VizRec 2.0: Towards a holistic workflow for visual data exploration

Doris Jung-Lin Lee, Aditya Parameswaran
{jlee782,adityagp}@illinois.edu
University of Illinois, Urbana-Champaign

Abstract

test abstract

1 Introduction

Visualization —¿ insights.

1. Visual analysis help discover insights, etc.

2. supporting cycle of visual data exploration.

3. In this paper, we introduce various challenges in the —- visual analysis, and discuss —- solutions —.

4. Challenge #1: Precise search

5. finding the right vis and relevant data is hard,

6. we discuss *zenvisage* as one use case and solution in this space (Section 2)

7. Challenge #2: Need for in-the-loop support. how do people query visually?

8. During our PD + others, hypothesis + in the loop considerations is important. We discuss relevant work and our findings in Section 3.

9. Need to formulate complex expressive queries, language is good, but need interface (e..g visual primitives) to support this. (Section 4)

10. top-down, bottom-up –¿ point to need for bottom-up recommendations (Section 5)

11. Challenge #3: Vague and intelligent search

12. Users might not always have something they want to start with, supporting vague querying (Section 4)

13. Challenge #4: Cold-start recommendation for data understanding. One aspect of this is to gain understanding of dataset, (e.g. representative and outliers in *zenvisage*, bottom up exploration)we discuss STORYBOARD as one example solution addressing this problem (Section 5).

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Precise Visual Querying

Visual analysis is valuable often reveal important ——. However, there are many ——.

2.1 Motivating Example

Astronomers from the The Dark Energy Survey (DES)[**?**] are interested in finding anomalous time series to discover astrophysical transients (objects whose brightness changes dramatically as a function of time), such as supernova explosions or quasars. When trying to find celestial objects corresponding to supernovae, which have a specific pattern of brightness over time, scientists need to individually inspect the visualizations of each object until they find ones that match the pattern. With more than 400 million objects in their catalog, each having their own set of time series brightness measurement, the process of manually exploring a large number of visualizations is not only error-prone, but also overwhelming for scientists who do not have extensive knowledge about their dataset.

2.2 Challenges

The astronomy use case highlights a common challenge in exploratory data analysis (EDA). There is often a large space of possible visualizations that could be generated from a given dataset and manual search through this large collection is inefficient.

There has been many related work in this ——- varying different space of possible visualizations, including visual encodings——data facets, . We will focus on

Data is agnostic to the user, intention —, by building tools—, Section 2 to 4 have focussed on extracting what user want from data. bridging together what user want from data, what data has to offer, supporting interactive discourse between the two.

Theres a large space of possibilities, manual search is tedious. Either using one-size-fits-all statistics, templates, heuristics as a solution or problem only applicable to a subset of analytic tasks[**?**, **?**]. Propose VQS as a solution[**?**].

2.2.1 Usage Scenario

2.3 Effortless Data Exploration with *zenvisage*

- ZQL offers a way to iterate over collections of visualizations[**?**]. Iterate over collections of visualizations [**?**]

-

2.4 Challenges Ahead

The goal here is to help novice submit precise queries without SQL background, easy to use interface. Our study found that VQS does more than just this, but still not enough.

- Precise Search Fail to understand intricacies of user need/intent, need more expressivity/flexibility for querying.

- No perfect training workload, real-world data + task is noisy and complex.

- towards more holistic model for insight discovery

3 Hypothesis Formation and the cycle of visual analysis

3.1 Visual Querying in the framework of Data Sensemaking

In developing *zenvisage*, we collaborated with with scientists in astronomy, genetics, and material science in a year-long participatory design process [**?**]. In particular, we study how various features impacts rapid visual querying and analysis, and how a visual query system (VQS) like *zenvisage* fit into a scientists analysis workflow. Our findings offered design guidelines for improving the usability and adoption of next-generation VQSs. More importantly, in this paper, we focus on applying our findings on visual querying to the context of supporting the full cycle of visual data exploration. We highlight two of the key findings related to this goal below: **Top-down and Bottom-up Querying Modalities** We employed Pirolli and Card's [**?**] information foraging framework for domain-experts to contextualize our study results. Pirolli and Card's notional model distinguishes between information processing tasks that are *top-down* (from theory to data) and *bottom-up* (from data to theory). In the context of visual querying, users employ top-down approaches by starting with a hypothesis on what patterns to look for and express it through sketching or inputting an equation (Figure 4.2b,d). On the other hand, bottom-up approaches originate from the data (or equivalently, the visualization). For example, the user may drag and drop a visualization of interest in the dataset as the input query or upload a visualization from an external dataset (Figure 4.2a,c).
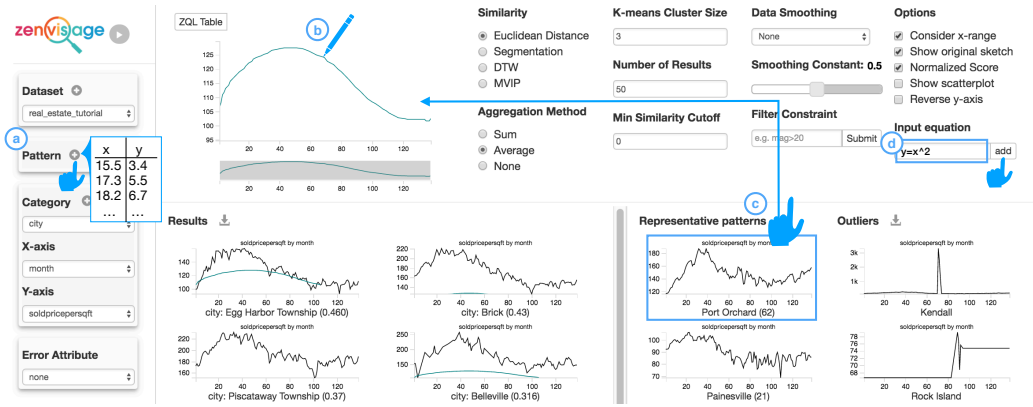

Figure 1: *zenvisage* offers a variety of querying modalities, including: a) uploading a sample pattern from an external dataset as a query, b) sketching a query pattern, c) dragging-and-dropping an existing pattern from the dataset, and d) inputting an equation as a query.

Our interactions with the scientists showed that different modalities for inputting a query can be useful for different problem contexts. To our surprise, despite the prevalence of sketch-to-query systems in literature, only two out of our nine users had a practical usage for querying by sketching. Overall, we found that bottom-up querying via drag-and-drop was more intuitive and more commonly used than top-down querying methods, such as sketching or input equations.

The main reason why participants did not find sketching useful was that they often do not start their analysis with a pattern in mind. Later, their intuition about what to query is derived from other visualizations that they see in the VQS, in which case it made more sense to query using those visualizations as examples directly (via drag-and-drop).

Both of these use cases suggest that while sketching is an useful analogy for people to express their queries, *the existing ad-hoc, sketch-only model for visualization querying is insufficient without data examples that can help analysts jumpstart their exploration*. We suspect that this is why existing sketch-to-query systems are not commonly adopted in practice. This result, however, points to a potential need for high-level query modification

interfaces that enable more precise visualization query specification or couple sketching with examples as a starting point in future VQSs.

Despite functional fitting being common in scientific data analysis, Figure **??** shows that querying by equation is also unpopular for similar reasons. In addition, the visualizations for both astronomy and biology exhibit complex processes that could not be written down as an equation analytically. However, even when analytical relationships do exist (in the case of material science), it is challenging to formulate functional forms in an prescriptive, ad-hoc manner.

In contrast, when we looked at bottom-up querying approaches, drag-and-drop was the most frequently used querying mechanism. Examples of practical uses of drag-and-drop includes inspecting the top-most similar visualizations that lie in a cluster and finding visualizations that are similar to an object of interest that exhibits a desired pattern. Likewise, many participants envisioned use cases for pattern loading. The ability to load in data patterns as a query would enable users to compare visualizations between different experiments, species, or surveys, query with known patterns from an external reference catalog (e.g. important genes of interest, objects labeled as supernovae), or verify the results of a simulation or downstream analysis by finding similar patterns in their existing dataset. In addition, users can also specify a more precise query that captures the essential shape features of a desired pattern (e.g. amplitude, width of peak), that cannot be precisely drawn as a sketched pattern.

While the usage of each querying feature may vary from one participant to the next, generally, drag-and-drop and pattern upload are considered bottom-up approaches that go from data to theory by enabling users to query via examples of known visualizations. For top-down approaches such as query by sketching and input equations, the user starts with an intuition about how their desired patterns should look like based on theory, then queries based on that. Our results indicate that *bottom-up querying approaches are preferred over top-down when the users have no desired patterns in mind*, which is commonly the case for exploratory data analysis.

Our *zenvisage* work ———- Bottom up and top down querying in VQS facilitates rapid insight discovery. More importantly pointed towards a need for vague querying. Give some examples of vague querying. **Construction of multi-step queries**

3.2 Supporting cycle of visual data exploration

Our participatory design findings points towards future —— in supporting a cycle of —. —-advocate —cycle of visual analysis.

- Essential ingredient in facilitating intelligent vague querying and exploration.

- This is a human process ([**?**, **?**])

- Iterative Hypothesis Exploration/Refinement : argue that the following properties is important to sustain this cycle of visual analysis

3.3 Towards 3Is of rapid hypothesis generation support

Given our observations from the participatory design study, we distill several desiderata for the next generation VQSs. Towards 3Is Interactive, Iterative, Informative (Give examples from the ZV-TVCG paper) **Interactive flow**: (how natural is it to move between analysis steps, facilitate fluid analysis and not get "stuck") : interactivity, feedback (latter is quite unexplored), and recommendation, expressivity (how easy is it to express what to do via interactions) and diversity of actions that could be performed. **Iterative**: query refinement, dialogue (not a one-shot query) Joining the flow: Section 4 focuses on the first two items . **Informative**: not just task-based interestingness but more explanation-based (causality, introduce distribution awareness notion in viz-sum), focussed on data understanding, which we will discuss in Section 5
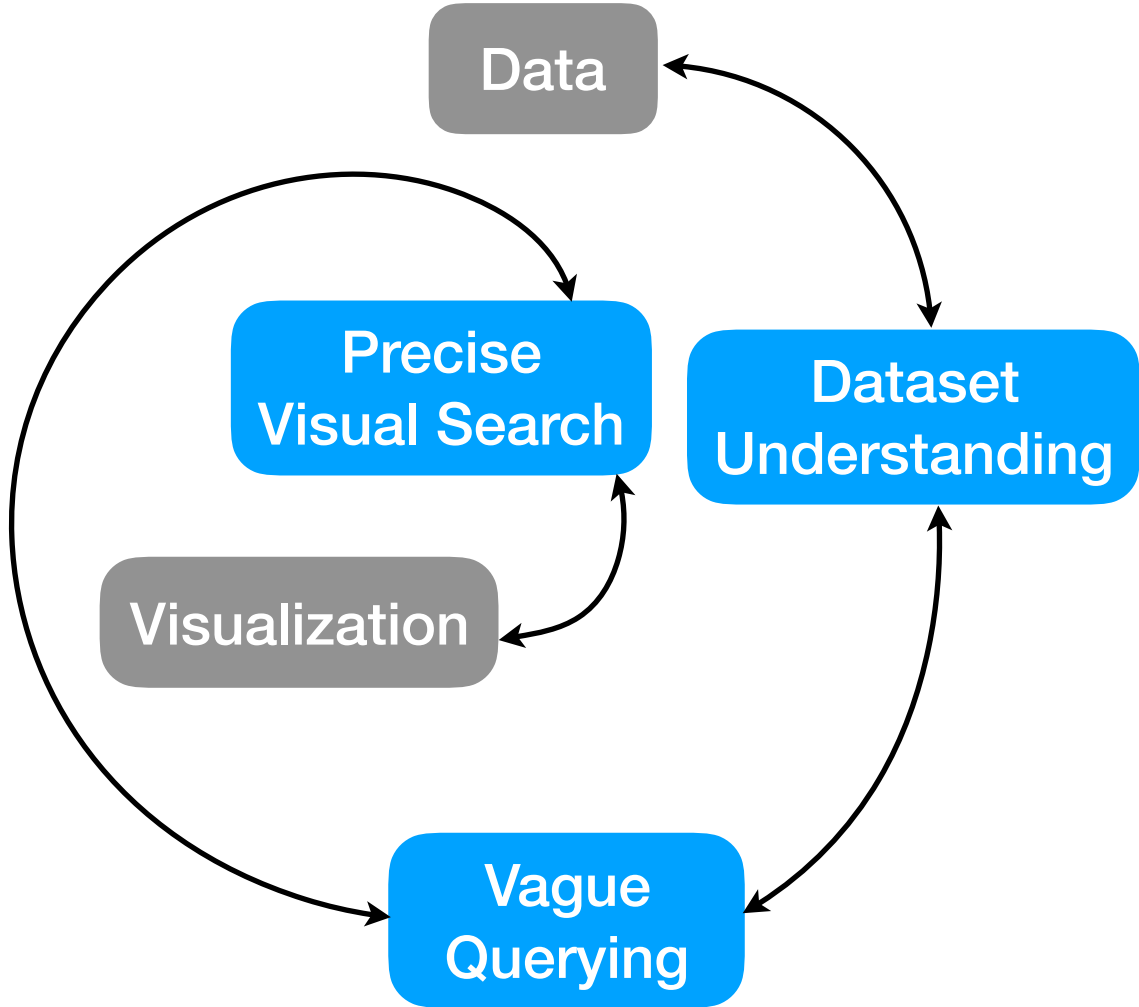
Figure 2: Cycle of visual data exploration.

4 Vague Intelligent Search

Accounting for user interaction, mental models. More global objective taking into account user with the goal of dataset understanding rather than task completion.

4.1 Challenges

- Inferring user intent in querying and context is important (both in terms of user input and what is recommended)

- tools can not assume user has querying intention. exploration without intention, user dont know what they are searching for –¿ Recommendation.

- The important thing here is identifying what should be done by the system v.s. requested from user. Inappropriate choice of these will result in lack of expressibility and user feeling lack of control of analysis, limiting exploration.

- Need for a unified framework of inference to take all of these into account (e.g. natural language, etc)

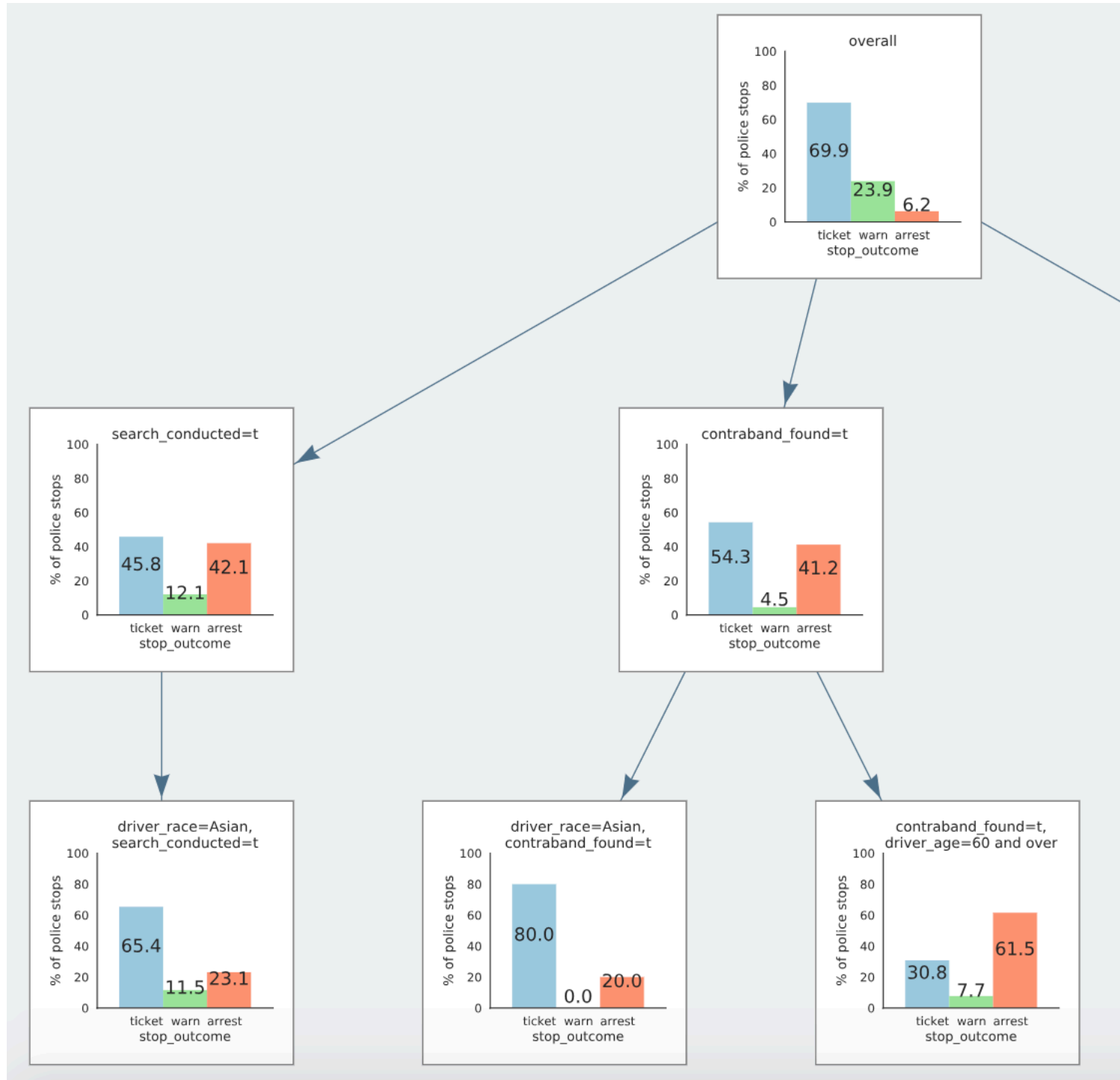## 4.2 STORYBOARD: Navigating Through Data Slices with Hierarchical Summary of Visualizations

Figure 3: test

5 Towards Dataset Understanding

5.1 Challenges

- Problem of cold-start recommendation (as discussed earlier use may not always know what to query for)

- related works have focussed on making specification easier , but not really trying to understnad user intent or what might the user want to see.

- Within a dataset, structure and provenance is essential to help users navigate and provide users with sense of coverage and completion. This is an important but underexplored area. (viz-sum, Sarvghad et al 2017)

- schema and attribute understanding (coverage, etc)

5.2 Examples

- understanding distributions (distribution awareness)

- providing overview recommendations (representative trends and outliers)

6 Concluding Remarks