Tasks to Tease Apart Scatterplot Design Decisions

Alper Sarikaya*
University of Wisconsin-Madison

Michael Gleicher†
University of Wisconsin-Madison

Abstract

Scatterplots are among the most common methods for exploring and presenting data, covering a wide range of tasks and designs. The variety of scatterplot designs has created a proliferation of potential design decisions to consider when constructing a scatterplot. However, there remain many unexamined assumptions in respect to the trade-offs between these decisions. In this short summary, we begin the process of synthesizing recent work to build descriptive knowledge of how design decisions affect the analysis tasks viewers perform with scatterplots. Through deriving twelve abstracted scatterplot tasks, we can start to tease apart the different affordances of design decisions, and begin to formulate a basis for prescriptive scatterplot design.

Index Terms: Scatterplot, task taxonomy, design space

1 Introduction

Scatterplots are a simple but often used visualization. Their use in a diverse range of applications and analysis scenarios has led to a wide variety of design variants. The diversity of available designs can make it difficult to make appropriate design decisions for specific scenarios. In this work, we seek more principled differentiation between design decisions for scatterplots by looking at how analysis tasks form a basis for the design space of scatterplots.

In order to attain this goal, we both gather analysis tasks concerning scatterplots and catalog visual designs of scatterplots from the visualization literature. Through the collection of analysis tasks that cross different analysis and data domains, we can arrive at a broadly applicable, abstract list of tasks specific to scatterplots. Surveying scatterplot designs allows us to catalog which visual strategies apply in which situations. With an understanding of the level of scatterplot task support for different designs, we can provide prescriptive guidance (collected in a descriptive manner) for choosing between visual techniques and identify areas within the design space with minimal task support, ripe for novel visual designs.

To bound our exploration of how tasks and visual designs are connected, we concentrate on scatterplots (see Friendly and Dinis [6] for a historical overview). Through our work presented herein, we demonstrate how different scatterplot design decisions can be motivated by thinking about the scatterplot-specific analysis tasks that each decision supports.

2 Related Work

The scatterplot has been studied in many domains, including statistics, geography, and perceptual psychology. In this work, we only consider the 2D scatterplot that encodes objects as points (or some other aggregate visual encoding) by position through two continuous, orthogonal dimensions, plus optional data conveyed through mark color, size, or shape. We extend this definition to capture scatterplot-like designs that may not explicitly represent the presence of an item by a single mark—such as a density plot. In statistics, Cleveland [5] provides a wide-ranging overview of the design decisions possible with a scatterplot. Research in geography focuses on similar issues in design (though with different names for concepts, e.g. "point spatializations" for scatterplots [11]), see MacEachren [8] for an overview. Perceptual psychology continues to study the methods of object and correlation detection, often in a scatterplot scenario (e.g., [2]).

Despite the large amount of research concerning scatterplots, there is considerably less research in how scatterplots *are used* to attain their goals—how do design decisions support analysis using scatterplots? A notable exception is Seldmair *et al.*'s work [12] on analysis of dimensionally-reduced data, a common application of scatterplots. Existing taxonomies for task are too general for application to scatterplots. Task taxonomies in the visualization literature are hierarchical, describing low-level tasks such as visual cognition [7], which are in turn utilized by low-level analytical tasks [1], up to high-level understanding tasks to generate knowledge [3]. In particular, Brehmer and Munzner [3] focus on the *why* and *how*, organizing the *why* into a spectrum from high-level rationales (e.g., consume, produce) to low-level tasks (e.g., identify, compare). In this work, we concentrate on low-level, analyst-focused tasks that scatterplots support, and use these tasks to help differentiate different design decisions.

3 Scatterplot Tasks

While many task taxonomies have been constructed for general information visualization, a complete range of analyst and viewer tasks has not been gathered for scatterplots. With good coverage of the task space around scatterplots, we can form a basis for mapping out the space of design decisions. To synthesize these tasks, we collected tasks at the level of analyst intent (*cf.* Meyer *et al.* [10]) from a variety of sources in the data visualization literature, including empirical evaluation of designs, technique papers, and position papers. These tasks' sources are detailed in the supplementary material. To abstract and cluster similar tasks, we asked four data visualization researchers (5–10 years of experience) to perform a card sort with the collected 29 tasks and arrived at the tasks in Table 1.

While not all of the derived tasks are specific to scatterplots, they have specific applicability to how viewers use scatterplots. As an example, *identify object* is a general concept across many data visualizations, but within the context of scatterplots specifically entails the act of reconciling and linking a mark's position to its represented object. In particular applicability to scatterplot designs, particular design decisions can have adverse support for object identification, as we discuss in the next section.

4 Designs

To start the collection of design decisions concerning scatterplots, we performed a systematic survey of the data visualization literature to cover the recent range of work on scatterplot design. We collected manuscripts by searching titles and abstracts of visualization conferences (Information Visualization, EuroVis, PacVis, EuroVis) for the keyword "scatter", finding 68 of 2499 that were relevant to scatterplots. For each paper (available in supplementary material), we recorded the explicit task support of each strategy. Through our exploration, we noticed that these particular analysis tasks were infrequently discussed and generally focused on a particular analysis

*e-mail: sarikaya@cs.wisc.edu
†e-mail: gleicher@cs.wisc.edu

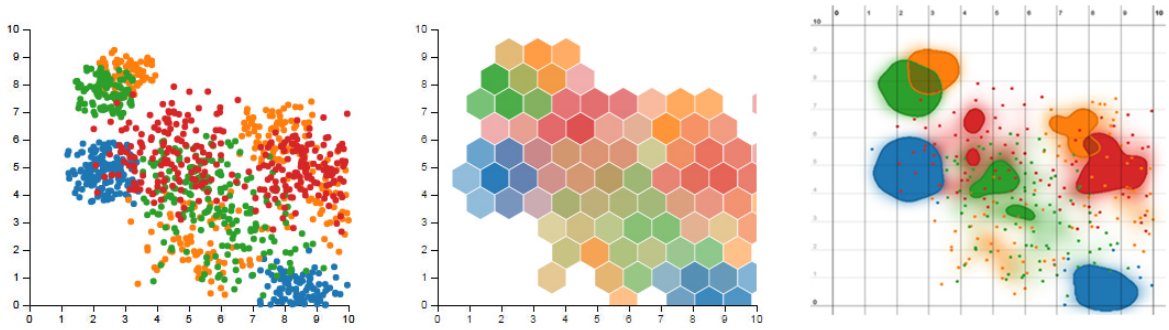

Figure 1: Three different designs of the same data (from left: traditional scatterplot, hexagonal binning [4], and Splatterplots [9]). Different methods of aggregation (binning vs. density) have detrimental effects on particular tasks (e.g. *identify object*, *identify outliers*).

#	Task	Description
1	Identify object	Identify the referent from the representation
2	Locate object	Find a particular object in its new spatialization
3	Verify object	Reconcile attribute of an object with its spatialization (or other visual encoding)
4	Search for known motif	Find a particular known pattern (cluster, correlation)
5	Browse data	Look for things that look unusual, global trends
6	Identify outliers	Find an object that doesn't match the 'modal' distribution
7	Characterize distribution	Do objects cluster? Part of a manifold? Range of values?
8	Identify correlation	Determine level of correlation
9	Explore neighborhood	Explore the properties of objects in a neighborhood
10	Numerosity comparison	Compare the numerosity/density in different regions of the graph
11	Object comparison	Do objects have similar attributes? Are these objects similar in some way?
12	Understand distances	Understanding a given spatialization (e.g. relative distances)

Table 1: The collected list of tasks performed with scatterplots.

scenario or addressing the issues of a domain problem.

Though we are early in our process of drawing distinctions between design decisions, there are a number of reconciliations readily apparent. As shown in Figure 1, different methods of aggregation have varying levels of support for our twelve scatterplot tasks. As a simple case study, if we consider the task of *identify outliers*, traditional scatterplots provide meager support, hexagonal binning [4] aggregates points together into bins that eliminates identification of points (poor support), and Splatterplots [9] offers differentiated identification of points lying outside of dense regions (technique-driven identification). By analyzing these aggregations from this perspective, we can start to tease apart the rationales to create prescriptive advice for effective scatterplot design.

As a second example, we can use the framework of tasks to dive into the subtleties between different aggregation mechanisms. As an example, it can be ambiguous to *characterize distribution* of the data series of the hexagonal bin plot (center, Fig. 1) due to aliasing based on bin placement. Other screen space-aware techniques such as Splatterplots may give additional fidelity to characterizing the distribution of densely packed points through local thresholding.

5 Discussion

Through this framework, we have sought to operationalize analysis tasks as a technique to compare the trade-offs between different design decisions. We are eager to use this framework to dissect why and how different design decisions support particular analyses, particularly to uncover subtleties between visual strategies. As an example, we can use this taxonomy to identify open problems, such as the lack of a succinct visual metaphor for supporting data with tens of classes (e.g. datasets with more than 20 series). We anticipate that this framework will help to scaffold discussion of scatterplot design going forward, and spur exploration of this design space.

Acknowledgements

We thank Deidre Stuffer for copy-editing. This work was supported by NSF award IIS-1162037.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117. IEEE, 2005.

[2] L. Best, A. Hunter, and B. Stewart. Perceiving relationships: A physiological examination of the perception of scatterplots. *Diagrams*, pages 244–257, 2006.

[3] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–85, 2013.

[4] D. B. Carr et al. Scatterplot Matrix Techniques for Large N. *Journal of the American Statistical Association*, 82(398):424, 1987.

[5] W. S. Cleveland. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software, Monterey, CA, USA, 1985.

[6] M. Friendly and D. Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103–130, 2005.

[7] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction*, 3(2):107–135, 1996.

[8] A. M. MacEachren. *How Maps Work: Representation, Visualization, and Design*. The Guilford Press, New York, New York, USA, 1995.

[9] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, 2013.

[10] M. Meyer, M. Sedlmair, and T. Munzner. The four-level nested model revisited. In *Proc. BELIV '12*, pages 1–6. ACM Press, 2012.

[11] D. R. Montello, S. I. Fabrikant, M. Ruocco, and R. S. Middleton. Testing the First Law of Cognitive Geography on Point-Display Spatializations. *Cosit*, pages 316–331, 2003.

[12] M. Sedlmair, T. Munzner, and M. Tory. Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013.