that for *small-n* studies (with a small number of participants), which are common in HCI, once a false-positive result is achieved, its effect size tends to be overestimated.

- *Questionable Research Practice*: Statistical analysis of study results allows significant leeway that can lead to false positives. For example, motivated to get a significant result ($p < 0.5$), some researchers may be susceptible to questionable research practices, *e.g.*, adding more participants, or changing the statistical test, etc.

- *Analysis Problems*: The data analysis is flawed through the application of the wrong statistics, incorrect comparisons, etc. For example, during data analysis, researchers may not check the assumptions of statistical tests, and thus pick the wrong method, or conduct the analysis incorrectly.

Additionally, some HCI practitioners have pointed out that confidence intervals in the current statistical paradigm can lead to "unjustified or arbitrary inferences" and cause confusion and misapplication even from those with a background in mathematical statistics [142]. Several have recommended that researchers should consider other statistical approaches, mentioning the Bayesian approach in particular as a valid alternative.

6.2.3 Potential Benefits and Trade-offs of Bayesian Statistics in dataVis and HCI

Using Bayesian statistics would address two of the three concerns presented by Kosara et al. [141]. Moving away from p-values prevents p-hacking and other questionable practices. Furthermore, the 5% false-positive rate from 95% confidence interval would no longer be a potential danger, as the credible intervals generated from the Bayesian approach explicitly integrates the uncertainty in the result.

In addition to addressing methodological concerns, Bayesian statistics are a better match for HCI given the nature of the field. Kay et al. [137] pointed out that small-n studies (those with a small number of participants) are common in HCI and sometimes cost-effective, e.g., when comparing a novel system to some existing approach. They suggest that in many fields, the current statistical approach requires meta-analysis to unify results and ensure that studies generating false positives are identified, but the nature of HCI has not shown to lead to meta-analyses. However, Bayesian statistics facilitate quantitative knowledge accrual that does not require conducting extra meta-analyses.

Kay et al. also note that a Bayesian approach will eliminate the dichotomization that characterizes most statistical analysis using NHST. Using NHST limits hypotheses to answering a binary question: "Is there a difference in mean visual elements?" There are a variety of other questions of possible interest, but NHST is not generally intended to answer them. However, a Bayesian approach allows for more nuanced consideration.

A significant obstacle to the application of Bayesian statistics in earlier years was the computational requirements for the calculations. Developments in computer technology, as well as statistical algorithms (e.g., Stan) have improved the situation.

On the other hand, the analytical problems mentioned by Kosara et al. is still likely to be relevant. As mentioned above, they suggested that improper statistical analysis can

compromise the validity of a study. Bayesian statistics are known for being statistically involved. For example, appropriate distributions must be chosen for a prior and for a potentially large amount of parameters.

6.2.4 Previous Studies with Open Datasets

In this work, we explore the effectiveness of Bayesian statistics by re-analyzing the interaction data collected from two previous studies from Feng *et al.* [12, 13]. ***SearchinVis*** [13] studied the effects of adding text-based search functionality to interactive visualizations on the web. Search enables a user to highlight elements in the visualization by typing keywords in a search box. ***HindSight*** [12] studied the effects of directly encoding people's personal interaction history in visualizations. During a user's exploration, visual elements the user interacts with are visually augmented to appear distinct from unexplored elements.

Participants and Conditions: Both studies include multiple experiments, one for each study-visualization pair, *e.g.*, *SearchinVis-255Charts*. All experiments had a between-subjects design, conducted on Amazon Mechanical Turk. Every participant was randomly assigned to interact with either the original or augmented version of a visualization.

Each participant was asked to analyze the visualization for as long as they liked before answering questions. The authors used an open-ended exploration task in order to simulate people's real-world explorations of web visualizations. Each participant went through several phases, including *introduction*, *exploration*, and *insight/strategy*. In the *introduction* phase, the participant was given instructions to interact with the visualization in any way they saw fit. Afterwards, the participant entered the *exploration* phase, where they could interact with the visualization without time limit. When the participant finished exploring, they entered the *insight/strategy* phase, where they answer questions about their findings aimed at highlighting possible differences in the control and experi-

ment conditions.

Interaction Data: During the *exploration* phase, each participant's interactions with visual elements were recorded. Each interaction includes the following properties: the start and end moments of the interaction, the interaction type (in this case, hovering), the set of interacted elements (there is one and only one element affected by each interaction), the duration of the interaction.

Specifically, to focus on the analysis approach, we select from these studies the visualizations that were adapted from published visualizations on the web, 255*Charts*, and exclude the remaining visualizations in these studies in the present analysis.

6.3   Re-examination of Previous Results

In this section, we present a re-examination of the results in the previous study *HindSight*. Specifically, we ask the same question: Does the interaction technique HindSight affect the number of visual elements people visit during exploration? We start with the previous analyses and results based on 95% confidence intervals, and then go through the Bayesian analysis process and derive the results. Finally, we compare the two approaches and discuss their differences, specifically, in assumptions and result interpretation.

6.3.1   Previous Analysis based on 95% CI

The analysis of the previous study *Hindsight* focus on confidence intervals and effect sizes, specifically, computing 95% confidence intervals using the bootstrap method, and effect sizes using *Cohens d*, which is the difference in means of the conditions divided by the pooled standard deviation. The significance testing and related statistics are also included to supplement the analyses.

The results show that the average participant in the hindsight condition visited more
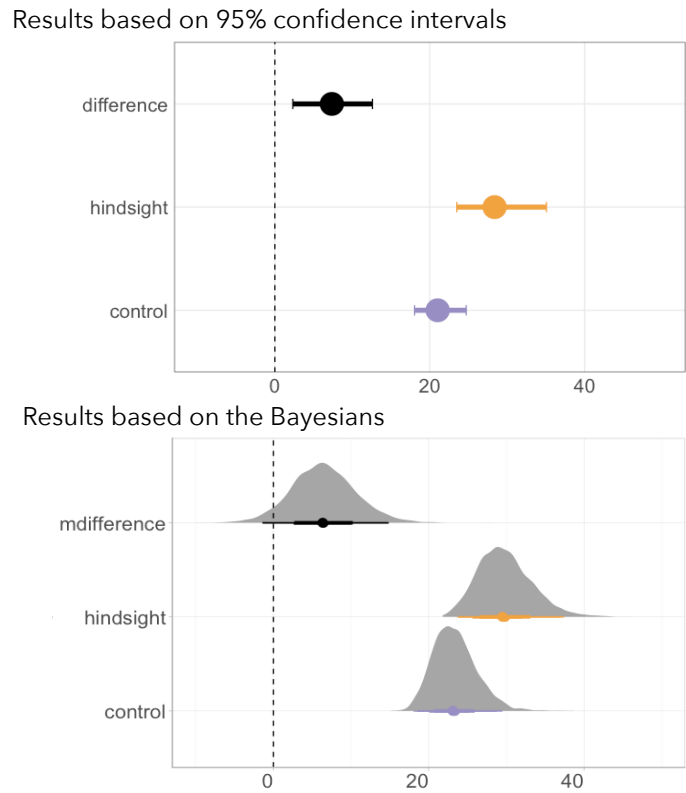
Figure 6.2: Comparison of the analyses based on 95% confidence intervals and on the Bayesian statistics.

visual elements (M = 28.4 95% CI [23.3,34.2]) than those in the control condition ((M = 20.7 95% CI [17.9,24.2]). Given the upper and lower limits of the confidence intervals, the average participant visits at least 1 additional chart with hindsight, and up to 16 more (d = 0.44 95% CI [0.12,0.75]).

6.3.2   Secondary Analysis based on the Bayesians

In the data, each observation describes a participant's exploration session, including the *condition* the participant was assigned to as well as the number of visited elements during exploration (*visits*).

We fit a log-normal regression to model the number of visited elements. Log-normal

regression has been shown to be better-suited to model long-tailed distributed data [143]. We describe the first part of the **model** as:

$$visits_i \sim Lognormal(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \times condition_i$$

(6.2)

These two lines describe the likelihood function *lognormal* and the generalized linear model. $visits_i$ is the response variable, the number of visited elements. $condition_i$ is the predictor variable, which means the condition a participant was assigned to (either hindsight or control). We have three parameters to be estimated, the intercept α (the global population effect), the slope β (the effect of condition), and the variance residual σ for the log-normal function.

To fit a Bayesian model, we must explicitly express initial belief of data by specifying prior distributions for each of the parameters, which in our case, are α, β and σ. Following the Bayesian workflow [144] suggested by Gabry et al., we specify weakly-informed priors based on the domain knowledge of people's interaction behavior with visualization. The prior distributions used in our analysis can be expressed as:

$$\alpha \sim Normal(2.5, 1)$$

$$\beta \sim Normal(0, 3)$$

(6.3)

$$\sigma \sim HalfCauchy(0.5)$$

The intercept α is given a Gaussian prior. We assume that a user visits at least 1 visual element, and at most 150 elements, which leads to the log-space intercept α being from 0 to $log(150) = 5$. A probability distribution of $Normal(2.5, 1)$ has the majority of probability in that range [0, 5] (plus/minus two standard deviations from the center 2.5). Thus we specify the prior $\alpha \sim Normal(2.5, 1)$. Then we assume the effect of condition being [-20, 20], i.e., people in the hindsight and control condition may differ in element

120

visits by -20 to 20. It is a wide range centered on 0 because of our uncertainty of the effect. Thus we specify the prior for the log-space slope $\beta \sim Normal(0,3)$, given that $log(20) = 3$. Finally, the residual variation is given a Half-Cauchy prior [145], restricting the possible values to be positive. We specify the scale parameter of the Half-Cauchy distribution $\gamma = 0.5$, and the σ will generally be in the range of [0, 1.2].

We used the brms R package [146] to compute the model. The model estimated an effect of *condition*: M = 6.37 95% CrI = [-1.66, 14.57], which has a large uncertainty. However, we found that there is a 94.3% probability that the participants in the hindsight condition visit more visual elements than those in the control condition. In addition, the posterior distribution (Figure 6.2) reveals a distributional difference between the means of the two conditions.

6.3.3 Comparison and Discussion

The two analysis approaches are different in two major aspects: the assumptions made for the data and the interpretation of results.

Assumptions. In the previous analysis, the data were assumed to be normally distributed. In the secondary analysis, we were allowed to explicitly make assumptions about both the distribution of the observed data (by specifying the *likelihood function*) and the probability distributions of the mean and other parameters (by specifying the *prior distributions*). We specified that the observations had a log-normal distribution, and that the prior probability distribution of the parameter μ (log-scaled mean) is normal.

Model	LOOIC	SE	ΔLOOIC	ΔSE	Model Difference
m.lognormal	956.33	18.37	0.00	0.00	$v_i = Lognormal(\mu_i, \sigma)$
m.normal	993.80	27.73	37.46	21.14	$v_i = Normal(\mu_i, \sigma)$

Table 6.1: Model comparison with LOOIC.

If we specified the likelihood function to be a Gaussian function in the Bayesian con-

text, and used weakly-informed priors, the analysis would be viewed similar to the previous analysis [140, 147]. However, informed by domain knowledge, we knew that the current assumptions we made can better reveal the experiment design (e.g., there is an upper limit for the number of visits since there were 255 elements in total in the visualization). Table 6.1 illustrates the comparison of different types of models. We observe that the model using log-normal as likelihood function has a lower LOOIC and SE.

Result interpretation. For the previous analysis, we could only say that there is a 95% probability that the mean difference between two conditions in the *observed data* is larger than zero, being somewhere between 1 and 16 (according to the effect size). For the secondary analysis, we could say there is a 94.6% probability that the means of the two conditions are different, and that there is a 66% probability that the difference is between 2 and 7. Basically, we had the entire probability distribution to describe the difference.

6.4   Modeling the Diverse Behavior Distributions

Bayesian approach makes it possible for us to avoid dichotomous analysis by providing us with *distributional parameter estimations*. By re-analyzing the data in Section 6.3, we obtain an estimation of the difference of the average number of visits between the two groups described by a posterior distribution. In the meantime, the distributional property of Bayesian approach opens up space for us to explore a broader range of research questions. Instead of estimating the effects on the *center* difference, we could potentially investigate other types of parameters, including but not limited to *variation*. For example, Kay et al. [138] explored the *preciseness* of people's estimations of uncertainty affected by different graphical displays, by comparing the *variations* of their estimations.

We use the previous experiment dataset *SearchinVis-255Charts*. Specifically, we want to examine whether or not the **visit duration** will be affected by the usage of the search

box, and if so, what will the effect look like. We are not only interested in the impact on the *scale*, but also on the *shape* of the probability distribution of visit duration.

In the data, each observation describes a participant's visit to a visual element, including visit duration (*duration*), the participant's identification number (*participant*), name of the visited element (*element*), and a boolean factor (*searching*) indicating whether the visit happened during the participant was using searching.

6.4.1 A Multi-level Model

We establish a multi-level model, by adding an intercept for each participant (described as $\alpha_{1,participant[i]}$ and $\alpha_{2,participant[i]}$ below). One advantage of the multi-level structure is that it facilitates *shrinkage* estimation, which can overcome the unbalanced group observations (different numbers of visits per participant in our case) by drawing them towards the global mean [148]. Herein we formally describe the **model**:

$$duration_i \sim Lognormal(\mu_i, \sigma_i)$$

$$\mu_i = \alpha_1 + \alpha_{1,participant[i]} + \beta_1 \times searching_i$$

$$\sigma_i = \alpha_2 + \alpha_{2,participant[i]} + \beta_2 \times searching_i \qquad (6.4)$$

$$\alpha_{1,participant[i]} \sim Normal(0, \sigma_{1,participant})$$

$$\alpha_{2,participant[i]} \sim Normal(0, \sigma_{2,participant})$$

As shown in the first line, *duration*$_i$ is the response variable, the time spent on visiting an individual visual element. We use *Lognormal* as likelihood model as it can model distributions with a long tail [143], which is likely to be the case of people's visit duration. In the second line, we form a linear formula for μ_i, the *scale* parameter. We select *searching*$_i$ as a predictor variable. $\alpha_{1,participant[i]}$ describes the by-participant varying intercepts, meaning that μ_i varies by participant. Each of these intercepts is described by a

normal distribution $\sigma_{1,participant}$.

In the third line, we form a linear formula for σ_i, the variation residual parameter as well. σ_i is the *shape* parameter of the log-normal distribution. Similarly, we select *searching$_i$* as a predictor variable, and specify by-participant varying intercepts $\alpha_{2,participant[i]}$, each of which has a normal distribution $\sigma_{2,participant}$.

We then specify a weakly-informed prior distribution for each parameter. Weakly-informed priors would let the model fitting to capture most information from observations instead of prior beliefs.

$$\alpha_1 \sim Normal(0,10)$$
$$\beta_1 \sim Normal(0,10)$$
$$\sigma_{1,participant} \sim HalfCauchy(10)$$
$$\alpha_2 \sim Normal(0,10) \tag{6.5}$$
$$\beta_2 \sim Normal(0,10)$$
$$\sigma_{2,participant} \sim HalfCauchy(10)$$

The first three lines describe the parameters to model the *scale* of visit duration. The intercept α_1 and the slope β_1 are given a commonly-used weakly informative Gaussian prior centered on 0 [149]. The standard deviations of the by-participant varying intercepts, are given a Half-Cauchy prior [145], restricting the possible values to be positive. We use the same setting for the intercept, slope and variation parameters for σ_i.

6.4.2 Model Fitting and Comparison

We used brms R package [146] to compute the model. The trace plots show that a wide range of possible values was explored for each parameter, and well-performed convergence of the chains. In addition, we compared the model (m4) with alternative mod-

els with fewer parameters (m1, m2, and m3) using Leave-One-Out Cross-Validation (LOOCV), as shown in Table 6.2. Results show that the current model m4 has a significant decrease in LOOIC.

Model	LOOIC	SE	ΔLOOIC	ΔSE	Model Difference (in general syntax)
m4	5164.09	113.52	0.00	0.00	$duration \sim 1 + searching + 1\|participant$, $sigma \sim 1 + searching + 1\|participant$
m3	5262.28	113.04	98.19	-24.80	$duration \sim 1 + searching + 1\|participant$, $sigma \sim 1 + searching$
m2	5294.21	113.90	130.12	27.39	$duration \sim 1 + searching + 1\|participant$
m1	5523.49	114.18	359.40	43.84	$duration \sim 1 + searching$

Table 6.2: Model (duration) comparison with LOOIC.

6.4.3   Inferences

Our research question is to know whether people's visit duration would be affected by the usage of the search box.
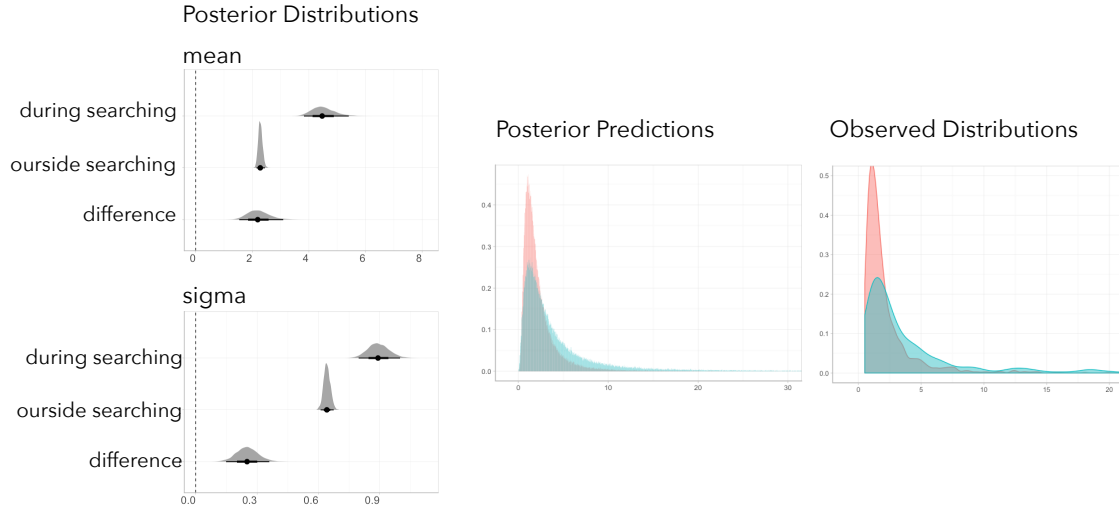
Figure 6.3: The results of the effect of searching on users' visit duration probability distribution. Each of the two diagrams on the left shows the posterior distributions of the two conditions (during search and outside searching) and their difference. The middle diagram shows the posterior predictions of the two conditions. The right diagram shows the distributions in real data.

The model estimated an effect of *searching$_i$* on the *scale* of duration distribution: M=2.22 95% CrI = [1.50, 3.03], which indicates that the user tends to spend longer time visiting a visual element when using search. The model also estimated an effect of *searching$_i$* on the *shape* of duration distribution: M = 0.25 95% CrI = [0.15, 0.36], which indicates that the probability distribution of duration tends to be right-skewed, having a heavier right tail.

The results indicate that the usage of search box changes both the *scale* and *shape* of the probability distribution of visit duration. This is further shown in the posterior predictive distributions (Figure 6.3), which is the distribution of possible unobserved values conditional on the observed values [150], that the usage of the search box shifts the duration distribution to become wider, to the right, and having a heavier tail. It indicates that when using search, a user may spend longer time visiting a visual element – the longer visit may not always happen, but the chance of longer visits increases.

6.5   Discussion

By re-examining the questions on the effect of condition on the number of visited elements, we demonstrated that Bayesian statistics (1) allow us to make reasonable assumptions on users' interaction behavior, instead of assuming normal distributions, and (2) provide rich information (posterior distributions) for us to describe the results, instead of drawing dichotomous conclusions.

By exploring the new question on the effect of using searching on visit duration, we demonstrated that (1) Bayesian multi-level modeling provides intuitive way to achieve shrinkage, which lowers the risk of over-estimation of effect size due to unbalanced group observations, and (2) it allows us to model and describe the effect from various perspectives (e.g., mean, variance, shape, etc.).

6.5.1 Potential Applications in Web Vis Interaction Analysis

Evaluate design alternatives. Researchers and practitioners base the evaluation of design alternatives mainly on dichotomous results. Online A/B testing on user interaction still focuses on p-values. Adopting the Bayesian statistics would lead to more informative behavior-driven design guidelines. Effects can be described by probability values instead of "yes-no". Creators would be able to better know how to choose from competing design choices.

Re-visualizing the predictions for creators. Real behavior data may be very noisy and it may be hard to identify underlying trends. It may be helpful to visualize posterior predictive distributions along with the real data (as shown in Figure 6.3).

6.5.2 Other Potential Benefits for Visualization Interaction Studies

Reasonable results for small-n analyses. Studying human behavior with visualizations has its own specific concerns. It is often quite necessary to collect enough data samples to detect a significant effect if it exists, but it is usually difficult to recruit enough participants to interact with the visualization tools.

The reasons may vary. Sometimes researchers want to conduct an *in-lab study* [3, 151], as the tool is not deployed online, or they want to better control the study environment. In this situation, time and other resources might be limited to conduct a large-scale study. In some other cases, participants of the study are required to have some specific expertise or characteristics as the tool is designed for a specialized population group (*e.g.*, when the target users of the novel technique are software engineers [152]). Recruiting many such participants is difficult.

Bayesian analysis may mitigate the negative impact of *small-n* and provide more reasonable results [137], from at least two aspects. First, a more precise result from the

probability distribution of effect sizes better informs the decision of further investigations, compared to a yes-no answer. Second, by setting an appropriate prior, one can effectively prevent the overspread data causing the CI bar to be too wide, or the worse case, the overestimation of effect sizes given an accidental significant result.

More accurate results for multiple comparisons. Problem occurs when we consider multiple statistical inferences simultaneously [153], *e.g.*, inferring the differences between elementary school students and high school students, *and* between male and female. Multiple comparisons exist when an experiment study involves multiple conditions or factors at once (*e.g.*, three interfaces and three levels of task complexity). Many prior works studying human behavior with visualizations involved multiple comparisons [106, 152].

However, the more comparisons there are, the more likely error will occur in the inferences made. NHST-based multiple comparisons require *corrections* when the underlying assumptions are violated (*e.g.*, the sphericity assumption), whereas the Bayesian approach to multiple comparisons, hierarchical modeling, does not need such corrections. "Bayesian hierarchical modeling diminishes false alarms by letting the data inform *shrinkage* of estimates."[147]

Knowledge accrual for (partially) replicated studies. Bayesian statistics enable quantitative knowledge accrual in scientific research [137]. If a prior study is fully or partially replicated in a new study, the quantitative knowledge from the prior study can inform the priors of the new study.

It is common to see fully or partially replicated studies on human behavior with visualizations. For example, when a novel interactive visualization technique is developed, to examine its effectiveness, researchers often evaluate the new technique against a baseline technique used in prior studies [3, 12, 152]. Sometimes different studies use the same

benchmark task [22, 154]. If adopting Bayesian techniques, researchers will be able to incorporate the knowledge in previous studies into the new study and derive more precise results.

6.5.3 Tradeoffs

Basically, it requires more statistics and domain expertise, especially when choosing the likelihood function, determining priors, select predictors, evaluate models.

Compared to Frequentists, Bayesian statistics require more expertise in *statistics*. The very basic knowledge requirements of doing Frequentist analysis include knowing what a t-test is, what a confidence interval is, etc. There are available software, *e.g.*, R and SPSS, that can automatically compute the results given a correct input: the dataset and some basic configurations. The input of Bayesian analysis include the dataset and the models. The analysis procedure is a non-trivial workflow [144] which requires several iterations to build models. If the models were incorrect, the analysis results would be wrong. People make mistakes even when performing the Frequentist analysis that is more straightforward. Kosara *et al.* [141] mentioned that one of the reasons for questionable study validity is that wrong statistical tests were selected for the analysis. The adoption of Bayesian statistics may lead to even more challenges on correctly performing the analysis. In addition to statistics, Bayesian analysis also requires domain expertise, which is not needed for Frequentist analysis. Domain experts need to involve in the analysis process to determine a prior that align with reality.

Researchers studying human behavior with visualizations have various backgrounds, *e.g.*, information visualization, human-computer interaction, and psychology, etc. They may or may not have strong expertise in statistics, and they may or may not have the domain knowledge to establish a probability model of people's analytical behavior. The benefits from Bayesian analysis mentioned above are based on that the analysis is correct.

The complex analysis process that requires statistics and domain expertise is a tradeoff of adopting Bayesian statistics.

6.6   Conclusion

We demonstrate some of the possible benefits of Bayesian statistics for examining visualization behavior, and point out the potential trade-offs through our demonstration, by re-analyzing interaction data from previous studies' open datasets regarding people's interactions with visualizations on the web [12, 13]. First, we use Bayesian statistics to re-examine the effects of interaction design choices on quantitative aspects of participants exploration sessions, and found that these analyses go beyond previously reported dichotomous results, and yield a more precise distribution of the overall effects of design choices in the studied interactive visualizations. Next, we demonstrate how these new statistical paradigms can address a new range of questions about interactive behavior. Specifically, Bayesian statistics facilitate precise and diverse representations of effects. We derive new findings on people's interactive explorations of visualizations, including the distributional effects of using search on peoples interaction duration.

Chapter 7

Towards Managing People's Interactive

Explorations of Visualizations on the Web
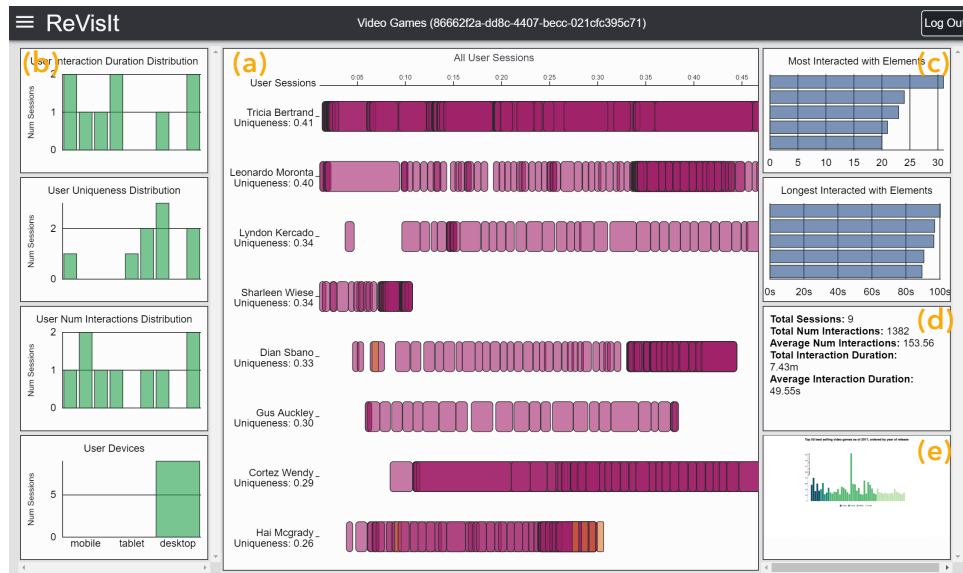
Figure 7.1: ReVisIt interactive visualization interface. (a) The timeline view of users' interaction traces. Each row represents a user session. (b) Statistical summaries of user sessions, e.g., distribution of exploration times, distribution of uniqueness values, etc. (c) Statistical summaries of visual elements, e.g., distribution of the elements' visit counts. (d) The other summaries regarding the deployed visualization, e.g., number of user sessions. (e) The overlay view of users' interactions. If an element is hovered over in the timeline view (a), a circle will display in the overlay view indicating the position of the visual element in the original visualization.

7.1 Introduction

Thanks to the development of authoring tools for data visualizations on the web (e.g., D3.js, HighCharts), more and more interactive visualizations are emerging online. The visualizations are from various kinds of creators, ranging from graphic editors of news organizations, to visualization hobbyists, who may have goals in mind during creation, such as to tell a story in the data, or to facilitate people to freely explore the dataset on their own. However, after deploying the visualization, these creators usually do not have effective means to get feedback from their audiences so as to know whether these goals are achieved. Although there are some common feedback mechanisms, including surveying and gathering comments, these processes require extra efforts from audiences, and may have selection bias, i.e., the possibly majority group of "silent audiences" are excluded.

One possible alternative is to gather and make sense of the *passive feedback* from audiences, i.e., their interaction logs. That way, feedback can be collected automatically with no extra effort from audiences. Such mechanisms already exist in the website development and deployment workflow, e.g., Google Analytics [155], that collect users' page visits and feeds them back to website creators. However, audiences' *page-level* interactions (e.g., button clicks) may not be sufficient to reveal their explorations of visualizations. In other words, it may be helpful to provide visualization creators with their audiences' *element-level* interaction logs (i.e., interactions with elements inside a visualization).

In the meantime, the visualization research focusing on people's analytical provenance also involves collecting and analyzing visualization interaction logs [6, 20, 37]. These works inspire us in how we can record and represent people's element-level interactions with visualizations. For example, Blascheck *et al.* [37] developed a visual ana-

lytics system to examine how analysts explore visual analytics applications (VA^2), where individual exploration sessions, including eye-movement, thinking-aloud and interaction data, are synced and displayed as time series. However, these visualization approaches may not be directly used to represent interactions with web visualizations, due to reasons such as the difference in scaling capacity (tens vs. possibly thousands of users), and the difference in creators' goals (domain-specific task performance vs. overall user engagement).

In this work, we explore how logging and re-visualizing people's element-level interactions with visualizations would inform the creators of web visualizations whether their goals are achieved. Specifically, we have the following two research questions:

1. What goals would web visualization creators have? Which of them can be informed by users' interactions?

2. How can we help web visualization creators learn whether their goals are achieved by logging and re-visualizing users' interactions?

To answer these questions, we take two steps. First, based on comprehensive literature review and iterative design steps, we design and implement a prototype visualization system *ReVisIt* that logs and re-visualizes people's interactions with visualizations. Second, we conduct interviews with the creators of visualizations on the web to learn about their goals and to explore how visualization systems such as *ReVisIt* would inform them about their goal achievement.

The **contributions** of this work are three-fold:

1. We propose a prototype interactive visualization *ReVisIt* that re-visualizes people's element-level interactions with visualizations on the web.

2. We establish a generalizable and scalable logger system that records people's interactions with various kinds of visualizations.

3. We conduct interviews with several visualization creators, learning their goals of

133

creation, and how re-visualizing their audiences' interactions would inform them about whether they have accomplished the goals.

7.2 Background

7.2.1 Logging Interactions in HCI

Several systems have been developed to capture users' interactions with websites and applications. One of the systems is UsaProxy [156, 157]. UsaProxy, as shown in Figure 3, sits in between the developers server and the user client, and modifies the HTML pages served to capture the user's interaction data. The data includes web interaction events for page loading and unloading, resizing, changes to focus, mouse clicks, mouse movements, scrolling, and key presses. While such logging mechanisms are automatic, i.e., the system does not require the developer to write code to track interactions, or require the client to run any specific software, the amount of data collected is very large, and more importantly, cannot be well tailored to the analyst's needs. That is to say, while collecting user interaction data is simple with UsaProxy, analyzing and visualizing it is not. The analyst would have to manually filter the data after it is collected based on what they are interested in learning. UsaProxy also does not provide visualizations of the data it collects, making it difficult for analysts to explore the data and draw insights without already having an idea of what to look for.

7.2.2 Visualizing Interactions in HCI

Various visual analytics systems and approaches have been developed in the field of general website design to analyze the collected website clickstreams.

One system that visualizes clickstream data is WebQuilt [158]. Similar to UsaProxy,

WebQuilt collects users' interaction data using a proxy-based approach, but the system goes beyond just logging the data. WebQuilt visualizes user interactions by constructing a user behavior graph. Each nodes represents a page visited by users, and each edge represents a transition between two pages. The graph visualization utilizes node size and edge color to indicate how long users spent on a page and how frequently a path was taken. Liu *et al.* [23] developed algorithms to extract sequence patterns from clickstreams. This work can be used to understand users' interaction sequences during web surfing. Zhao *et al.* [3] created a visualization called MatrixWave to compare two clickstreams datasets, and found it to scale better than commonly used Sankey diagrams. This work can be used to compare alternative website designs. These visual analytics systems and approaches showing users' explorations of websites inspire our proposed research regarding how we can effectively display users' explorations of a visualization.

With the interactive visualization approaches, analysts are able to examine users' interaction data at an aggregate level, inspect a specific group or individual users through filtering. Analysts are thus allowed to easily discover people's navigation patterns, and identify areas of interest, such as which pages most captured users attention. While these are meaningful insights, these systems only visualizes people's page-level interactions. It remain non-trivial for analysts to effectively make sense of people's interactions with some specific elements on the page, such as the usage of a slider, which might allow for a more detailed analysis of user behavior.

7.2.3 Visualizing Interactions with Visualizations

In the field of visual analytics, several systems have been developed to visualize people's interactions with visualizations.

Han et al. developed a visual analytics approach to help analysts make sense of people's explorations of visualizations [159]. They first identified requirements of the system

regarding flexibility and practicality. First, the system needs to be able to capture relevant interaction events. Next, the raw events must be categorized either automatically or manually (by analysts). Then the categorized events need to be visualized. The IntiVisor [160] system provides visualizations that can be congured to fit various types of visualizations, since all visualizations are not the same. Additionally, the system has the ability to be applied to a variety of visualization engines.

Blascheck *et al.* [37] developed a visual analytics system to examine how analysts explore visual analytics applications (VA^2). Each user's exploration data, including eye-movement, thinking-aloud and interaction data, are synced and displayed as time series. Users' exploration sessions are hierarchically clustered according to the similarity of their eye movement scanpaths. Analysts can also search for specific patterns in the eye-movement or interaction data. However, this visual analytics approach may not be directly used to address the challenge regarding user explorations of web visualizations. Because, for instance, VA^2 focuses on representing the interaction traces from individual users, while on the web, where usually thousands of people interact with a visualization, individual traces may not be sufficient to reveal the overall trends. Nevertheless, VA^2 has inspired our research in multiple ways. For example, VA^2 displays an individual user's interaction data in a 2D time-space form, with one dimension for time, and the other dimension for area of interest (AOI). In addition, VA^2 automatically clusters the users' eye-movement paths by their similarity. This inspires the design of our system regarding how we can organizing the users by their interaction data.

7.3 An Initial Design Space for Re-visualizing Interaction Data

We developed an initial design space for visualizing interactions with visualizations. We used a snowballing approach to collect relevant papers. We first defined 15 seed papers, and then used forward and backward referencing for more papers. These papers include techniques that visualize three data types: event sequences, website clickstreams and visualization interactions. Figure 7.2 shows the initial design space. We extracted several dimensions that can be used to categorize the papers: visual representations, interactivity, views and event types.

- **User Tasks:** The visualizations support various types of tasks. Two of them are inter-record and intra-record examinations [116], i.e., examining individual events and overview of events. A third task is to compare two or more individual events.

- **Event Types:** There are two types of events: point events and interval events [116]. For example, button click is a point event, and hovering is an interval event.

- **Visual Representations:** We extracted several visual representations used in these approaches, including timeline displays, icicle trees, dentograms, superimposed paths and histograms, etc.

- **Interactivity:** Most systems involve interactivity, which includes zooming, filtering and inspection.

7.4 ReVisIt: A Visualization System Prototype

In this section, we present *ReVisIt*, the visualization system prototype that we designed and implemented. We first introduce the architecture of the system, and then describe the interactive visualization interface.

Figure 7.2: An initial design space for visualizing interaction data.

The table below transcribes the rotated design-space matrix. Rows are the surveyed papers (with Year and Data Type); columns are the visualization features, grouped under **Visual Representations**, **Interactivity**, **User Tasks**, and **Event**. Cell marks are shown as in the original (`e` = event sequences, `w` = website clickstreams, `<` = visualization interactions).

Year	Paper	Data Type	Histogram	Pie Chart	Timeline / Order	Pixel Map	Heatmap	Superimposed paths	Icicle Tree	General Tree / Graph	Treemap	State/Transition Diagram	(Clustered) Matrix	Table / List (Details)	Dentogram	Visual Clusters	Transition Matrix	Geographical Display	Other	General	Zooming	Filtering	Inspection	Searching / Quering	Exploration	Annotation	Inter-Record (Overview)	Intra-Record (Individual)	Similarity / Difference	Point Events	Interval Events
1951	NN and FA [Kleene]	event sequences							e			e																			
1983	Icicle Plots [Kruskal et al.]	event sequences			e																						e				
1994	Computer Log [Eick et al.]	event sequences			e														e	e	e						e	e	e	e	
1994	Abstract events [Kunz]	event sequences			e					e																					
2005	Rainstorm [Abdullah et al.]	event sequences			e	e														e	e			e		e	e	e	e	e	e
2005	CoCo [Malik et al.]	event sequences			e													e	e	e				e			e	e			
2006	Pattern Finder [Fails et al.]	event sequences										e						e		e					e		e	e		e	e
2007	Compound Glyphs [Pearlman et al.]	event sequences							e	e				e					e	e					e	e	e			e	e
2011	LifeFlow [Wongsuphasawat et al.]	event sequences	e		e	e			e		e	e		e						e				e	e		e	e		e	e
2012	User-Driven [Tominski]	event sequences	e		e				e	e										e				e			e				
2013	EventFlow [Monroe et al.]	event sequences		e																e											
2014	DecisionFlow [Gotz et al.]	event sequences	e		e				e			e								e							e	e		e	e
2015	Cohort Analysis [Bernard et al.]	event sequences	e		e					e										e				e			e			e	
2015	(S)qulieries [Zgraggen et al.]	event sequences								e														e							
2016	Cohort [Krause et al.]	event sequences	e		e	e			e		e									e				e			e	e		e	
2016	NetworkAlerts [Cappers et al.]	event sequences	e		e					e		e							e	e				e			e			e	e
2018	EventPad [Cappers et al.]	event sequences																													
1996	[Gray et al.]	website clickstreams	w		w															w											
2003	Nav Patterns	website clickstreams											w																		
2007	Session Viewer [Lam et al.]	website clickstreams			w							w		w						w				w							
2011	[Card et al.]	website clickstreams	w	w																											
2012	[Shen et al.]	website clickstreams	w				w	w												w											
2012	Visual Cluster [Wei et al.]	website clickstreams	w				w	w	w	w										w				w							
2013	Web Video [Chorianopoulos et al.]	website clickstreams														w															
2014	[Pachidi et al.]	website clickstreams										w																			
2014	Lecture Videos [Kim et al.]	website clickstreams	w		w															w									w	w	
2015	MatrixWave [Zhao et al.]	website clickstreams	w																w	w											
2015	VisMOOC [Shi et al.]	website clickstreams							w	w										w	w			w	w		w			w	
2016	SensePath [Nguyen et al.]	website clickstreams																													
2017	Clickstreams [Liu et al.]	website clickstreams			w				w			w								w	w						w				
2008	Graphical Histories [Heer et al.]	visualization interactions			<					<		<								<											
2008	Financial [Jeong et al.]	visualization interactions			<						<																				
2009	Reasoning [Dou et al.]	visualization interactions			<						<									<	<										
2011	Investigative Analysis [Kang et al.]	visualization interactions			<			<																							
2014	[Reda et al.]	visualization interactions																													
2014	Finding Waldo [Brown et al.]	visualization interactions																													
2014	Insight [Guo et al.]	visualization interactions			<					<		<								<	<										
2016	History [Pachuilo et al.]	visualization interactions					<	<									<			<											
2016	VA2 [Blascheck et al.]	visualization interactions			<		<								<					<	<			<	<			<	<	<	<
2016	**ReVisit**	visualization interactions	<		<		<					<								<	<	<			<	<	<	<		<	<

7.4.1 System Architecture

The ReVisIt system aims to aid in the feedback mechanism of the current visualization creation process. Figure 7.3 shows the process of visualization creation, which has two parts. In the first part, the visualization creator uses data to create a visualization and deploys it on the web. In the second part, *ReVisIt* enables the visualization creator to gather passive feedback from the audience, by (1) logging the audiences' interactions with the visualization, (2) processing the raw interaction logs to automatically measure and analyze interaction behavior, and (3) visualizing the raw and processed interaction data.
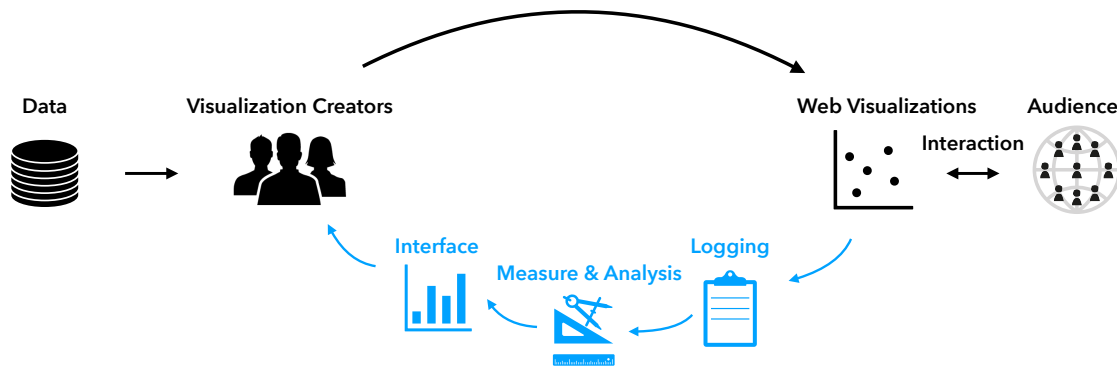
Figure 7.3: ReVisIt system architecture.ReVisIt enables the visualization creator to gather passive feedback from the audience, by (1) logging people's interactions with the visualization, (2) processing the raw interaction logs to automatically measure and analyze interaction behavior, and (3) visualizing the raw and processed interaction data.

Herein we introduce the key technical details of how the audiences' interactions will be logged and pre-processed:

Interaction Tracking: Users interactions can be tracked through the event listeners added to the SVG-based visualization elements, and stored as event streams. First, we track users' interactions with visualizations by leveraging the Scalable Vector Graphics (SVG) approaches commonly used for developing web visualizations, and the event sys-

tem in web pages. SVG approaches manipulate text-based rules which dene the resulting page, with human-readable text tags relating to lines, rectangles, text, etc. The modern web (Web 2.0) uses an event-based system to permit and respond to user interactions with elements in the page. An event refers to a click, hover, key-press, or mouse movement, etc. An event listener can be attached to an element (in this case, the SVGs) on the web page, and detect specific interaction events (e.g., clicks) with the element. To track interactions with different elements of a web page with minimal code, we utilize the HTML element tree structure that is present in each web page. Every HTML page can be considered as a text document with elements (tags) in a tree-like structure, where the root is the document element. With the HTML tree structure, we can not only attach an event listener to a specific element, but also create multiple listeners and attach them to the element's children and the entire tree branch.

Interaction Logging: After the listeners are properly attached to elements, we log the captured events using the logging server. The detected events can be organized in chronological order as an event stream. Each raw event is either directly recorded as or transformed to a point event (e.g., clicking) or an interval event (e.g., hovering) [116]. A point event has properties such as interaction *type*, the *moment* it occurred, and the *visual element* involved. An internal event has an additional property: interaction *duration*. Since we cannot directly obtain an internal event from event listeners, we form an interval event by combining two point events that mark the start and end moments, and computing the duration of the events. For example, we combine two successive interactions with a visual element – a *mouseover* event and a *mouseout* event, and form an interval event *hovering* that has duration as property. Once data are collected into an event stream, we record it using the logger server.

Interaction Processing and Serving: After the user interaction data has been logged, we apply analyses to the raw interaction events and serve them to the visualization interface. The data is saved in JSON schema, a standard format for storing and transferring data, and served using Node.js [161]. The analysis processing includes calculating metrics based on raw interaction events. These metrics are overall metrics (e.g., number of sessions), session-based metrics (e.g., uniqueness scores [162]) and element-based metrics (e.g., most-visited elements).

7.4.2   Interactive Visualization Interface

The interactive visualization interface feeds the raw and processed interaction data back to visualization creators. The interactive visualization includes several components: a central timeline view showing individual and aggregated interaction traces, an overlay view highlighting the visited elements in the original visualization, and several side visualization charts that provide summaries of user sessions and visual elements (Figure 7.1).

Linked Timeline View and Overlay View: As shown in Figure 7.1(a), the central **timeline view** aggregates individual interaction traces into timelines. Each row represents a user session. Each rectangular element represents an interaction event, either clicking (orange) or hovering (purple). The width of a hovering event represents the amount of time the user spent on the visual element. Individual user sessions are listed with a randomly-generated name for identification purposes, and are sorted based on the uniqueness metric [162], from most to least. Sorting by uniqueness allows visualization creators to identify the common interaction patterns (through the sessions with lower uniqueness values), as well as to inspect various unique interaction patterns (through the sessions with higher uniqueness values).

We implemented several interactive features that enable visualization creators to explore their audiences' interaction data. First, the creator can pan and zoom the timeline view to obtain an overview of all the interaction traces or to focus on certain parts of the timelines. Second, clicking on a randomly-generated name of the user session will expand the timeline. As shown in Figure 7.5, the expanded view shows a single user session, where each row represents a visual element, which shows an individual user's time distribution on the visualization elements. Clicking on "View All Sessions" collapses the individual session and resumes to the aggregated timeline view. Third, as shown in Figure 7.4, hovering over a rectangular element displays a tooltip with details about the interaction event, including the visual element the user interacted with, the start and end moments, and the duration of the interaction event (hovering only). Besides a tooltip, hovering over an event highlights (through increased size) the other events involving the same visual element. Finally, the visual elements involved in the interaction event are highlighted in the overlay view.

As shown in Figure 7.1(b), The **overlay ovew** includes a screenshot of the original visualization. When the creator user hovers over an interaction event in the timeline view, a circle flashes in front of the screenshot, showing where that interaction event occurred on the original visualization. The creator user can also hover over a sequence of interaction events to "replay" an exploration process and to inspect where these events occurred in the original visualization.

The linked timeline view and overlay view enable visualization creators to examine how their audiences spend time on different elements in the visualization. However, there are several limitations to the design. First, the overlay view is based on the screenshot of the original visualization, which means that it can only represent a fixed view of the visualization. The other views such as where the visual elements are rearranged have to be discarded. Second, if the creator user wants to "replay" an exploration process, he/she

has to hover over the entire timeline sequence, which takes numerous efforts. The future design of the system will improve the flexibility and lower the interaction cost [163].

(a) Timeline View   (a) Overlay View
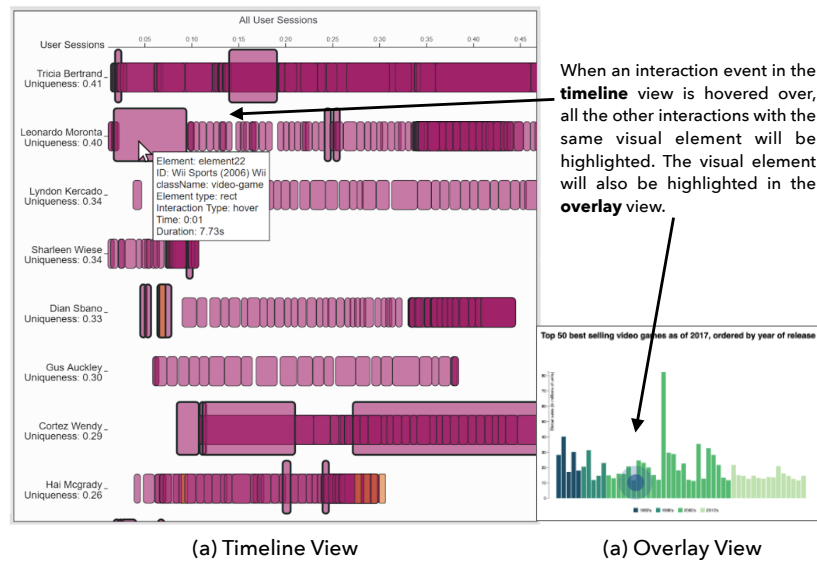
Figure 7.4: The timeline view (a) and overlay view (b) are linked together. In the overlay view, there is a screenshot of the original visualization, and the visual elements can be mapped to their original positions in the visualization.

Summary Charts / Filters: To help visualization creators better capture the overall trends of their audiences' behavior, we provide statistical summaries regarding both the user sessions and visual elements. These summary charts are interactive and can be used to filter sessions and elements in the timeline view.

To the left of the timeline view there are four **user session summary charts** that summarize different aspects of the user sessions (Figure 7.1(b)): *User Interaction Duration Distribution* is a histogram showing the exploration times of user sessions. *User Uniqueness Distribution* is a histogram showing users' uniqueness scores. *User Number of Interactions Distribution* is a histogram showing the number of interactions performed by users. *Devices Used* is a bar chart showing the number of users by device type: phone, tablet, or desktop.
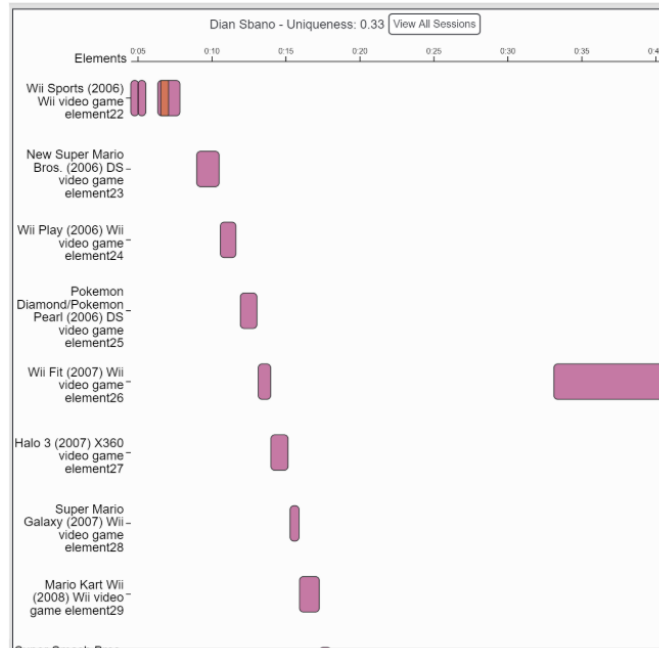
Figure 7.5: Every user session in the timeline view can be expanded or collapsed. The expanded view shows a single user session, where each row represents a visual element.

These summary charts allow visualization creators to identify trends in their audiences' interaction traces. In addition, the charts also support filtering the user sessions, i.e, when clicking a bin in a chart, one can filter the user sessions in the timeline view. The filtering feature enables visualization creators to compare the behavioral patterns from different segments of their audiences (e.g., those spent a long time exploring the visualization vs. those spent a short time).

To the right of the timeline view there are two **visual element summary charts** that summarize interactions based on the elements in the visualization (Figure 7.1(c)). *Most Visited Elements* is a bar chart showing the top five visual elements that users visited the most. *Longest Visited Elements* is a bar chart showing the top five visual elements that users spent the longest time interacting with. Hovering over a bar highlights the visual element in both the timeline and overlay views.

Figure 7.1(d) shows the additional summary statistics that visualization creators might

144

be interested in. They include the total number of users who explored the visualization, the total number of interactions performed by all users, the average number of interactions per user, and the total and average amount of time the users spend exploring the visualization.

7.5 Interviews with Visualization Creators

We interviewed four data visualization creators to learn their goals of creating visualizations and to evaluate how ReVisIt could help them measure the success of their goals. In the interviews, each creator answered questions regarding their work and then was given a chance to interact with ReVisIt and give their feedback. This section explains how we went about preparing for the interviews and the results we obtained from them.

7.5.1 Interview Methodology

Stimuli. We selected three interactive visualizations and collected usage data. The users who interacted with the visualizations were college students. As shown in Figure 7.6, the three visualizations are:

- 255*Charts*: The visualization is from the New York Times. It displays 255 small line charts showing the growth of various industries following the Great Recession [71].
- *Colleges*: It displays a bubble chart, where each bubble represents a college. Hovering over a college shows information about it, including its admission rate, annual cost, and median of earnings. The underlying dataset is College Scorecard [109].
- *Games*: It displays a bar chart of the top 50 best selling video games ordered by year of release. Hovering over a bar shows detailed information about the video game, including the exact number of units sold, year of release, genre, platform,

developer, publisher, and rating [164].
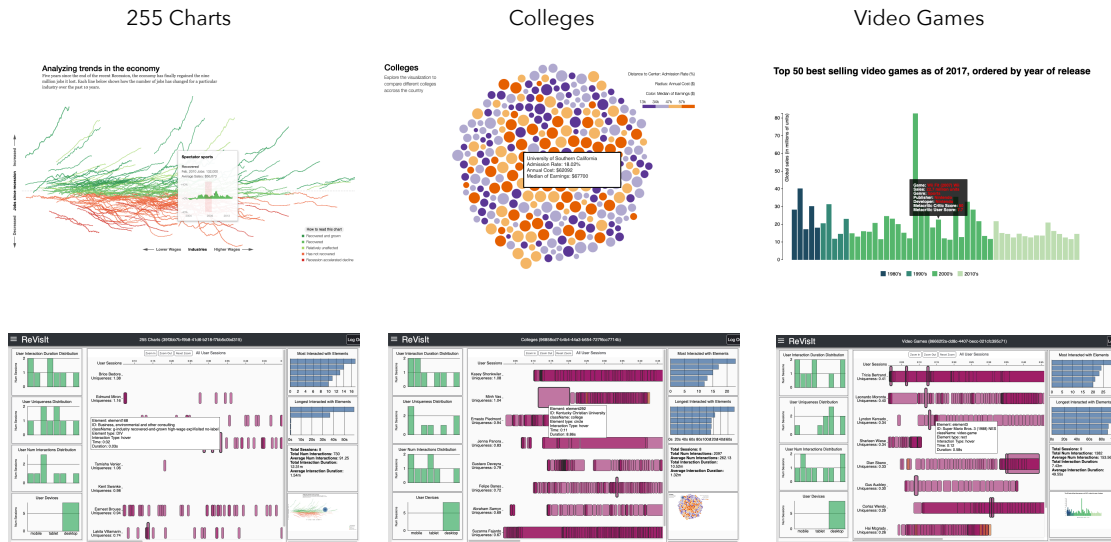
Figure 7.6: ReVisIt Stimuli. Three visualizations *255Charts*, *Colleges* and *Games*, together with the collected and re-visualized usage data.

We selected these visualizations for the following reasons. First, *255Charts* and *Colleges* have been used in previous studies [12, 13]. Second, for each visualization, there is information hidden that needs people to drill-down and discover. We used ReVisIt logger to record users' interactions with the elements inside these visualizations.

Participants. We interviewed four visualization creators (2 male 2 female). Two of them (P1 and P2) were graduate students whose research involving creating visualization systems. The other two (P3 and P4) were professionals who developed interactive data visualizations on the web.

Procedure. Herein we give an overview of the interview structure, including the questions we asked each participant:

1. *Introduction and consent*.

2. *Creators' goals and working context*: The participant was asked to provide and

describe one or more visualizations she had created. Next, the participant explained the goals she had in mind of creating the visualizations. Then we asked about how she currently measured the success of these goals.

3. *Introduction to ReVisIt*: The participant first interacted with one of the stimulus visualizations: *Colleges*, 255*Charts* or *Games*, assuming herself as the creator of the visualization, and listed the user behavior she would like to observe. Then she was introduced to *ReVisIt*, shown the features, and then allowed to explore freely.

4. *Task-oriented questions*: The participant answered a series of task-oriented questions:

    (a) Can you identify which user had the most interactions?

    (b) Compare two users. Can you describe their similarities and differences in how they interacted with the visualization?

    (c) Which user session appears to be the most unique?

    (d) When viewing all of the user sessions, how would you begin to describe how this audience is interacting with the visualization?

    (e) Which element was hovered over the most?

5. *General feedback*: The participant was asked to provide some general feedback by answering the following questions:

    (a) Through ReVisIt, were you able to answer those questions that you had before about the user interactions? What questions could not be answered?

    (b) Do you think the visualizations presented in the ReVisIt dashboard capture the information you would look for if you were analyzing the audience of a visualization?

    (c) Do you find it easy to understand how users were interacting with this visualization when you look through the information presented by ReVisIt?

    (d) What features of ReVisIt did you find helpful? Were there any that were not

helpful?

(e) What features could we add to ReVisIt to help you learn more about your visualization audience?

(f) Now that you have seen that logging user interactions for visualization like this is possible, how do you think this could help you in your projects?

7.5.2 Interview Results

Goals from Creators: Enabling Users to See Overview, to Identify Visual Patterns and to Freely Explore

What are the goals that you were aiming to achieve through creating the visualization?

In general, the visualization creators wanted their audience to have a *holistic view* of data through the visualization, to identify *visual patterns* in data, and to freely manipulate data. Some of the patterns are identified by creators, and they design the visualization in a way that their audience would quickly discover these patterns.

P3: *The sparseness of the table (part of the creator's visualization) was intentional. I want people to draw eye to the important things, sparseness.*

P4: *[Want the audience] to have a holistic view, and see patterns.*

I expect them to do filtering.

Interaction is always the key.

Current Measure of Success: Zero Data, Conversations, or Google Analytics

How do you currently measure whether your visualization was successful in achieving your goals?

Oftentimes visualization creators measure success through conversations with target users, e.g., when the visualization is created for a certain group of clients (P1, P4). They

148

may have casual conversations with surrounding people to get feedback for the design (P3). Sometimes they use Google Analytics to get a general idea of page visits (P2, P4). Some other times they do not collect feedback from their audience (P3).

Questions about Audience: Why are people coming to this? Where do they look at? How long do they spend?

What are some questions you have regarding the audience of the visualization and how they react and interpret it?

After showing the interviewees one of the stimuli (Figure 7.6), we asked the interviewees about the questions on the audience if they were the creators of the visualization. The answers mainly focused on two parts: people's *motivation* and *action*. Specifically, for motivation, P3 was interested in "Why are people coming to this?" and "How engaged are they [to the topic]?". P1 and P4 were interested in "What questions are the users trying to answer?". In terms of users' action, the creators were interested in those that may indicate where a person's attention goes, e.g. what elements people interact with (P2, P3, P4), how long they spend on exploration (P3), and in what order these interactions take place (P2, P3).

Can ReVisIt Answer These Questions?

Through ReVisIt, were you able to answer the questions you had before about regarding user interactions? What questions could not be answered?

The interviewees reported that some questions could be answered through ReVisIt. For example, ReVisIt could tell creators the visual elements people interacted with. P3 mentioned the preference of the *overlay* display that shows a person's interaction on top of the visualization. The *timeline* display was also liked by the interviewees (P1, P3). When seeing the timeline view, P3 expressed some new questions worth investigating

149

by exploring ReVisIt in-depth, e.g., "The first three to five seconds may be crucial to detect users' attention." P2 and P4 favored the feature of highlighting the records of the same visual element when hovering over one of them, or hovering over the element in the summary histograms.

There were some other questions could not be answered through ReVisIt. For example, people's interactions couldnt really tell their motivation (P3). Without additional information from users, it remained unclear who the users were (P2). Some interviewees also mentioned that ReVisIt might have a steep learning curve, whereas creators would benefit more once familiar with the system (P2, P3).

7.6 Discussion and Conclusion

Nowadays, data visualizations are being used on the web to communicate information to a broad range of audiences. While these visualizations aid in the understanding of complex data, visualization creators do not have effective means for understanding how their diverse audiences are reacting to their work, especially for understanding how a user engages with a visualization.

To bridge this gap, we propose the concept of passive evaluation, where a visualization user's interactions are logged without interfering with the users process, and then the data get processed before feeding back to the creator to visually explore. To realize this concept, we have created a prototype system ReVisIt, that captures and aggregates peoples interactions with a visualization. ReVisIt contains three key components, the logger which collects the audience interactions, the server that analyzes and serves the data, and the interface that visualizes the interaction data. To evaluate the value of ReVisIt, we conducted several interviews with data visualization creators to learn their goals of creating visualizations and how ReVisIt would help them measure success and understand

their audiences. Through these interviews, we found that ReVisIt make creators aware of individual and overall trends in their audiences interaction patterns. We also identified rising challenges regarding the future development of the feedback system, and the needs for new metrics characterizing people's visualization interactions. For example, the creators' unfilled needs of learning their audiences' motivations might be approached through characterizing the orders of people's interactions with visual elements. The successful implementation of this work would lead to visualization creators being aware of their audiences' engagement, and evaluating or improving their designs accordingly.

Chapter 8

Discussion and Conclusion

8.1 Research Contributions

Herein we summarize the research contributions of this dissertation, centering around the three high-level research questions.

8.1.1 Considering RQ1: Quantifying the Behavioral Effects of Design Alternatives

We present two interaction techniques, HindSight and text-based search, that can be widely applied to visualizations on the web, and quantify their effects on people's interactions with visualizations.

The first technique is HindSight – a representation of personal interaction history that directly encodes interaction history as a visual variable in the data [12]. We applied the HindSight technique to three existing visualizations on the web. In controlled experiments of over 400 participants, we found that HindSight designs encouraged people to visit more data and recall different insights after interaction.

The second technique is text-based search functionality in visualization [13]. We

observed the behavioral impact of integrating text-based search with five visualizations across 830 online participants, that search can alter the data people explore and how they engage with it.

Through these studies, we contribute two evidence-based design techniques that can be used as guidance when creating visualizations. In the meantime, our studies also open up larger design spaces with regards to how to integrate these techniques into existing visualizations (e.g., the visual channels used to encode interaction history, and the position of the search box) and how different design choices would impact people's interaction behavior.

8.1.2 Considering RQ2: Characterizing and Modeling Diverse Explorations

We move beyond existing methods to characterize and model diverse explorations, by deriving novel metrics and adopting alternative statistical paradigms (i.e., Bayesian statistics).

First, we propose metrics that capture novel aspects of people's open-ended explorations from their interaction traces. The metrics include *exploration-uniqueness*, measuring how unique a user's exploration is compared to others, and *exploration-pacing*, measuring temporal strategies by users. We evaluate these metrics along with four other metrics recently proposed in visualization literature by applying them to interaction data from our prior studies. The results suggest that the new metrics (1) can reveal new characteristics of people's use of visualizations, (2) can be used to evaluate statistical differences between visualization designs, and (3) are statistically independent of prior metrics used in visualization research.

Second, we re-analyze interaction data from previous studies using Bayesian statis-

tics, demonstrating some of the possible benefits of Bayesian statistics for examining visualization behavior. We first re-examined the effects of interaction design techniques (HindSight) on quantitative aspects of participants' explorations (the number of visited elements), and found that these analyses go beyond previously reported dichotomous results, and yield a more precise distribution of the overall effects. We then utilize Bayesian techniques to address a new range of questions about interaction behavior, and found that Bayesian statistics facilitate diverse representations of behavioral effects. We also derive new findings on people's interactive explorations of visualizations, including the distributional effects of using search on people's interaction duration.

8.1.3 Considering RQ3: Re-visualizing the Low- and High-level Statistics of Interaction Data

We explore how logging and re-visualizing people's interactions with visualizations would inform the creators of web visualizations whether their goals are achieved. We make the following contributions. First, based on comprehensive literature review and iterative design steps, we propose a prototype interactive visualization *ReVisIt* that re-visualizes people's element-level interactions with visualizations on the web. Second, We establish a generalizable and scalable logger system that records people's interactions with various kinds of visualizations. Finally, we conduct interviews with the creators of visualizations on the web to learn about their goals and to explore how visualization systems such as *ReVisIt* would inform them about their goal achievement. Results show that visualization creators would like to learn their audiences' intents, which could be learned by examining how they distribute time on different visual elements.

8.2 General Discussion and Future Opportunities

8.2.1 Supporting User Exploration of Visualization

Both interaction techniques investigated in this dissertation, HindSight and text-based search, aim at supporting people to navigate through visualization with many visual elements. These techniques we propose, while evidence-based, are general and basic, and they can be further advanced in the following two directions.

Examining different design choices. For each technique, we can evaluate different design options according to the initial design space we suggested, to create detailed design guidelines and uncover the underlying mechanisms of the behavioral effect. In the dissertation, we defined and evaluated the techniques on a general basis, and in the meantime we suggested an initial design space for each technique.

For example, for HindSight, there are two different ways to "trigger" encoded interaction history, i.e., the visited elements could either be "highlighted" or "faded", as shown in Figure 8.1. We used the first choice for the designs in our study, and detected behavioral effects, whereas it remains unclear whether the "faded" design would lead to the same effect. Investigating this question might advance the understanding of the "underlying mechanism" that triggers the engaging behavior.

Similarly, for text-based search, we also suggested a design space, and one dimension is "What attributes of data should be searched?" In our experiments, we examined the effect of searching by the *name* attribute, but searching can also be enabled for other attributes, e.g., the numerical attributes. It is worth examining if being able to search by number (e.g., "mpg $<$ 30") would lead to different behavioral effect compared with searching by name.

In sum, exploring different design choices for the techniques can potentially detail the
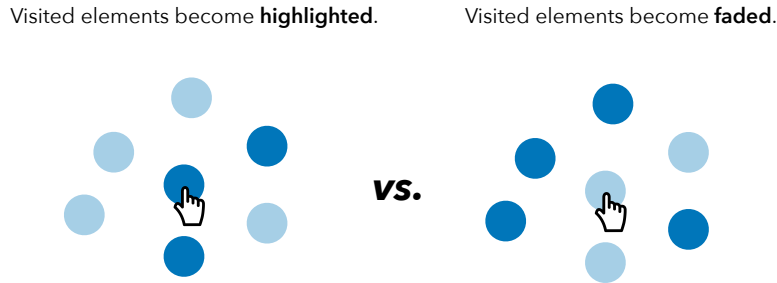
Visited elements become **highlighted**.   Visited elements become **faded**.

vs.

Figure 8.1: There are two design choices for "triggering" HindSight techniques, i.e., visited elements can either be "highlighted" or "faded".

design guidelines as well as deepen the understanding of "what factors of these techniques work indeed".

Combining the advantages of multiple interaction techniques.   We could also combine the interaction techniques: HindSight and text-based search, together with other generally-applicable techniques (e.g., social interaction techniques [9]), to further support people to navigate through and engage the data.

One example is to enable users to search from their own interaction history. People may want to revisit old visual elements, but may not remember their locations, or maybe the elements have been moved to other locations (e.g., due to sorting). It might be helpful to enable people to retrieve visited elements by name. This can be achieved by re-designing the auto-completion of the search box, i.e., marking the names of visited elements differently than others.

Another example is to provide users with social information of elements while they are searching. The social information can be elements' popularity among other users, as shown in Figure 8.2. This design is similar to the Scented Widgets from Willett et al. [26]. However, we suggest that social information could be *hidden*, i.e., showing up only during searching. It might prevent users from being overwhelmed and distracted by

other's interaction history.
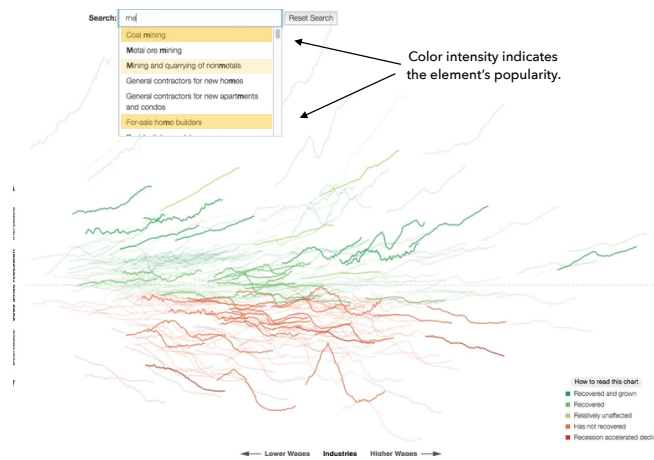
Figure 8.2: Provide users with social information of elements while they are searching.

8.2.2 Interaction Behavior as a Proxy to Insight Generation

We may utilize people's interaction logs to infer their insight generation. While knowing user behavior is informative, visualization creators may also want to know about their audiences' insights gained through exploration. As stated by North, *the ultimate purpose of visualization is insight* [165]. However, it is not trivial to gather insights from diverse online audiences, whereas collecting interaction logs is much more practical. One possible way to bridge this gap is to study the connection between people's interaction behavior and their insight generation, so that some of their insights could be inferred from specific interaction patterns.

Research in the visual analytics field has made advances in connecting analysts' interactions with their insights. Guo *et al.* [22] studied how analysts arrive at insights through the analysis process. In their study, each analyst participant was asked to analyze a dataset by interacting with a visual analytic system, and identify possible explanations

157

for an event. By examining the participants' interaction logs and insights, the authors found several correlations between interaction features and insight metrics. Dou et al. [6] also found that people's insights could partially be recovered from their interaction patterns. Yi *et al.* [166] also studied peoples' insight generation process. They discussed the factors that could affect insight generation, including those that might be revealed in users' interaction traces. *One of the most important factors to help users gain insight might be the degree of users' engagement into the dataset.* This work inspires us particularly regarding what aspects of users' interaction (engagement-related features) might be able to infer their insights.

While people's insight generation during online exploration might be different from the strict analytical tasks, we did observe some connections between interactions and insights. For example, from our HindSight study, through manual inspections, we found that participants who had different interaction patterns also appeared to have different recalled insights. For example, as shown in Figure 8.3, the user on the left visited many elements with a rapid pace, and the recalled insights were mostly global findings. The user on the right visited fewer elements with a lower pace, and their recalled insights included more specific findings. Was it due to coincidence or underlying relations between interaction and insights? Besides, where would insights likely to occur along the exploration process? The trends indicate that it is possible to predict users' insights with certain features from the interaction data.

The knowledge of how interaction data is related to users' insights in the web context would have both practical and theoretical values. From a practical perspective, the knowledge enables visualization creators to learn the ultimate impact of the visualization – on their users' possible insight generation, based on their interaction data. From a theoretical perspective, the knowledge derived might help researchers in visualization and cognitive science further understand peoples' insight generation process, especially re-
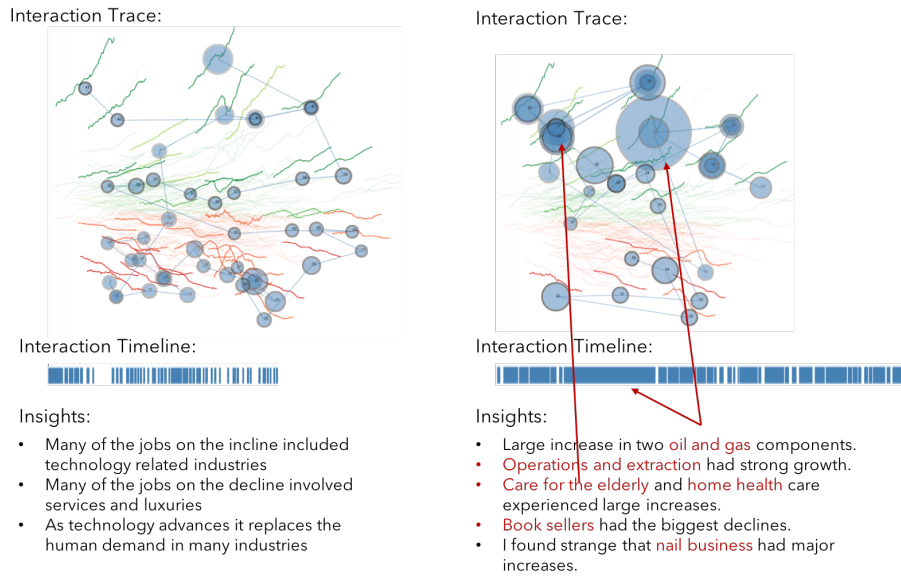
Figure 8.3: Three users' interaction traces and recalled insights.

garding user engagement. Yi *et al.* [166] discussed that engagement might be an essential factor in insight generation. The usage of Baby Name Voyager [31] also showed evidence of connections between engaging behavior and insights:

> *"...Perhaps more important is that evidence suggests many people are engaging deeply with the visualization, spending considerable time and discovering for themselves facts and insights about name trends..."*

These hypotheses may be confirmed through the future study of the relationship between interactions and insights.

8.2.3 Generalized Interaction Logging with Formalized Visualization Context

One problem that prevents interaction logging from being fully generalized is that it remains unclear how to include the visualization *context* in interaction events. Such context

could reveal the effect of the interaction on the visualization and data space. Currently, the context in an interaction event is either less descriptive or too system-specific. The following example shows a typical format of an interaction event with *less descriptive context*:

$$I = (t, a, e, d) \tag{8.1}$$

where t denotes the moment when the event occurs, a denotes the type of action (*e.g.*, hovering), e denotes the target visual element, and d denotes the duration of the interaction. This format can be used to describe events such as "clicking on element E", or "hovering on element F for 3 seconds". This format, although generalizable, will potentially cause ambiguity in interaction analysis. For instance, in Figure 8.4.a, at a low level, "clicking element A" is the same type of event as "clicking element B", but at a high level, they are different types of interactions because element A is a point in a scatterplot and element B is a button in the legend. The two interactions lead to different kinds of results in the visualization – self-highlighting versus highlighting a group of other points. Figure 8.4.b shows another example where clicking a bar in a bar chart reveals a single point in the underlying data, while clicking a bar in a *histogram* reveals multiple points at once. The two interactions are different in how they reveal the data space. Without knowing how interactions manipulate the visualization and data space, one can hardly analyze the interaction logs in-depth.

On the other hand, some interactions include context that is *system-specific*. For example, in our HindSight study, when logging people's interactions with the "Storytelling" visualization from Boy et al. [4], to distinguish interactions with the "year" and "country" elements, we added extra information to every interaction event to specify the context. Such system-specific logging strategy was also used broadly in the studies involving interactions with complex visualizations [20, 160]. However, logging system-specific
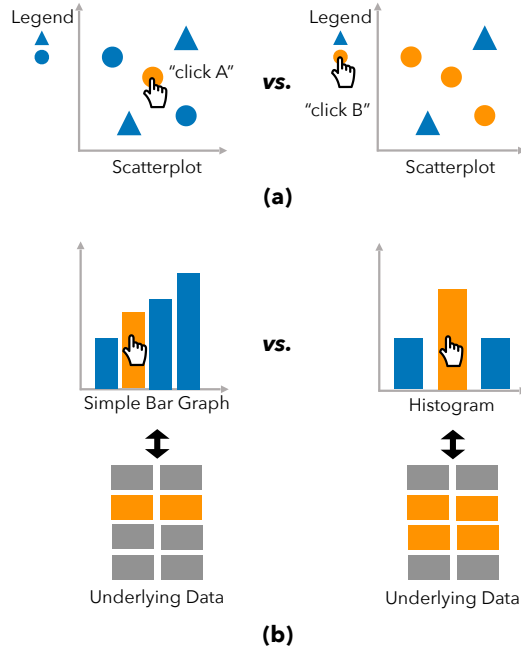
Figure 8.4: The examples illustrating the ambiguity from the low-level interaction events.

context makes the interaction analysis techniques less generalizable. In sum, we need a formalized way to record visualization context.

To address this gap, we could potentially leverage the research that aims at formalizing the visualization states, e.g., the P-Set model [167] and Vega-Lite [168]. In the future years, we might be able to log interactions with formalized context, by utilizing the visualization states developed from these frameworks:

$$I = (t, a, e, d, s) \tag{8.2}$$

where t denotes the moment when the event occurs, a denotes the type of action (*e.g.*, hovering), e denotes the target visual element, d denotes the duration of the interaction, and s) denote the state of the visualization after the interaction event.

Figure 8.5 shows an example representation based on Vega-Lite grammar. There are

four interaction events, each of which has a *visualization state*, indicating the state of the visualization after the interaction event. A state is comprised of the original visualization parameters and a *selected* parameter marking the selected portions of the data. As shown in Figure 8.5, when no interaction event occurs, the *selected* parameter is empty. When the user clicks on a point in the scatterplot where a point interaction is defined, the *selected* parameter contains the *id* of that point. When an interactive element in the legend is selected, the *selected* parameter contains the selected *origin*. When an element in a histogram is selected, the *selected* parameter contains the selected value of $X1$.
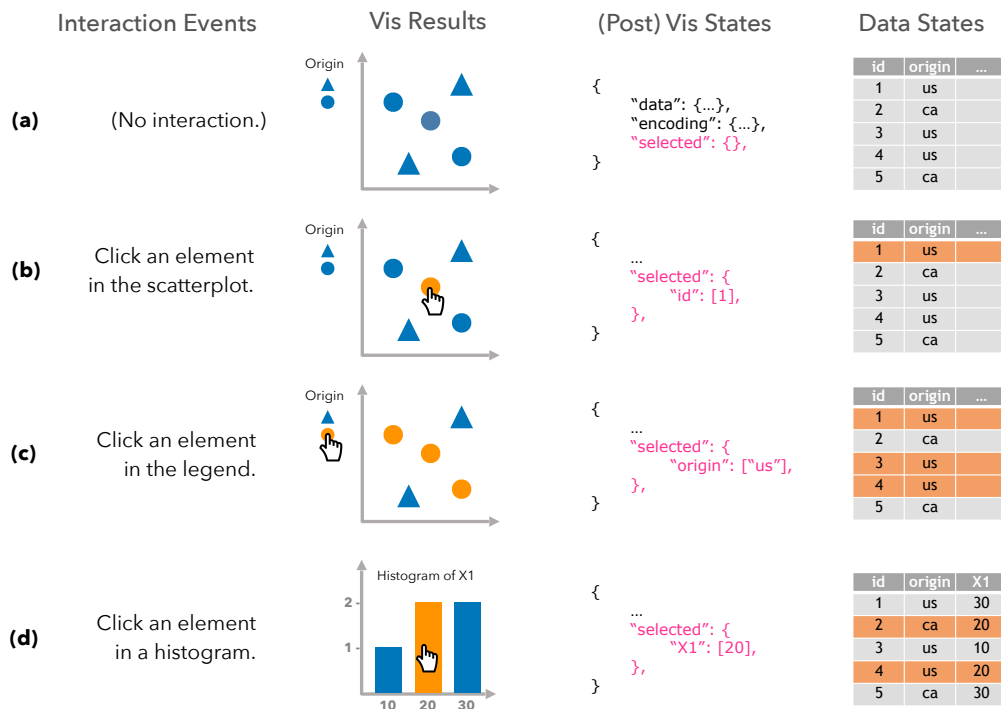

Figure 8.5: Four interaction events with visualization states, visualization results, and changes in data space.

With formally and precisely defined visualization states, we could fully utilize the interaction events as described in the following.

Derive the effects of interactions on data space. Often the analysts want to know how users examined the *underlying data* through visualization exploration. For example, Wall *et al.* [36] proposed that the coverage and distribution of data points interacted by the user could infer his/her level of bias during exploration. It would be trivial to implement the bias metrics in any visual analytics systems, if the affected data points could be automatically derived upon every interaction (as shown on the right column of Figure 8.5.bcd). Similarly, the uniqueness metric proposed in our previous work [162] could be extended to a broader range of visualizations, to measure how unique a user's exploration is given his/her coverage of underlying data points.

Derive the effects of interactions on visualization space. Upon each interaction, the visualization changes. Sometimes a single visual element is highlighted, and sometimes the whole view focus of the visualization is shifted [167]. When analyzing user's interaction traces, it may be critical to learn how the visualization changed with the interactions, especially for the study of analytic provenance [41, 46]. In some studies, researchers captured a screenshot of the visualization after every user interaction [37, 41, 167], but storing a large number of images requires lots of extra space. However, if these visualization results were reproducible from the recorded visualization states, the space cost of interaction logging would decrease. The analysts would be able to replay users' actions in the original application.

8.3   Closing Remarks

The rapid growth of interactive visualizations on the web opens a window for people to make sense of complex datasets, and it also makes it challenging for visualization creators to understand and engage their diverse audience.

In this dissertation, we explore and evaluate components of a design-feedback loop

that would enable visualization creators to better accommodate their audiences as they explore the visualizations. We quantify, model and create tools that manage peoples open-ended explorations of visualizations on the web. In particular, we quantify the effects of design alternatives on peoples interaction patterns in visualizations, develop new metrics that characterize facets of peoples exploration processes and create tools that manage and analyze an audiences interaction data for a given visualization.

Successful results of the research may have a tangible impact on the creation of visualizations on the web. With the design-feedback loop facilitated by this work, creators would be able to learn from people's behavior about how well the design can support their audiences to freely explore and to arrive at personal insights.

Bibliography

[1] Mike Bostock, Shan Carter, and Matthew Ericson. At the national conventions, the words they used. *New York Times*, 2012. URL `https://archive.nytimes.com/www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html`.

[2] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. Leading people to longer queries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3019–3022. ACM, 2013.

[3] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 259–268. ACM, 2015.

[4] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete. Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1449–1458. ACM, 2015.

[5] Alvitta Ottley, Huahai Yang, and Remco Chang. 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*, pages 3251–3254. ACM, 2015.

[6] Wenwen Dou, Dong Hyun Jeong, Felesia Stukes, William Ribarsky, Heather Richter Lipford, and Remco Chang. Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications*, 29(3), 2009.

[7] Alex Endert, Remco Chang, Chris North, and Michelle Zhou. Semantic interaction: Coupling cognition and computation through usable interactive analytics. *IEEE Computer Graphics and Applications*, 35(4):94–99, 2015.

[8] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[9] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. Voyagers and voyeurs: supporting asyn-

chronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1029–1038. ACM, 2007.

[10] Jeremy Boy, Louis Eveillard, Françoise Detienne, and Jean-Daniel Fekete. Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE transactions on visualization and computer graphics*, 22(1):639–648, 2016.

[11] Lyle Friedman, Matt Daniels, and Ilia Blinderman. Hollywoods gender divide and its effect on films. `https://pudding.cool/2017/03/bechdel/index.html`, 2016. [Online; accessed 01-Sep-2017].

[12] Mi Feng, Cheng Deng, Evan M Peck, and Lane Harrison. Hindsight: encouraging exploration through direct encoding of personal interaction history. *IEEE transactions on visualization and computer graphics*, 23(1):351–360, 2017.

[13] Mi Feng, Cheng Deng, Evan M Peck, and Lane Harrison. The effects of adding search functionality to interactive visualizations on the web. 2018.

[14] Jeremy Ashkenas and Alicia Parlapiano. How the recession shaped the economy, in 255 charts.

[15] Finviz. S p 500 map. `https://finviz.com/map.ashx/`, 2014. [Online; accessed 01-Sep-2017].

[16] Renee Lightner and Theo Francis. Inside america's boardrooms. `http://graphics.wsj.com/boards-of-directors-at-SP-500-companies/`, 2016. [Online; accessed 01-Sep-2017].

[17] Mike Barry and Brian Card. Visualizing mbta data. `http://mbtaviz.github.io/`, 2014. [Online; accessed 01-Sep-2017].

[18] Dominikus Baur. The death of interactive infographics? 2017.

[19] Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, pages 922–929, 2017.

[20] Heidi Lam, Daniel Russell, Diane Tang, and Tamara Munzner. Session viewer: Visual exploratory analysis of web session logs. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 147–154. IEEE, 2007.

[21] Omar ElTayeby and Wenwen Dou. A survey on interaction log analysis for evaluating exploratory visualizations. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 62–69. ACM, 2016.

[22] Hua Guo, Steven R Gomez, Caroline Ziemkiewicz, and David H Laidlaw. A case study using

visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1):51–60, 2016.

[23] Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, and Alan Wilson. Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):321–330, 2017.

[24] Zhicheng Liu and Jeffrey Heer. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2122–2131, 2014.

[25] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014.

[26] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6): 1129–1136, 2007.

[27] Zhicheng Liu, Bernard Kerr, Mira Dontcheva, Justin Grover, Matthew Hoffman, and Alan Wilson. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, volume 36, pages 527–538. Wiley Online Library, 2017.

[28] Zeqian Shen, Jishang Wei, Neel Sundaresan, and Kwan-Liu Ma. Visual analysis of massive web session data. In *Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on*, pages 65–72. IEEE, 2012.

[29] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497. ACM, 2001.

[30] Jeffrey Heer and Ed H Chi. Separating the swarm: categorization methods for user sessions on the web. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 243–250. ACM, 2002.

[31] Martin Wattenberg and Jesse Kriss. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, 2006.

[32] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 33–36, 2011.

[33] William A Pike, John Stasko, Remco Chang, and Theresa A O'connell. The science of interaction.

Information Visualization, 8(4):263–274, 2009.

[34] Remco Chang, Mohammad Ghoniem, Robert Kosara, William Ribarsky, Jing Yang, Evan Suma, Caroline Ziemkiewicz, Daniel Kern, and Agus Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 155–162. IEEE, 2007.

[35] Purvi Saraiya, Chris North, and Karen Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005.

[36] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[37] Tanja Blascheck, Markus John, Kuno Kurzhals, Steffen Koch, and Thomas Ertl. Va 2: a visual analytics approach for evaluating visual analytics applications. *IEEE transactions on visualization and computer graphics*, 22(1):61–70, 2016.

[38] Eric D Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics*, 22(1):31–40, 2016.

[39] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 473–482. ACM, 2012.

[40] Jeffrey Michael Heer. *Supporting asynchronous collaboration for interactive visualization*. University of California, Berkeley, 2008.

[41] Jeffrey Heer, Jock Mackinlay, Chris Stolte, and Maneesh Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE transactions on visualization and computer graphics*, 14(6), 2008.

[42] Jeffrey Heer and George Robertson. Animated transitions in statistical data graphics. *IEEE transactions on visualization and computer graphics*, 13(6):1240–1247, 2007.

[43] OECD. Oecd regional well being. `https://www.oecdregionalwellbeing.org/US25.html`, 2016. [Online; accessed 01-Sep-2017].

[44] The BBC News. The world at seven billion. `http://www.bbc.com/news/world-15391515`, 2011. [Online; accessed 01-Sep-2017].

[45] George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[46] Louis Bavoil, Steven P Callahan, Patricia J Crossno, Juliana Freire, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. Vistrails: Enabling interactive multiple-view visualizations. In *Visualization, 2005. VIS 05. IEEE*, pages 135–142. IEEE, 2005.

[47] TJ Jankun-Kelly, Kwan-Liu Ma, and Michael Gertz. A model for the visualization exploration process. In *IEEE Visualization, 2002.*, pages 323–330, 2002.

[48] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[49] Carl Gutwin. Traces: Visualizing the immediate past to support group interaction. In *Graphics interface*, pages 43–50, 2002.

[50] Carl Gutwin and Diana Anton. Does location come for free? the effects of navigation aids on location learning. *Technical Report HCI-TR-06-03: Computer Science Department, University of Saskatchewan*, 2006.

[51] Carl Gutwin and Andy Cockburn. Improving list revisitation with listmaps. In *Proceedings of the working conference on Advanced visual interfaces*, pages 396–403, 2006.

[52] M Tory A Sarvghad. Exploiting analysis history to support collaborative data analysis. *Graphics Interface*, pages 123–130, 2015.

[53] J Kriss M Wattenberg. Designing for social data analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 2006.

[54] Alan Wexelblat and Pattie Maes. Footprints: history-rich tools for information foraging. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 270–277, 1999.

[55] Ben Shneiderman. Direct manipulation: A step beyond programming languages. In *ACM SIGSOC Bulletin*, volume 17, page 1993, 1981.

[56] Peter Pirolli and Stuart Card. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–58, 1995.

[57] William C Hill, James D Hollan, Dave Wroblewski, and Tim McCandless. Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–9. ACM, 1992.

[58] Jason Alexander, Andy Cockburn, Stephen Fitchett, Carl Gutwin, and Saul Greenberg. Revisiting read wear: analysis, design, and evaluation of a footprints scrollbar. In *Proceedings of the SIGCHI*

169

Conference on Human Factors in Computing Systems, pages 1665–1674. ACM, 2009.

[59] Amy Skopik and Carl Gutwin. Improving revisitation in fisheye views with visit wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 771–780, 2005.

[60] Yedendra Babu Shrinivasan and Jarke J van Wijk. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1237–1246, 2008.

[61] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1663–1672, 2012.

[62] M Agrawala W Willett, J Heer. Scented widgets: Improving navigation cues with embedded visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 2007.

[63] Colin Ware. *Information visualization: Perception for design*. Elsevier, 2012.

[64] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212, 2010.

[65] Winter Mason and Siddharth Suri. Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[66] Geoff Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge Publishing, 2013.

[67] Ronald L Wasserstein and Nicole A Lazar. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 2016.

[68] Fanny Chevalier, Pierre Dragicevic, and Steven Franconeri. The not-so-staggering effect of staggered animated transitions on visual tracking. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2241–2250, 2014.

[69] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378–382, 1971.

[70] Jim Vallandingham. The rise and decline of ask metafilter. `http://projects.flowingdata.com/tut/linked_small_multiples_demo/`. Accessed: 2016-03-01.

[71] Jeremy Ashkenas and Alicia Parlapiano. How the recession shaped the economy, in 255 charts. `http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-`

`recession-reshaped-the-economy-in-255-charts.html`, 2014. [Online; accessed 01-Sep-2017].

[72] Jakob Nielsen and Kara Pernice. *Eyetracking web usability*. New Riders, 2010.

[73] Shneiderman Ben Shneiderman and Catherine Plaisant. Designing the user interface 4 th edition. *ed: Pearson Addison Wesley, USA*, 2005.

[74] Max L Wilson. Search user interface design. *Synthesis lectures on information concepts, retrieval, and services*, 3(3):1–143, 2011.

[75] Max L Wilson, Bill Kules, Ben Shneiderman, et al. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends® in Web Science*, 2(1):1–97, 2010.

[76] Max L Wilson, m.c. schraefel, and Ryen W White. Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology*, 60(7):1407–1422, 2009.

[77] Edward Clarkson, Krishna Desai, and James Foley. Resultmaps: Visualization for search interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1057–1064, Nov 2009.

[78] Ben Shneiderman, David Feldman, Anne Rose, and Xavier Ferré Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the 5th ACM Conference on Digital Libraries*, pages 57–66. ACM, 2000.

[79] Lucy Terry Nowell, Robert K France, and Deborah Hix. Exploring search results with envision. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Extended Abstracts*, pages 14–15. ACM, 1997.

[80] Christoph Kunz and Veit Botsch. Visual representation and contextualization of search results-list and matrix browser. In *Dublin Core Conference*, pages 229–234, 2002.

[81] Jock D Mackinlay, George G Robertson, and Stuart K Card. The perspective wall: Detail and context smoothly integrated. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 173–176. ACM, 1991.

[82] Paul André, Max L Wilson, Alistair Russell, Daniel A Smith, Alisdair Owens, et al. Continuum: Designing timelines for hierarchies, relationships and scale. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, pages 101–110. ACM, 2007.

[83] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 2012.

[84] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[85] Christopher Williamson and Ben Shneiderman. 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*, pages 338–346. ACM, 1992.

[86] Daniel A Keim and H-P Kriegel. Visdb: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 14(5):40–49, 1994.

[87] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM, 2016.

[88] Catherine Plaisant, Jesse Grosjean, and Benjamin B Bederson. Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *IEEE Symposium on Information Visualization, 2002*, pages 57–64. IEEE, 2002.

[89] Jonathan Harris. Wordcount. `http://www.wordcount.org/`, 2003. [Online; accessed 01-Sep-2017].

[90] Ben Fry. Zipdecode. `http://benfry.com/zipdecode/`, 2004. [Online; accessed 01-Sep-2017].

[91] S Blair Hedges, Joel Dudley, and Sudhir Kumar. Timetree: A public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972, 2006.

[92] Mohammad Ghoniem, Dongning Luo, Jing Yang, and William Ribarsky. Newslab: Exploratory broadcast news video analysis. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 123–130. IEEE, 2007.

[93] Christopher Collins and Sheelagh Carpendale. Vislink: Revealing relationships amongst visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1192–1199, 2007.

[94] Haeyong Chung, Seungwon Yang, Naveed Massjouni, Christopher Andrews, Rahul Kanna, and Chris North. Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114. IEEE, 2010.

[95] Anastasia Bezerianos, Pierre Dragicevic, Jean-Daniel Fekete, Juhee Bae, and Ben Watson. Ge-

neaquilts: A system for exploring large genealogies. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1073–1081, 2010.

[96] Jean-François Im, Michael J McGuffin, and Rock Leung. Gplom: The generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2606–2614, 2013.

[97] Ellen Isaacs, Kelly Damico, Shane Ahern, Eugene Bart, and Mudita Singhal. Footprints: A visual search tool that supports discovery and coverage tracking. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1793–1802, 2014.

[98] Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, 2014.

[99] Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda, and William Ribarsky. Vairoma: A visual analytics system for making sense of places, times, and events in roman history. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):210–219, 2016.

[100] Nicolas Fernandez. D3 clustergram with zooming and searching. `https://bl.ocks.org/cornhundred/1b9404efc32d6b135fab`, 2015. [Online; accessed 01-Sep-2017].

[101] Chris Elsden, Sebastian Mellor, Patrick Olivier, Pete Wheldon, David Kirk, and Rob Comber. Resviz: Politics and design issues in visualizing academic metrics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 5015–5027. ACM, 2016.

[102] Adam Pearce and Dorothy Gambrell. This chart shows who marries ceos, doctors, chefs and janitors. `https://www.bloomberg.com/graphics/2016-who-marries-whom`, 2016. [Online; accessed 01-Sep-2017].

[103] Gregor Aisch and Kevin Quealy. Stephen curry's 3-point record in context: off the charts. `https://www.nytimes.com/interactive/2016/04/16/upshot/stephen-curry-golden-state-warriors-3-pointers.html`, 2016. [Online; accessed 01-Sep-2017].

[104] Jeffrey Heer, Stuart K Card, and James A Landay. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 421–430. ACM, 2005.

[105] Tableau Software. `https://www.tableau.com`, 2003. [Online; accessed 01-Sep-2017].

[106] Jagoda Walny, Samuel Huron, Charles Perin, Tiffany Wun, Richard Pusch, and Sheelagh Carpendale.

Active reading of visualizations. *IEEE transactions on visualization and computer graphics*, 24(1): 770–780, 2018.

[107] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. Narratives in crowdsourced evaluation of visualizations: A double-edged sword? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2017.

[108] Steve Haroz, Robert Kosara, and Steven L Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200. ACM, 2015.

[109] US Department of Education. College scorecard data. `https://collegescorecard.ed.gov/data/`, 2016. [Online; accessed 01-Sep-2017].

[110] Pierre Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pages 291–330. Springer, 2016.

[111] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[112] W3 Consortium. W3 accessibilty guidelines. `https://www.w3.org/WAI/WCAG20/`, 2016. [Online; accessed 01-Sep-2017].

[113] Sharon Oviatt and Philip Cohen. Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3):45–53, 2000.

[114] Kenneth Cox, Rebecca E Grinter, Stacie L Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4(3-4):297–314, 2001.

[115] David Gotz. Soft patterns: Moving beyond explicit sequential patterns during visual analysis of longitudinal event datasets. In *Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis*, 2016.

[116] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12): 2227–2236, 2013.

[117] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1747–1756. ACM, 2011.

[118] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 38–49. ACM, 2015.

[119] Emanuel Zgraggen, Steven M Drucker, Danyel Fisher, and Robert DeLine. (s— qu) eries: Visual regular expressions for querying and exploring event sequences. 2015.

[120] Margit Pohl, Sylvia Wiltner, and Silvia Miksch. Exploring information visualization: describing different interaction patterns. In *Proceedings of the 3rd BELIV'10 Workshop: Beyond time and errors: novel evaluation methods for information visualization*, pages 16–23. ACM, 2010.

[121] Ji Soo Yi, Youn ah Kang, and John Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6): 1224–1231, 2007.

[122] Theresa A O'Connell and Yee-Yin Choong. Metrics for measuring human interaction with interactive visualizations for information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1493–1496. ACM, 2008.

[123] Kerry Rodden, Hilary Hutchinson, and Xin Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2395–2398. ACM, 2010.

[124] Ali Sarvghad, Melanie Tory, and Narges Mahyar. Visualizing dimension coverage to support exploratory analysis. *IEEE transactions on visualization and computer graphics*, 23(1):21–30, 2017.

[125] Narges Mahyar, Sung-Hee Kim, and Bum Chul Kwon. Towards a taxonomy for evaluating user engagement in information visualization. In *Workshop on Personal Visualization: Exploring Everyday Life*, volume 3, page 2, 2015.

[126] Tanja Blascheck, Steffen Koch, and Thomas Ertl. Logging interactions to learn about visual data coverage. *LIVVIL: Logging Interactive Visualizations & Visualizing Interaction Logs*, 2016.

[127] Jeffrey Heer. Capturing and analyzing the web experience. In *Proc. CHI 2002*, 2002.

[128] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[129] Jeffrey Heer and Ed H Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining*, pages 51–58, 2001.

[130] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. Acm, 2011.

[131] Angi Rösch and Harald Schmidbauer. Waveletcomp 1.1: A guided tour through the r package. *URL: http://www. hsstat. com/projects/WaveletComp/WaveletComp_guided_tour. pdf*, 2016.

[132] Enrico Bertini, Andrada Tatu, and Daniel Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[133] Birsen Donmez, Patricia E Pina, and ML Cummings. Evaluation criteria for human-automation performance metrics. In *Performance evaluation and benchmarking of intelligent systems*, pages 21–40. Springer, 2009.

[134] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1153–1160. IEEE, 2013.

[135] Alistair Croll and Benjamin Yoskovitz. *Lean analytics: Use data to build a better startup faster.* " O'Reilly Media, Inc.", 2013.

[136] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 2019.

[137] Matthew Kay, Gregory L Nelson, and Eric B Hekler. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532. ACM, 2016.

[138] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5092–5103. ACM, 2016.

[139] Matthew Kay and Jeffrey Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics*, 22(1):469–478, 2016.

[140] Ben Lambert. *A Students Guide to Bayesian Statistics*. Sage, 2018.

[141] Robert Kosara and Steve Haroz. Skipping the replication crisis in visualization: Threats to study validity and how to address them. *OSF Preprints. August*, 26, 2018.

[142] Richard D Morey, Rink Hoekstra, Jeffrey N Rouder, Michael D Lee, and Eric-Jan Wagenmakers.

The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23(1): 103–123, 2016.

[143] Dror G Feitelson. Workload modeling for performance evaluation. In *IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation*, pages 114–141. Springer, 2002.

[144] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182 (2):389–402, 2019.

[145] Nicholas G Polson, James G Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[146] Paul-Christian Bürkner et al. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[147] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[148] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2015.

[149] Ladislas Nalborczyk, Cédric Batailler, Hélène Lœvenbruck, Anne Vilain, and Paul-Christian Bürkner. An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research,(in press)*.

[150] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[151] Yunhai Wang, Yanyan Wang, Haifeng Zhang, Yinqi Sun, Chi-Wing Fu, Michael Sedlmair, Baoquan Chen, and Oliver Deussen. Structure-aware fisheye views for efficient large graph exploration. *IEEE transactions on visualization and computer graphics*, 25(1):566–575, 2019.

[152] Kirsten Risden, Mary P Czerwinski, Tamara Munzner, and Daniel B Cook. An initial examination of ease of use for 2d and 3d information visualizations of web content. *International Journal of Human-Computer Studies*, 53(5):695–714, 2000.

[153] G Rupert Jr et al. *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

[154] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE transactions on visualization and computer graphics*, 24(1):340–350, 2018.

[155] Beatriz Plaza. Google analytics for measuring website performance. *Tourism Management*, 32(3): 477–481, 2011.

[156] Richard Atterer and Albrecht Schmidt. Tracking the interaction of users with ajax applications for usability testing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1347–1350. ACM, 2007.

[157] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proceedings of the 15th international conference on World Wide Web*, pages 203–212. ACM, 2006.

[158] Jason I Hong, Jeffrey Heer, Sarah Waterson, and James A Landay. Webquilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Information Systems*, 19(3):263–285, 2001.

[159] Yi Han, Gregory D Abowd, and John Stasko. Requirements for visual interaction analysis systems. In *IEEE VIS 2016 Workshop on Logging Interactive Visualizations & Visualizing Interaction Logs,(Baltimore, MD, USA)*, 2016.

[160] Yi Han, Gregory D Abowd, and John Stasko. Flexible organization, exploration, and analysis of visualization application interaction events using visual analytics. In *IEEE VIS 2016 Workshop on The Event Event: Temporal & Sequential Event Analysis,(Baltimore, MD, USA)*, 2016.

[161] Stefan Tilkov and Steve Vinoski. Node. js: Using javascript to build high-performance network programs. *IEEE Internet Computing*, 14(6):80–83, 2010.

[162] Mi Feng, Evan Peck, and Lane Harrison. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics*, 2018.

[163] Heidi Lam. A framework of interaction costs in information visualization. *IEEE transactions on visualization and computer graphics*, 14(6), 2008.

[164] Rush Kirubi. Video game sales with ratings. `https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings`, 2016. [Online; accessed 28-Feb-2019].

[165] Chris North. Toward measuring visualization insight. *IEEE computer graphics and applications*, 26 (3):6–9, 2006.

[166] Ji Soo Yi, Youn-ah Kang, John T Stasko, and Julie A Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization,*

page 4. ACM, 2008.

[167] TJ Jankun-Kelly, Kwan-Liu Ma, and Michael Gertz. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):357–369, 2007.

[168] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23 (1):341–350, 2017.