Title
Three Lessons from Accelerating Scientific Insight Discovery via Visual Querying

Journal
Patterns, 1(7)

ISSN
2666-3899

Authors
Lee, Doris Jung-Lin
Siddiqui, Tarique
Karahalios, Karrie
et al.

Publication Date
2020-10-01

DOI
10.1016/j.patter.2020.100126

Peer reviewed

Patterns

CellPress
OPEN ACCESS

Opinion

Three Lessons from Accelerating Scientific Insight Discovery via Visual Querying

Doris Jung-Lin Lee,[1,*] Tarique Siddiqui,[2] Karrie Karahalios,[2] and Aditya Parameswaran[1]
[1]University of California, Berkeley, Berkeley, California, USA
[2]University of Illinois at Urbana-Champaign, Urbana and Champaign, Illinois, USA
*Correspondence: dorislee@berkeley.edu (D.J.-L.L.), tsiddiq2@illinois.edu (T.S.), kkarhal@illinois.edu (K.K.), adityagp@berkeley.edu (A.P.)
https://doi.org/10.1016/j.patter.2020.100126

Exploratory data analysis is a crucial part of data-driven scientific discovery. Yet, the process of discovering insights from visualization can be a manual and painstaking process. This article discusses some of the lessons we learned from working with scientists in designing visual data exploration system, along with design considerations for future tools.

Exploratory data analysis (EDA) is a crucial part of the scientific process and often informs subsequent modeling and hypothesis testing. Scientists perform EDA to identify research questions and uncover insightful patterns and trends in their data—with information visualizations often serving as a gateway to understanding complex scientific phenomena. However, the current process of discovering visualizations with actionable insights requires scientists to manually construct and examine large numbers of visualizations one-at-a-time. This painstaking process of combing through a considerable number of visualizations in search of desired patterns is a common challenge across many domains. We reflect on some of the challenges and lessons learned from working with three groups of end-user scientists in designing visual data exploration systems for scientific discovery.[1]

Lesson #1: Managing Large Collections of Visualizations Is a Common Challenge

The current process of manually exploring a large number of visualizations is tedious, overwhelming, and error-prone for scientists who are not already intimately familiar with their datasets. One group of scientific researchers we worked with study how genes relate to phenotypes expressed during early embryonic development.[2,3] Their data consist of a collection of gene expression profiles over time for mouse stem cells, aggregated over multiple experiments. The expression profile for each gene is plotted as a line chart, with time on the x axis and expression level on the y axis. During EDA, a common task that these researchers perform is correlating gene function with their expression profiles, to gain a high-level overview of the expression profile patterns. However, researchers often have to write custom data pipelines for preprocessing, clustering, and visualizing hundreds and thousands of line charts (one for each gene). While existing systems such as STEM[4] provide a graphical user interface for clustering, comparing, and visualizing time series from microarray experiments, they do not support users' ability to search for visually similar patterns to a given trend or visualization. For instance, researchers are often interested in finding patterns that look similar to the profile of a known gene regulated by the estrogen receptor, since these visually similar patterns suggest that the gene may be affected by the same factors. The challenge of generating and searching through large numbers of visualizations is not only specific to the genetics setting, but similar needs also arise in other scientific domains, such as searching for supernovae signatures in astronomical light curves or searching for electrolytes with desirable physical property trends for battery design.

To mitigate the aforementioned challenge, there is a need for a "search engine" for visualizations, enabling scientists to quickly find desired visual patterns in their datasets via an intuitive querying interface. *Visual query systems* are a class of visualization search interfaces that allow users to specify the desired pattern via a high-level specification (e.g., show me a line chart where values first increase then decrease), with the system returning recommendations of visualizations that match the specified pattern. This is in stark contrast with traditional visualization construction interfaces, such as ggplot or Tableau, where users have to exactly specify the visual encoding (chart type, mark properties) and data aspects (what to visualize on the x and y axis, what subsets of data to visualize). Early work in visual query systems focused on interfaces to search for time series with specific patterns. For example, TimeSearcher[5] supports a rectangular box as a query specification mechanism, with the system filtering out all of the time series that do not pass through it. In QuerySketch[6] and Google Correlate,[7] the query is sketched as a pattern on a canvas, with the system filtering out all the time series that have a different shape. Subsequent work recognized the ambiguity in sketching by studying how humans rank the similarity in patterns[8] and improving the expressiveness of sketched queries through finer-grained specification interfaces and pattern-matching algorithms.[9] While these systems are promising in addressing the challenge of searching through large collections of visualizations, despite almost two decades of research, these visual query systems have not been widely adopted by the end-users they were designed for. The subsequent lessons that we learned from working with scientists points to some fundamental design flaws that may have led to the lack of the adoption of these systems.

Lesson #2: Proactive Recommendation and Guidance Jumpstarts Exploration

In exploratory data analysis, scientists often do not have an exact inquiry in mind but instead make use of the insights discovered from their data to inform their hypotheses in a "bottom-up" manner. For example, genetics researchers often first cluster their gene expression data and get an overview of the patterns in their dataset before composing a visual query for genes similar to the common patterns. Often, these hypotheses are under-specified or ambiguous, which can be challenging to even articulate and sketch in a visual query interface. For instance, an astronomer that we worked with wanted to look for patterns that are slightly irregular, indicating the presence of a pulsating star. Given the challenge of cold-start exploration, there is a need for visual data exploration systems to provide recommendation and guidance. The idea of incorporating recommendations into visual exploration systems is not new. There has been work in designing *visualization recommendation* systems that suggest interesting visualizations to users to accelerate data exploration. These systems often leverage statistical and perceptual properties[10,11] to surface relevant patterns and trends in the underlying dataset.

From our experience in working with scientists, we find that users frequently leveraged a recommended list of common visual patterns and outliers to jumpstart their inquiries. For example, one scientist identified that the three representative patterns recommended by a visual query system with proactive recommendations corresponded to the same three groups of genes discussed in a recent publication:[3] induced genes (), repressed genes (), and transient genes (). They then used the visual query system to search for other genes that also exhibited a transient pattern by dragging and dropping the recommended pattern into the query canvas. Based on the query results, they learned that among the transient patterns, there was a group of genes that all transitioned at the same timestep, while others transitioned at different timesteps. These examples demonstrate how incorporating even basic recommendations during visual querying can be helpful to users in providing data examples that inspire further exploration.

Lesson #3: Integrative Workflow Encourages Experimentation

By observing how our scientific collaborators interacted with a visual query system, we discovered that scientists engaged in three primary ways of making sense of their data—a finding that echoes a classic cognitive model of how analysts perform information processing tasks.[12] These three sensemaking processes include *top-down pattern search* (where users leverage their intuition about what the pattern should look like to perform the pattern search), *bottom-up inquiry* (where user queries are driven by something that they observe in the data), and *context creation* (where users navigate across different visualization collections to understand the pattern in their data). Further analysis revealed that all groups of participants engage with all three sensemaking processes to some degree, but in different proportions, depending on their analytical needs. In fact, we find that *all three sensemaking processes are essential* for visual query systems to flexibly support a diverse range of analytical goals.

In addition to integration across all three sensemaking processes, we found that integration across the different parts of the analysis workflow is also crucial for encouraging rapid and sustained experimentation with data. In particular, we observed that many scientists switched between parameter specification, code execution, and visualization comparison in their existing data analysis workflow. The non-interactive nature of these segmented workflows incurs a substantial cognitive barrier during exploratory data analysis, especially in collaborative settings. For example, a geneticist would specify several clustering and visualization parameters in a custom data-processing script, then visually inspect whether the patterns in each cluster look "clean" in a graphical user interface (GUI). The geneticist would then iteratively tune parts of the clustering script and rerun the analysis until the visualized clusters look "clean enough." While modifying and rerunning the pipeline took no more than 15 min every time, the multi-step, segmented workflow meant that the genetics team had to make all the changes offline, so that their valuable meeting time is not wasted on regenerating results. When the team switched to a more integrative system that allowed both cluster specification and visual querying in a single interactive window, the new tool dramatically sped up their collaborative analysis process and encouraged them to experiment with previously unexplored parts of their pipeline.

Our findings point to the need for *integrative* systems that support seamless transitions between analytic activities. These lessons have inspired our most recent project Lux (https://github.com/lux-org/lux), which combines the benefits of interactive visualization interfaces and programmatic specification of analysis and visualizations. Visualizations are displayed as a widget within an integrated, reproducible computational notebook environment, providing a seamless transition between code and interaction. More importantly, by bringing interactive visualization into notebooks, this approach supports intelligent visual data discovery alongside other data science activities, such as data cleaning and modeling, without the cost of switching between different tools and interfaces.

Conclusion

The aforementioned lessons that we learned from working with multiple groups of scientists, such as the importance of an integrative visual query system across all sensemaking processes, open up a number of new research directions in providing proactive and tailored assistance during visual data exploration. Indeed, visual querying systems can be incredibly powerful and effective, but need to be mindfully designed in a manner that encourages holistic, proactive, and efficient exploration of patterns in data. We hope our lessons can inform the next generation of tools for data-driven scientific discovery.

REFERENCES

1. Lee, D.J., Lee, J., Siddiqui, T., Kim, J., Karahalios, K., and Parameswaran, A. (2020). You can't always sketch what you want: Understanding Sensemaking in Visual Query Systems. IEEE Trans. Vis. Comput. Graph. *26*, 1267–1277.

2. Peng, P.C., and Sinha, S. (2016). Quantitative modeling of gene expression using DNA shape features of binding sites. Nucleic Acids Res. *44*, e120.
3. Gloss, B.S., Signal, B., Cheetham, S.W., Gruhl, F., Kaczorowski, D.C., Perkins, A.C., and Dinger, M.E. (2017). High resolution temporal transcriptomics of mouse embryoid body development reveals complex expression dynamics of coding and noncoding loci. Sci. Rep. *7*, 6731.
4. Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics *7*, 191.
5. Hochheiser, H., and Shneiderman, B. (2004). Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. Information Visualization *3*, 1–18.
6. Wattenberg, M. (2001). Sketching a graph to query a time-series database. In CHI'01 Extended Abstracts on Human factors in Computing Systems (ACM), pp. 381–382.
7. Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., and Kumar, S. (2011). Google correlate whitepaper. Google https://research.google/pubs/pub41695/.
8. Correll, M., and Gleicher, M. (2016). The semantics of sketch: Flexibility in visual query systems for time series data. In Visual Analytics Science and Technology (VAST), 2016 IEEE Conference (IEEE), pp. 131–140.
9. Ryall, K., Lesh, N., Lanning, T., Leigh, D., Miyashita, H., and Makino, S. (2005). Querylines: approximate query for visual browsing. In CHI'05 Extended Abstracts on Human Factors in Computing Systems (ACM), pp. 1765–1768.
10. Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. (2016). Voyager: Exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans. Vis. Comput. Graph. *22*, 649–658.
11. Vartak, M., Rahman, S., Madden, S., Parameswaran, A., and Polyzotis, N. (2015). Seedb: efficient data-driven visualization recommendations to support visual analytics. Proceedings VLDB Endowment *8*, 2182–2193.
12. Pirolli, P., and Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proceedings of International Conference on Intelligence Analysis *5*, 2–4.

About the Authors

Doris Lee is a graduate student in the School of Information at UC Berkeley.

Tarique Siddiqui is a senior researcher at Microsoft Research.

Aditya Parameswaran is an assistant professor in the School of Information and Department of Electrical Engineering and Computer Sciences at UC Berkeley.

Karrie Karahalios is a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign.