Accelerating Scientific Data Exploration via Visual Query System

ABSTRACT

The increasing availability of rich and complex data in a variety of scientific domains poses a pressing need for tools to enable scientists to rapidly make sense of and gather insights from data. One proposed solution is to design visual query systems (VQSs) that allow scientists to interactively search for desired patterns in their datasets. While many existing VQSs promise to accelerate exploratory data analysis by facilitating this search, they are not widely used in practice. Through a year-long collaboration with scientists in three distinct domains—astronomy, genetics, and material science—we study the impact of various features within VQSs that can aid rapid visual data analysis, and how VQSs fit into scientists' analysis workflow. Our findings offer design guidelines for improving the usability and adoption of next-generation VQSs, paving the way for VQSs to be applied to a variety of scientific domains.

KEYWORDS

Visual analytics, visualization, exploratory data analysis, visual query, scientific data.

ACM Reference Format:
. 1997. Accelerating Scientific Data Exploration via Visual Query System. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97).* ACM, New York, NY, USA, 5 pages. **https://doi.org/10.475/123_4**

1 INTRODUCTION

2 RELATED WORKS

3 METHODS

We adopted a mixed methods research methodology that draws inspiration from ethnographic methods, iterative and participatory design, and controlled studies [9, 10, 16] to understand how VQSs can be used for scientific data analysis. Working with researchers from three different scientific research groups, we identified the needs and challenges of scientific data analysis and the potential opportunities for VQSs, via interviews and cognitive walkthroughs.

We recruited participants by reaching out to research groups via email and word of mouth, who have experienced challenges in dealing with large amounts of data. We initially spoke to analysts from 12 different potential application areas and narrowed down to three use cases in astronomy, genetics, and material science for our participatory design study. Six scientists from three research groups participated in the design of *zenvisage*. On average, the participants had more than 8 years of research experience working in their respective fields. We list the participants in Table **??**, and will refer to them by their anonymized ID as listed in the table throughout the paper.

Our initial inspiration for building a VQS came from informal discussions with academic and industry analysts. Their current workflows required analysts to manually examine large numbers of visualizations to derive insights from their data. Given these early conversations with the participants, we built a basic VQS to serve as the functional prototype in the design study. As shown in Figure 1, this early version of *zenvisage* allowed users to sketch a pattern or drag-and-drop an existing visualization as a query,then the system would return visualizations that had the closest Euclidean distance from the queried pattern. The details of the system is described in our previous work [17, 18], which focused on the systems and scalability aspects of the VQSs.

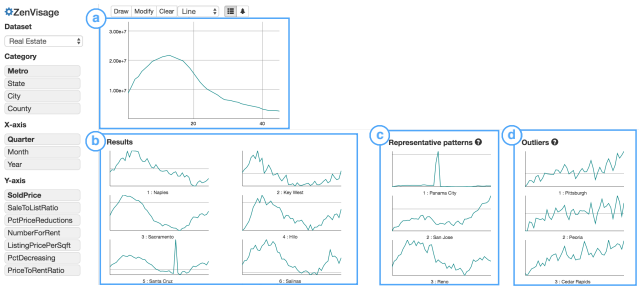

Figure 1: The *zenvisage* prototype allowed users to sketch a pattern in (a), which would then return (b) results that had the closest Euclidean distance from the sketched pattern. The system also displays (c) representative patterns obtained through K-Means clustering and (d) outlier patterns to help the users gain an overview of the dataset.

Visualization systems are often evaluated using controlled studies that measure the user's performance against an existing visualization baseline [15]. Techniques such as artificially inserting "insights" or setting predefined tasks for example datasets work well for objective tasks, such as debugging data errors [6, 13], but these contrived methods are unsuitable for

WOODSTOCK'97, July 1997, El Paso, Texas USA
2016. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

trying to learn about the types of real-world queries users may want to pose on VQSs. Due to the unrealistic nature of controlled studies, many have proposed using a more multi-faceted, ethnographic approach to understand how analysts perform visual data analysis and reasoning [7, 8, 11, 15, 16]. In order to make the user study more realistic, we opted for a qualitative evaluation where we allowed participants to bring datasets that they have vested interests in to address unanswered research questions. Participatory design has been successfully used in the development of interactive visualization systems in the past [1, 2]. Sedlmair et al. [8] advocate that design study methodology is suitable for use cases in which the data is available for prototyping, but the task is only partially known and the information is partially in the user's head. In that regard, our scientific use cases with VQS is well-suited for a design study methodology, as we learn about the scientist's data and analysis requirements and design interactions that helps users translate their "in-the-head" specifications into actionable visual queries.

The use of functional prototypes is common in participatory design to provide a starting point for the participants. For example, Ciolfi et al.[3] studied two different alternatives to co-design (starting with open brief versus functional prototype) in the development of museum guidance systems and found that while both approaches were equally fruitful, functional prototypes can make addressing a specific challenge more immediate and focused. Our motivation for providing a functional prototype at the beginning of the participatory design sessions is to showcase capabilities of VQSs. Especially since VQSs are not common in the existing workflows of these scientists, participants may not be able to imagine their use cases without a starting point.

During the participatory design process, we collaborated with each of the teams closely with an average of two meetings per month, where we learned about their datasets, objectives, and how VQSs could help address their research questions. A detailed timeline of our engagement with the participants and the features inspired by their use cases can be found in Figure 2. Participants provided datasets they were exploring from their domain, whereby they had a vested interest in using a VQS to address their own research questions. Through this process, we identified and incorporated more than 20 desired features into the VQS prototype over the period of a year. Finally, we conducted a realistic, qualitative evaluation to study how analysts interact with different VQS components in practice. The evaluation study participants included the six scientists from the participatory design study, along with three additional "blank-slate" participants who had never encountered *zenvisage* before. While participatory design subjects actively provided feedback on *zenvisage* with their data, they only saw us demonstrating their requested features and explaining the system to them, rather than actively using the system on their own. So the evaluation study was the first time that all nine of the participants used *zenvisage* to explore their datasets.

4 PARTICIPANTS AND DATASETS

During the design study, we observed the participants as they conducted a cognitive walkthrough demonstrating every component of their current data analysis workflow. Cognitive walkthroughs highlight the existing workflows and behavior that participants have adopted for conducting certain tasks [12].

discuss, based on cognitive walkthrough

We summarize the common properties of and differences between these three groups of researchers in Figure **??**. brevity, we have diverted to our technical report.

Astronomy (*astro*)

The Dark Energy Survey (DES) is a multi-institutional project with over 400 scientists. Scientists use a multi-band telescope that takes images of 300 million galaxies over 525 nights to study dark energy[4]. The telescope also focuses on smaller patches of the sky on a weekly interval to discover astrophysical transients (objects whose brightness changes dramatically as a function of time), such as supernova explosions or quasars. The output is a time series of brightness observations associated with each object extracted from the images observed.

For over five months, we worked closely with an astronomer on the project's data management team working at a supercomputing facility. The scientific goal is to identify a smaller set of potential candidates that may be astrophysical transients in order to study their properties in more detail. These insights can help further constrain physical models regarding the formation of these objects.

While an experienced astronomer who has examined many transient light curves can often distinguish an interesting transient object from noise by sight, they must visually examine and iterate through large numbers of visualizations of candidate objects. Manual searching is time-consuming and error prone as the large majority of the objects are not astronomical transients. Participant A1 was interested in *zenvisage* as he recognized how specific pattern queries could help scientists directly search for these rare objects.

Genetics (*genetics*)

Gene expression is a common data type used in genomics and is obtained via microarray experiments. The data used in the participatory design sessions was the gene expression data over time for mouse stem cells aggregated over multiple experiments, downloaded from an online database[1].

[1] `ncbi.nlm.nih.gov/geo/`

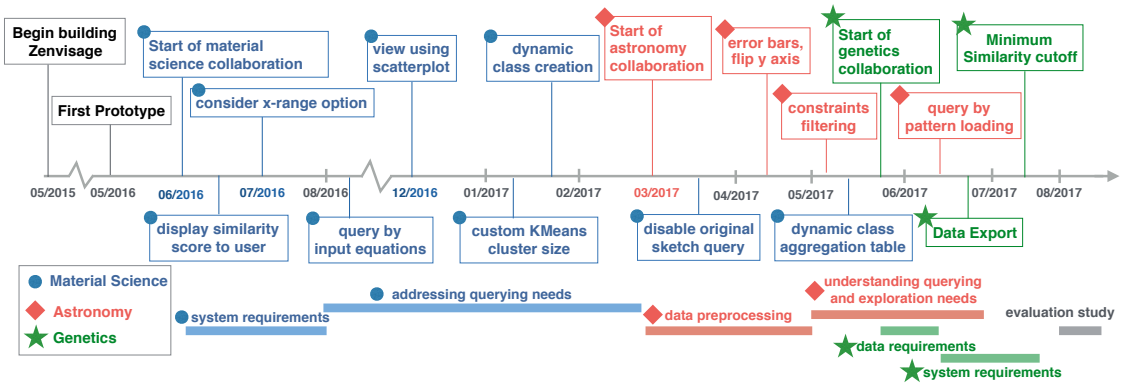

Figure 2: Participatory design timeline for the scientific use cases.

We worked with a graduate student and a PI at a research university over three months who were using gene expression data to better understand how genes are related to phenotypes expressed during early development [5, 14]. They were interested in using *zenvisage* to cluster gene expression data before conducting analysis with a downstream machine learning workflow.

Participant G1 processes the raw microarray data by using a preprocessing script written in R. To analyze the data, the preprocessed data is loaded into a desktop application for visualizing and clustering gene expression data. G1 sets several clustering and visualization parameters on the interface before pressing a button to execute the clustering algorithm. The cluster visualizations are then displayed as overlaid time series for each cluster, as shown in the visualization in Figure **??**b. G1 visually inspects that all the patterns in each cluster looks "clean" and checks the number of outlier genes that do not fall into any of the clusters. If the number of outliers is high or the visualizations look unclean, she reruns the analysis by increasing the number of clusters. When the visualized clusters look "good enough", G1 exports the cluster patterns into a csv file to be used as features in their downstream regression tasks.

Prior to the study, the student (G1) and PI (G3) spent over a month attempting to determine the best number of clusters for their upstream analysis based on a series of static visualizations and statistics computed after clustering. While regenerating their results took no more than 15 minutes every time they made a change, the multi-step, segmented workflow meant that all changes had to be done offline, so that valuable meeting time was not wasted trying to regenerate results. The team had a vested interest in participating in the design of *zenvisage* as they saw how the interactive nature of VQSs and the ability to query other time series with clustering results could dramatically speed up their collaborative analysis process.

Material Science (*matsci*)

We collaborated with material scientists at a research university who are working to identify solvents that can improve battery performance and stability. These scientists work with large datasets containing over 25 chemical properties for more than 280,000 different solvents obtained from simulations.

We worked closely with a graduate students, a postdoctoral researcher, and a PI for over a year to design a sensible way of exploring their data using VQSs. Each row of their dataset represents a unique solvent, and consists of 25 different chemical attributes. They wanted to use *zenvisage* to identify solvents that not only have similar properties to known solvents but also are more favorable (e.g. cheaper or safer to manufacture), as well as to understand how changes in certain chemical attributes affects them.

Participant M1 starts his data exploration process with a list of known and proven solvents as a reference. For instance, he would search for solvents which have boiling point over 300 Kelvins and the lithium solvation energy under 10 kcal/mol using basic SQL queries. This helps him narrow down the list of solvents, and move on to the other properties for similar processing. The scientist also considers the availability and the cost of the solvents while exploring the dataset. When the remaining list of the solvents is sufficiently small, he drills down to more detail (e.g., such as looking at the chemical structure of the solvents to consider the feasibility of conducting experiments with the solvent).

5 DESIGN GUIDELINES

6 CONCLUSION

REFERENCES

[1] Cecilia R Aragon, Sarah S Poon, Gregory S Aldering, Rollin C Thomas, and Robert Quimby. 2008. Using visual analytics to maintain situation awareness in astrophysics. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*. IEEE, 27–34. **https://doi.org/10.1088/1742-6596/125/1/012091**

[2] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human*

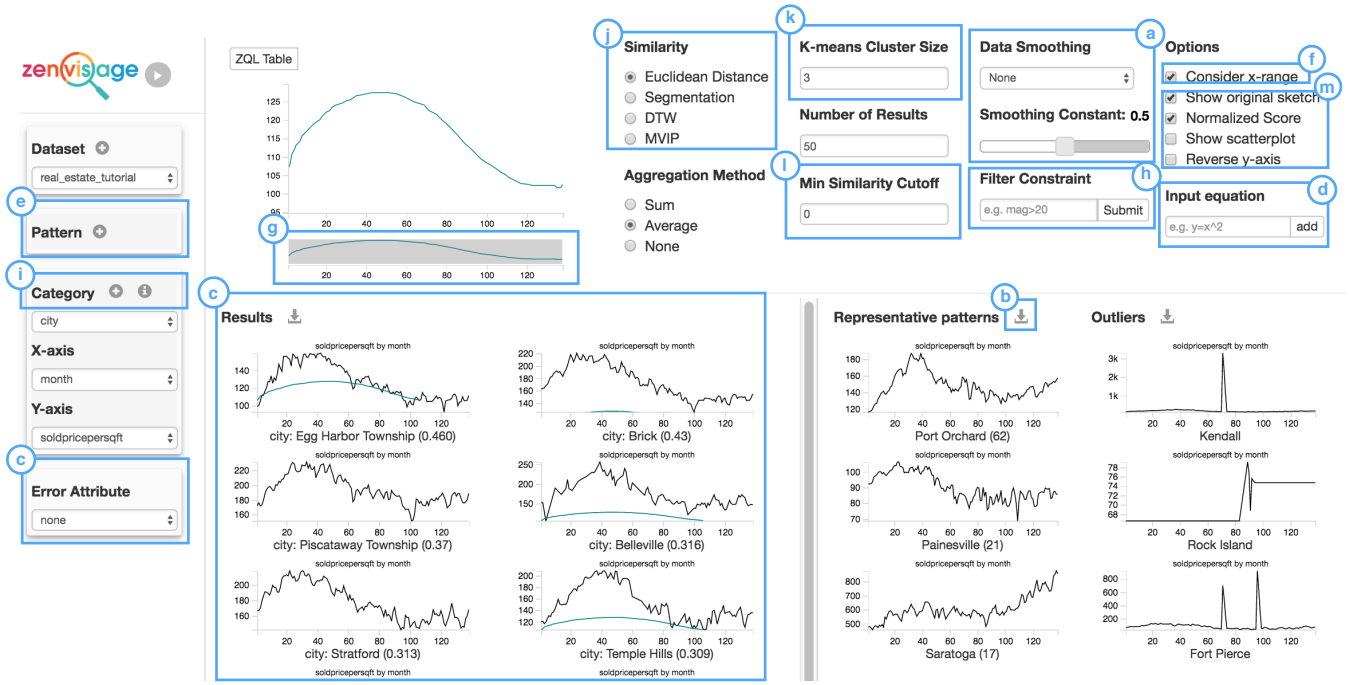

Figure 3: Our VQS after participatory design, which includes: the ability to preprocess via (a) interactive smoothing; (b, c) the ability to export data outputs ; querying functionalities via (d) equations and (e) patterns; query specification mechanisms including (f) x-range invariance, (g) x-range selection and filtering, (h) Filtering, and (i) Dynamic class creation; (j, k, l) system parameter options; (m) visualization display options. Prior to the participatory design, *zenvisage* only included a single sketch input with no additional options. *zenvisage* also displayed representative patterns and outlier patterns, as shown in Figure 1.

Factors in Computing Systems. ACM, 443–452. https://doi.org/10.1145/2207676.2207738

[3] Ciolfi et al. 2016. Articulating Co-Design in Museums: Reflections on Two Participatory Processes. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16* (2016), 13–25. https://doi.org/10.1145/2818048.2819967

[4] Drlica Wagner et al. 2017. Dark Energy Survey Year 1 Results: Photometric Data Set for Cosmology. (2017). arXiv:1708.01531

[5] Brian S. Gloss, Bethany Signal, Seth W. Cheetham, Franziska Gruhl, Dominik C. Kaczorowski, Andrew C. Perkins, and Marcel E. Dinger. 2017. High resolution temporal transcriptomics of mouse embryoid body development reveals complex expression dynamics of coding and noncoding loci. *Scientific Reports* 7, 1 (2017), 6731. https://doi.org/10.1038/s41598-017-06110-5

[6] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3363–3372. https://doi.org/10.1145/1978942.1979444

[7] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

[8] Tamara Munzner Michael Sedlmair, Miriah Meyer. 2012. Design Study Methodolgy: Reflections from the Trenches and the Stacks. 1, 12 (2012), 2431–2440.

[9] Delbert C. Miller, Neil J. Salkind, and Delbert C. Miller. 2002. *Handbook of research design and social measurement*. SAGE.

[10] Michael J. Muller and Sarah Kuhn. 1993. Participatory Design. *Commun. ACM* 36, 6 (June 1993), 24–28. https://doi.org/10.1145/153571.255960

[11] Tamara Munzner. 2009. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009). https://doi.org/10.1109/TVCG.2009.111

[12] Jakob Nielsen. 1994. Usability Inspection Methods. In *Conference Companion on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 413–414. https://doi.org/10.1145/259963.260531

[13] Patel et al. 2010. Gestalt: integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd annual ACM symposium on User Interface Software and Technology*. ACM, 37–46. https://doi.org/10.1145/1866029.1866038

[14] Pei Chen Peng and Saurabh Sinha. 2016. Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Research* 44, 13 (2016), e120. https://doi.org/10.1093/nar/gkw446

[15] Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM, 109–116. https://doi.org/10.1145/989863.989880

[16] Ben Shneiderman and Catherine Plaisant. 2006. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM, 1–7. https://doi.org/10.1145/1168149.1168158

[17] Tarique Siddiqui, John Lee, Albert Kim, Edward Xue, Chaoran Wang, Yuxuan Zou, Lijin Guo, Changfeng Liu, Xiaofo Yu, Karrie Karahalios,

and Aditya Parameswaran. 2017. Fast-Forwarding to Desired Visualizations with zenvisage. (2017). https://doi.org/10.1145/1235

[18] Siddiqui et al. 2016. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *Proceedings of the VLDB Endowment* 10, 4 (2016), 457–468. https://doi.org/10.14778/3025111.3025126