Towards an Integrated Solution for Intelligent Visual Data Discovery

Doris Jung-Lin Lee
University of California, Berkeley
Berkeley, CA 94709, USA
dorislee@berkeley.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3375035

Abstract

Visual data exploration enables users to identify trends and patterns, generate and verify hypotheses, and detect outliers and anomalies. However, the growing scale and complexity of data exploration present a barrier to discovering useful, actionable insights from data. Often, users may not know which visualizations would lead to desirable insights, resulting in wasted efforts in manual searching. Users can also be overwhelmed or lose track of where to look across a large potential space of attributes and filter combinations, leading to missed insights or even potentially erroneous conclusions. My thesis research involves designing systems that democratize data science by providing automated guidance to analysts in visual data exploration. My current work includes accelerating manual data exploration tasks and supporting user exploration in analytical workflows. Finally, I will describe future work in developing a high-level language for addressing analytical inquiries.

Author Keywords

visual analytics; data science tools; recommendation

Introduction

Context and Motivation: From healthcare to business to education, data-driven insights play an increasing role in the way we understand the world and make decisions. However, despite their significance, the task of discovering relevant insights from data remains out-of-reach to most end-users without programming or data expertise. In particular, during data exploration, analysts are often overwhelmed by the large number of design choices that they have to make in order to find relevant insights, such as what attributes and values to visualize or which data instance yields interesting trends and patterns. The current ad-hoc exploration process often involves manually searching through a large space of potential visualizations, without a clear notion of where the visualization instance that would lead to desirable insights may lie.

Research Objective: The goal of my research is to empower users from diverse backgrounds and levels of expertise with tools that support their analytical inquiries, in order to enable the discovery of valuable insights. My thesis statement is that *providing computational assistance to users in the process of data analysis can lead to more effective and higher-quality exploration sessions*.

Current Contribution: My current research involves designing systems that provide automated guidance to analysts in visual data exploration. My existing research on visualization recommendation systems (ZENVISAGE [7] and VISPILOT [6]) aims to accelerate users to relevant insights, avoiding the tedious and manual parts of the data exploration process. My subsequent work on FRONTIER investigated how to better design organized collections of related visualization recommendations to support user's analytical workflow. A more detailed outline of this research agenda can be found in [8].

Next Steps and Expected Contribution: The proposed next steps in my dissertation is designing a language that bridges the gap between users' high-level analytical intent and low-level system actions for data science. My dissertation contributes towards a vision for an intelligent visual data exploration assistant that anticipates user intent, proactively seeks opportunities to accelerate users towards insights, and offers feedback and guidance based on user's analytical needs.

Background: I am a fourth-year PhD student at the School of Information at UC Berkeley, working with Professor Aditya Parameswaran. My research investigates how to design visual analytic systems that support users in the process of data exploration. My current contributions include the research and design of three different visualization querying and recommendation systems (published at IUI, VAST and one under submission). I am currently in the process of iterating on my thesis direction and expect to finish in 2021.

Related Work

My dissertation draws from related work in HCI, databases, and visualization research. We note that this section covers only a limited subset of the existing research landscape relevant to my thesis work, a more comprehensive review of the visualization recommendation literature can be found in [10, 12].

Most existing visualization specification tools require users to manually specify the exact data content and visual encoding of what to visualize, either through direct manipulation (Tableau, Excel) or programming (ggplot, D3, Vega, and matplotlib). Yet, the task of finding useful insights from a dataset using individually-generated charts from these manual tools can be tedious and overwhelming for analysts without visualization design expertise [7, 13, 4]. To

this end, many research have focussed on recommending interesting insights from data through either automated or mixed-initiative approaches [5, 7, 13, 4, 1, 11, 2, 6].

Automated visualization recommendation systems take in minimal inputs from the users and perform a single predefined analytical task by selecting interesting visualizations ranked based on some statistical data property. For example, Anand and Talbot [1] used randomized permutation tests to automatically select categorical partitioning variables that results in the most interesting conditional data patterns. Vartak et al. [11] used a deviation-based measure to find trends that differ the most from a reference dataset.

Recently, there has been a shift towards mixed-initiative recommendation systems that combine manual specification with recommendations [4, 5, 13, 2]. These systems takes in partially specified data content and visual encoding and suggest visualizations that are both relevant to their selection and interesting. For instance, both Voyager [13] and DIVE [4] allow users to select data attributes of interest. Voyager suggests visualizations based on iterating through possible attributes or encodings via the notion of wildcards, while DIVE creates groups of visualizations that cover some subset of the user-specified fields.

Research Progress

My current work has been contributing to a vision towards a *visual discovery assistant* [8] that enables users to expressively convey their high-level discovery goals, while the system automatically searches through visualization collections to suggest useful insights. I employ a mix of design studies, system prototyping, and observational user studies to identify design principles for building such systems. This research includes designing systems that help users accelerate a single visual analytical task, such as pattern searching [7] and drill-down analysis [6], and support effective exploration in an analytical workflow.

Accelerating tasks through search & recommendation
Line charts are one of the most commonly used chart types for illustrating complex data processes. However, searching for line charts corresponding to a desired pattern is often a manual and error-prone process, requiring the analyst to examine large numbers of visualizations. Visual query systems (VQSs) have been developed to enable analysts to interactively search for line charts with desired visual patterns. However, despite decades of prior work in this area, these systems have not been adopted in practice. In our work, we employed a user-centered design process to bridge the gap between research and practice on VQSs.

To identify challenges associated with how users make use of such systems, we worked with scientists from material science, astronomy, and genetics, in a year-long participatory design process to develop an integrative end-to-end visual query system, ZENVISAGE (Figure 1). By studying how various visual querying capabilities are used in practice, we identified three sensemaking processes and developed a taxonomy of capabilities that support each process. Our evaluation study indicated that sketching a pattern for querying is often not as useful as prior work suggests, partly because participants often did not have a clear idea of what patterns to sketch and search for. Our study led to design guidelines for visual query systems, highlighting need for querying mechanisms beyond sketching and pointing towards a future VQS that seamlessly integrates all three sensemaking processes.

One common challenge that emerged from working with real-world analysts was that, oftentimes, users struggle to figure out where to start their analysis or even what to search for. Our subsequent work examines the challenges

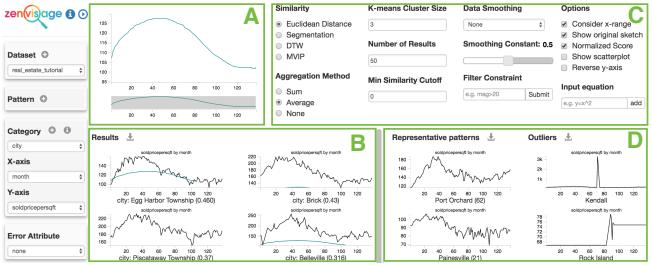

Figure 1: ZENVISAGE allows users to specify a pattern of interest through either sketching, inputting an equation, or dragging and dropping an existing visualization to the query canvas (A). The system performs shape-matching between the queried pattern and other possible visualizations and returns a ranked list of visualizations that are most similar to the queried pattern (B). Users can adjust various system-level settings (C) or browse through the recommended list of common patterns and outliers in their dataset (D).

that users face when examining data subsets by progressively adding filters, in a process commonly known as *drill-down*. As the number of data subsets grows exponentially with the number of attributes and values in the dataset, manually generating and examining each visualization in this space presents a major bottleneck to exploration. In particular, we identify a phenomenon known as the *drill-down fallacy*, where users can erroneously attribute an over-specific cause to an effect due to missed comparisons against relevant data subsets.

To prevent users from falling prey to the drill-down fallacy and guide users towards key insights in the dataset, we developed a system, VISPILOT, that automatically traverses through the space of all possible data subsets to recommend the set of most informative and interesting data subsets to the users. Compared to two baseline conditions that simulated conventional approaches to multidimensional data analysis, our user study showed that VISPILOT is more interpretable and helped users contextualize the recommended visualizations, leading to higher task performance.

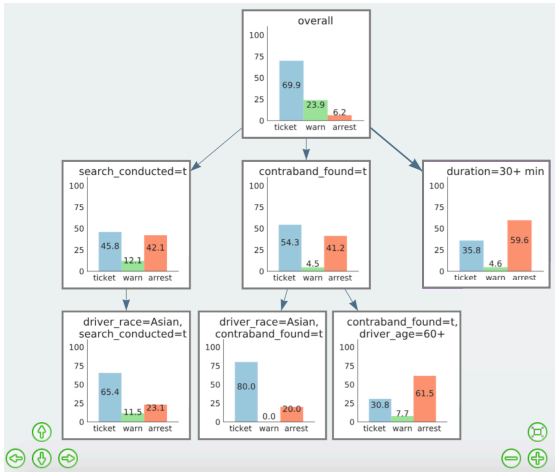

Figure 2: VISPILOT automatically selects a small set of informative and interesting visualizations to convey key distributions within a dataset. Starting from the overall distribution (root node), the selected visualizations are displayed as a network of connected visualizations with each edge representing a drill-down from the parent population to provide users with the appropriate context for comparison.

Supporting recommendations in analytical workflows

While ZENVISAGE and VISPILOT help analysts accelerate a single class of visual exploration task, real-world user inquiries often consist of a diverse range of analytical tasks, each requiring different types of visualization recommendation and guidance. To help users make sense of a heterogeneous collection of recommendations, we set out to understand the role of *analytical actions* (organized collec-

tion of related visualization recommendations) in analytical workflows. Our goal was to understand the efficacy of actions, how these actions influence users' exploration strategies, and what types of actions are most useful for different tasks and data characteristics.

By synthesizing visualization recommendations and online analytical processing (OLAP) literature, we develop a taxonomy of common action categories in visual analytics. We develop a system, FRONTIER, that implements ten action categories as a design probe to explore their usage and effectiveness. FRONTIER enables users to navigate across the visualization space by browsing and selecting visualizations organized across different action categories as potential "next steps" in their analysis (Figure 3). We conducted a mixed-methods user study comparing FRONTIER with an uncategorized baseline condition to understand how users employ visualization recommendations with and without action-based categorizations.

Our study shows that actions improve both the efficiency and utility of exploratory analysis by helping users establish a mental framework to reason about recommendation results. We found that participants preferred action categories that adds an attribute or displayed data subpopulations, since these actions facillitated rapid comparisons across the visualizations. Participants exhibited diverse, unexpected workflows that leveraged and intermixed action-based recommendations with manual control. Our study findings point to design principles of individual action categories, as well as design implications towards personalized and adaptive recommendations.

Future Directions

As part of future work, I am designing a high-level language for data science to support analytical inquiries. The motivation of this work comes from the limitations that I observed in visualization recommendation systems in its lack of flexibility and support for the diversity of tasks present in real-world workflows. Given the myriad of analytical tasks, there is a need for a consolidated, extensible language that enables users to ask questions about their data, such as searching for outliers, explaining anomalies, comparing across visualization, or getting an overview summary of the dataset. The goal of this research is two-folds: 1) the development of a high-level language bridges the current gulf of execution [9] between **how users think about data** (in terms of natural-language inquiries involving domain concepts) and **how data science is actually performed** (through a series low-level operations involving data cleaning, statistics, visualization), and 2) as Heer [3] describes a domain-specific language "*provides a shared medium in which both people and machines reason about and formulate actions*". Such a high-level language can serve as an intermediate layer between end-users and intelligent ap-

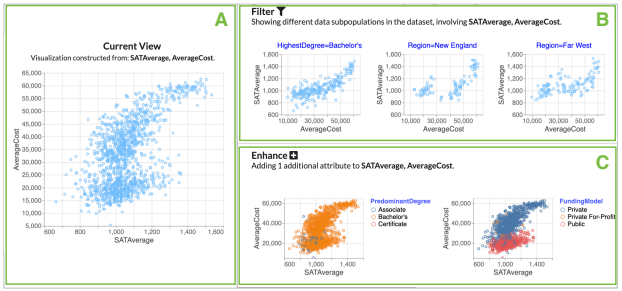

Figure 3: FRONTIER displays organized collections of visualization recommendations as actions. The Current View (A) displays the visualization constructed based on the attributes and values that the user have selected. The `Filter` action (B) displays how the Current View changes across different data subsets and the `Enhance` action (C) displays the Current View with an additional attribute.

plications for assisting users in the process of visual data discovery, such as in natural-language conversational interfaces. I hope that this work will enable novel user experiences that empower, engage, and educate a diverse set of end-users to interact with data—unlocking unbounded potential for bringing effective data-driven decision-making to the masses.

Acknowledgment

I would like to thank my collaborators and colleagues from University of Illinois, Tableau Research, and UC Berkeley, and my advisor Aditya Parameswaran. My PhD work is supported by funds from the National Science Foundation and Toyota Research Institute.

REFERENCES

[1] Anushka Anand and Justin Talbot. 2015. Automatic Selection of Partitioning Variables for Small Multiple Displays. 2626, c (2015). DOI: http://dx.doi.org/10.1109/TVCG.2015.2467323

[2] Google. 2015. Explore in Google Sheets. (2015). https://www.youtube.com/watch?v=9TiXR5wwqPs

[3] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850. DOI: http://dx.doi.org/10.1073/pnas.1807184115

[4] Kevin Hu, Diana Orghian, and César Hidalgo. 2018. DIVE: A Mixed-Initiative System Supporting Integrated Data Exploration Workflows. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 5.

[5] Alicia Key, Bill Howe, Daniel Perry, and Cecilia Aragon. 2012. VizDeck. *Proceedings of the 2012 international conference on Management of Data - SIGMOD '12* (2012), 681. DOI: http://dx.doi.org/10.1145/2213836.2213931

[6] Doris Jung-Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya Parameswaran. 2019a. Avoiding the Drill-Down Fallacy with Storyboard: Assisted and Accelerated Data Exploration. *ACM 24th International Conference on Intelligent User Interfaces (IUI '19)* (2019).

[7] Doris Jung-Lin Lee, John Lee, Tarique Siddiqui, Jaewoo Kim, Karrie Karahalios, and Aditya Parameswaran. 2019b. You can't always sketch what you want: Understanding Sensemaking in Visual Query Systems. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2019). DOI: http://dx.doi.org/10.1109/TVCG.2019.2934666

[8] Doris Jung-Lin Lee and Aditya Parameswaran. 2018. The Case for a Visual Discovery Assistant: A Holistic Solution for Accelerating Visual Data Exploration. *IEEE Bulletin of Technical Committee on Data Engineering* (2018).

[9] Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[10] Manasi Vartak, Silu Huang, Tarique Siddiqui, Samuel Madden, and Aditya Parameswaran. 2017. Towards Visualization Recommendation Systems. *ACM SIGMOD Record* 45, 4 (2017), 34–39. DOI: http://dx.doi.org/10.1145/3092931.3092937

[11] Manasi Vartak, Samuel Madden, and Aditya N Parmeswaran. 2015. SEEDB : Supporting Visual Analytics with Data-Driven Recommendations. (2015).

[12] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Towards a general-purpose query language for visualization recommendation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 4.

[13] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2017. Voyager 2 : Augmenting Visual Analysis with Partial View Specifications. (2017). DOI: http://dx.doi.org/10.1145/3025453.3025768