Eilbert, J., Peters, Z., Eliott, F., Stassun, K., and Kunda, M. (2018). Shapes in scatterplots: Comparing human visual impressions and computational metrics. 40th Annual Meeting of the Cognitive Science Society, Madison, WI. p. 330-335.

# Shapes in Scatterplots: Comparing Human Visual Impressions and Computational Metrics

**Joseph Eilbert**[a,b], **Zameese Peters**[b,c],
**Fernanda Monteiro Eliott**[a], **Keivan Stassun**[b], **Maithilee Kunda**[a]
[a]Department of Electrical Engineering and Computer Science, [b]Department of Physics and Astronomy
Vanderbilt University, Nashville, TN 37235-1679, USA
[c]Present affiliation: Norfolk State University

## Abstract

We are currently in the process of designing and implementing a computational cognitive system that combines perception, memory, attention, and domain-specific semantic knowledge to perform data visualization tasks. While this work is still in early stages, we report here on one subset of this larger project that involves building a "visual long term memory" for the system. To constrain the problem, we assume a domain of astronomy, and we focus exclusively on scatterplot visualizations. In this paper, we present three of our initial steps along this path. First, we collected and analyzed a catalog of 74 scatterplots from real astronomy sources (papers, books, etc.), which we consider to be typical data visualizations that astronomers would frequently encounter during their education. Second, we asked a team of human raters to rate all 74 scatterplots along nine dimensions describing shape categories, taken from a computational approach originally suggested by John and Paul Tukey called scagnostics. Third, we calculated computer-based scagnostics for a subset of the scatterplots. We measured inter-rater agreements among the human raters and between the calculated and human ratings.

**Keywords:** astronomy; data visualization; scagnostics.

## Introduction

As observational astronomy, among other research fields, has encountered increasingly large datasets, rigorous analysis by experts has become increasingly time-consuming and difficult (Abello, Pardalos, & Resende, 2013). Computational tools including interactive visualization software (Burger et al., 2013) and other cognitive supports (Honavar, Hill, & Yelick, 2016) are in high demand to address the need to visualize large, multivariate datasets and aid researchers in isolating plots or other data views of interest. Crucial to the design of such systems is a rigorous understanding of the human cognitive processes that are at work during different data visualization tasks, from exploration to interpretation.

One of the big-picture research questions that drives our work is: What is the role played by *memory* in data visualization and interpretation? Many studies look at immediate perceptual properties of visualizations, but the field of information visualization as a whole emphasizes that data visualization is an interactive process that unfolds over time. Within the context of a single data visualization episode, studies have looked at systems that provide histories or bookmarks to previously seen plots, to aid the human user in remembering their prior interactions later in the episode (Callahan et al., 2006).

In addition to this kind of within-episode short term memory, long-term memory must also play crucial roles in data visualization. For example, a scientist's semantic and mathematical knowledge will heavily influence their interpretation of a particular visualization.

We hypothesize that *long-term visual memory* also plays an important role in data visualization. When scientists learn their domain, they often learn about important data relationships by looking at figures. To what extent do the visual properties of these figures stay with scientists later in their career? Not all figures are likely to be remembered in perfect detail, but probably every economist can quickly draw the same supply and demand curves, and probably every astronomer can produce a rough sketch of the Hertzsprung-Russell diagram. If an astronomer is later looking at any other plot of spectral star classifications, might it be that they are not only performing semantic comparisons but also visual comparisons?

Studying these kinds of questions in people is a tall order. Another approach is to try to build computational models that enact similar types of memory, reasoning, and visualization processes. Such models can then be studied to learn more about the task of data visualization itself and about what kinds of representations and reasoning processes might be sufficient for supporting certain levels or kinds of performance.

We are currently in the process of designing and implementing a computational cognitive system that combines perception, memory, attention, and domain-specific semantic knowledge to perform data visualization tasks. While this work is still in early stages, we report here on one subset of this larger project that involves building a "visual long term memory" for the system. To constrain the problem, we assume a domain of astronomy, and we focus exclusively on scatterplot visualizations which, while a relatively simple form of visualization, are still very widely used in virtually all data-related domains.

In essence, we want our cognitive system to have access to "visual memories" of many of the same scatterplots that human astronomers might remember from their education. We are also interested in how humans perceive and remember these plots, e.g., which plot properties are remembered, which might be forgotten or even mis-remembered, etc.

In this paper, we present three of our initial steps along this path. First, we collected and analyzed a catalog of 74 scatterplots from real astronomy sources (papers, books, etc.), which we consider to be typical data visualizations that astronomers would frequently encounter during their education. Second, we asked a team of human raters to rate all 74 scatterplots along nine dimensions describing shape categories,

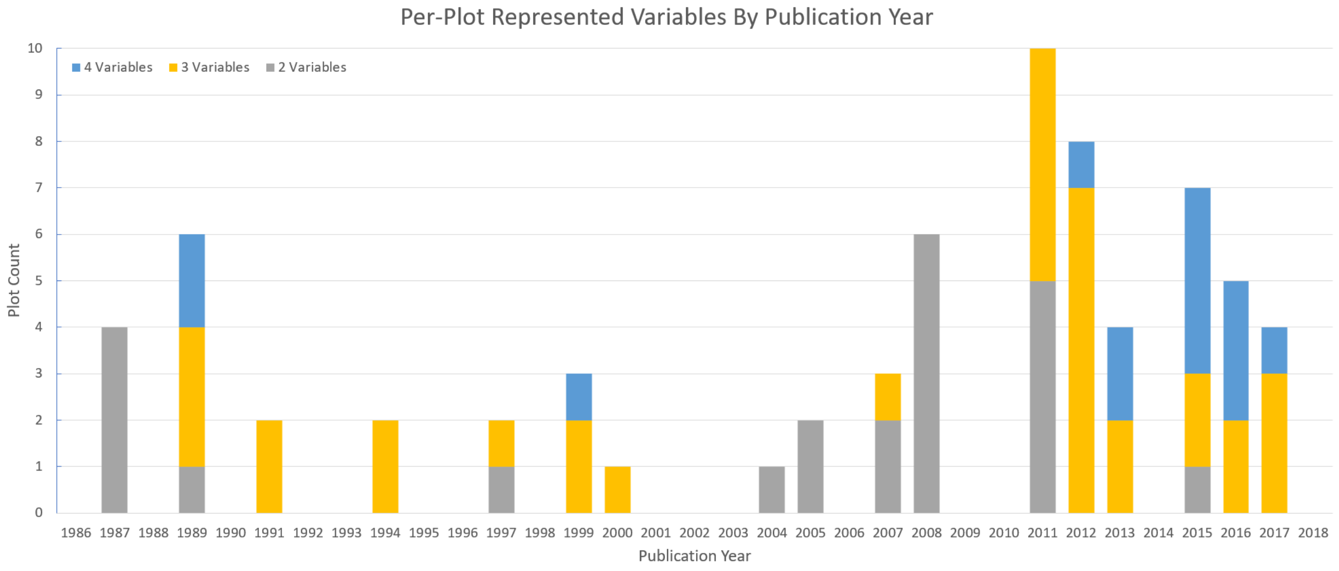

Figure 1: Number of variables represented in each scatterplot by publication year of source. Two variables correspond to each axis, and additional variables are represented by extra axes, color scales, text labels, and point-glyph sizes and shapes.

taken from an computational approach for obtaining scatterplot quality metrics originally suggested by John and Paul Tukey called scagnostics (Wilkinson, Anand, & Grossman, 2005). Third, we calculated computer-based scagnostics for a subset of the scatterplots. We then measured inter-rater agreements among the human raters and between the calculated and human ratings.

## Related Work

Several studies have categorized and provided guidelines for implementing various scatterplot visualization techniques (Etemadpour, Linsen, Paiva, Crick, & Forbes, 2015; Sarikaya & Gleicher, 2018). These techniques address certain visualization goals guiding the presentation of large, multivariate datasets on traditional 2D or 3D scatterplots. Other studies have examined differences between human visual impressions of scatterplots according to scagnostics shape categories and computed scagnostics (Lehmann, Hundt, & Theisel, 2015; Sedlmair & Aupetit, 2015). Unlike these, which typically focus on visualizations of real or synthetic high-dimensional datasets, we focus on scatterplots that have been published. Published scatterplots differ from “dataset-generated” scatterplots because presumably someone, somewhere, has specifically adjusted the visual properties of these scatterplots using certain visualization techniques with specific communication goals in mind. Underlying our work is our desire to understand how people perceive, interpret, and remember these “intentional” types of scatterplots.

## Part 1: Scatterplot Dataset

This collection of scatterplots is **not** intended to represent a systematic sampling of the astronomy literature, but rather a representative collection of the types of scatterplots that astronomers are likely to find engaging and familiar. Thus, the work we present here on these scatterplots should be considered more like a case study than a generalizable sampling.

To that end, scatterplots were sourced using two ad-hoc methods: First, ’visually interesting’ scatterplots were solicited by email from faculty of the Vanderbilt University Department of Physics and Astronomy. Special requests were made for series of scatterplots which show a visual progression through a larger dataset.

Second, scatterplots within Binney and Merrifield’s definitive *Galactic Astronomy* textbook were collected (Binney & Merrifield, 1998). *Galactic Astronomy* was chosen as a canonical textbook with which astronomers are likely to be familiar. We collected a total of 74 scatterplots from both collection methods.

## Results

Scatterplots in the catalog range from simple bivariate scatterplots to complex multivariate scatterplots. Of 31 total sources, 15 sources yielded one scatterplot each, five sources yielded two and three scatterplots each, and two sources yielded four, six, and seven scatterplots.

Point-glyph refers to a unique class of visual mark with a certain size, color, shape, and fill. Thirty-four scatterplots use one point-glyph, 17 use two different point-glyphs, six scatterplots use three and four point-glyphs, four scatterplots use five and six different point-glyphs, two scatterplots use seven different point-glyphs, and one scatterplot uses fifteen different point-glyphs.

Continuous variables are most often represented visually with perpendicular axes, a gradient color scheme, or variable

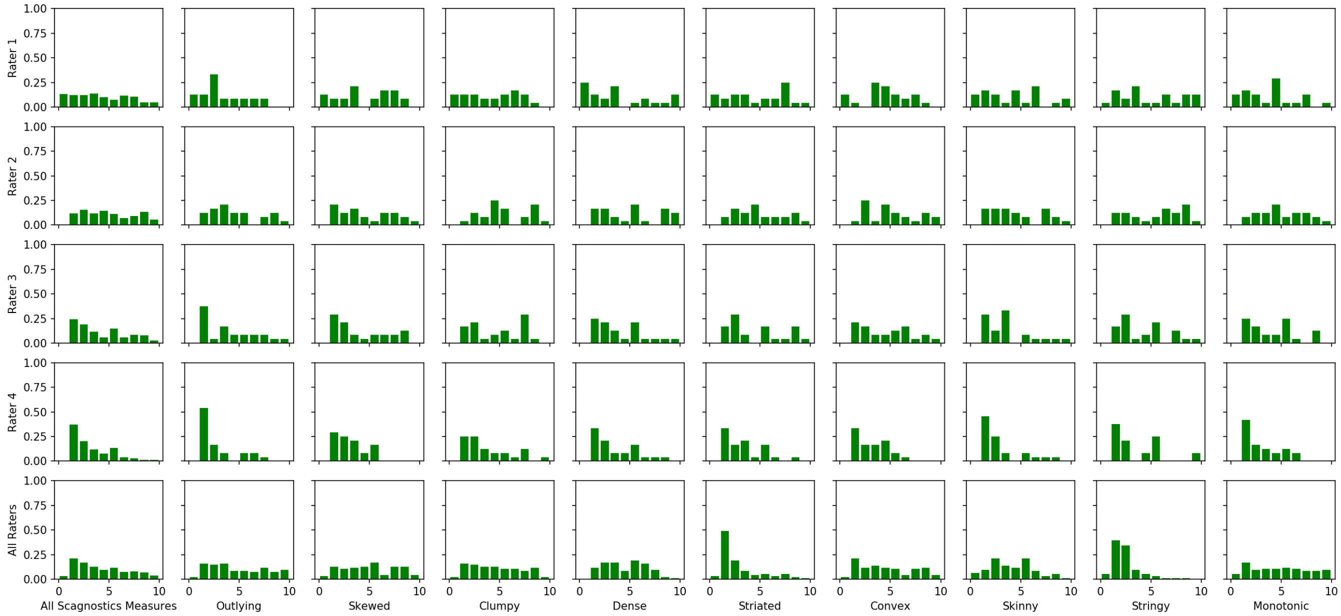

Figure 2: Human ratings for the 24 scatterplots in Set 1.

point-glyph size. Categorical variables are most often represented with discrete coloring schemes, different point-glyphs, and text labels.

Figure 1 shows the distribution of number of variables represented in each scatterplot as a function of publication year. The total height of each bar represents the total number of scatterplots collected from publications in that year. The colored sections divide the bar into the number of scatterplots containing two variables (grey), three variables (yellow), or four variables (blue). As we collected these scatterplots, we wondered if scatterplots published in earlier years might have less visual complexity than later scatterplots due to the increases in ease of plotting, graphics, printing, etc. We constructed this figure to qualitatively inspect for this kind of trend. Sure enough, many of the newer scatterplots have four variables, while hardly any of the earlier scatterplots do. (Note again that this result is specific to our scatterplot dataset; drawing general conclusions about scatterplots in general, or even astronomy-specific scatterplots, would require a larger and more systematic review of published scatterplots.)

## Part 2: Scagnostics by Human Raters

Nine undergraduate students took part in a scatterplot-rating study. Raters received a document containing brief instructions and 24 or 25 scatterplots selected from the catalog. The instructions included a scagnostics diagram from (Dang & Wilkinson, 2014b), as well as the following descriptions of each scagnostics measure:

- Outlying: Degree that a small number of points are separated away from a dense majority of points
- Skewed: Degree that the relative density of points is developed uniformly across the graph
- Clumpy: Degree that points are gathered in condense area with no excess points around
- Dense: Degree that points are heavily dispersed[1]
- Striated: Degree that points align into a low frequency wave with parallel lines
- Convex: Degree that points form the perimeter of a circle
- Skinny: Degree that plot resembled a concise/tight convex hyperbola
- Stringy: Degree that points align to a concise wave that has a positive slope
- Monotonic: Degree that plot only increases or decreases and is densest along that path

Note that these descriptions are quite loose, and likely do not match the "true" computational interpretations of each scagnostics measure. Because the raters were untrained in visual quality measures, the ratings obtained are based on interpretations of this single set of instructions and raters' visual perception rather than previous knowledge of scagnostics or related quality measures. Unlike professional astronomers, raters were unlikely to recognize any scatterplots from the catalog, avoiding bias from previously-formed impressions.

For each scatterplot, raters were instructed to assign a rating from one to ten for each of the scagnostics measures.

The scatterplot catalog was split into three sets of scatterplots, with Set 1 containing 24 scatterplots and Sets 2 and 3 containing 25 scatterplots each. Sets 1 and 2 received rat-

[1] Eight measures are consistently identified throughout scagnostics publications. One other measure is identified either as Straight (Wilkinson et al., 2005), Sparse (Wilkinson & Wills, 2008; Dang, Anand, & Wilkinson, 2013; Dang & Wilkinson, 2014a), or Dense (Dang & Wilkinson, 2014b) The Dense measure was chosen for this study as the most visually indicative to human raters.

ings from four raters each, and Set 3 received ratings from three raters. Rater 1, the experimenter who organized this part of the study, rated all three sets of scatterplots. This experimenter-rater was included as a practical means of acquiring one individual's rating of all 74 plots in this preliminary study and would be excluded from a larger, systematic study.

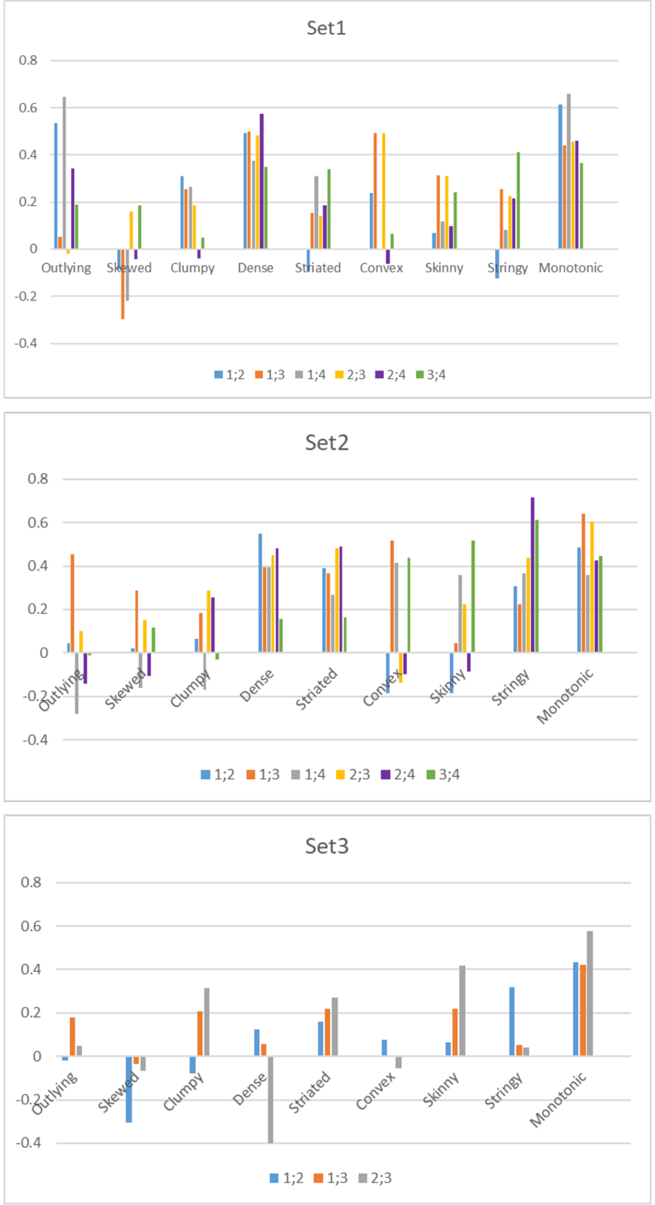

Figure 3: Kendall's rank correlation coefficient across ratings given by different raters in sets 1, 2, and 3. The coefficient was calculated over the ordered lists of each scagnostics rating given to each set of plots.

## Results

To look at overall rating patterns, we first examined distributions of the ratings themselves, without regard to agreement. Figure 2 shows histograms with the proportions of particular ratings (i.e., 1-10) given by different raters. We just show this histogram for Set 1; results for Sets 2 and 3 are similar.

It is clear that absolute ratings vary greatly for each individual. Rater 4 in particular seemed to assign primarily low ratings, while several distributions for each of the other raters show a roughly bimodal distribution. These strategies can be seen in the Clumpy results for each rater. The bottom row of Figure 2 shows that the ratings compiled from all raters are uniformly distributed across the 1-10 scoring scale for all scagnostics measures except for Striated and Stringy, which show a marked preference for lower scores. This roughly uniform rating distribution for most measures may reflect the raters' expectation of a full range of each measure to be present within the set of scatterplots shown to them.

Several raters, including Rater 1, used ratings of zero though they were not instructed to do so. These raters evaluated scagnostics measures on an 11-point scale rather than the expected 10-point scale, but their unmodified results were still considered in comparisons to other human raters and calculated scores.

We then computed Kendall's rank correlation coefficient for ratings within each scagnostic category (Kendall, 1938), using R (McLeod, 2011).

Here, despite the differences in absolute rating ranges, almost all raters showed positive agreement with one another for each scagnostics measure, as shown in Figure 3. Notably, raters tended to show the most disagreement in the Skewed measure and the most agreement in the Monotonic measure. Overall, no raters consistently agreed or disagreed with one another across all scagnostics measures.

## Part 3: Human vs. Calculated Scagnostics

Because the scatterplots in our catalog were obtained as images directly from a wide range of sources, the datasets used to construct them were unavailable. We used the image processing package Fiji to facilitate the generation of two-dimensional point location values for a subset of scatterplot images from the catalog (Schindelin et al., 2012; Rueden et al., 2017).

Scagnostics values for these point location values were calculated using a scagnostics package for R (Wilkinson & Anand, 2012). The locations of point-glyphs which overlap one another could not be determined using this method, so the criterion for inclusion in the subset for scagnostics measure calculation was that no point-glyphs overlap. Twenty-six scatterplots met this criterion: eight from Set 1, six from Set 2, and twelve from Set 3. Point location values were determined for all 26 plots.

Overlapping point-glyphs present a challenge to human visual analysis not captured by quality metrics calculated on their data values. There exist methods to represent to human viewers point-glyphs which would otherwise be hidden by overlapping, but these methods do not affect calculated quality metrics (Mayorga & Gleicher, 2013). Comparison be-

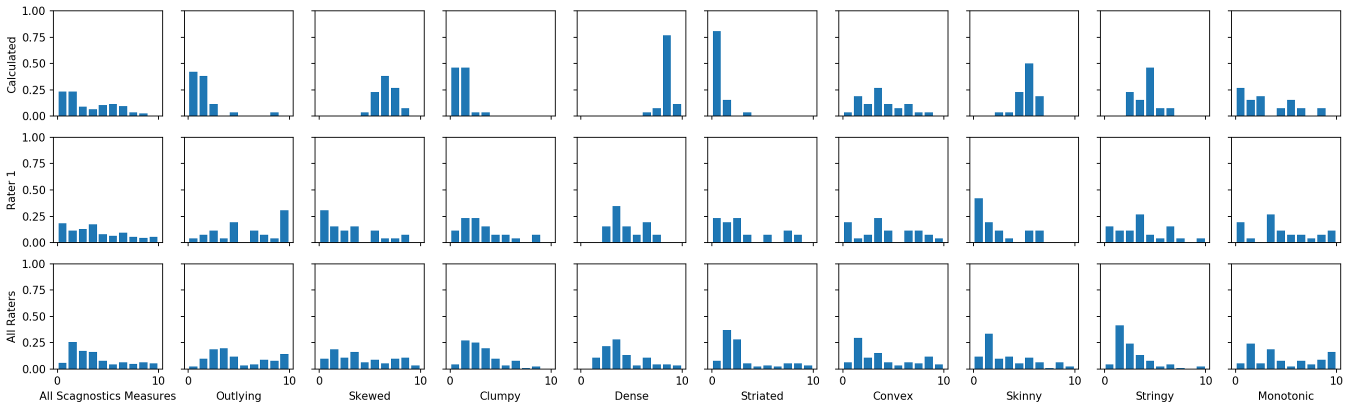

Figure 4: Calculated and human ratings for the 26 plots in the set of scatterplots with non-overlapping point-glyphs.

tween human-rated and calculated scagnostics values in this study is restricted to scatterplots without overlapping point-glyphs to avoid discrepancy in the visual information available to human raters and the data used for calculation.

The aforementioned R package calculated a value for the Sparse scagnostics measure, the opposite of the Dense measure used for human ratings (see Footnote 1 above). Values of the calculated scagnostics measure Sparse were subtracted from 1 to transform them into values of the Dense measure for comparison to human ratings. In order to directly compare calculated values on the range [0,1] and human ratings on the range [0,10], the calculated values for all measures were multiplied by 10.

### Results

As we did in Part 2, we first examined distributions of the ratings themselves, without regard to agreement. Figure 4 shows histograms with the proportions of particular ratings (i.e., 1-10) given by the computer-based scagnostics calculations (top), Rater 1 (middle), and over all raters (bottom), just to give a flavor of the comparisons.

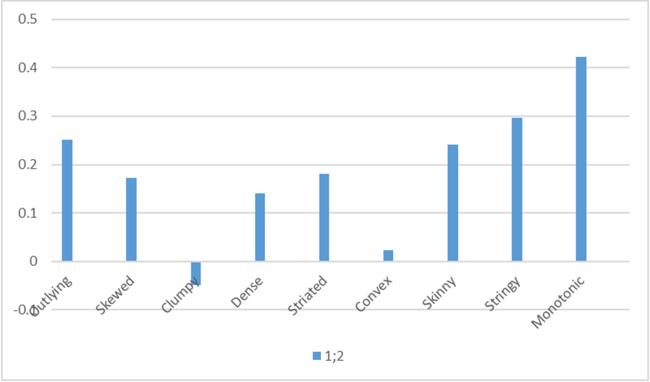

Figure 5: Kendall's rank correlation coefficient across calculated ratings and those of Rater 1. The coefficient was calculated over the ordered lists of each scagnostics rating given to the set of plots for which the data was able to be digitized.

Calculated scores were much less evenly distributed than human ratings. While agreement between human raters and calculated values can be seen in the Striated and Clumpy measures, overall it appears that human ratings of scagnostics measures is poorly matched by the calculated scagnostics values. Nonetheless, calculated and human ratings for all measures combined seem to agree on a skew towards lower values.

We then computed Kendall's rank correlation coefficient for ratings within each scagnostic category, as before, with results shown in Figure 5.

Positive agreement was seen overall between Rater 1 and calculated scagnostics values. Similar to between individual human raters, it appears agreement between the calculated and Rater 1's scagnostics ratings is strongest for the Monotonic measure. While Rater 1 and the calculated values had similar distributions of ratings for the Clumpy measure visible in Figure 4, it appears that it yielded the most disagreement of any measure when comparing individual scatterplots.

## Contributions and Next Steps

We have conducted an open-ended exploration of the visual properties of a set of 74 published astronomy scatterplots. Contributions of this work include:

- Creation of a dataset of 74 real-world, non-synthetic scatterplots used in astronomy
- Characterization of individual variations in the visual perception of scagnostics quality measures in scatterplots
- Comparison of human interpretations and calculated values of scagnostics measures in non-synthetic astronomy scatterplots

This exploratory study suggests quantifiable trends in human perception of visual qualities of real-world scatterplots which should be considered in the development of calculated visual quality metrics such as scagnostics.

Continued work will build on the results presented here to investigate how the visual properties of scatterplots held in long-term memory, as part of semantic, domain-specific knowledge, can help a computational cognitive system perform data visualization tasks. Future work will address this study's small sample size and unsystematic plot collection process to produce more robust, generalizable results. We expect that in the long term, findings from this work will not only help to uncover the cognitive processes that people use during data visualization but also will inform the design of innovative interactive data visualization systems and other cognitive support tools.

## Acknowledgments

This work was made possible by funding and support from the Vanderbilt Initiative for Data-intensive Astrophysics (VIDA), NSF REU grant PHY-1560035, and the Vanderbilt University School of Engineering (VUSE) Summer Research Program.

## References

Abello, J., Pardalos, P. M., & Resende, M. G. (2013). *Handbook of massive data sets* (Vol. 4). Springer.

Binney, J., & Merrifield, M. (1998). *Galactic astronomy*. Princeton University Press.

Burger, D., Stassun, K. G., Pepper, J., Siverd, R. J., Paegert, M., De Lee, N. M., & Robinson, W. H. (2013). Filtergraph: An interactive web application for visualization of astronomy datasets. *Astronomy and Computing*, *2*, 40–45.

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. (2006). Vistrails: visualization meets data management. In *Proceedings of the 2006 acm sigmod international conference on management of data* (pp. 745–747).

Dang, T. N., Anand, A., & Wilkinson, L. (2013). Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, *19*(3), 470–483.

Dang, T. N., & Wilkinson, L. (2014a). Scagexplorer: Exploring scatterplots by their scagnostics. In *Visualization symposium (pacificvis), 2014 ieee pacific* (pp. 73–80).

Dang, T. N., & Wilkinson, L. (2014b). Transforming scagnostics to reveal hidden features. *IEEE transactions on visualization and computer graphics*, *20*(12), 1624–1632.

Etemadpour, R., Linsen, L., Paiva, J. G., Crick, C., & Forbes, A. G. (2015). Choosing visualization techniques for multidimensional data projection tasks: A guideline with examples. In *International joint conference on computer vision, imaging and computer graphics* (pp. 166–186).

Honavar, V., Hill, M., & Yelick, K. (2016). Accelerating science: A computing research agenda. *arXiv preprint arXiv:1604.02006*.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Lehmann, D. J., Hundt, S., & Theisel, H. (2015). A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *it-Information Technology*, *57*(1), 11–21.

Mayorga, A., & Gleicher, M. (2013). Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, *19*(9), 1526–1538.

McLeod, A. I. (2011). Kendall rank correlation and mann-kendall trend test [Computer software manual].

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (2017). Imagej2: Imagej for the next generation of scientific image data. *BMC bioinformatics*, *18*(1), 529.

Sarikaya, A., & Gleicher, M. (2018). Scatterplots: Tasks, data, and designs. *IEEE transactions on visualization and computer graphics*, *24*(1), 402–412.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... others (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, *9*(7), 676–682.

Sedlmair, M., & Aupetit, M. (2015). Data-driven evaluation of visual quality measures. In *Computer graphics forum* (Vol. 34, pp. 201–210).

Wilkinson, L., & Anand, A. (2012). scagnostics: Compute scagnostics - scatterplot diagnostics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=scagnostics` (R package version 0.2-4)

Wilkinson, L., Anand, A., & Grossman, R. (2005). Graph-theoretic scagnostics.

Wilkinson, L., & Wills, G. (2008). Scagnostics distributions. *Journal of Computational and Graphical Statistics*, *17*(2), 473–491.