Hierarchy of Visual Features

Susan VanderPlas

March 30, 2015

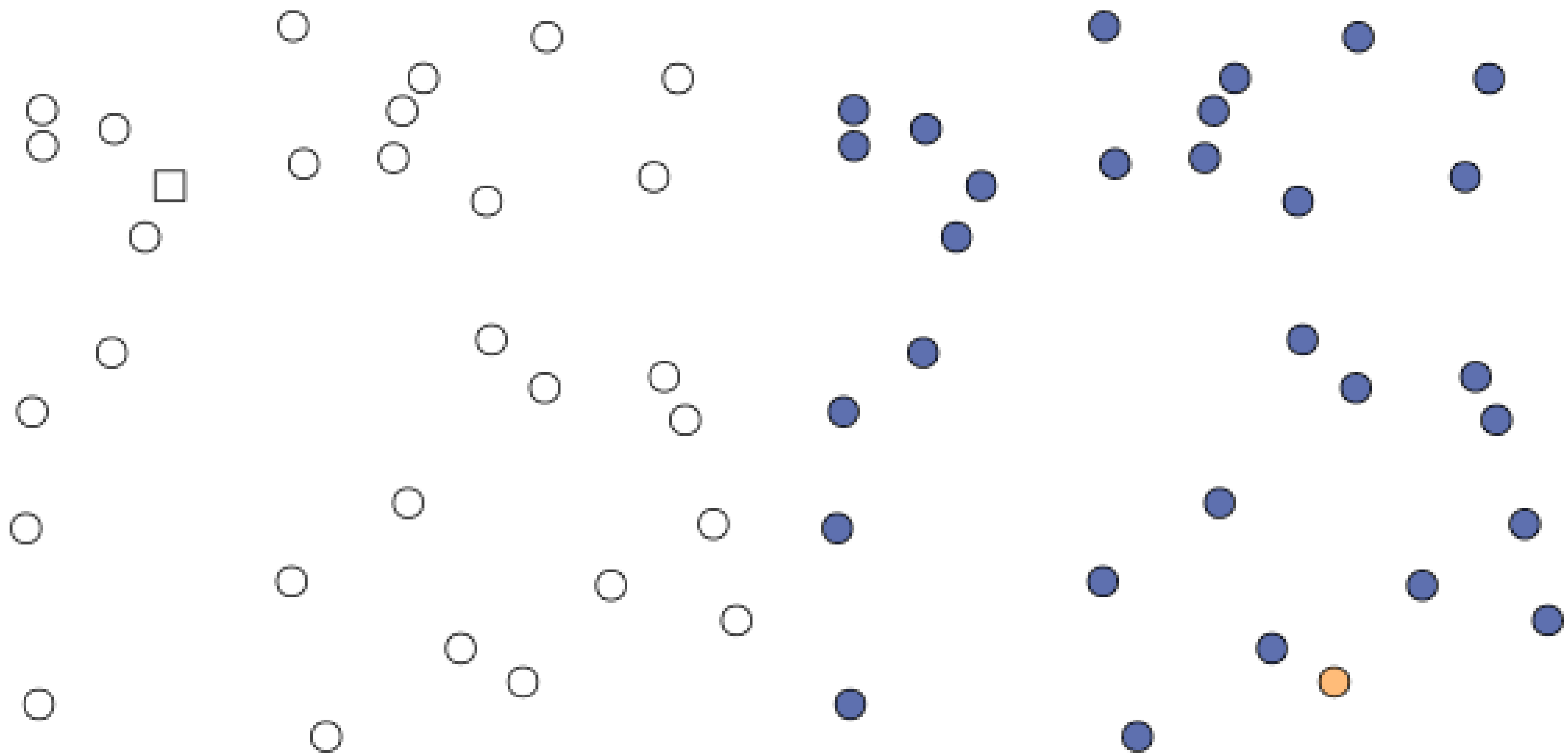
Outline

- Human Perception & Statistical Graphics
- Experimental Structure
  - Model
  - Parameter Structure
  - Plot Aesthetics
- Results
  - Accuracy
  - Participant Reasoning
  - Response Time
- Conclusions & Future Work

Human Perception and Statistical Graphics

Preattentive Feature Detection

Preattentive perception occurs before conscious attention is focused on the stimulus, within the first 200 ms

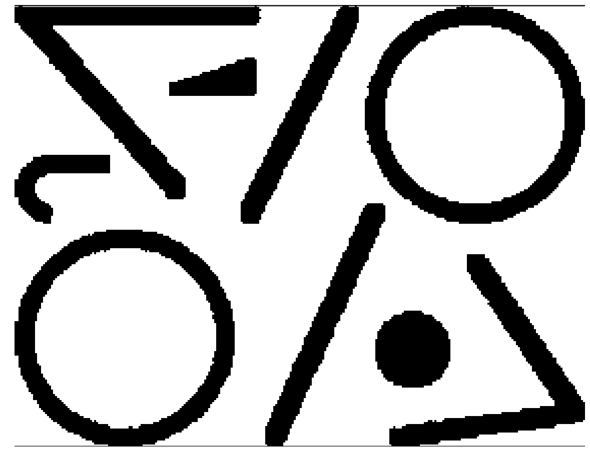


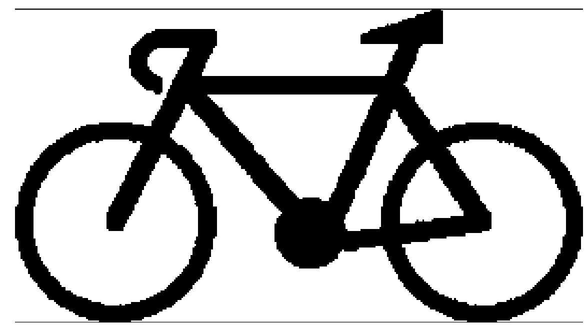
Perception of Statistical Plots

- We don't perceive plots preattentively, but some research has studied preattentive plot perception

  (Healy & Enns, 1999, 2012)

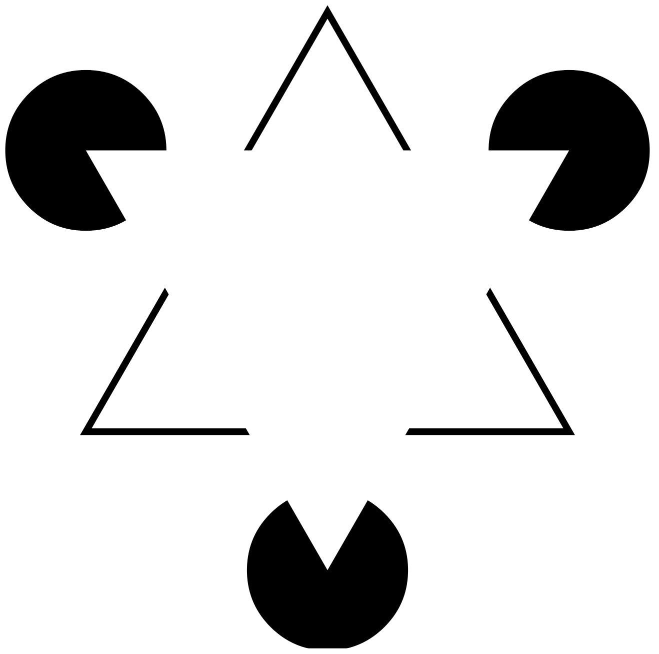
- Preattentive feature detection impacts ease and accuracy of reading statistical plots





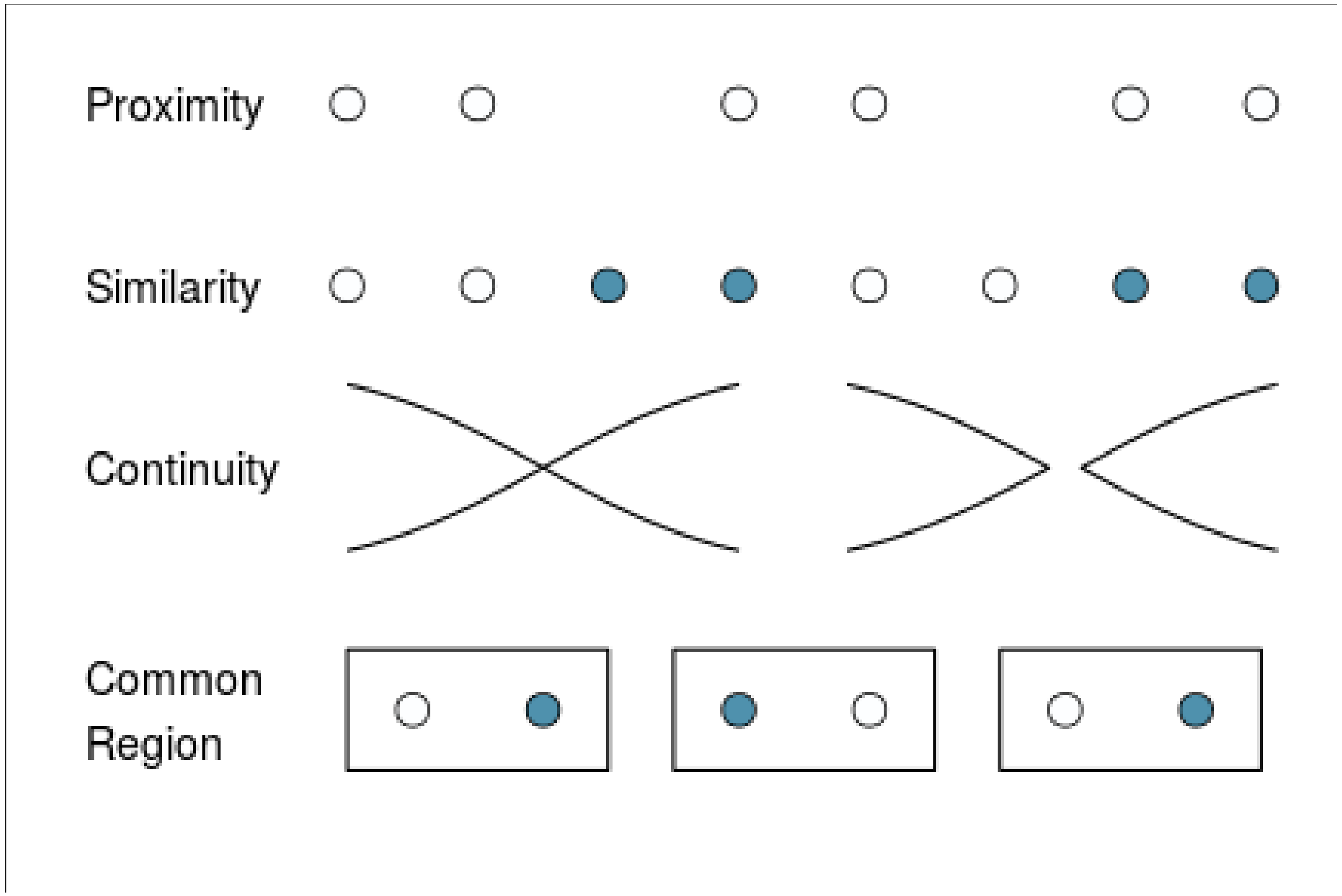
Gestalt Laws of Perception

- Rules that make sense of complex visual information through experience
- "Top-down" rules organize information hierarchically
- Subconsciously order and group visual input



"The whole is different from the sum of the parts"

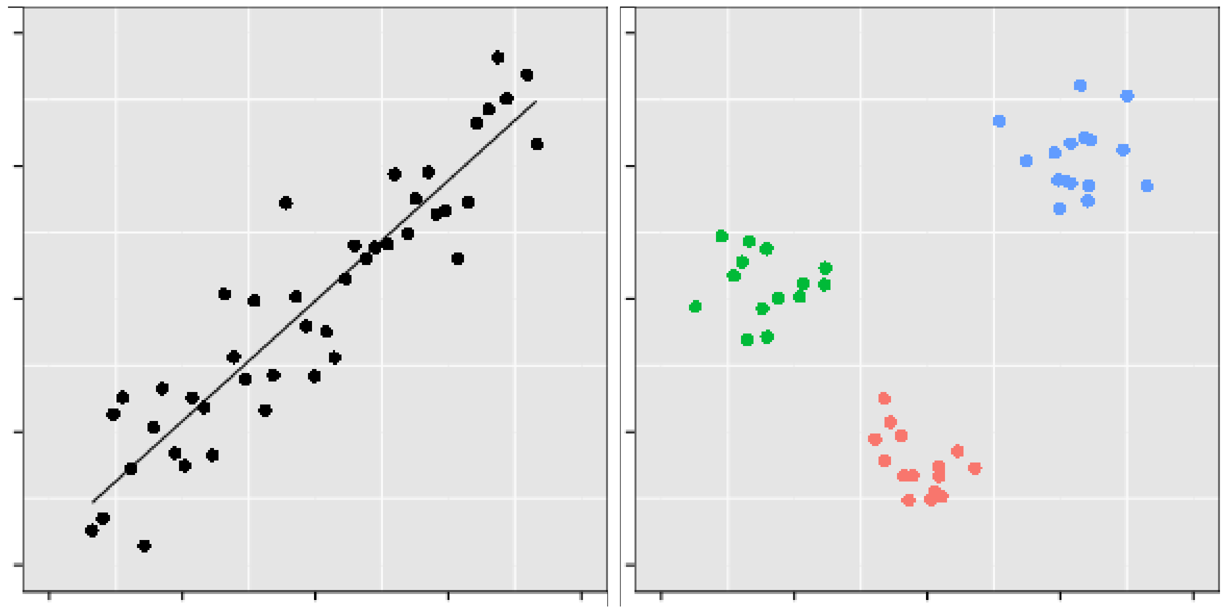
Gestalt Laws



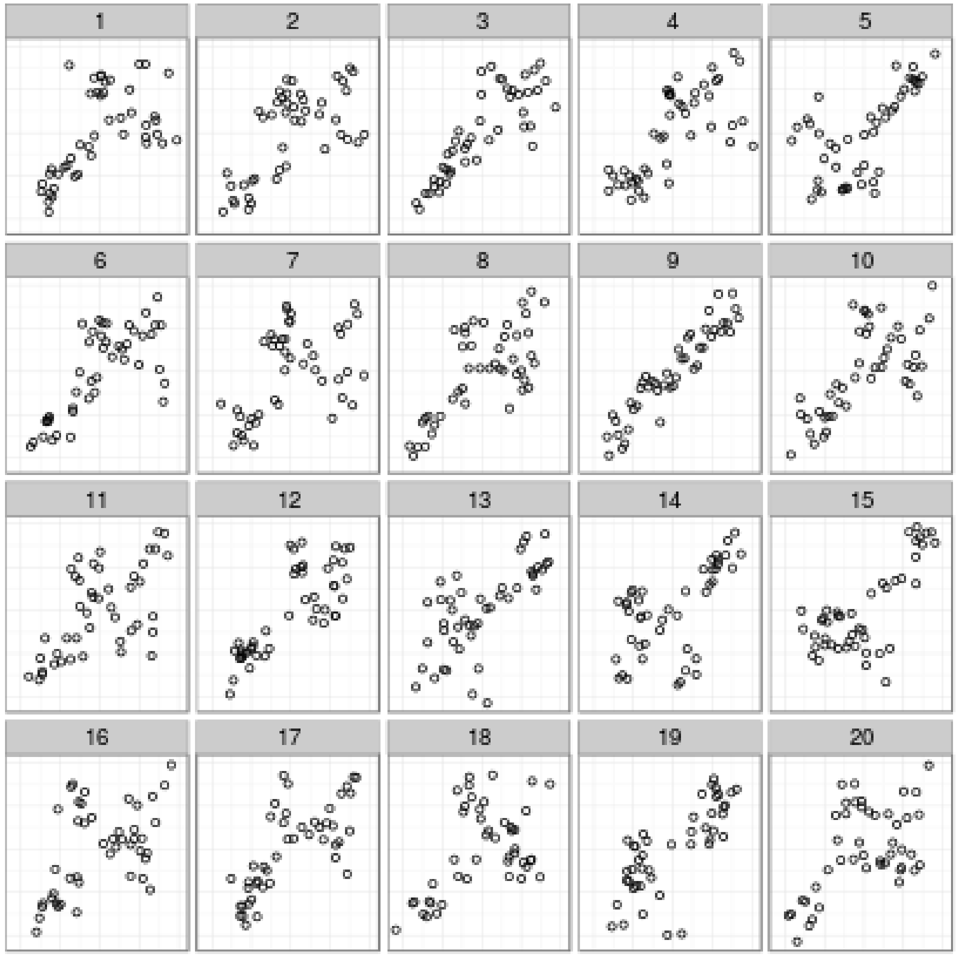
- 3 groups of 2 white dots
- 4 groups of 2 white or blue dots
- 2 continuous s-shape curves, 2 v-shaped curves
- 3 groups of 1 white and 1 blue dot

Experiment:

How do plot aesthetics affect perception of statistical plots?



Lineups: "Which plot is the most different?"



Answer: 9

Standard design:

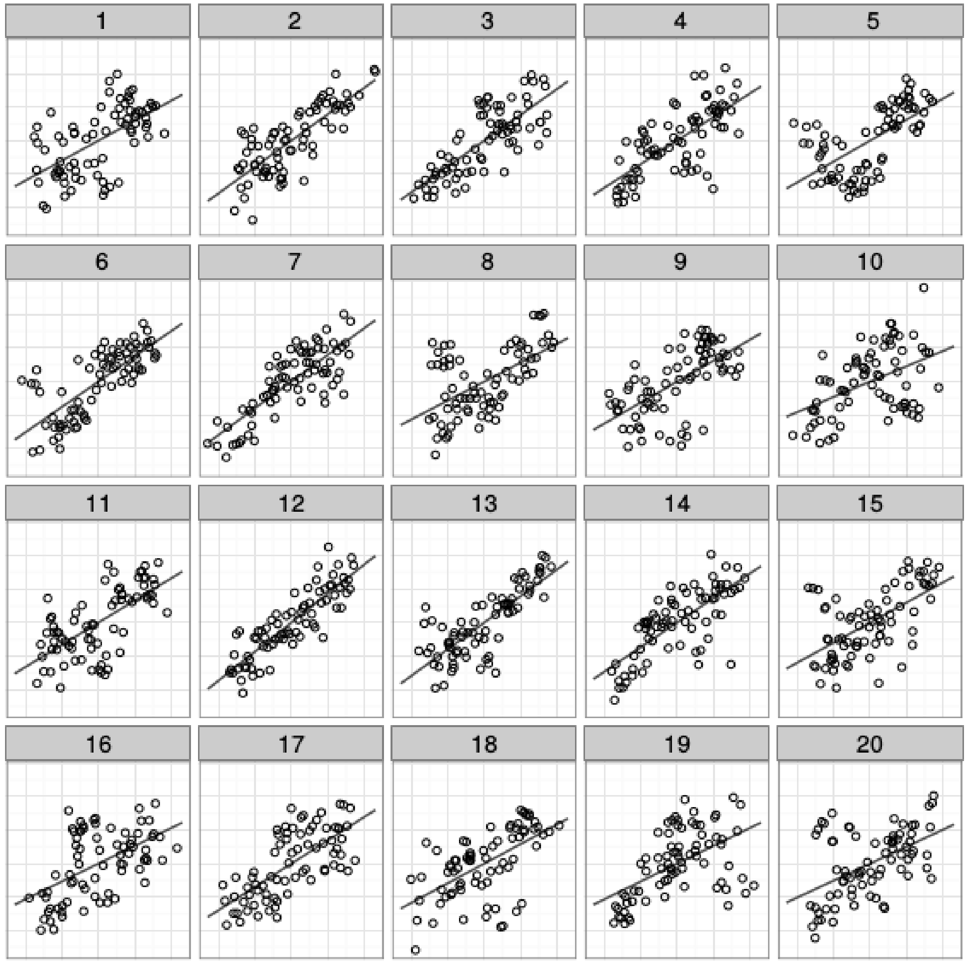
- One "target" plot
  (real data or generated from H_A)
- 19 null plots generated from H_0
- $P(\text{select target} | H_0) = 0.05$
- Allows quantification of significance for graphical findings

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, *Phil. Transac.*

Majumder, M., Hofmann, H., and Cook, D. (2013). Validation of visual statistical inference, applied to linear models, *JASA*

Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical inference for infovis, *TVCG*

Lineups: "Which plot is the most different?"



Modification:

- Two targets, each from a different model
- Null plots: 50% mixture of the two models

Participant Responses

Plot 12: 59.1% (Trend target)

Plot 5: 9.1% (Cluster target)

Other: 31.7%

Sample size: 22

Trend target: 12, Cluster target: 5

Lineups: "Which plot is the most different?"

Modification:

- Two targets, each from a different model
- Null plots: 50% mixture of the two models

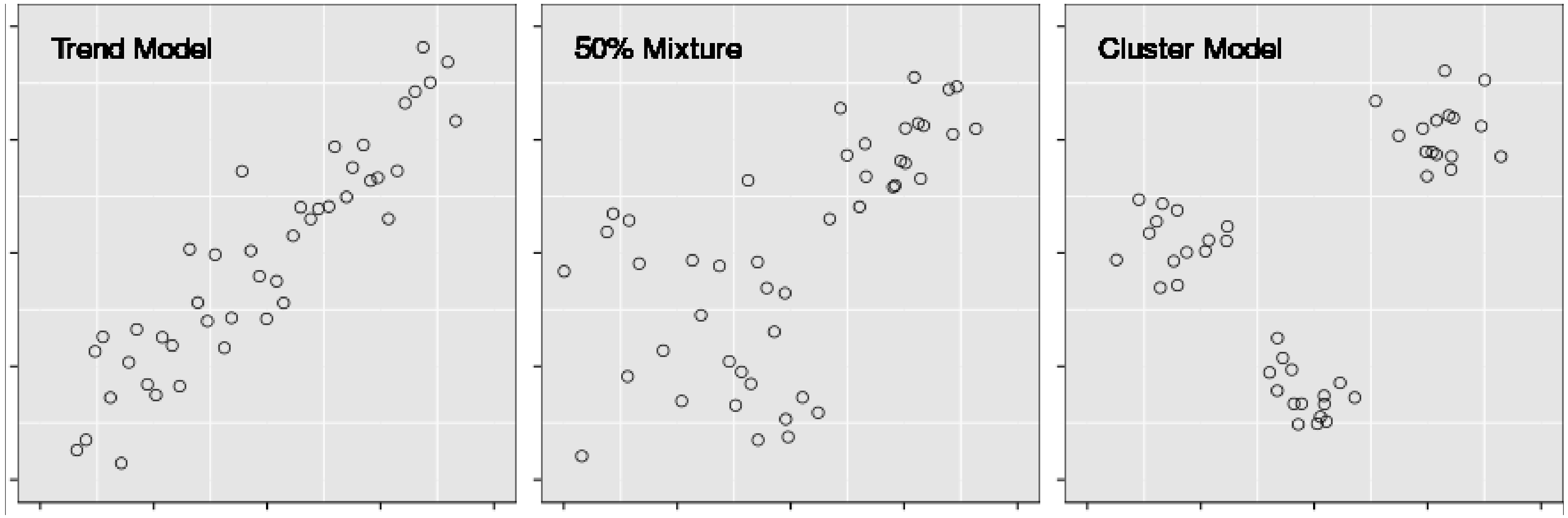
Participant Responses

Plot 12:	9.7%	(Trend target)
Plot 5:	29.0%	(Cluster target)
Plot 18:	32.3%	
Other:	29.1%	

Sample size: 31

Trend target: 12, Cluster target: 5

Data-Generating Models



Parameters

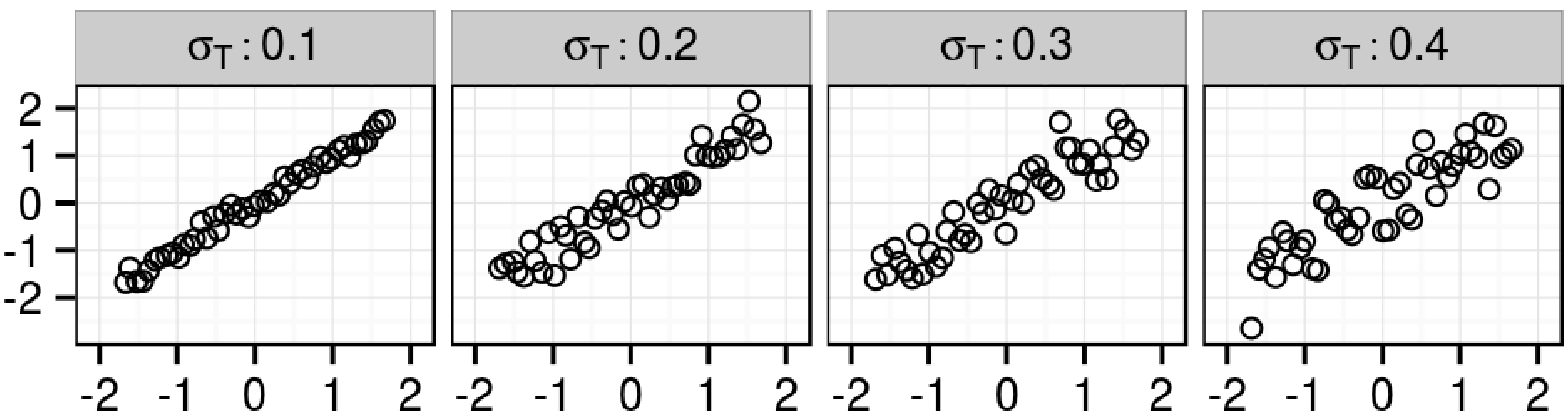
σ_T: Variability in y	λ: Mixing parameter	K: # clusters σ_C: Variability around cluster centers

Trend Model M_T

Input: sample size N, σ_T standard deviation from the line $y = x$

Output: vectors x and y with N observations

1. Generate $\tilde{x}_i, i = 1, \ldots, N$, as a seq. of evenly spaced points from $[-1, 1]$.
2. Jitter $\tilde{x}_i$ by adding small perturbations
3. Generate y_i as a linear function of x_i: $y_i = x_i + e_i, e_i \sim N(0, \sigma_T^2)$.
4. Center and scale x_i, y_i



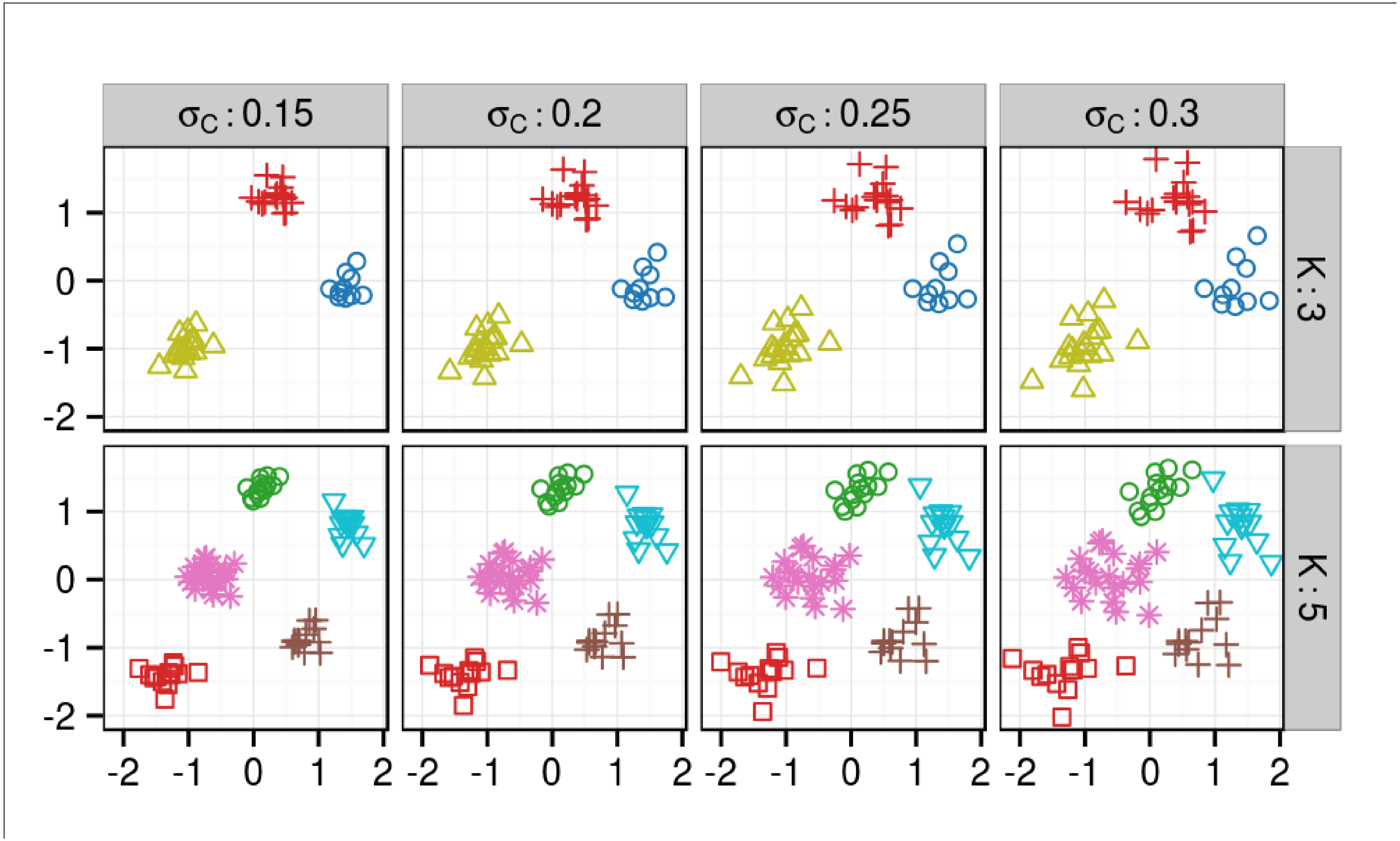
Cluster Model M_C

Input: sample size N, number of clusters K, cluster std. dev. σ_C

Output: vectors $\mathbf{x}$ and $\mathbf{y}$ with N observations

1. Generate K cluster centers (c^x, c^y) on a $K \times K$ grid; center and scale
2. Sample group sizes $g = (g_1, \ldots, g_K)$ with $N = \sum_{i=1}^{K} g_i$
3. Jitter points around cluster centers by $N(\mathbf{0}, \sigma_C^2)$
4. Center and scale x_i, y_i

Cluster Model M_C



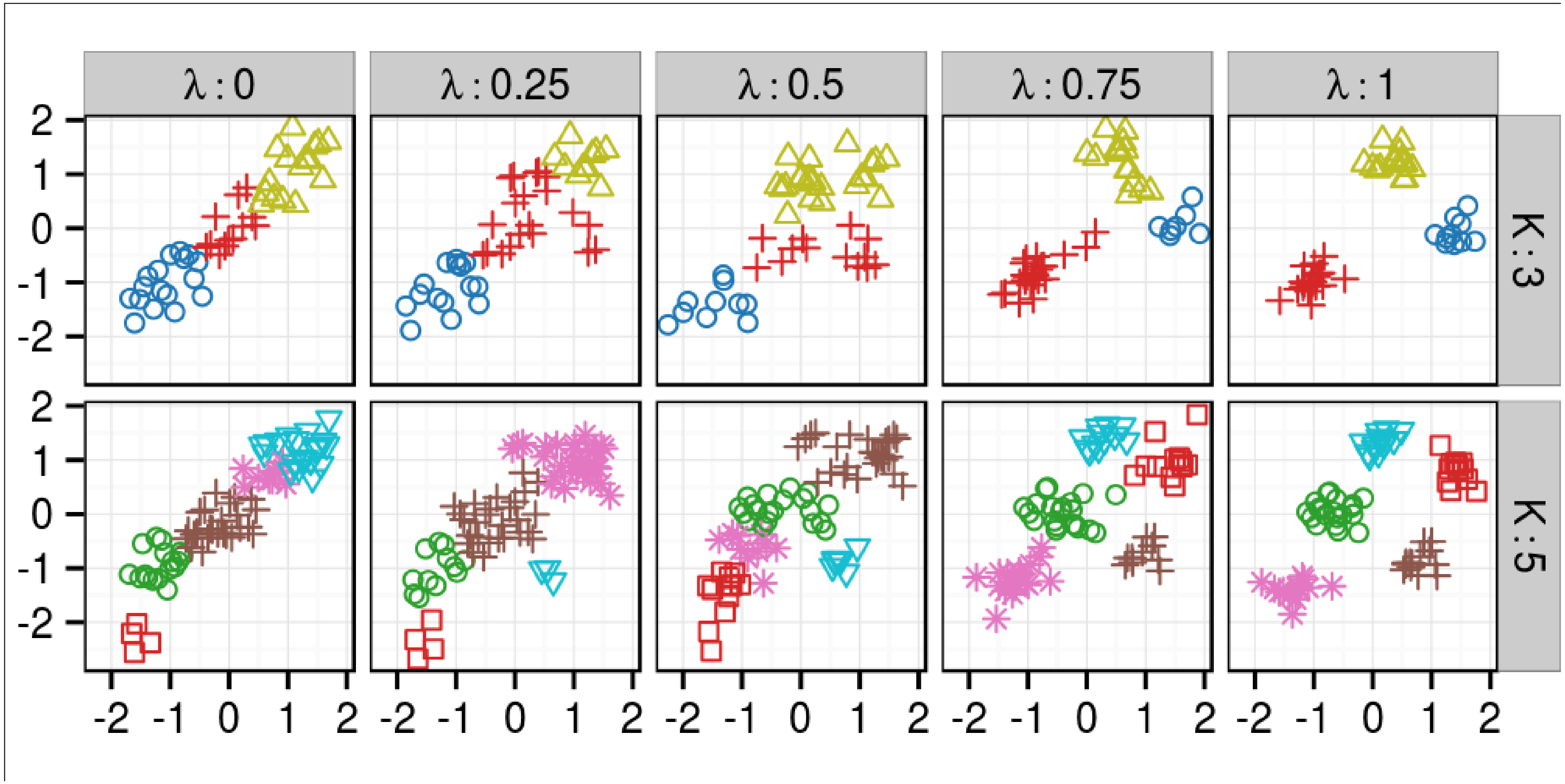
Null Model M_0

Input: sample size N, number of clusters K, σ_C, σ_T, and mixing parameter λ

Output: vectors $\mathbf{x}$ and $\mathbf{y}$ with N observations

1. Generate datasets from M_C and M_T
2. Select $n_c \sim \text{Binomial}(N, \lambda)$ points from the data generated by M_C
3. Select $n_T = N - n_c$ points from the data generated by M_T
4. Center and scale the points in x and y
5. Assign groups using hierarchical clustering

Null Model M_0



Measuring Signal Strength

Trend $$R^2 = \frac{SS_{Reg}}{SS_{Tot}}$$

For (x_{ij}, y_{ij}) the jth point in cluster i

Cluster

$$SS_C = \sum_{i=1}^{K} \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)^2 + (y_{ij} - \overline{y}_i)^2$$

$$SS_{Tot} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} (x_{ij} - \overline{x})^2 + (y_{ij} - \overline{y})^2$$

Define $C^2 := \frac{SS_C}{SS_{Tot}}$ to measure cluster cohesion.

Parameter Values: Simulation

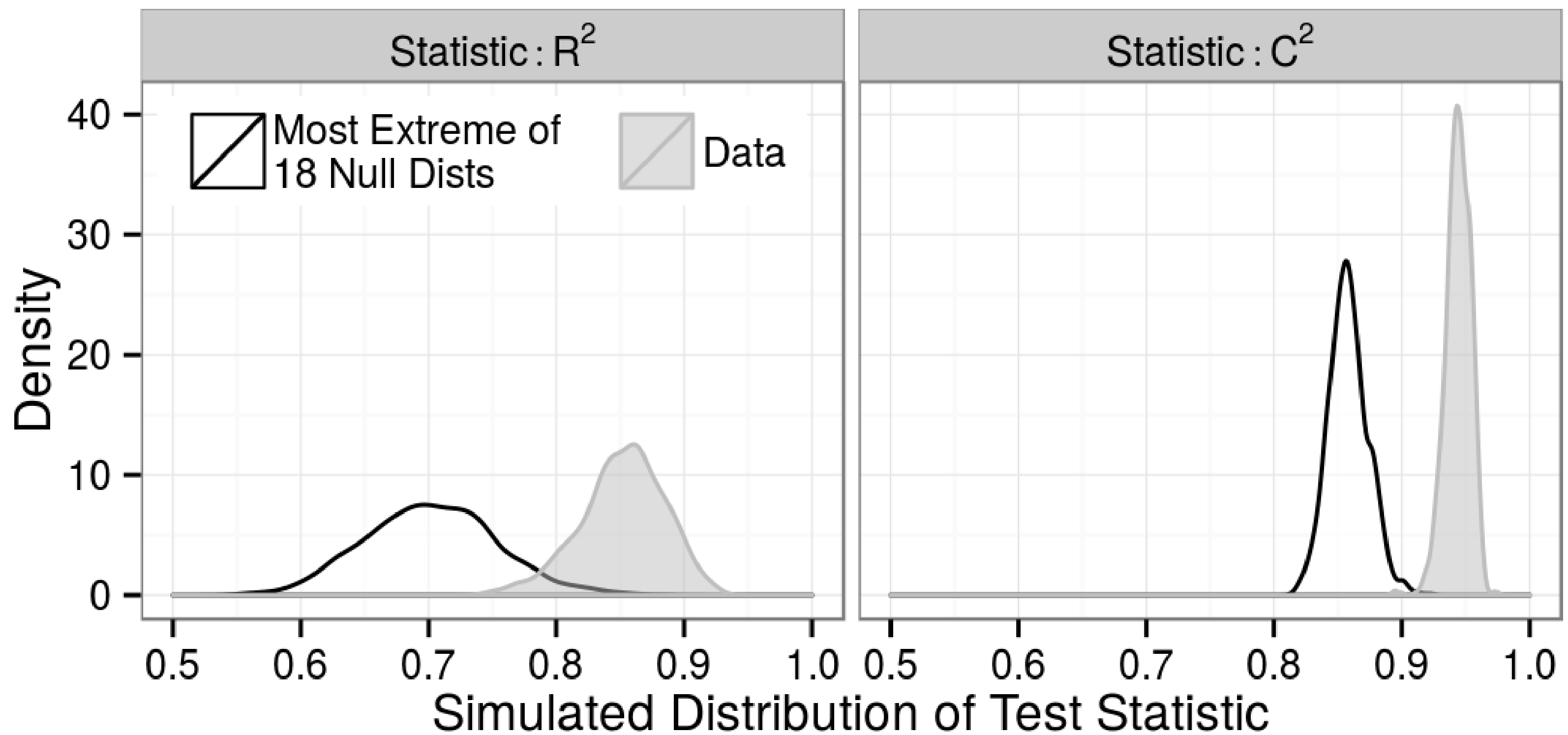
For all combinations of

$$\sigma_T \in \{0.2, 0.25, \ldots, 0.5\}$$
$$\sigma_C \in \{0.1, 0.15, \ldots, 0.4\}$$
$$K \in \{3, 5\}$$

1. Generate 1000 datasets consisting of the following sub-plot datasets: 1 from M_T, 1 from M_C, and 18 from M_0
2. For each lineup dataset, calculate the following:
    1. Trend target R^2
    2. Maximum null plot R^2
    3. Cluster target C^2
    4. Maximum null plot C^2

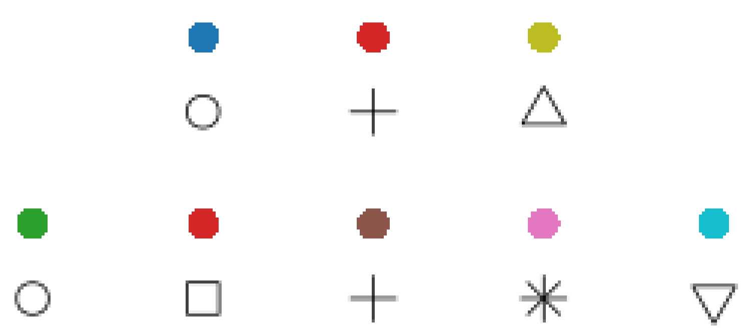
Simulation:

$\sigma_T = 0.25$, $\sigma_C = 0.20$, and $K = 3$



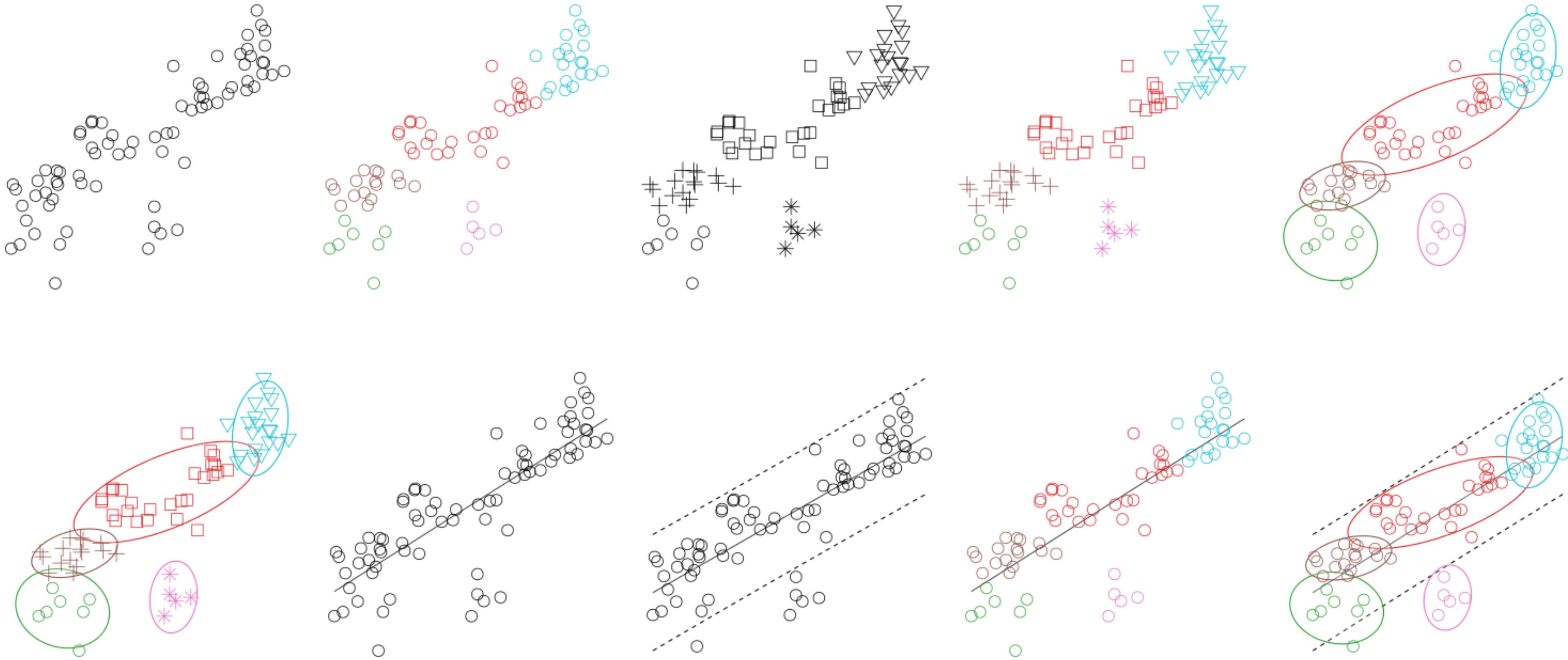
Plot Aesthetic Combinations

		Trend Emphasis		
	Strength	0	1	2
Cluster Emphasis	0	Plain	Line	Line + Pred. Interval
	1	Color Shape	Color + Line	
	2	Color + Shape Color + Ellipse		Color + Ellipse + Line + Pred. Interval
	3	Color + Shape + Ellipse		



Palettes selected to provide maximum perceptual distance (Ç. Demiralp, et al., 2014). Shapes conform to guidelines in Robinson (2003) and Lewandowsky & Spence (1989).

Plot Aesthetic Combinations



Experimental Structure

Model Parameters

- Trend Strength $\sigma_T =$ easy, med., hard
- Cluster Strength $\sigma_C =$ easy, med., hard
- Number of Clusters $K = 3, 5$

Plot Aesthetics

- Plain
- Trend
- Trend + Pred. Int.
- Color + Trend
- Color + Ellipse + Trend + Pred. Int.
- Color
- Shape
- Color + Shape
- Color + Ellipse
- Color + Shape + Ellipse

Plot Level

- 18 parameter combinations
- 3 datasets/parameter combination
- 10 plot types for each dataset
  = 540 total plots

Evaluation Level

- Participants evaluate 10 plots:
  - 1 of each aesthetic
  - 1 of each combination of σ_T and σ_C randomized over K

Data Collection

- Participants recruited through Amazon Mechanical Turk
- Experiment ran for `23.8` hours
- `1356` individuals completed `13519` evaluations

- Data removed:
  - Participants who did not complete 10 trials:
    `159` participants, `1060` trials
  - Any trials in excess of 10 for each participant (`421` trials)

Final dataset: `12010` trials completed by `1201` participants.

Data Collection

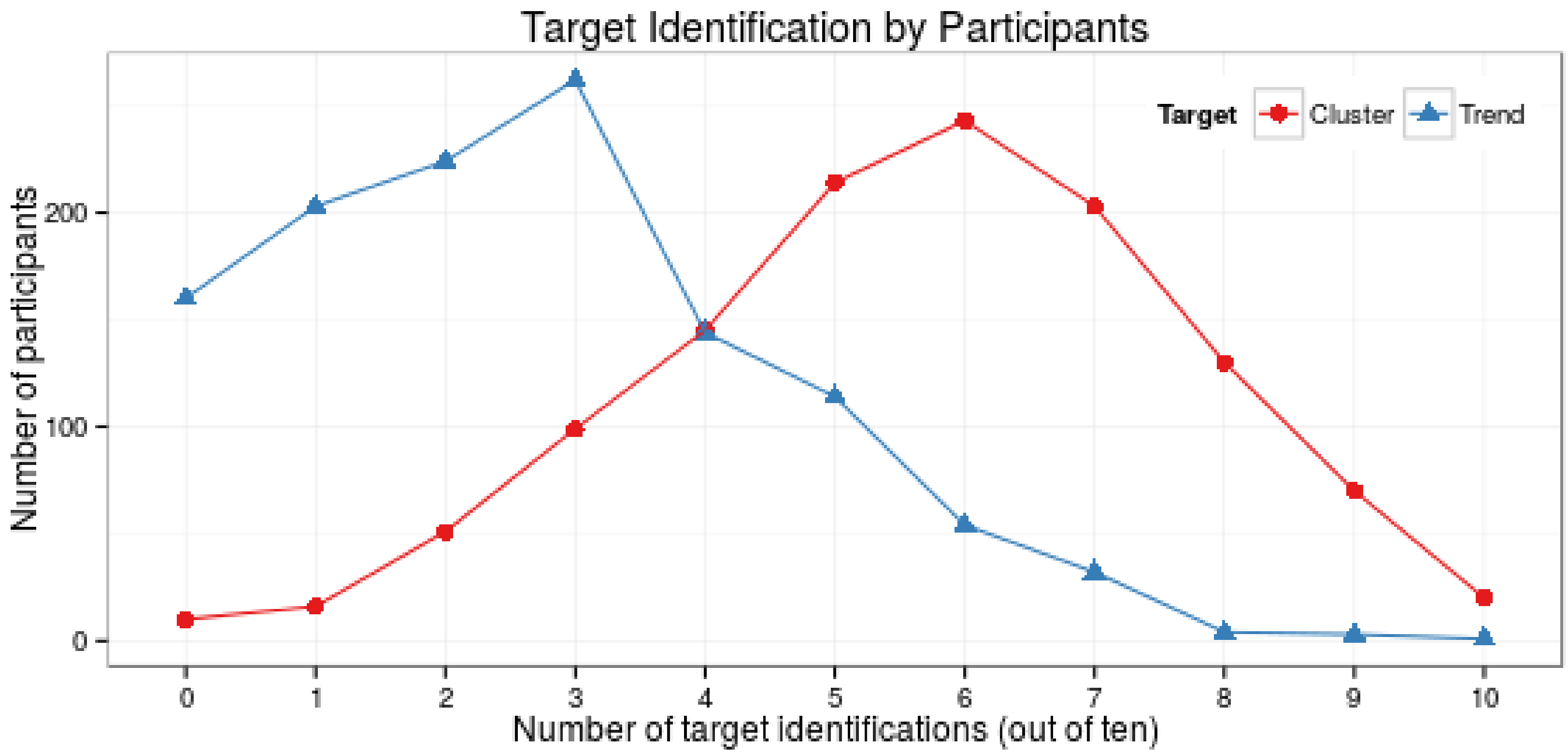
Participants provided:

- Demographic information: age range, gender, education level
- 10 plot evaluations
  - Target plot identification (one or more sub-plots)
  - Level of confidence in their answer (1 = least, 5=most)
  - Reasoning
    (i.e. "Strongest linear relationship", "Clustered points", "Odd shape")

Results

- Accuracy
  - Single Target Selection
  - Cluster vs. Trend Target Selection
- Participant Reasoning
- Response Time

Target Identification



Participants selected more cluster targets than line targets, however, 5 plot types were expected to emphasize clustering, and only 2 plot types were expected to emphasize trends.

Modeling Target Selection

$$\text{logit}\, Y \;\; = \mathbf{X}\beta \quad + \mathbf{J}\gamma \quad + \mathbf{K}\eta + \epsilon$$

where fixed effects β_i describe
the effect of specific plot types

$$\gamma_j \overset{iid}{\sim} N\left(0, \sigma^2_{\text{dataset}}\right)$$

$$\eta_k \overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right)$$

$$\epsilon_{ijk} \overset{iid}{\sim} N\left(0, \sigma^2_e\right)$$

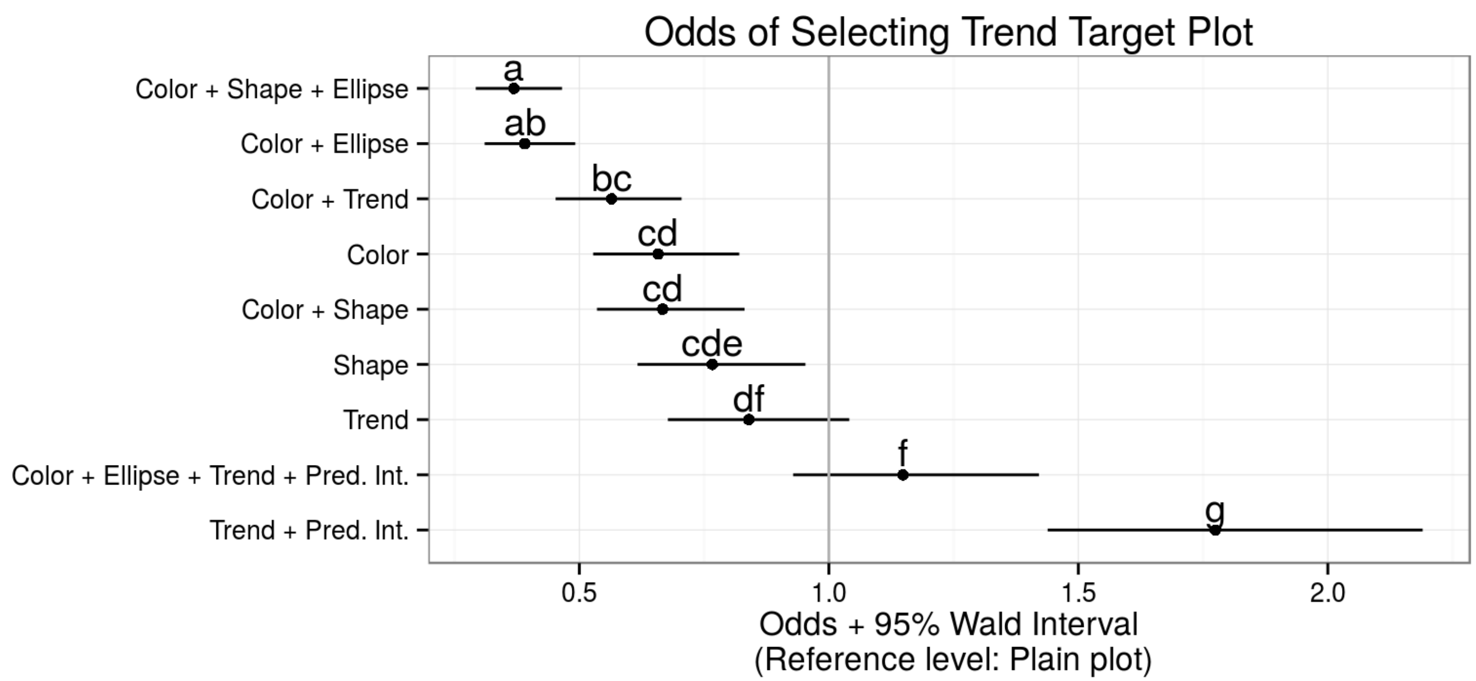
Dataset and participant effects are orthogonal by design

Variability due to model parameters is contained within the random effects for dataset

Modeling Single Target Selection

Target: Trend Plot M_T

$$Y = \text{Participant selected the plot generated by } M_T$$



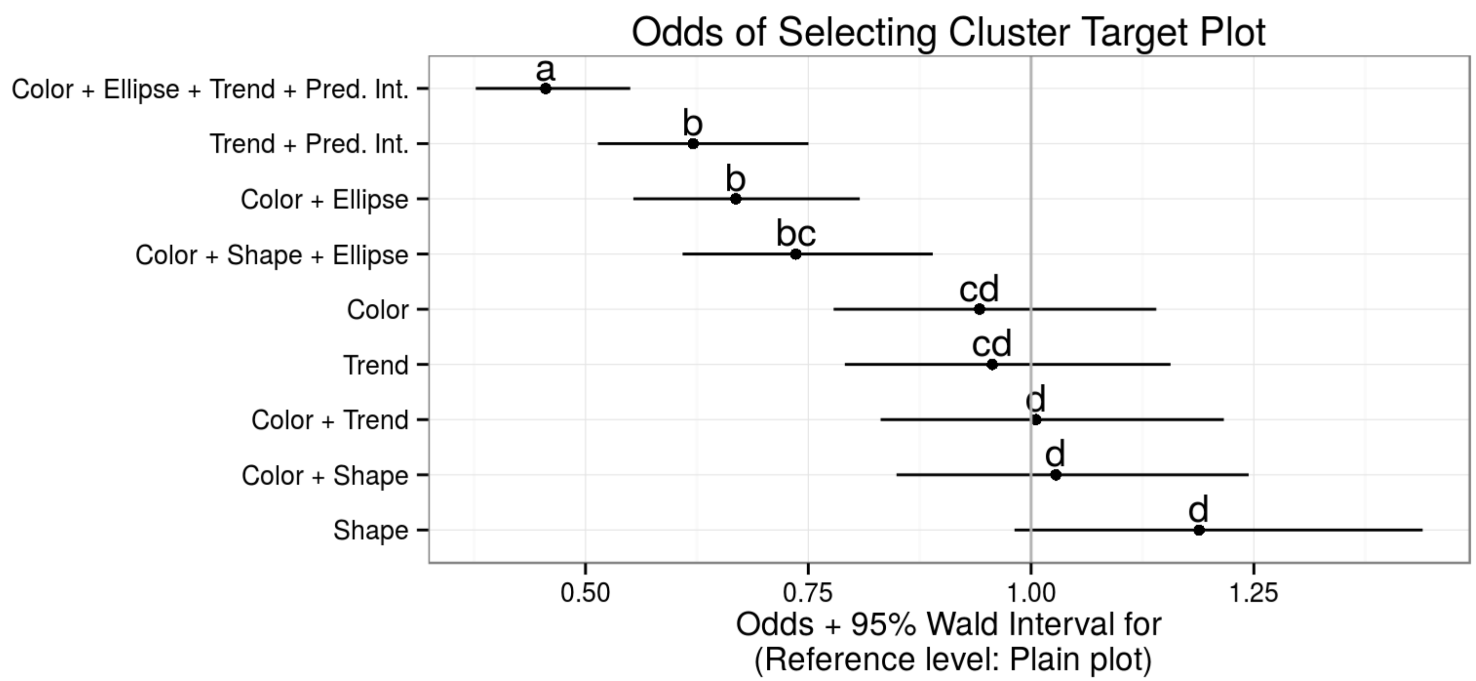
Plot types are significantly different if they do not share a letter

Participants are `0.37` times as likely to select trend targets when plots have color, shape, and ellipse aesthetics.

Participants are `1.77` times as likely to select trend targets when plots have trend lines and prediction intervals.

Target: Cluster Plot M_C

$$Y = \text{Participant selected the plot generated by } M_C$$

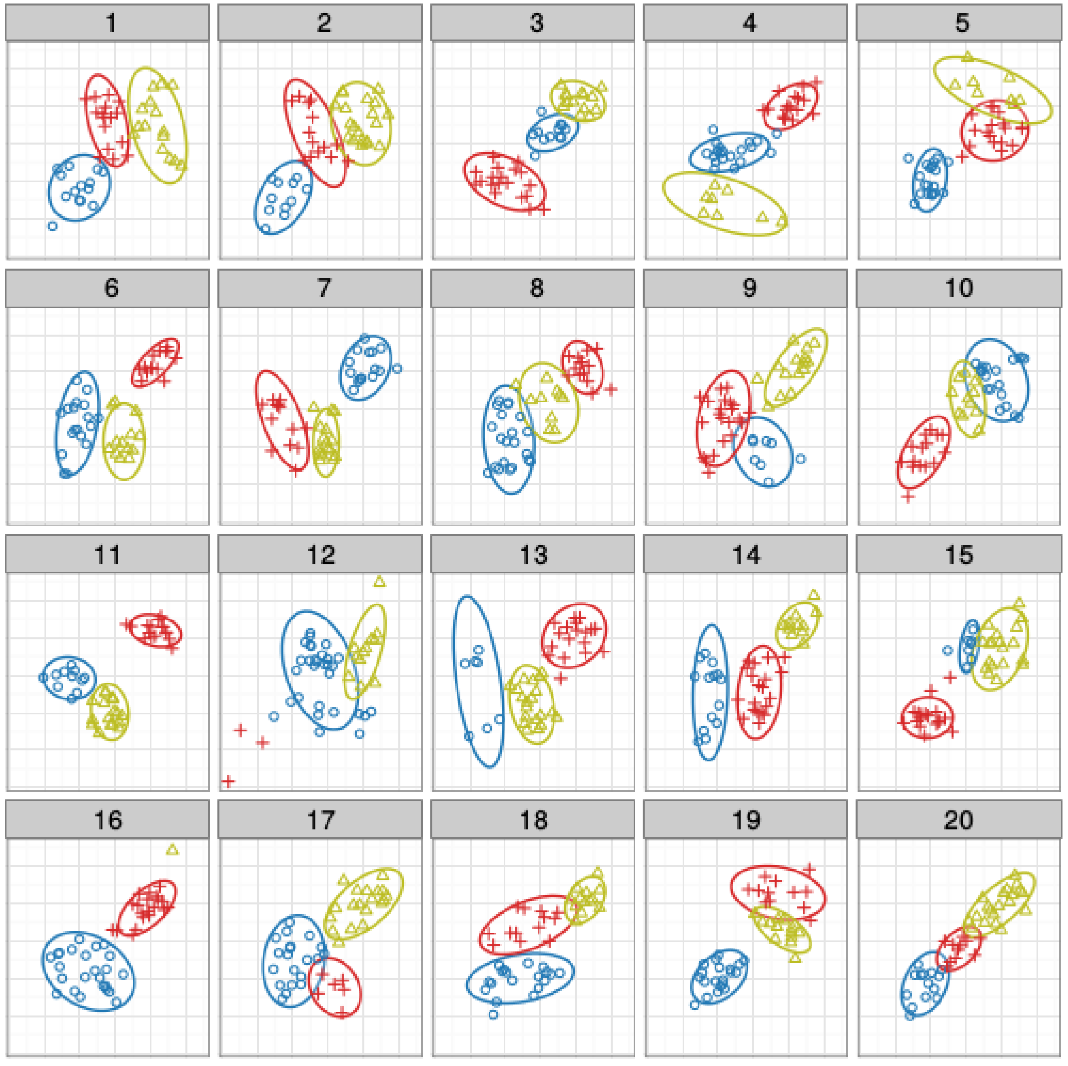


Plot types are significantly different if they do not share a letter

Participants are `0.67` times as likely to select cluster targets when plots have color and ellipse aesthetics. What happened?

Participants are `1.19` times as likely to select cluster targets when plots have shape aesthetics.

Cluster Target Identification



Participant Responses

Plot 20: 38.1% (Trend target)

Plot 11: 19.0% (Cluster target)

Plot 16: 23.8%

Plot 12: 14.3%

Plot 13: 4.8%

Sample size: 21

Trend target: 20, Cluster target: 11

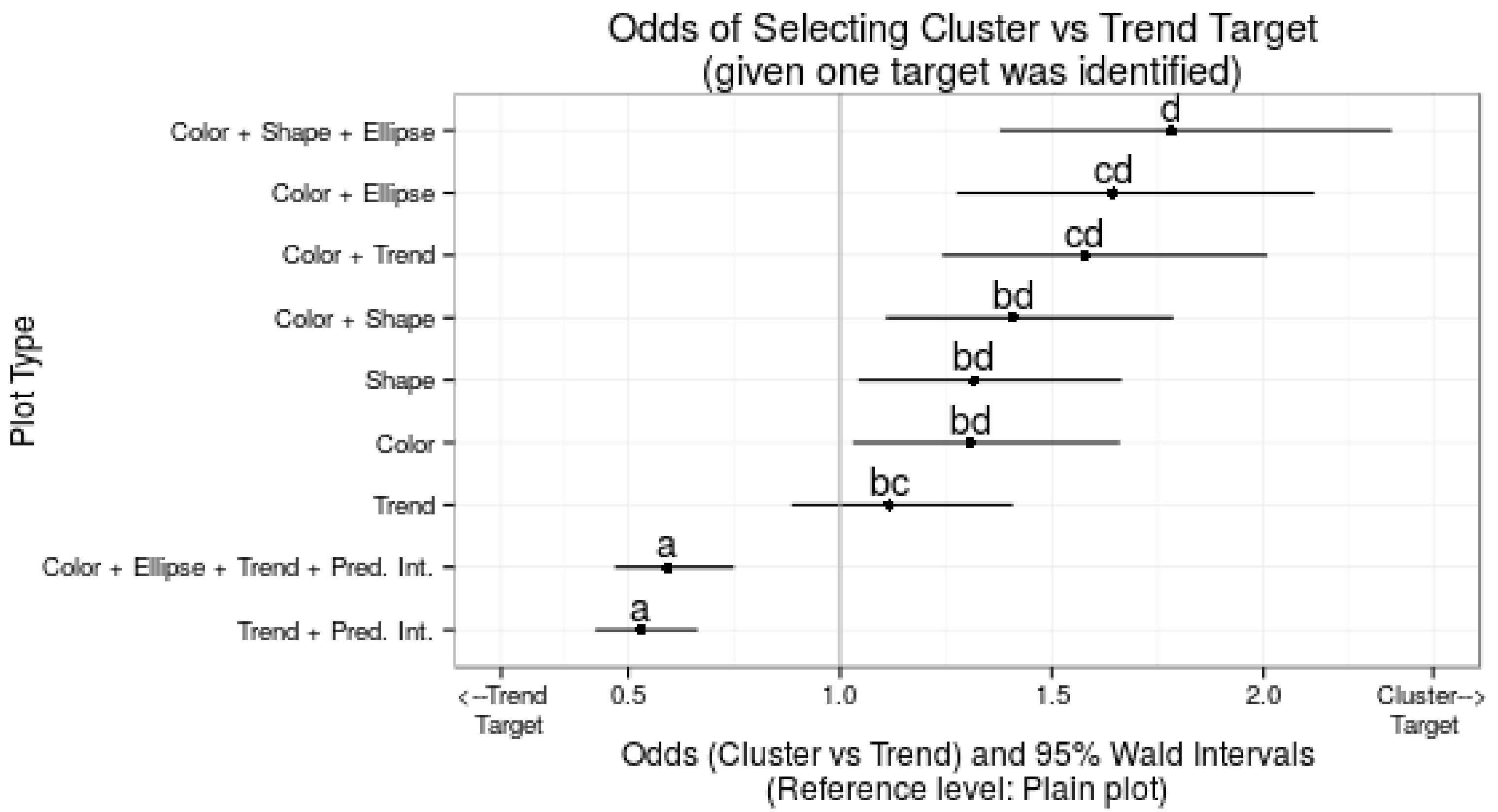
Faceoff: Cluster vs. Trend?

Cluster vs. Trend

Define C_{ijk} to be the event
{Participant k selects the cluster target for dataset j with aesthetic set i},
and T_{ijk} to be the analogous selection of the trend target.

$$\operatorname{logit} P(C_{ijk} | C_{ijk} \cup T_{ijk}) = \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta + \epsilon$$

Cluster vs. Trend



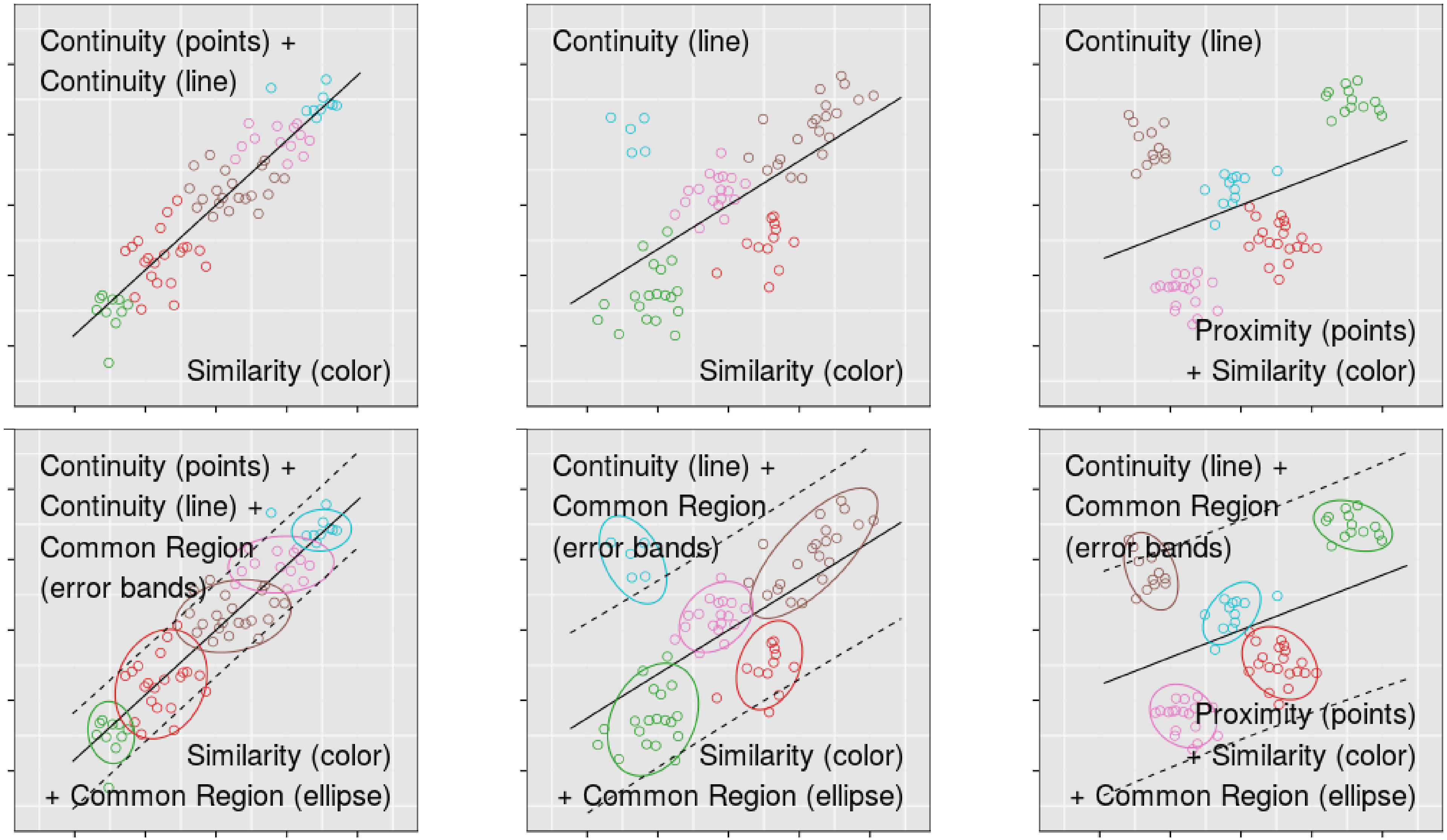
Plot types are significantly different if they do not share a letter

Participants are `0.53` times as likely to select cluster targets when plots have trend line and prediction interval aesthetics.

Participants are `1.78` times as likely to select cluster targets when plots have color, shape, and ellipse aesthetics.

Mixed Signals?

Exploring the gestalt of plots with conflicting aesthetics



Participants are `1.58` times as likely to select the cluster target with color + trend line, relative to the plain plot.

Participants are `0.59` times as likely to select the cluster target with color + ellipse + trend line + pred. int., relative to the plain plot.

Participant Reasoning

Participant Reasoning

Plain Plots

Neither Target
(N=127)

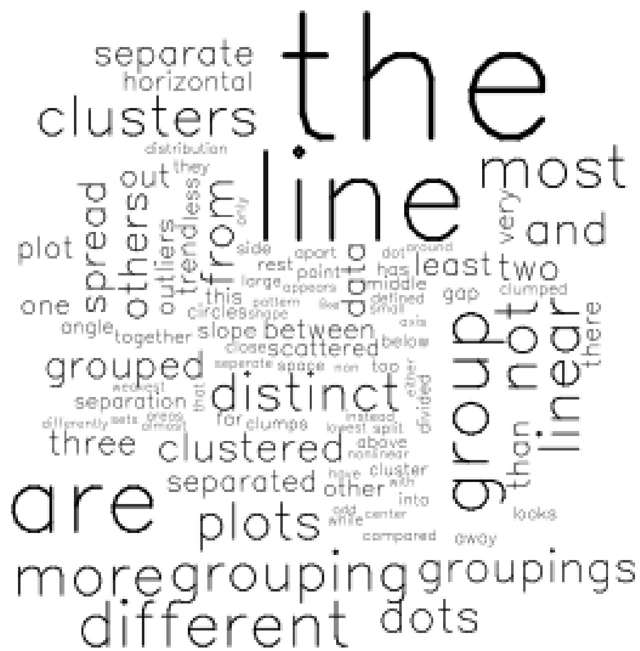
Cluster Target
(N=712)

Trend Target
(N=355)

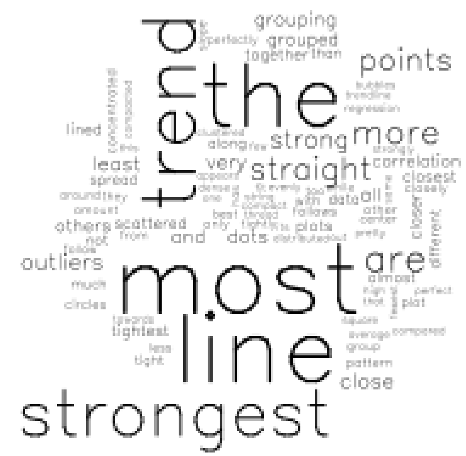
Participant Reasoning

Trend line

Neither Target
(N=159)



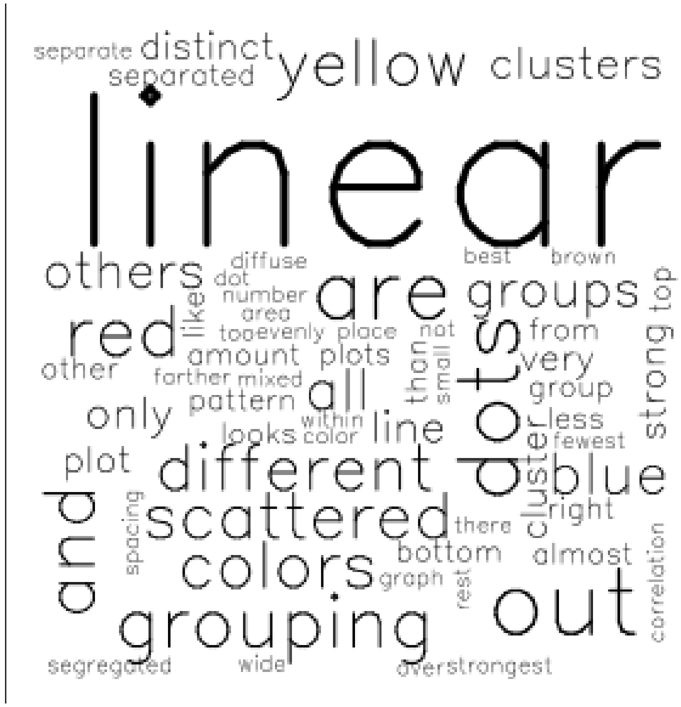
Cluster Target
(N=694)



Trend Target
(N=333)

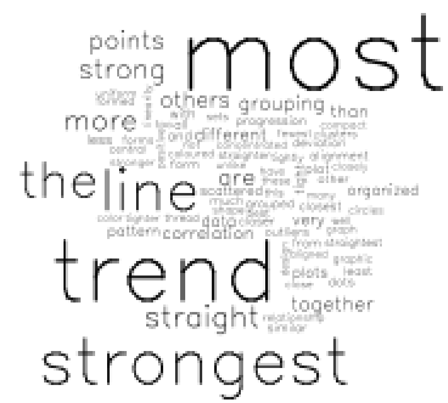
Participant Reasoning

Color Plots



Neither Target
(N=188)

Cluster Target
(N=715)



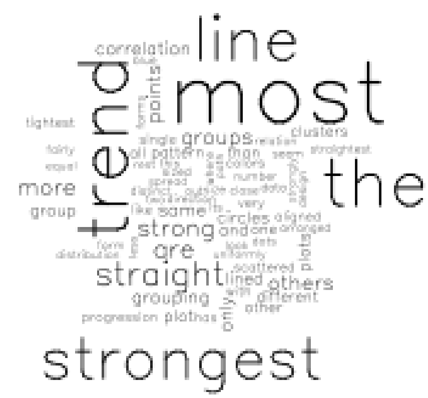
Trend Target
(N=292)

Participant Reasoning

Color + Ellipse Plots

Neither Target
(N=347)

Cluster Target
(N=621)

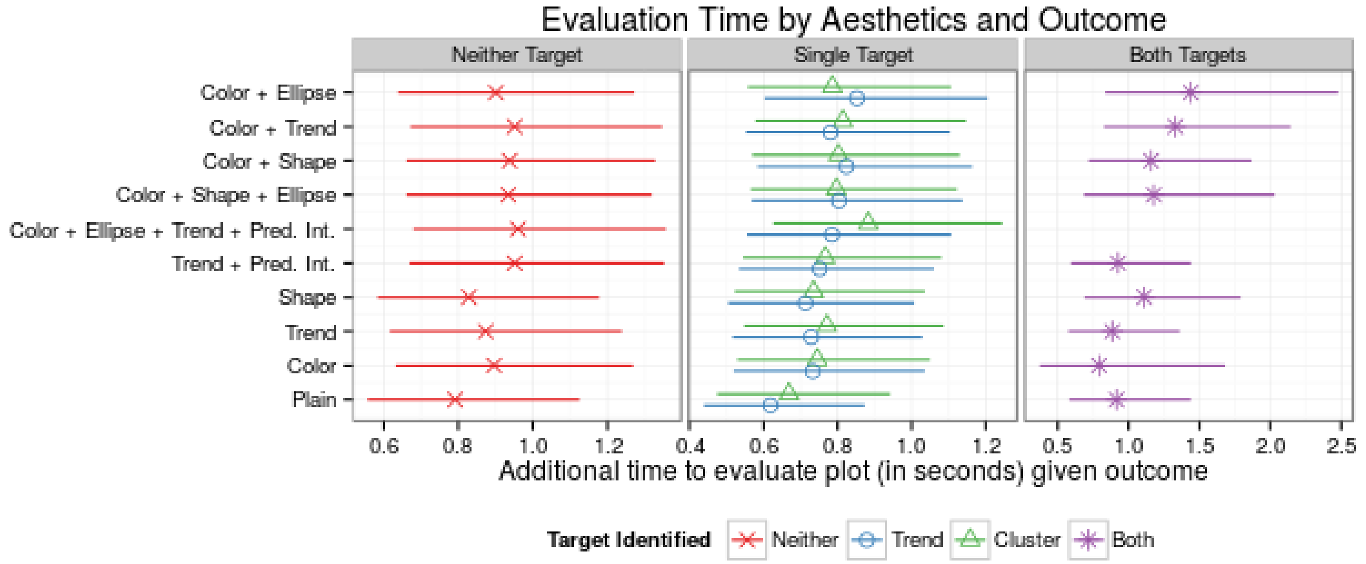


Trend Target
(N=222)

Response Time

Summary: Response Time

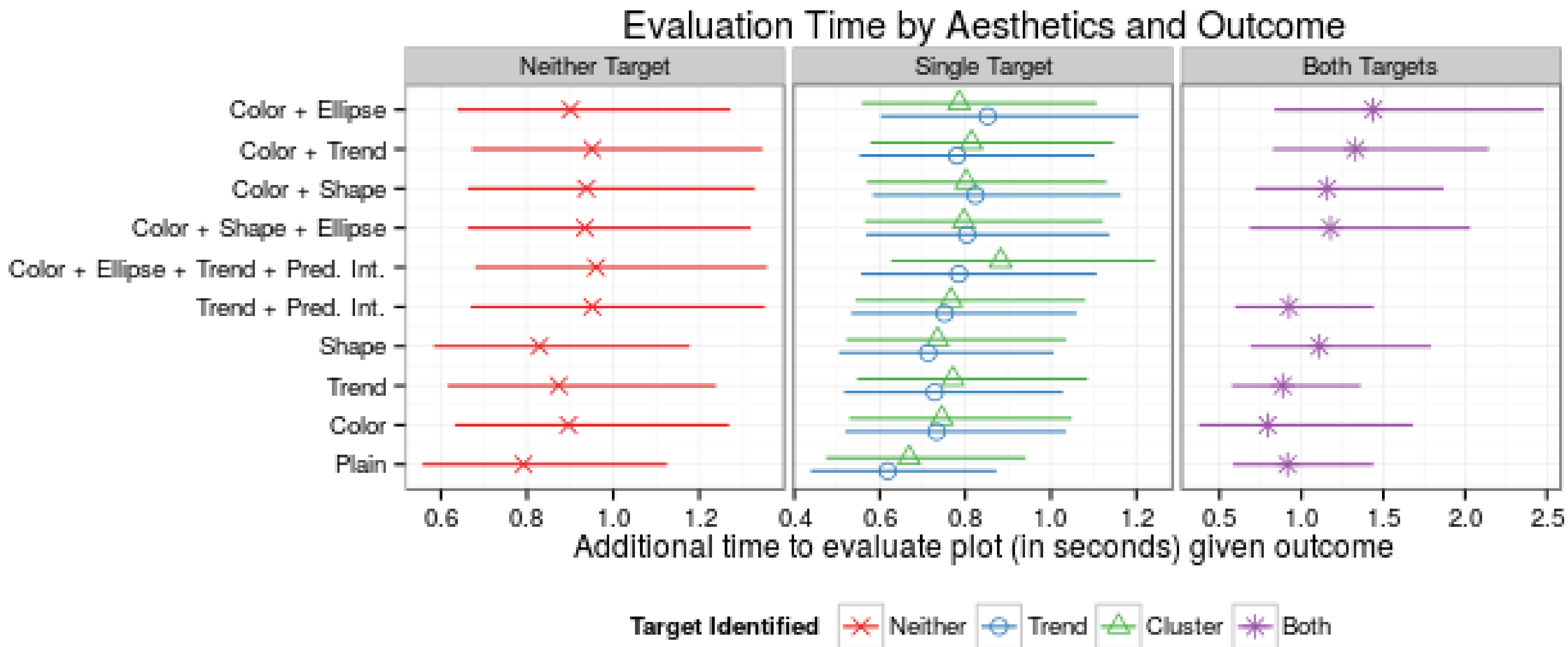
- First trials take more time than subsequent trials
- Participants take more time to evaluate plots with more aesthetics
- Participants who identify a single target are faster than those who did not successfully identify a target; Identifying both targets takes the most time



Response Time

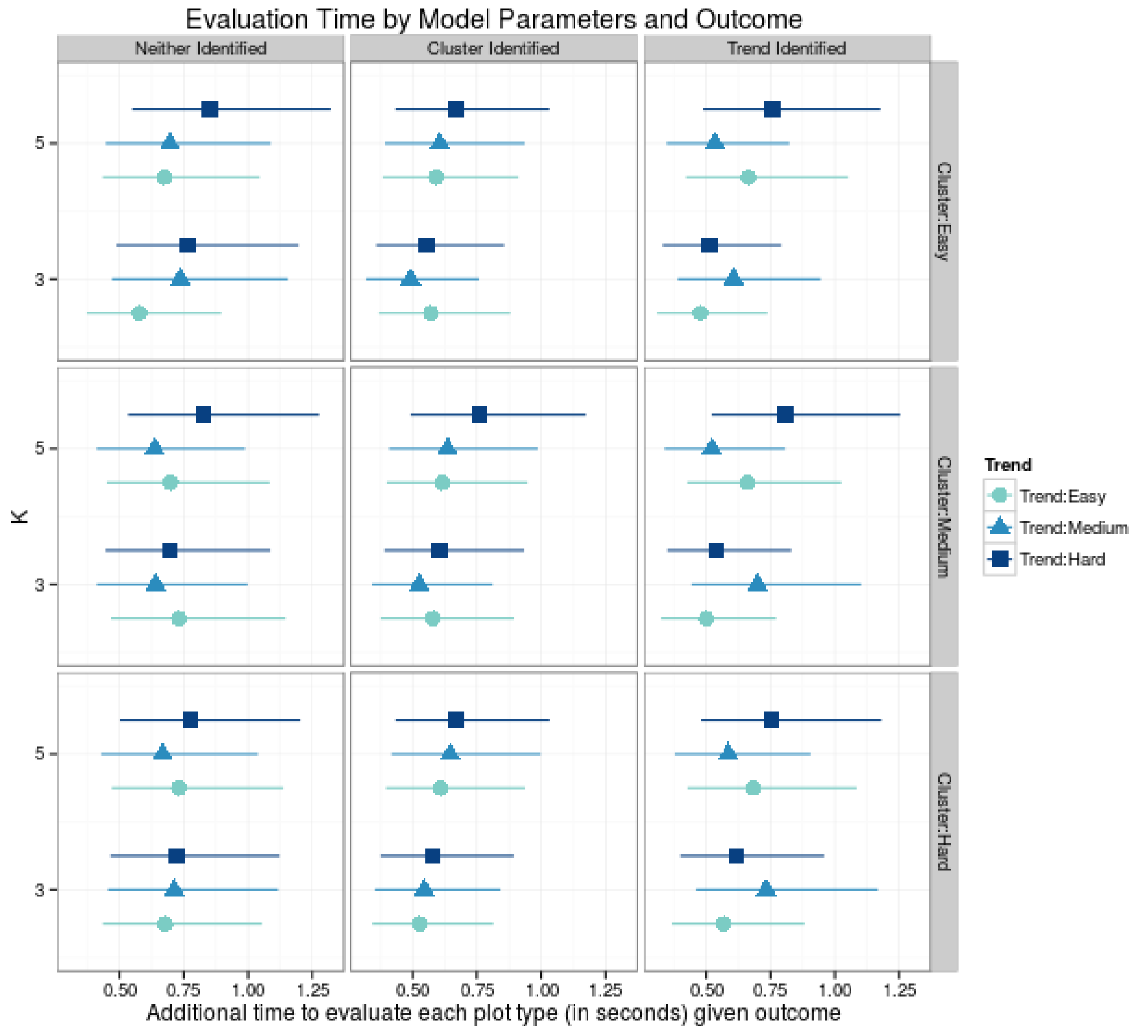
$$
\begin{aligned}
\log(\text{Trial Time}) &\sim \mathbf{X}\beta + \mathbf{J}\gamma + \mathbf{K}\eta + \epsilon \\
\text{where } \beta_i &\text{ describe outcome and aesthetic combinations,} \\
&\text{ plus an initial trial effect} \\
\gamma_j &\overset{iid}{\sim} N\left(0, \sigma^2_{\text{dataset}}\right) \\
\eta_k &\overset{iid}{\sim} N\left(0, \sigma^2_{\text{participant}}\right) \\
\text{and } \epsilon_{ijk} &\overset{iid}{\sim} N\left(0, \sigma^2_e\right)
\end{aligned}
$$

Response Time



Participants take more time to evaluate plots with more aesthetics.
Participants who identified a single target plot were faster than participants who could not identify a target plot.

Response Time



Discussion

Conclusions

- Plot aesthetics influence perception of ambiguous data displays
- Aesthetic effects are not additive:
  Conflict conditions don't show similar/neutral results
- Aesthetics which recruit new gestalt heuristics have more influence, and we can quantify the size of that influence
- Uneven groups in null plots emphasize different features
  (but still show the importance of aesthetic/heuristic interactions)
  - The lineup method allows us to examine <u>why</u> participants switched hypotheses from "Linear relationship" or "Clusters of points" to "Uneven groups"

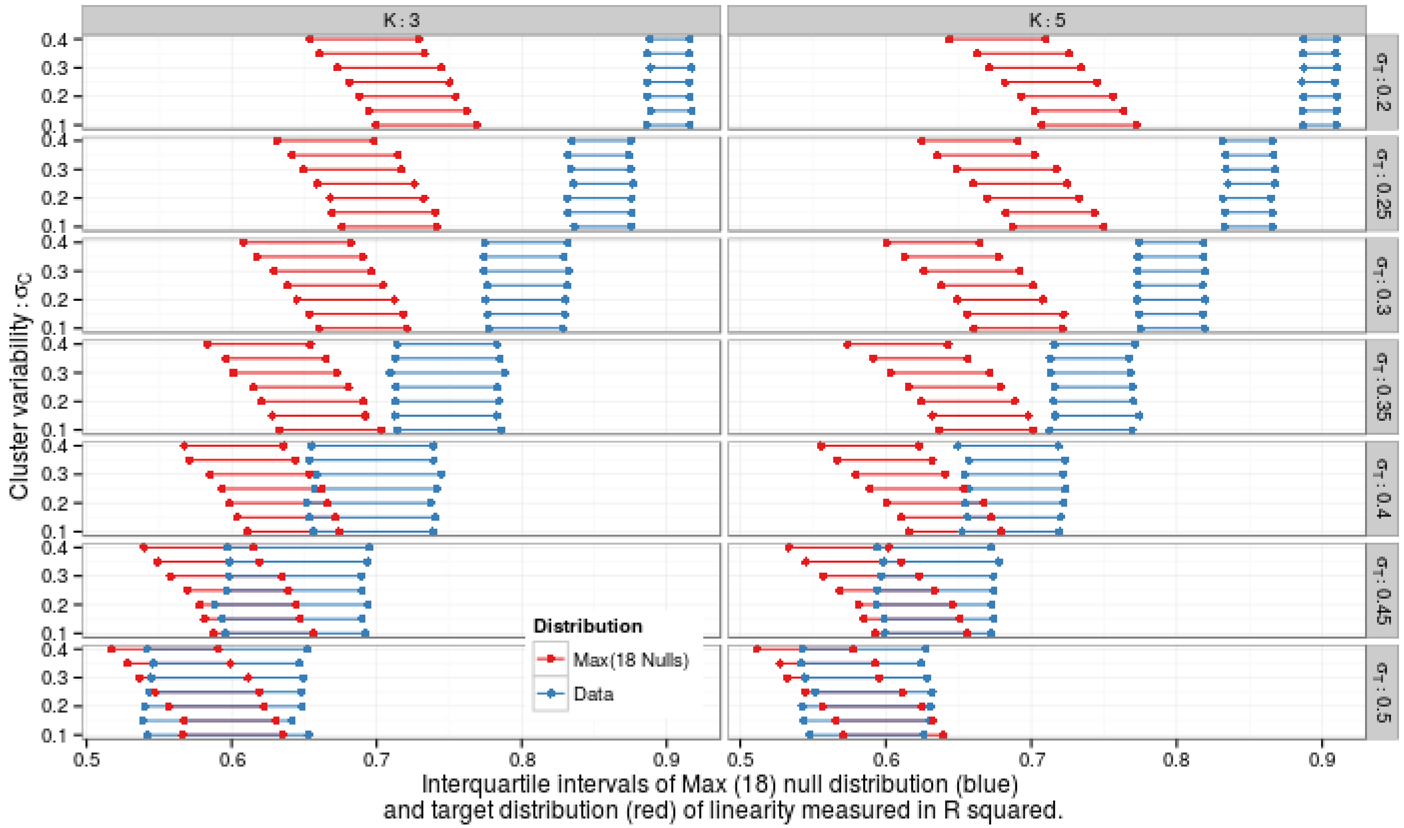
    Similar to a Type III error: the error of giving the right answer to the wrong problem (A.W. Kimball)

Future Work

- Restrict group sizes so null plots have the same objects as target plots
- Explore the effect of different types of common region for error bands and ellipses - shading, bounding boxes, etc.
- Test ellipse and error band aesthetics alone and with trend lines and color to examine interaction effects
- Test plotted statistics (trend line, ellipses, error bands) with and without data points to examine interactions between heuristics from the data and heuristics from summary statistics

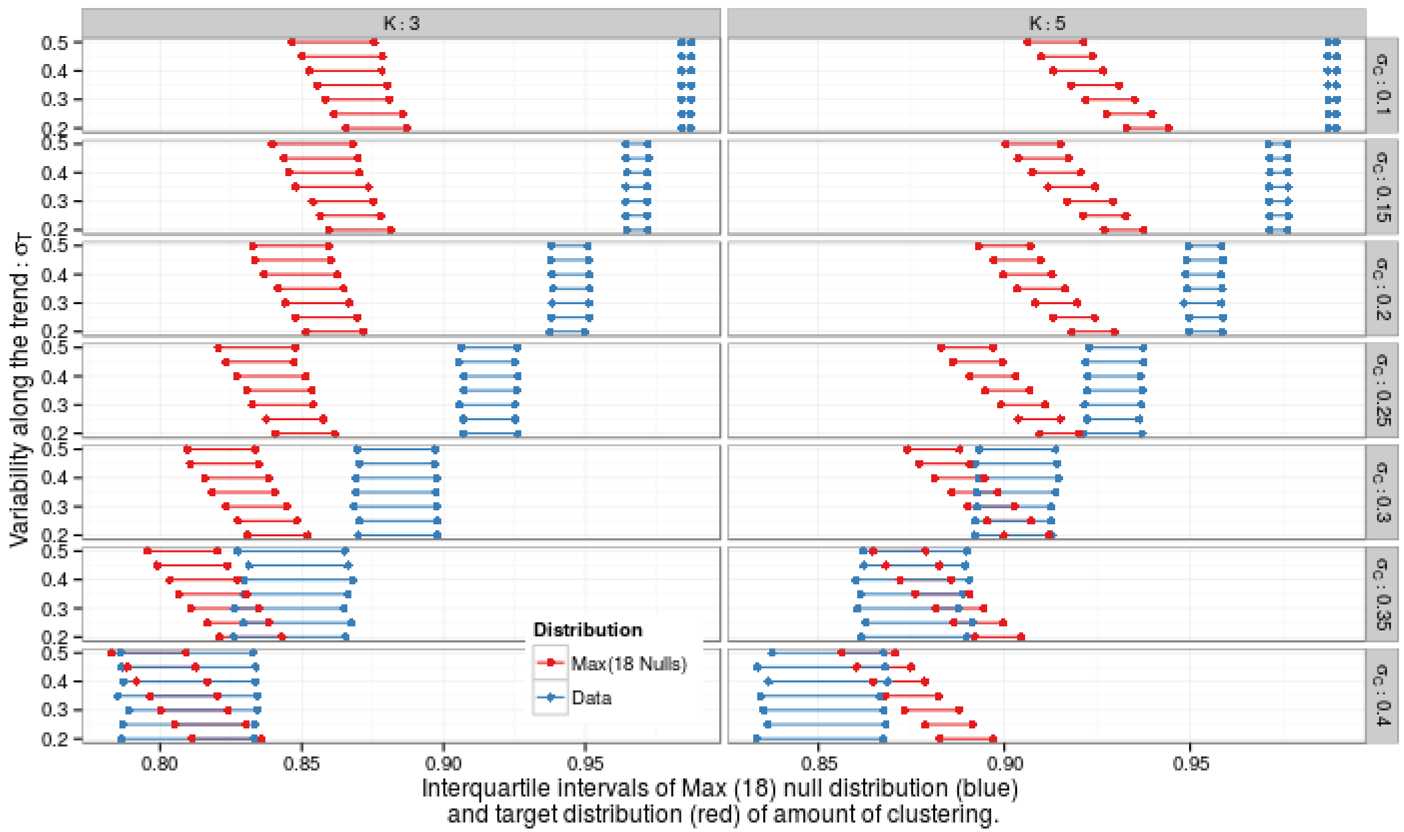
Simulation Details

Simulation: Parameter Space



Values of $\sigma_T = .25, .35, .45$ will provide easy, medium, and hard difficulty levels for trend target identification.

Simulation: Parameter Space



Values of $\begin{matrix} K = 3 : \sigma_C = \{.25, .3, .35\} \\ K = 5 : \sigma_C = \{.2, .25, .3\} \end{matrix}$ will provide easy, medium, and hard difficulty levels for cluster target identification.