Clusters beat Trend!? Testing feature hierarchy in statistical graphics

Susan VanderPlas*
Department of Statistics and Statistical Laboratory, Iowa State University
and
Heike Hofmann
Department of Statistics and Statistical Laboratory, Iowa State University

June 24, 2016

Abstract

Graphics are very effective for communicating numerical information quickly and efficiently, but many of the design choices we make are based on subjective measures, such as personal taste or conventions of the discipline rather than objective criteria. We briefly introduce perceptual principles such as preattentive features and gestalt heuristics, and then discuss the design and results of a factorial experiment examining the effect of plot aesthetics such as color and trend lines on participants' assessment of ambiguous data displays. The quantitative and qualitative experimental results strongly suggest that plot aesthetics have a significant impact on the perception of important features in data displays.

Keywords: Visual inference, Lineup protocol, Preattentive Features, Saliency of Plot Aesthetics, User Study.

*The authors gratefully acknowledge funding from the National Science Foundation Grant # DMS 1007697. All data collection has been conducted with approval from the Institutional Review Board IRB 10-347

1 Introduction and Background

Limits on attention span, short term memory, and information storage mechanisms within the human brain make it difficult for us to process numerical information in raw form effectively. (Well-designed) data displays are much better suited for this kind of communication, as they serve as a form of external cognition (Zhang, 1997; Scaife and Rogers, 1996), ordering and visually summarizing data and thereby invoking our higher-bandwidth visual system.

One fantastic example of this phenomenon is the Hertzsprung-Russell (HR) diagram, which was described as "one of the greatest observational syntheses in astronomy and astrophysics" because it allowed astronomers to clearly relate the absolute magnitude of a star to its spectral classification, facilitating greater understanding of stellar evolution (Spence and Garrison, 1993). The data it displayed was previously available in several different tables, but when plotted within the same chart, information that was invisible in a tabular representation became immediately clear (Lewandowsky and Spence, 1989b). Graphical displays more efficiently utilize cognitive resources by reducing the burden of storing, ordering, and summarizing raw data. This frees bandwidth for higher levels of information synthesis and allows observers to note outliers, understand relationships between variables and form new hypotheses.

Statistical graphics are powerful because they efficiently and effectively convey numerical information, but there exists relatively sparse empirical information about how the human perceptual system processes these displays. Our understanding of the perception of statistical graphics is informed by general psychological and psychophysics research as well as more specific research into the perception of data displays (Cleveland and McGill, 1984). The approach taken in this paper is to utilize principles of preattentive perception, gestalt heuristics, and statistical lineups in order to better understand the effect of additional aesthetics on the perception of statistical graphics.

One relevant focus of psychological research is preattentive perception, that is, perception which occurs automatically in the first 200 ms of exposure to a visual stimulus (Treisman, 1985). Research into **preattentive perception** provides us with some information about the temporal hierarchy of graphical feature processing. Color, line orientation, and

shape are processed pre-attentively; that is, within 200 ms, it is possible to identify a single target in a field of distractors, if the target differs with respect to color or shape (Goldstein, 2009). Research by Healey and Enns (1999) extends this work, demonstrating that certain features of three-dimensional data displays are also processed pre-attentively. However, neither target identification nor three-dimensional data processing always translate into faster or more accurate inference about the data displayed, particularly when participants have to integrate several preattentive features to understand the data.

Feature detection at the attentive stage of perception has also been examined in the context of statistical graphics. Researchers have evaluated the perceptual implications of utilizing color, fill, shapes, and letters to denote categorical or stratified data in scatter plots. Cleveland and McGill (1984) ranked the optimality of these plot aesthetics based on response accuracy, preferring colors, amount of fill, shapes, and finally letters to indicate category membership. Lewandowsky and Spence (1989a) examined both accuracy and response time, finding that color is faster and more accurately perceived (except by individuals with color deficiency). Shape, fill, and discriminable letters (letters which do not share visual features, such as HQX) were identified as less accurate than color, while confusable letters (such as HEF) result in significantly decreased accuracy.

Gestalt psychology is another area of psychological research which examines perception as a holistic experience, establishing and evaluating mental heuristics used to transform visual stimuli into useful, coherent information. Gestalt rules of perception can be easily applied to statistical graphics, as they describe the way we organize visual input, focusing on the holistic experience rather than the individual perceptual features.

For example, rather than perceiving four legs, a tail, two eyes, two ears, and a nose, we perceive a dog. This is due to certain perceptual heuristics, which provide a "top-down" method of understanding visual stimuli by taking into account past experience.

The rules of perceptual organization relevant to graph perception in this experiment are:

- **Proximity**: two elements which are close together likely belong to a single unit.
- **Similarity**: the more aesthetics two elements share, the more likely they belong to a single unit.

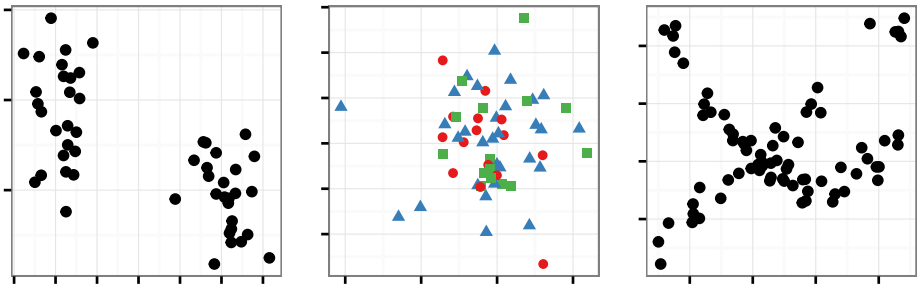


Figure 1: *Proximity* renders the fifty points of the first scatter plot as two distinct (and equal-sized) groups. Shapes and colors create three groups of points in the middle scatter plot by invoking the Gestalt principle of *Similarity*. *Good Continuation* renders the points in the scatter plot on the right hand side into two groups of points on curves: one a straight line with an upward slope, the other a curve that initially decreases and at the end of the range shows an uptick.

- **Good continuation**: elements which blend together smoothly are likely one unit.
- **Common region**: elements contained within a common region likely belong together.

A complete list of the rules of perceptual grouping can be found in Goldstein (2009).

The plots in Figure 1 demonstrate several of the gestalt principles which combine to order our perceptual experience from the top down. These laws help to order our perception of charts as well: points which are colored or shaped the same are perceived as belonging to a group (similarity), points within a bounding interval or ellipse are perceived as belonging to the same group (common region), and regression lines with confidence intervals are perceived as single units (continuity and common region).

The processing of visual stimuli utilizes low-level feature detection, which occurs automatically in the preattentive perceptual phase, and higher-level mental heuristics which are informed by experience. Both types of mental processes utilize physical location, color, and shape to organize perceptual stimuli and direct attention to graphical features which stand out.

Research on preattentive perception is important because features that are perceived pre-attentively require less mental effort to process from raw visual stimuli than non-preattentive features. Top-down gestalt heuristics are subsequently applied to the categorized features in order to make sense of the visual scene once the attentive stage of perception is reached.

Statistical graphics can be difficult to examine experimentally; qualitative studies rely on descriptions of the plot by participants who may not be able to articulate their observations precisely, while quantitative studies may only be able to examine whether the viewer can accurately read numerical information from the chart, instead of exploring the overall utility of the data display holistically. Here, we are describing the setup and results of a study using statistical lineup methodology to provide quantitative and qualitative information.

Statistical lineups are an important experimental tool for quantifying the significance of a finding in a graphical display. Lineups fuse commonly used psychological tests (target identification, visual search) (Vanderplas and Hofmann, 2016) with statistical hypothesis tests, thereby enabling formal experimental evaluation of statistical graphics.

Lineups are an experimental tool designed to serve as a visually conducted hypothesis test, separating significant effects from those that would be expected under a null hypothesis (Buja et al., 2009; Majumder et al., 2013; Hofmann et al., 2012; Wickham et al., 2010). A statistical lineup consists of (usually) 20 sub-plots, arranged in a grid (ten examples for lineups are shown in Figure 6). Of these sub-plots, or panels, one is the "target", generated from either real data or an alternate model (equivalent to H_A in hypothesis testing); the other 19 panels are generated either using bootstrap samples of the real data or by generating "true null" plots from the null distribution H_0. If a participant can identify the target from the field of distractors, this counts as evidence against the null hypothesis. Based on the number of evaluations and the number of target identifications, significance of a finding is then determined in the same sense as for a conventional hypothesis test. Performance on lineups has been shown to depend primarily on logical reasoning ability, and does not depend significantly on statistical training (Vanderplas and Hofmann, 2016).

Apart from the hypothesis testing construct, the use of statistical lineups to test statistical graphics conforms nicely to psychological testing constructs such as visual search (DeMita et al., 1981; Treisman and Gelade, 1980), where a single target is embedded in a field of distractors and response time, accuracy, or both are used to measure the complexity of the underlying psychological processes leading to identification.

In this paper we **modify the lineup protocol** by introducing a second target to each

lineup. The two targets represent two different, competing signals. An observer's choice then demonstrates empirically which signal is more salient. The probability distribution which underlies the hypothesis testing framework changes slightly with the addition of a second target (see Appendix B for more details). Single-target lineups rely on the hypothesis testing framework; the dual-target lineup is more similar to a (Bayesian) comparison of the relative strengths of two competing models with respect to a common null.

Cognitively, the presence of two targets leads to a dual-target search scenario (Fleck et al., 2010), which introduces a 'masking' effect where the more salient target is selected and the search for more targets stops ("satisfaction of search"), i.e. people tend to pick the more salient target and not notice the second target, even though in the absence of the more salient target they would have picked it out.

In the present study, participants were allowed to submit multiple selections to prevent any forced-choice scenario which might skew the results. However, only 0.6% of the evaluations resulted in an identification of both targets.

The search for multiple targets is a more demanding cognitive task (Cain et al., 2011) that is more sensitive to contextual effects (Adamo et al., 2015), but without any time constraints imposed by the experimental protocol, participants can be expected to identify at least one of the plots with accuracy comparable to a single-target search task.

Using this testing framework, we apply different aesthetics, such as color and shape, as well as plot objects which display statistical calculations, such as trend lines and bounding ellipses. These additional plot layers, discussed in more detail in the next section, are designed to emphasize one of the two competing targets and affect the overall visual signal of the target plot relative to the other target and the null plots. We expect that in a situation similar to the third plot of Figure 1, the addition of two trend lines would emphasize the "good continuation" of points in the plot, producing a stronger visual signal, even though the underlying data has not changed. Similarly, the grouping effect in the first plot in the figure should be enhanced if the points in each group are colored differently, as this adds similarity to the proximity heuristic. In plots that are ambiguous, containing some clustering of points as well as a linear relationship between x and y, additional aesthetic cues may "tip the balance" in favor of recognizing one type of signal over the other.

The study in this paper is designed to inform our understanding of the perceptual implications of these additional aesthetics, in order to provide guidelines for the creation of data displays which provide visual cues consistent with gestalt heuristics and preattentive perceptual preferences.

The next section discusses the particulars of the experimental design, including the data generation model, plot aesthetics, selection of color and shape palettes, and other important considerations. Experimental results are presented in Section 3, and implications and conclusions are discussed in Section 4.

2 Experimental Setup and Design

In this section, we discuss the models generating data for the two types of signal plots and the null plots, the selection of plot aesthetic combinations and aesthetic values, and the design and execution of the experiment.

2.1 Data Generation

Conventional lineups require a single "target" data set and a method for generating null plots. When utilizing real data for target plots, null plots are often generated through permutations.

Here, it is possible to generate true null plots from a null model that do not depend on the data used in the target plot. This experiment measures two competing gestalt heuristics, proximity and good continuation, using two data-generating models. Both models provide data in the same range of values in X and Y; M_C generates data with K clusters, while M_T generates data with a positive linear relationship between X and Y. Null data sets are created using a mixture model M_0 which combines M_C and M_T. In order to facilitate mixing these two models, controls on cluster centers generated by M_C ensure that X and Y have a positive linear relationship with a correlation $\rho \in (0.25, 0.75)$, similar to the linear relationship between data sets generated by M_0.

These constraints provide some assurance that participants who select a plot with data generated from M_T are doing so because of visual cues indicating a linear trend (rather than

a lack of clustering compared to plots with data generated from M_0), and participants who select a plot with data generated from M_C are doing so because of visual cues indicating clustering, rather than a lack of a linear relationship relative to plots with data generated from M_0.

2.1.1 Regression Model M_T

This model has the parameter σ_T to reflect the amount of scatter around the trend line. It generates N points $(x_i, y_i), i = 1, ..., N$ where x and y have a positive linear relationship. The data generation mechanism is as follows:

Algorithm 2.1

Input Parameters: sample size N, σ_T standard deviation around the linear trend
Output: N points, in form of vectors x and y.

1. *Generate $\tilde{x}_i$, $i = 1, ..., N$, as a sequence of evenly spaced points in $[-1, 1]$.*
   This step ensures that the full range in x is used, which in turn keeps the ratio of x to y range constant.

2. *Jitter $\tilde{x}_i$ by adding small uniformly distributed perturbations to each of the values: $x_i = \tilde{x}_i + \eta_i$, where $\eta_i \sim Unif(-z, z)$, $z = \frac{2}{5(N-1)}$.*

3. *Generate y_i as a linear regressand of x_i: $y_i = x_i + e_i$, $e_i \sim N(0, \sigma_T^2)$. Several values of σ_T^2 are shown in Figure 2.*

4. *Center and scale x_i, y_i.*

We compute the coefficient of determination for all of the plots to assess the amount of linearity in each panel, computed as

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{1}$$

where TSS is the total sum of squares, $TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{N} e_i^2$, the residual sum of squares. The expected value of the coefficient of determination $E[R^2]$ in this scenario is

$$E\left[R^2\right] = \frac{1}{1 + 3\sigma_T^2},$$

because $E[RSS] = N\sigma_T^2$ and $E[TSS] = \sum_{i=1}^{N} E\left[y_i^2\right]$ (as $E[Y] = 0$), where

$$E\left[y_i^2\right] = E\left[x_i^2 + e_i^2 + 2x_i e_i\right] = \frac{1}{3} + \sigma_T^2.$$

The use of R^2 to assess the strength of the linear relationship (rather than the correlation) is indicated because human perception of correlation strength more closely aligns with R^2 (Bobko and Karren, 1979; Lewandowsky and Spence, 1989b).

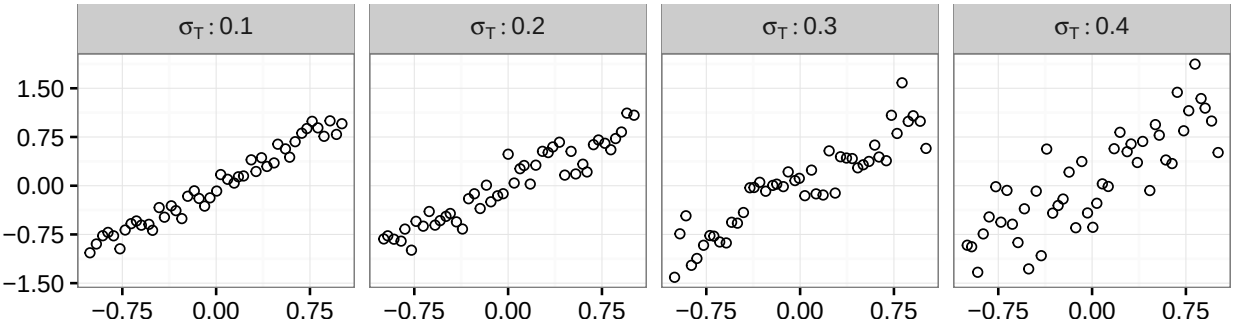


Figure 2: Set of scatter plots showing one draw each from the trend model M_T for parameter values of $\sigma_T \in \{0.1, 0.2, 0.3, 0.4\}$.

2.1.2 Cluster Model M_C

We begin by generating K cluster centers on a $K \times K$ grid, then we generate points around selected cluster centers.

Algorithm 2.2

Input Parameters: N points, K clusters, σ_C cluster standard deviation

Output: N points, in form of vectors x and y.

1. *Generate cluster centers (c_i^x, c_i^y) for each of the K clusters, $i = 1, ..., K$:*

   (a) *in form of two vectors c^x and c^y of permutations of $\{1, ..., K\}$, such that*

   (b) *the correlation between cluster centers $cor(c^x, c^y)$ falls into a range of $[.25, .75]$.*

2. *Center and standardize cluster centers (c^x, c^y):*

$$\tilde{c}_i^x = \frac{c_i^x - \bar{c}}{s_c} \quad \text{and} \quad \tilde{c}_i^y = \frac{c_i^y - \bar{c}}{s_c},$$

where $\bar{c} = (K+1)/2$ and $s_c^2 = \frac{K(K+1)}{12}$ for all $i = 1, ..., K$.

3. *For the K clusters, we want to have nearly equal sized groups, but allow some variability. Cluster sizes $g = (g_1, ..., g_K)$ with $N = \sum_{i=1}^{K} g_i$, for clusters $1, ..., K$ are therefore determined as a draw from a multinomial distribution:*

$$g \sim Multinomial\ (K, p)\ where\ p = \tilde{p} / \sum_{i=1}^{K} \tilde{p}_i,\ for\ \tilde{p} \sim N\left(\frac{1}{K}, \frac{1}{2K^2}\right).$$

4. *Generate points around cluster centers by adding small normal perturbations:*

$$\begin{aligned} x_i &= \tilde{c}^x_{g_i} + e^x_i,\ where\ e^x_i \sim N(0, \sigma_C^2), \\ y_i &= \tilde{c}^y_{g_i} + e^y_i,\ where\ e^y_i \sim N(0, \sigma_C^2). \end{aligned}$$

5. *Center and scale x_i, y_i.*

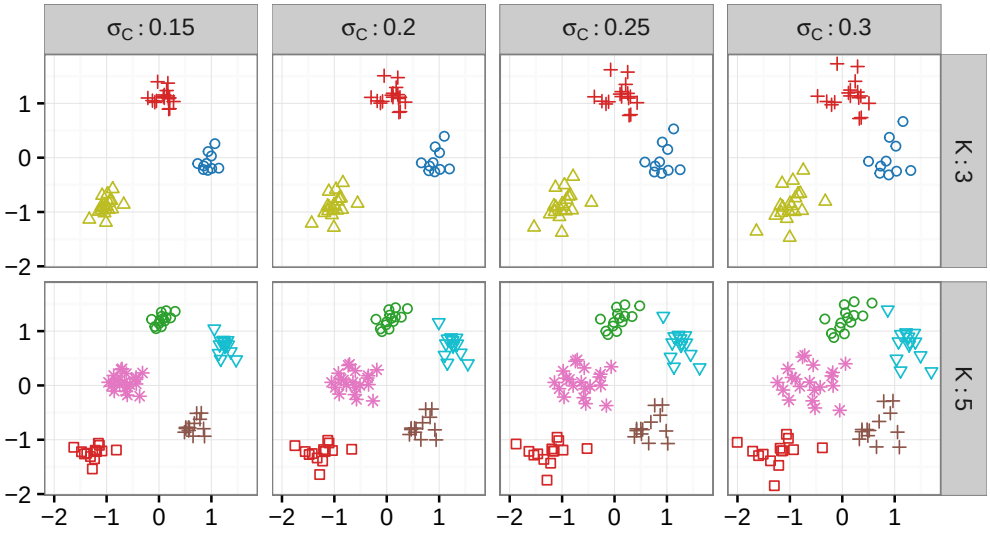


Figure 3: Scatter plots of clustering output for different inner cluster spread σ_C (left to right) and different number of clusters K (top and bottom), generated using the same random seed at each parameter setting. The colors and shapes shown are those used in the lineups for $K = 3$ and $K = 5$.

As a measure of cluster cohesion we use a coefficient to assess the amount of variability within each cluster, compared to total variability. Note that for the purpose of clustering, variability is measured as the variability in both x and y from a common mean, i.e. we implicitly assume that the values in x and y are on the same scale. This ensures that $\sigma_C$is a scaling parameter that regulates the amount of cluster cohesion (see Figure 3).

For two numeric variables x and y and grouping variable g with $g_i \in \{1, ..., K\}, i = 1, ..., n$, we compute the *cluster index* C^2 as follows: let $j(i)$ be the function that maps index $i = 1, ..., n$ to one of the clusters $1, ..., K$ given by the grouping variable g. Then for each level of g, we find a cluster center as $\bar{x}_{j(i)}$ and $\bar{y}_{j(i)}$, and we determine the strength of the clustering by comparing the within cluster variability with the overall variability:

$$\begin{aligned} C^2 &= 1 - \frac{CSS}{TSS}, \\ CSS &= \sum_{i=1}^{n} \left(x_{j(i)} - \overline{x}_{j(i)}\right)^2 + \left(y_{j(i)} - \overline{y}_{j(i)}\right)^2, \\ TSS &= \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 + \left(y_i - \bar{y}\right)^2. \end{aligned} \tag{2}$$

The cluster index C^2, which is approximately inversely linear in σ_C^2, measures the actual amount of clustering in the generated data.

2.1.3 Null Model M_0

The generative model for null data is a mixture model M_0 that draws $n_c \sim \text{Binomial}(N, \lambda)$ observations from the cluster model, and $n_T = N - n_c$ from the regression model M_T. Observations are assigned to specific clusters using hierarchical clustering, which creates groups consistent with any structure present in the generated data. This provides a plausible grouping for use in aesthetic and statistics requiring categorical data (color, shape, bounding ellipses).

Null data in this experiment is generated using $\lambda = 0.5$, that is, each point in a null data set is equally likely to have been generated from M_C and M_T to ensure maximal distance of the null plots from either target.

2.1.4 Parameters used in Data Generation

Models M_C, M_T, and M_0 provide the foundation for this experiment; by manipulating cluster standard deviation σ_C and regression standard deviation σ_T for varying numbers of clusters $K = 3, 5$, we systematically control the statistical signal present in the target plots and generate corresponding null plots that are mixtures of the two distributions. For

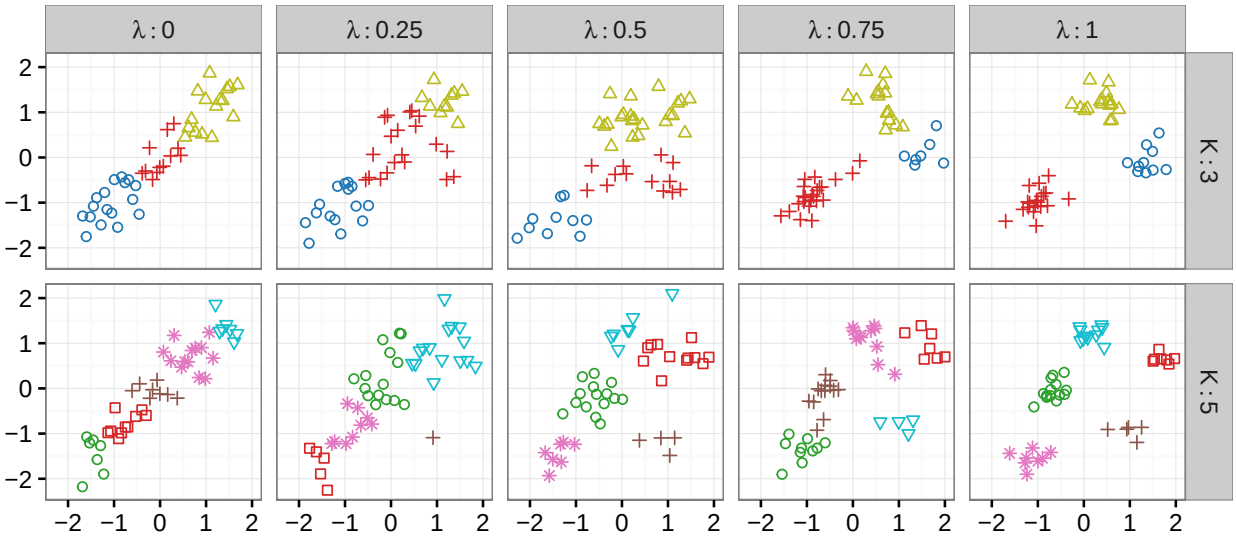


Figure 4: Scatter plots of data generated from M_0 using different values of λ, generated using the same random seed at each λ value.

each parameter set $\{K, N, \sigma_C, \sigma_T\}$, as described in Table 1, we generate a lineup data set consisting of one set drawn from M_C, one set drawn from M_T, and 18 sets drawn from M_0.

Parameter	**Description**	**Choices**
K	# Clusters	3, 5
N	# Points	$15 \cdot K$
σ_T	Scatter around trend line	.25, .35, .45
σ_C	Scatter around cluster centers	.25, .30, .35 ($K = 3$) .20, .25, .30 ($K = 5$)

Table 1: Parameter settings for generation of lineup data sets.

The parameter values were chosen in an approach similar to that taken in Roy Chowdhury et al. (2014): for each combination of $\sigma_T \in \{0.2, 0.25, ..., 0.5\}$, $\sigma_C \in \{0.1, 0.15, ..., 0.4\}$, and $K \in \{3, 5\}$ we simulated 1000 lineup data sets. Then trend and cluster strength indices, R^2 and C^2, were computed for the simulated target plots, and compared to the most extreme value for each of the 18 null plots of the same lineup data.

The resulting distributions allow us to objectively assess the difficulty of detecting the target data sets computationally (without relying on human perception) within the full parameter space. That is, a target plot with $R^2 = 0.95$ is very easy to identify when surrounded by null plots with $R^2 = 0.5$, while null plots with $R^2 = 0.9$ make the target plot more difficult to identify.

Figure 5 shows densities of each measure computed from the maximum of 18 null plots compared to the measure in the signal plot for one combination of parameters. There is some overlap in the distribution of R^2 for the null plots compared to the target plot displaying data drawn from M_T. As a result, the distribution of the cluster statistic values are more easily separated from the null data sets than the distribution of the trend statistic, e.g. $\sigma_C = 0.20$ is producing cluster target data sets that are a bit easier to identify numerically than trend targets with a parameter value of $\sigma_T = 0.25$.

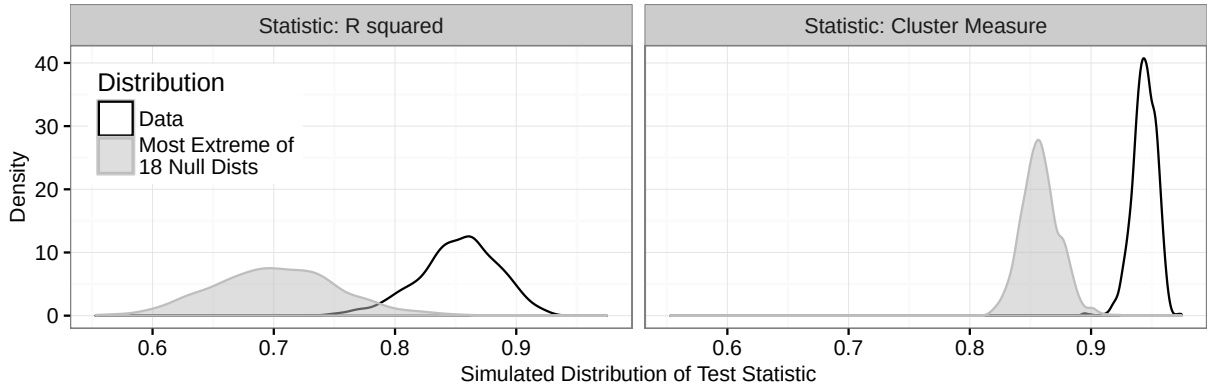


Figure 5: Density of test statistics measuring trend strength and cluster strength for target distributions and null plots based on 1,000 draws of lineup data with $\sigma_T = 0.25, \sigma_C = 0.20$ and $K = 3$.

Graphical summaries of simulation results for a range of values for σ_C and σ_T are provided in Appendix A. Using information from the simulation, we identify values and generate lineup data sets for each σ_T and σ_C (as shown in Table 1) corresponding to "easy", "medium" and "hard" numerical comparisons between corresponding target data sets and null data sets. It is important to note that the numerical measures we have described in equations (1) and (2) only provide information on the numerical discriminability of the target data sets from the null data sets; the simulation cannot provide us with exact infor-

mation on the perceptual discriminability. It has been established that human perception of scatter plots does not replicate statistical measures exactly (Bobko and Karren, 1979; Mosteller et al., 1981; Lewandowsky and Spence, 1989b).

Each of the generated data sets is then plotted as a lineup using aesthetics which emphasize clusters and/or linear relationships, in order to experimentally determine how these aesthetics change a participant's preference and ability to identify each target plot.[1] The next section describes the aesthetic combinations and their anticipated effect on participant responses.

		Trend Emphasis		
	Strength	0	1	2
Cluster Emphasis	0	None	Trend	Trend + Error
	1	Color Shape	Color + Trend	
	2	Color + Shape Color + Ellipse		Color + Ellipse + Trend + Error
	3	Color + Shape + Ellipse		

Table 2: Plot aesthetics and statistical layers which impact perception of statistical plots, according to gestalt theory.

2.2 Lineup Rendering

2.2.1 Plot Aesthetics

Gestalt perceptual theory suggests that perceptual features such as shape, color, trend lines, and boundary regions modify the perception of ambiguous graphs, emphasizing clustering in the data (in the case of shape, color, and bounding ellipses) or linear relationships (in the case of trend lines and prediction intervals), as demonstrated in Figure 1. For each data set we examine the effect of the plot aesthetics (color, shape) and statistical layers (trend line,

[1]Code for stimuli generation, simulation code and results are provided in the github repository at `http://github.com/srvanderplas/FeatureHierarchy`.

boundary ellipses, prediction intervals) shown in Table 2 on target identification. Examples of these plot aesthetics are shown in Figure 6.

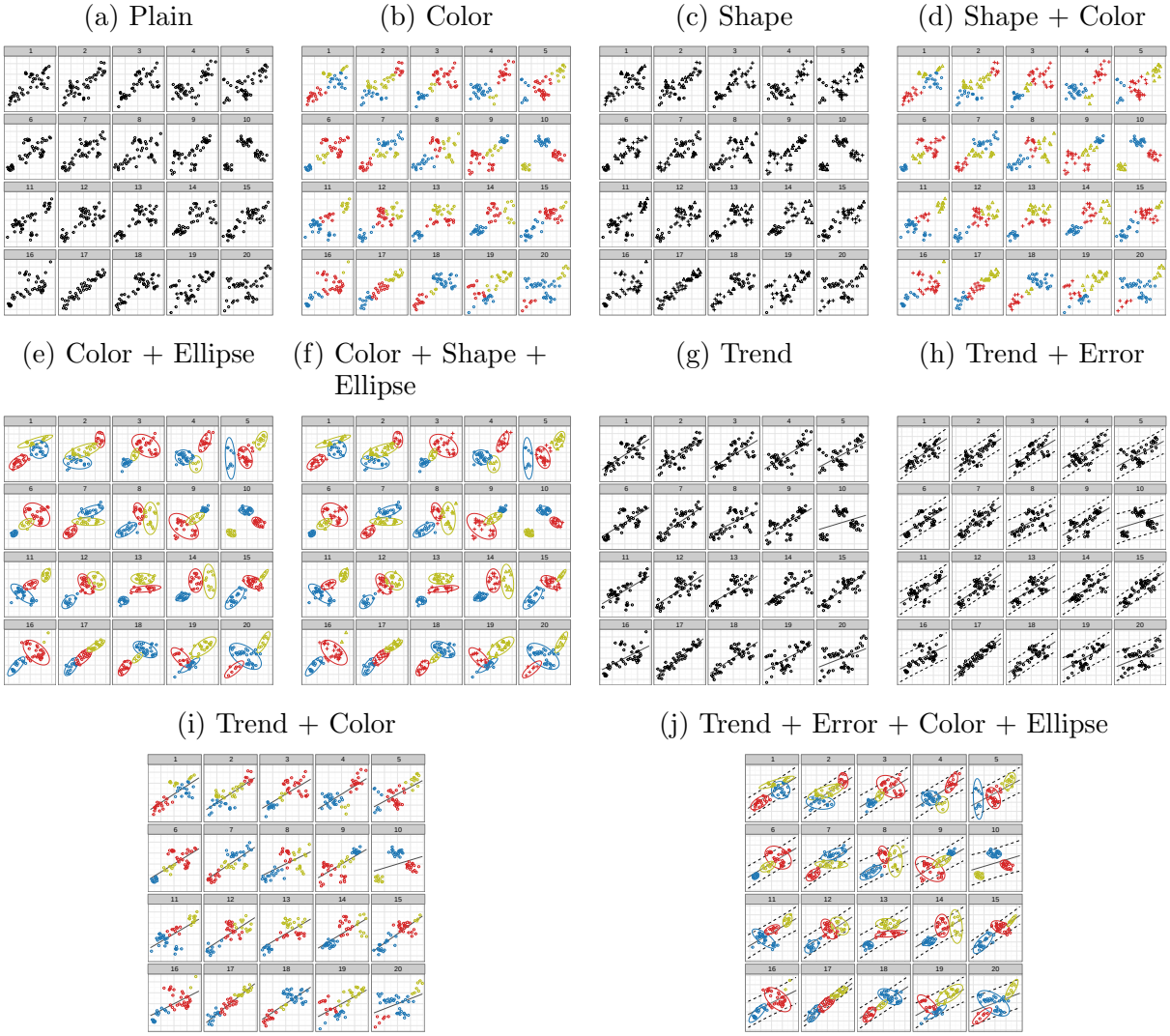


Figure 6: Set of lineups for $K = 3$, $\sigma_T = 0.25$ and $\sigma_C = 0.20$. Each lineup shows one of the 10 designs comprised of feature combinations compared in this study.

We expect that relative to a plot with no extra aesthetics or statistical layers, the addition of color, shape, and 95% boundary ellipses increases the probability of a participant selecting the target plot with data generated from M_C, the cluster model, and that the addition of these aesthetics decreases the probability of a participant selecting the target plot with data generated from M_T, the trend model. In addition, we expect that with multiple aesthetics which emphasize the clustering of data, participants will have a higher

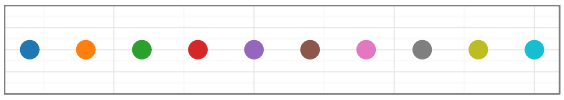
probability of selecting the cluster target plot (Eriksen and Hake, 1955).

Similarly, we expect that relative to a plot with no extra aesthetics or statistical layers, the addition of a trend line and prediction interval ("error band") increases the probability of a participant selecting the target plot with data generated from M_T, the trend model, and decreases the probability of a participant selecting the target plot with data generated from M_C, the cluster model.

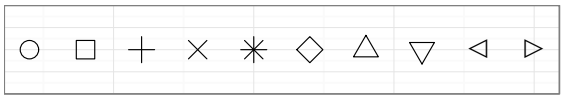
2.2.2 Color and Shape Palettes

Colors and shapes used in plot designs were selected based on the results of previous studies (Demiralp et al., 2014; Robinson, 2003; Healey et al., 1996) in order to maximize preattentive feature differentiation. Demiralp et al. (2014) provide sets of 10 colors and 10 shapes, with corresponding distance matrices, determined by user studies. Using these perceptual kernels for shape and color, we identified a maximally differentiable set of 3 and 5 colors each.

Figure 7: Color and shape palettes investigated for differentiability in Demiralp et al. (2014).



(a) Color Palette. For the present study gray was removed from the palette to make the experiment more inclusive of participants with color deficiency.



(b) Shape palette. Due to varying point size between Unicode and non-Unicode characters, the last two shapes were not used in this study.

The color palette used in Demiralp et al. (2014) and shown in Figure 7a is derived from colors available in the visualization software Tableau (Hanrahan, 2003).

In order to produce experimental stimuli accessible to the approximately 4% of the population with red-green color deficiency (Gegenfurtner and Sharpe, 2001), we removed the gray hue from the palette, as gray is often difficult to distinguish from red and green for those with protanopia and deuteranopia, the most common types of colorblindness. This modification also resulted in maximally different color combinations that did not include red-green combinations, which would also impact the ability of color-deficient individuals

to participate fully in this experiment.

Software compatibility issues led us to exclude two shapes used in Demiralp et al. (2014) and shown in Figure 7b. The left and right triangle shapes (available only in unicode within R) were excluded from our investigation due to size differences between unicode and non-unicode shapes. After optimization over the sum of all pairwise distances, the maximally different shape sequences for the 3 and 5 cluster data sets also conform to the guidelines in Robinson (2003): for $K = 3$ the shapes are from Robinson's group 1, 2, and 9, for $K = 5$ the shapes are from groups 1, 2, 3, 9, and 10. Robinson's groups are designed so that shapes in different groups show differences in preattentive properties; that is, they are easily distinguishable. In addition, all shapes are non-filled, making them consistent with one of the simplest solutions to over-plotting of points in the tradition of Tukey (1977); Cleveland (1994) and Few (2009). For this reason we abstained from the additional use of alpha-blending of points to diminish the effect of over-plotting in the plots.

2.3 Experimental Design

The study is designed hierarchically, as a factorial experiment for combinations of σ_C, σ_T, and K, with three replicates at each parameter combination. These parameters are used to generate lineup data sets which serve as blocks for the plot aesthetic level of the experiment; each data set is rendered with every combination of aesthetics described in Table 2. Participants are assigned to generated plots according to an augmented balanced incomplete block scheme: each participant is asked to evaluate 10 plots, which consist of one plot at each combination of σ_C and σ_T, randomized across levels of K, with one additional plot providing replication of one level of $\sigma_C \times \sigma_T$. Each of a participant's 10 plots presents a different aesthetic combination.

2.4 Hypotheses

The primary purpose of this study is to understand how visual aesthetics affect signal detection in the presence of competing signals. We expect that plot modifications which emphasize similarity and proximity, such as color, shape, and 95% bounding ellipses, increase the probability of detecting the clustering relationship, while plot modifications

which emphasize good continuation, such as trend lines and prediction intervals, increase the probability of detecting the linear relationship.

A secondary purpose of the study is to relate signal strength (as determined by data set parameters σ_C, σ_T, and K) to signal detection in a visualization by a human observer.

2.5 Participant Recruitment

Participants were recruited using Amazon's Mechanical Turk service(Amazon, 2010), which connects interested workers with "Human Intelligence Tasks" (HITs), which are (typically) short tasks that are not easily automated. This service has been found to produce results with reasonably high data quality that alos compare well to laboratory studies Crump et al. (2013). In our study, only workers with at least 100 previous HITs at a 95% successful completion rate were allowed to sign up for completing the task. These restrictions reduce the amount of data cleaning required by ensuring that participants have experience with the Mechanical Turk system, as well as a vested interest in performing well.

Participants had to complete a pre-trial before being able to access the experiment. The lineups used in the pre-trial contained only a single target,and participants had to correctly identify the target in at least two lineups. The webpage used to collect data from Amazon Turk participants is available at `https://erichare.shinyapps.io/lineups/`. No data was recorded from the pre-trial because participants had not provided informed consent at this point.

Once participants completed the example task and provided informed consent, they could accept the HIT through Amazon and were directed to the main experimental task.

2.6 Task Description

Participants were required to complete ten lineups, answering "Which plot is the most different from the others?". Participants were asked to provide a short reason for their choice, such as "Strong linear trend" or "Groups of points", and to rate their confidence in their selection from 0 (least confident) to 5 (most confident). After the first question, basic demographic information was collected: age range, gender, and highest level of education.

Throughout the experiment, participants were not informed about the inclusion of a

second target into the lineup plots. The small number of participants choosing multiple plots in their answer suggests that most participants did not discover that two target plots were present in each lineup and were thus naive to the true purpose of the experiment.

3 Results

3.1 General results & Demographics

Data collection was conducted over a 24 hour period, during which time 1356 individuals completed 13519 unique lineup evaluations. Participants who completed fewer than 10 lineups were removed from the study (159 participants, 1060 evaluations), and lineup evaluations in excess of 10 for each participant were also removed from the study (421 evaluations). After these data filtration steps, our data consist of 12010 trials completed by 1201 participants.

Of the participants who completed at least 10 lineup evaluations, 61% were male, relatively younger than the US population and relatively well educated (see Figure 8). Each plot was evaluated by between 11 and 36 individuals (Mean: 22.24, SD: 4.64). 82.9% of the participant evaluations identified at least one of the two target plots successfully (Trend: 26.8%, Cluster: 56.8%).

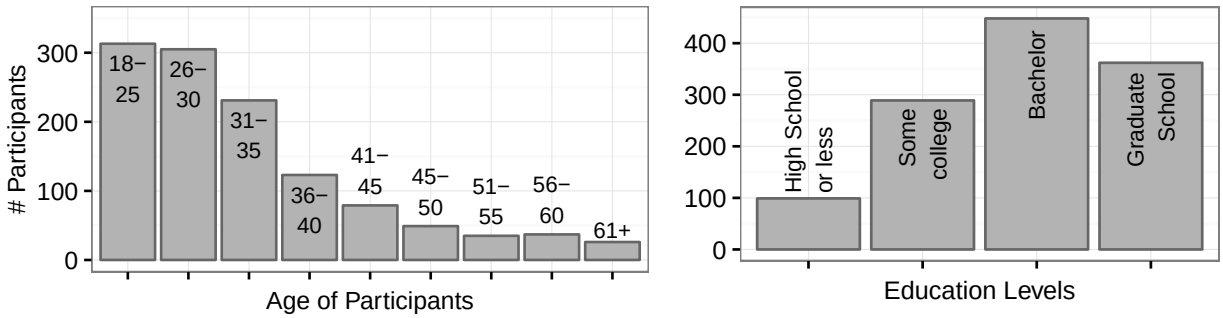


Figure 8: Basic demographics of participants.

From Figure 9 we see that participants identified more cluster targets than trend targets (there were more aesthetics expected to emphasize clustering in the data), but also did not primarily identify one target type over the other. Generally a participant selected both targets over the course of ten lineup trials.

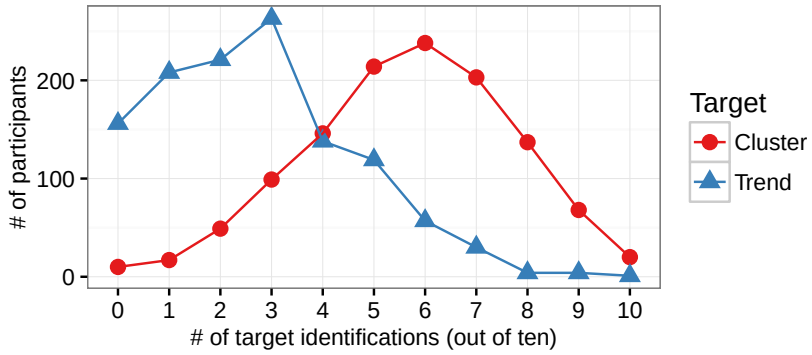


Figure 9: Target identifications by participants. Generally, participants are not primed for one target over the other.

In the evaluation of the collected data, we modeled accuracy rates, response times, participant confidence, and the selection of cluster vs. trend targets in detail. A thorough discussion of these models and the corresponding results is provided in Appendix C. Only the models describing target plot selection are discussed here.

For each design (aesthetic combination), we first consider the probability that a participant selects one of the two target plots, and then we consider the conditional probability of selecting the cluster target over the trend target.

3.2 Target Plot Identifications

Lineup evaluations provide us with a simple to assess measure of accurateness: we can say, that a lineup is evaluated 'accurately', if the data plot in the lineup was identified (or here, if at least one of the target plots in the lineup was identified). Accuracy is therefore a measure of both a participant's skill and the strength of the signal in the data/target plot. In a model, we accommodate for different skills of participants by including a random intercept, because in this situation we are much more interested in a further investigation of a target's signal strength. As planned, an increase of either within cluster variability (s_C) or scatter around the trend line (s_T) leads to a significant drop in accuracy. Over the course of the experiment, accuracy drops by a small, but significant amount with each additional evaluation, probably due to a fatigue effect. The different designs have significantly different accuracies associated with them: Figure 10 shows an overview of the

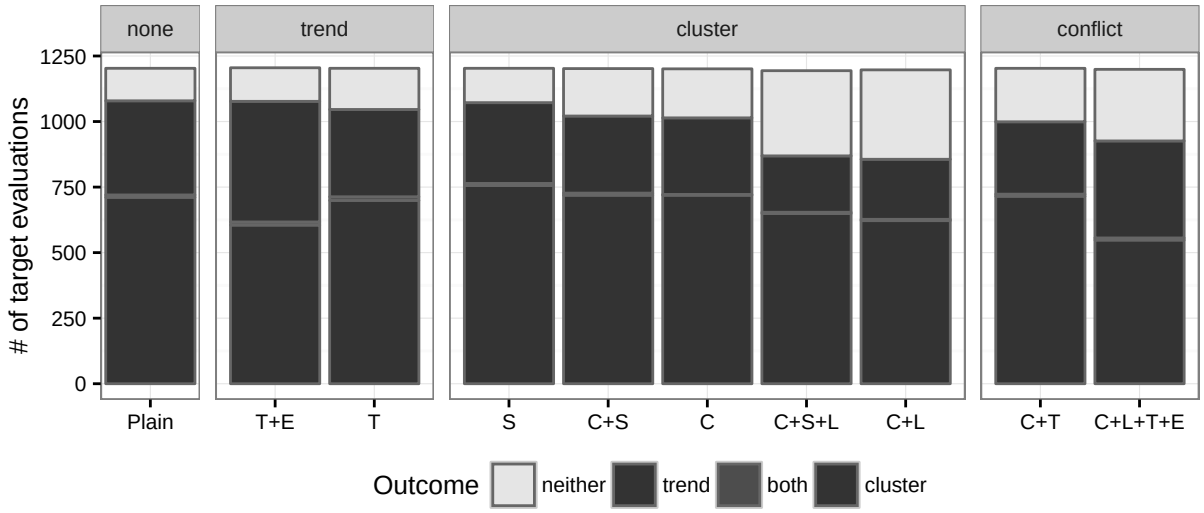


Figure 10: In dark: number of evaluations by design, in which at least one of the targets was identified. Each of the dark areas is split into two, according to the type of target, with evaluations where both targets were identified between the two. Due to the design of the experiment, each design was evaluated almost the same number of times (between 1195 and 1208 times, outlined rectangles). T = Trend, E = Error, S = Shape, C = Color, and L = Ellipse.

number of evaluations by design (outlines) and the number of times participants chose at least one of the targets (dark shaded areas). Designs associated with clustering as shown in Table 2 lead to significantly fewer correct evaluations ($\chi^2_9 = 389$, p-value < 0.0001). This might have been due to an unintended signal in the lineups because of imbalances in the group allocation of clusters, which are emphasized strongly by color and ellipses. We discuss this in more detail in Section 3.4 based on participants' reasoning.

Additional information about the model structure, numeric values of the parameter estimates, and more interpretation are provided in Appendix C.1.

3.3 Face-Off: Trend versus Cluster

In order to assess which of the two stimuli dominated in each of the designs, we first model the probability that a participant identified at least one of the targets (9959 trials).

For all trials in which at least one of the targets was correctly identified, we compare the probability of selecting the cluster target generated by M_C with the probability of selecting

the trend target generated by M_T. Define C_{ijk} to be the event

$$\{\text{Participant } k \text{ selects the cluster target for data set } j \text{ with aesthetic set } i\}$$

and T_{ijk} to be the analogous selection of the trend target. We model the cluster versus trend decision using a logistic regression with a random effect for each data set to account for different difficulty levels in the generated data, and a random effect for participant to account for skill level, as shown in equation 3.

$$\text{logit } P(C_{ijk}|C_{ijk} \cup T_{ijk}) = \mathbf{W}\alpha + \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta, \tag{3}$$

where

- α is a vector of fixed effects $(\mu, \alpha_T, \alpha_C, \alpha_K)$. μ is a baseline average of the probability to pick the cluster target over the trend target. α_T and α_C are parameters which describe the effect of the standard error around trend lines $s_T \in \{0.25, 0.35, 0.45\}$ and within cluster variability $s_C \in \{0.2, 0.25, 0.3, 0.35\}$, and α_K is the effect of the number of clusters $K \in \{3, 5\}$.
- β_i describe plot aesthetics,
- $\gamma_j \overset{iid}{\sim} N(0, \sigma^2_{\text{data}})$, random effect for data set specific characteristics,
- $\eta_k \overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right)$, random effect for participant characteristics.

We assume that the errors associated with a single trial evaluation, ϵ_{ijk}, are normally distributed with variance σ^2_e. We also assume that random effects for data set and participant are orthogonal. Additional model details, estimated fixed effects, and significance tests are provided in Appendix C.4.

The estimated log odds of a decision in favor of cluster over trend target for each of the designs are shown in Figure 11. From left to right the (log) odds of selecting the cluster target over the trend target increase. As hypothesized, the strongest signal for identifying groups, is color + shape + ellipse, while trend + error results in the strongest signal in favor of trends. Most of the effects are not significantly different, as seen in the letter values (Piepho, 2004) based on Tukey's Post Hoc difference tests on the left hand side of

the figure, representing pairwise comparisons of all of the designs, adjusted for multiple comparison. The estimates for parameters α_C and α_T, quantifying the effect of increased variability within clusters (s_C) and around the trend line (s_T), are highly significant and work as hypothesized: with an increase in the variability, the strength of the target's signal decreases and correspondingly the probability for detecting the corresponding target decreases significantly (see Table 5 in Appendix C.4 for exact numbers and a discussion of further co-variates).

The addition of interaction effects between plot design and control parameters s_C or s_T does not significantly improve the fit of the model. That is, the selection of the cluster or trend target does not appear to be influenced by the combination of a plot design and the difficulty of a lineup (as controlled by the parameters s_C and s_T), at least within the ranges of parameters examined in this experiment.

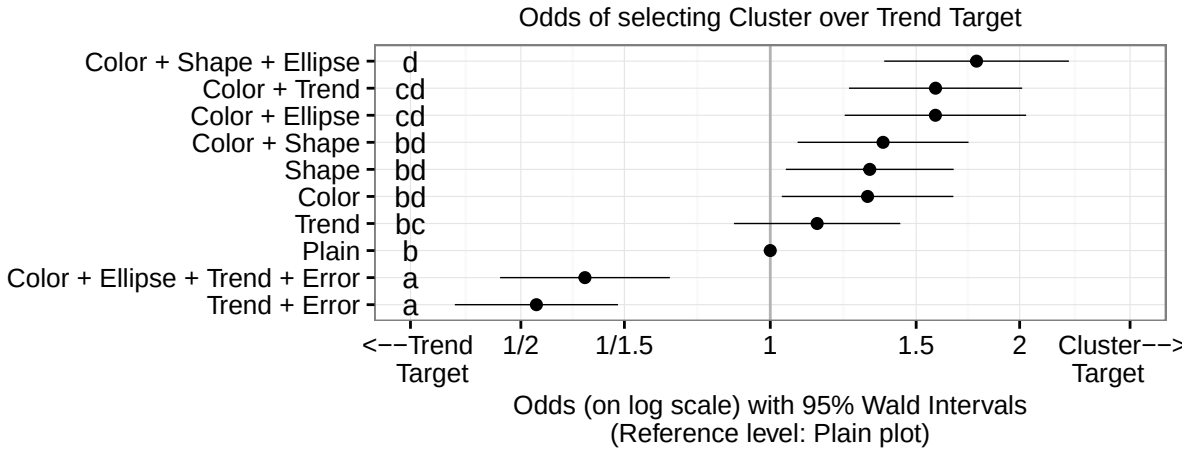


Figure 11: Estimated odds of decision for cluster versus trend target based on evaluations that resulted in the identification of one of these targets. Designs are significantly different if they do not share a letter as given on the left hand side of the plot.

Examining the model results from the perspective of Gestalt heuristics, it is clear that the similarity/proximity effect, as indicated by spatial clustering and aesthetics such as color and shape, dominates the equation, including dominating the color + trend (similarity vs. continuity) condition.

When trend line and prediction intervals ("error bands", or "error" as an aesthetic description) are present in the same plot, the additional Gestalt principle of common region

is recruited, in addition to the continuity heuristic present due to the trend line and the linear relationship between x and y. The interaction between these heuristics dominates the perceptual experience, decreasing the probability that a participant will select the cluster target plot in favor of the trend target.

This interaction effect explains the different outcomes seen by the two conditions with conflicting aesthetics: the color + trend condition is more likely to result in cluster plot selection, while the color + ellipse + trend + error condition is more likely to result in trend plot selection, because the combined effect of the gestalt heuristics present in the trend and prediction interval elements is stronger than the effect of color and ellipse elements, which only invoke Gestalt heuristics of similarity and common region.

In summary, the results from this experiment show that in order to gain a significant difference from a plain representation and visually emphasize groups or trends, we need to make use of a statistical layer associated with a statistical interval/probability region in the form of an error band or an ellipse.

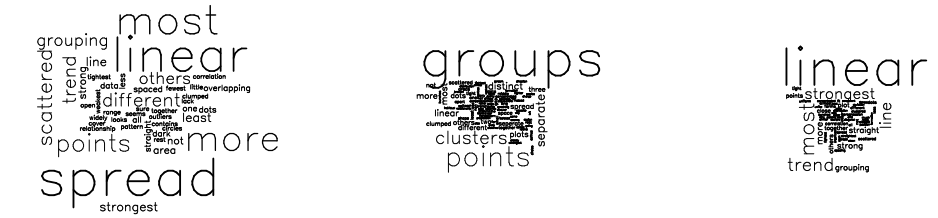
The lineup experimental protocol allows us to collect participant justifications for their target selection. These short explanations provide some additional insight into participant reasoning, and further support the gestalt explanation for the experimental results.

3.4 Participant Reasoning

As part of each trial, participants were asked to provide a short justification of their plot choice. Figure 12 gives an overview of summaries of participants' reasoning in form of word clouds. In the word clouds, stopwords are excluded from participants' reasons, unless they refer to quantities, such as 'none', 'all', 'some', 'few', etc. Reasons are also stemmed, so that words such as 'group', 'groups', 'grouping', 'grouped', and so on, all appear as the same (most prevalent) word in the cloud. What can be seen is a strong focus in terms of the reasoning depending on the outcome. If the participant chose one of the targets, the reasoning reflects this choice. When neither of the targets is chosen, there is less focus in the response. The word clouds look surprisingly similar independently of design - with the exception of the Ellipse + Color plot: here, specific colors are mentioned, which might be an indication that participants were distracted from the intended target by an imbalance in the

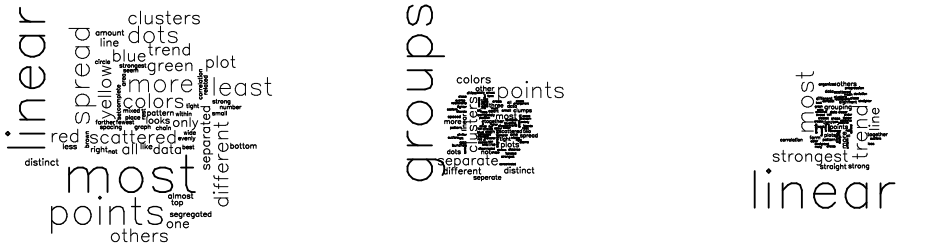
group allocation in one of the null panels. This is further investigated in Appendix C.1.2.

(a) Plain, neither target (b) Plain, cluster target (c) Plain, trend target



(d) Trend, neither target (e) Trend, cluster target (f) Trend, trend target

(g) Color, neither target (h) Color, cluster target (i) Color, trend target



(j) Color + Ellipse, neither (k) Color + Ellipse, cluster (l) Color + Ellipse, trend

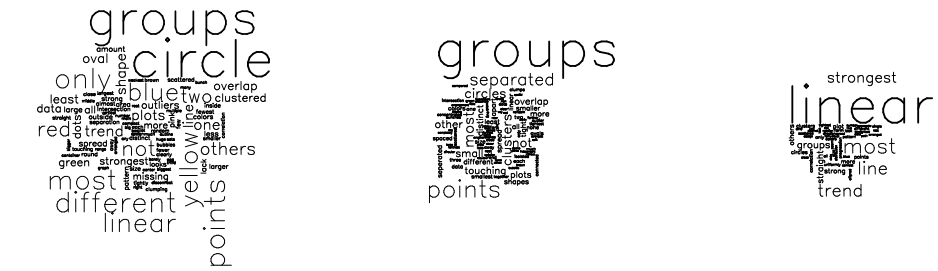


Figure 12: Wordclouds of participants' reasoning by outcome for a selected number of designs. Mostly, the reasoning and the choice of the target are highly associated. For the Color + Ellipse plot, participants were distracted from either target by an imbalance in the cluster/color distribution, as can be seen from the reasoning in the bottom left wordcloud.

Further examination of individual participants' responses illustrates that group size (and thus, missing ellipses in the target plot) was a factor in the decision to identify specific null plots as different. Participants provided responses such as "There is no circle highlighting the yellow symbols in this plot" and "Lack of a circle around the red symbols", which highlight the visual cues which were missing from the identified null plots.

This was due to an unintended side-effect of using the k-means algorithm for cluster allocation in the null plots: due to an imbalance in the size of clusters, an additional cue was introduced into the null plots. The estimation of bounding ellipses fails for clusters with fewer than three points and in these cases, ellipses were not drawn. Visually, the conspicuous absence of an ellipse led participants to select null plots with that feature. When we include a variable encoding the absence of one of the ellipses in a lineup into model (3) it is highly significant (χ^2_1=9.6, P-value=0.002), with an estimate of -0.8075, indicating that if one of the ellipses in a lineup is missing, the probability of picking the cluster target is reduced to less than half (44.6%) of the trend target. If we additionally consider the effect of a missing ellipse on individual designs, it is also significant (χ^2_9=20.4, P-value=0.0155). Appendix C.1.2 contains more details on this model and its effects.

4 Discussion and Conclusions

Taken together, the results presented suggest that plot aesthetics influence the perception of the dominant effect in the displayed data. This effect is not simply additive (otherwise, the two conflicting aesthetic conditions would result in similarly neutral effects); rather, the effect is consistent with layering of gestalt perceptual heuristics. Plot layers which add additional heuristics show larger effects than plot layers which duplicate heuristics that are already in play. For example, adding ellipses to a plot which has color aesthetics increases cluster recognition by recruiting the common region heuristic in addition to the point similarity heuristic recruited by color. Adding shape to a plot which has color aesthetics increases cluster recognition only slightly, but does not add additional gestalt heuristics (though point similarity is emphasized through two different mechanisms).

Statistically, this is important because the addition of ellipses or prediction intervals provides important statistical context, while reinforcing the visual emphasis by addition of

the common region heuristic. Graphics which more effectively convey the statistical results are composed of aesthetic layers which recruit multiple gestalt heuristics in order to present a unified message. This represents a departure from the "show the data" mentality, but is still consistent with the goal of good graphics, that is, to convey the data in a way that is easily understandable while still providing appropriate detail.

While more studies are necessary to fully explore the non-additive mechanism of additional gestalt heuristics and understand their effect in other types of plots, these results demonstrate the importance of carefully constructing graphs to convey the most important aspects of the displayed data.

5 Supplementary Materials

Appendix: Contains additional details about the methodology and results presented in the paper.

- Appendix A provides details about simulations of the data-generating model described in Section 2.1.4.
- Appendix B discusses the probability distribution underlying hypothesis testing for dual-target lineups.
- Appendix C contains a thorough discussion of models fit to participant target selections, response time, and self-reported confidence. Variations on these models address participant reasoning as discussed in Section 3.4, the relationship between participant response time and accuracy, as well as additional features of the data.

(Appendix.pdf, pdf file)

R code R code to re-produce figures of the paper as well as fit models used in the paper and the online appendix. (JCGS_VanderPlas_Hofmann.R, R file)

Data Anonymized responses from the turk study to investigate feature hierarchy. Each line corresponds to one lineup evaluation by a participant. (modeldata.csv, csv file)

References

Adamo, S. H., Cain, M. S., and Mitroff, S. R. (2015), "Targets Need Their Own Personal Space: Effects of Clutter on Multiple-Target Search Accuracy," *Perception*, 0301006615594921.

Amazon (2010), "Mechanical Turk," `https://www.mturk.com/mturk/welcome`.

Bobko, P. and Karren, R. (1979), "The perception of Pearson product moment correlations from bivariate scatterplots," *Personnel Psychology*, 32, 313–325.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009), "Statistical inference for exploratory data analysis and model diagnostics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367, 4361–4383.

Cain, M. S., Dunsmoor, J. E., LaBar, K. S., and Mitroff, S. R. (2011), "Anticipatory anxiety hinders detection of a second target in dual-target search," *Psychological Science*, 22, 866–871.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, 1st ed.

Cleveland, W. S. and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, pp. 531–554.

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013), "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research," *PLoS ONE*, 8, 1–18.

Demiralp, C., Bernstein, M., and Heer, J. (2014), "Learning Perceptual Kernels for Visualization Design," *Visualization and Computer Graphics, IEEE Transactions on*, 20, 1933–1942.

DeMita, M. A., Johnson, J. H., and Hansen, K. E. (1981), "The validity of a computerized visual searching task as an indicator of brain damage," *Behavior Research Methods & Instrumentation*, 13, 592–594.

Eriksen, C. W. and Hake, H. W. (1955), "Multidimensional stimulus differences and accuracy of discrimination." *Journal of Experimental Psychology*, 50, 153.

Few, S. (2009), *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Burlingame, CA: Analytics Press, 1st ed.

Fleck, M. S., Samei, E., and Mitroff, S. R. (2010), "Generalized satisfaction of search: Adverse influences on dual-target search accuracy." *Journal of Experimental Psychology: Applied*, 16, 60.

Gegenfurtner, K. R. and Sharpe, L. T. (2001), *Color vision: From genes to perception*, Cambridge University Press.

Goldstein, E. B. (2009), *Encyclopedia of perception*, Sage Publications.

Hanrahan, P. (2003), "Tableau software white paper - visual thinking for business intelligence," *Tableau Software, Seattle, WA*.

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

Healey, C. G. and Enns, J. T. (1999), "Large datasets at a glance: Combining textures and colors in scientific visualization," *Visualization and Computer Graphics, IEEE Transactions on*, 5, 145–167.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical tests for power comparison of competing designs," *Visualization and Computer Graphics, IEEE Transactions on*, 18, 2441–2448.

Lewandowsky, S. and Spence, I. (1989a), "Discriminating strata in scatterplots," *Journal of the American Statistical Association*, 84, 682–688.

— (1989b), "The perception of statistical graphs," *Sociological Methods & Research*, 18, 200–242.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of visual statistical inference, applied to linear models," *Journal of the American Statistical Association*, 108, 942–956.

Mosteller, F., Siegel, A. F., Trapido, E., and Youtz, C. (1981), "Eye fitting straight lines," *The American Statistician*, 35, 150–152.

Piepho, H.-P. (2004), "An algorithm for a letter-based representation of all-pairwise comparisons," *Journal of Computational and Graphical Statistics*, 13, 456–466.

Robinson, H. (2003), "Usability of Scatter Plot Symbols," *ASA Statistical Computing & Graphics Newsletter*, 14, 9–14.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., and Zhao, Y. (2014), "Utilizing Distance Metrics on Lineups to Examine What People Read From Data Plots," *arXiv.org*.

Scaife, M. and Rogers, Y. (1996), "External cognition: how do graphical representations work?" *International journal of human-computer studies*, 45, 185–213.

Spence, I. and Garrison, R. F. (1993), "A remarkable scatterplot," *The American Statistician*, 47, 12–19.

Treisman, A. (1985), "Preattentive processing in vision," *Computer Vision, Graphics, and Image Processing*, 31, 156 – 177.

Treisman, A. M. and Gelade, G. (1980), "A feature-integration theory of attention," *Cognitive psychology*, 12, 97–136.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Lebanon, IN: Addison Wesley.

Vanderplas, S. and Hofmann, H. (2016), "Spatial Reasoning and Data Displays," *IEEE Transactions on Visualization and Computer Graphics*, 459–468.

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010), "Graphical inference for infovis," *Visualization and Computer Graphics, IEEE Transactions on*, 16, 973–979.

Zhang, J. (1997), "The nature of external representations in problem solving," *Cognitive science*, 21, 179–217.