Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics

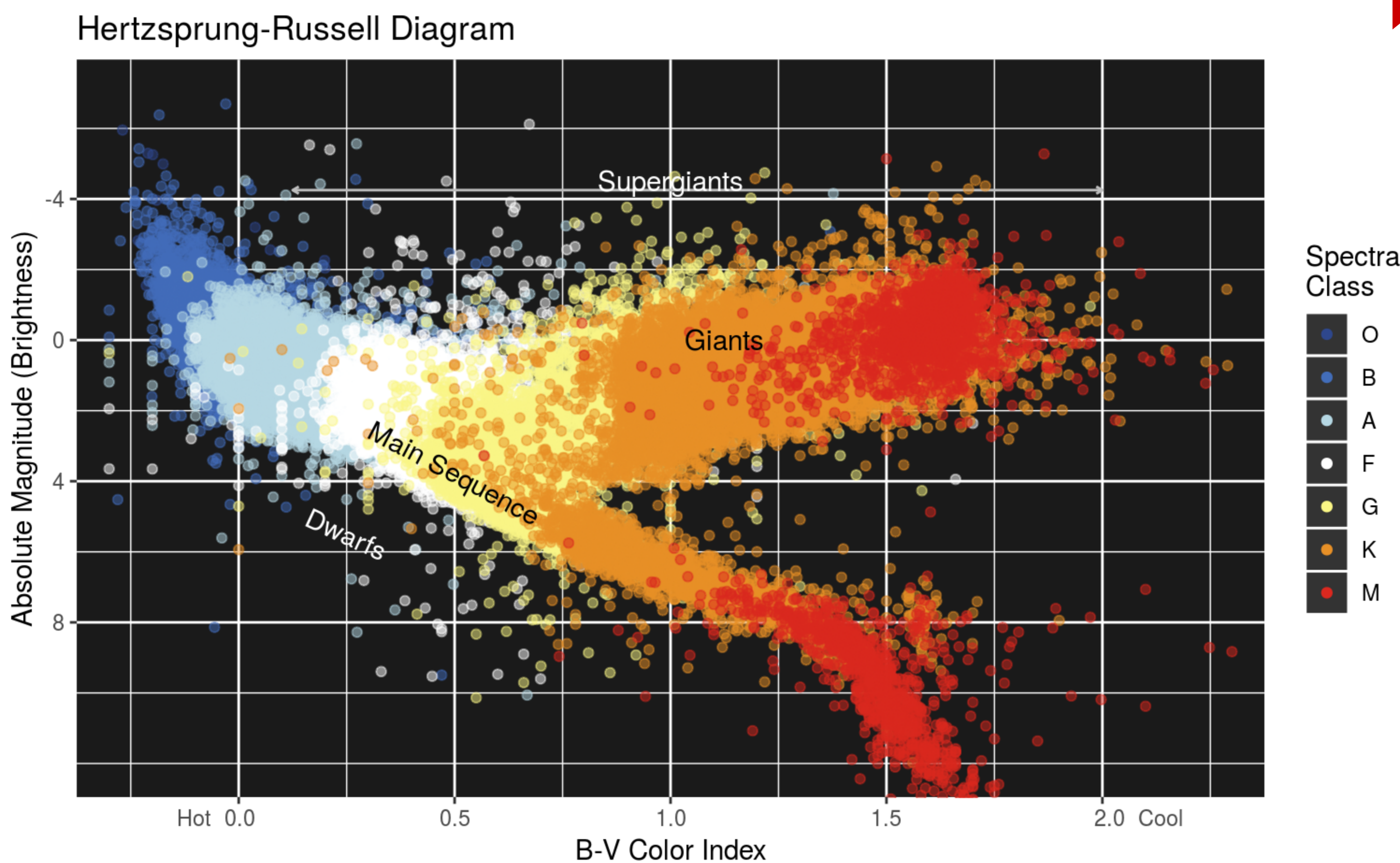
Susan VanderPlas & Heike Hofmann

Iowa State University

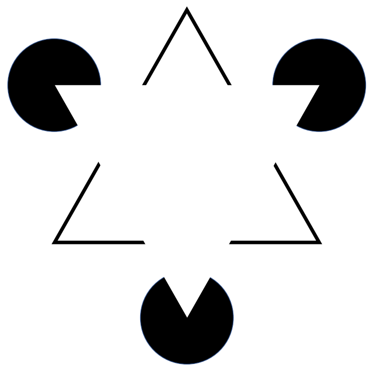
Graphics and Perception

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*

John Tukey


Hertzsprung-Russell Diagram
Supergiants
Giants
Main Sequence
Dwarfs
Absolute Magnitude (Brightness)
-4
0
4
8
Hot 0.0
0.5
1.0
1.5
2.0 Cool
B-V Color Index
Spectral Class
O
B
A
F
G
K
M

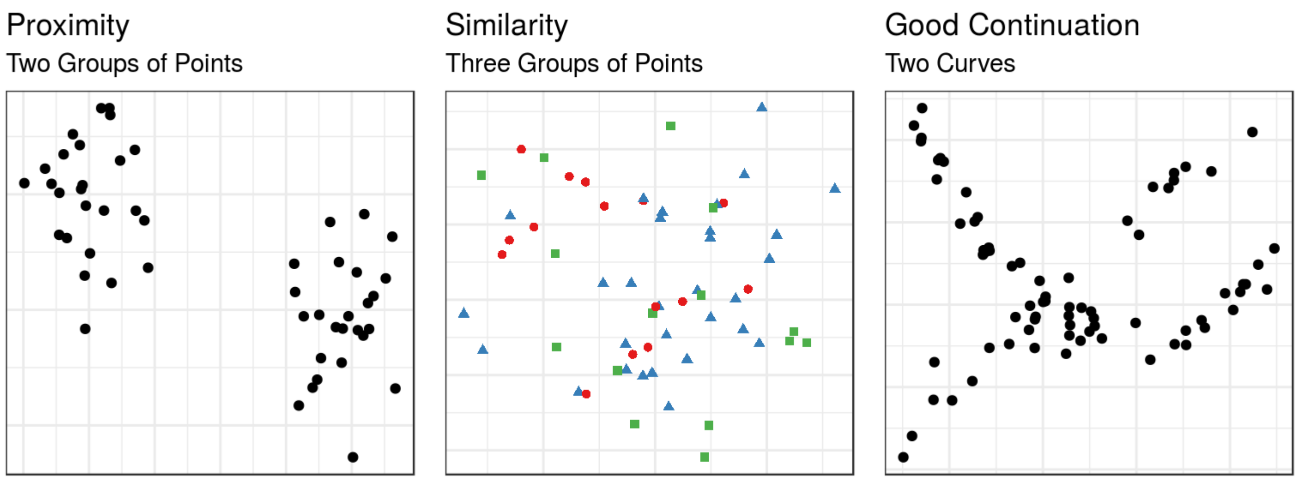
Gestalt Laws of Perception



> *The whole is different than the sum of the parts*

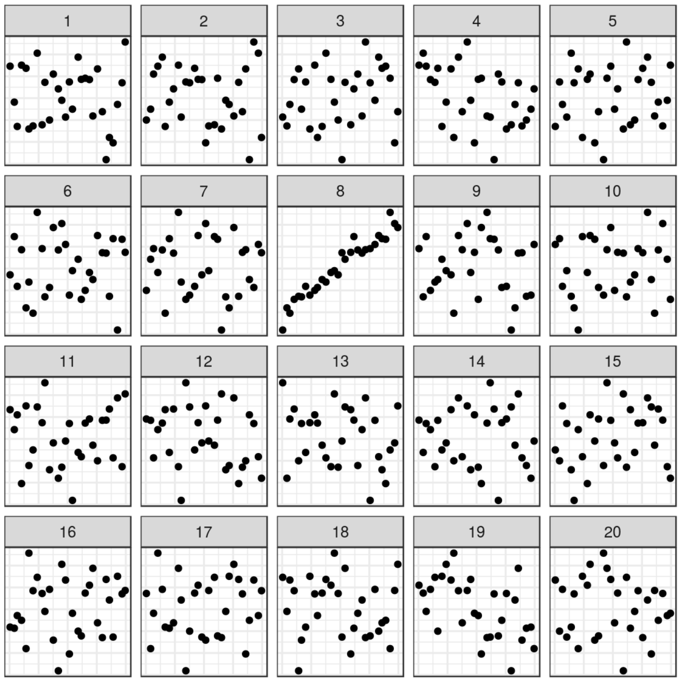
- Rules that make sense of complex visual information using experience
- Information organized hierarchically
- Subconscious process to order and group visual input

Gestalt Plots



How do plot aesthetics change our perception of the plotted data?

Statistical Lineups

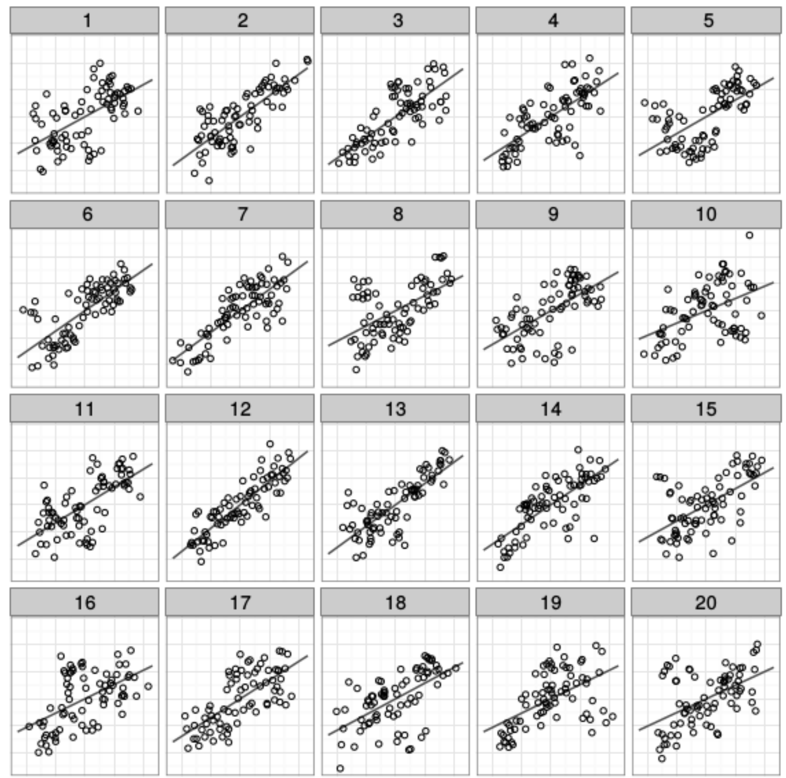


Which plot is the most different?

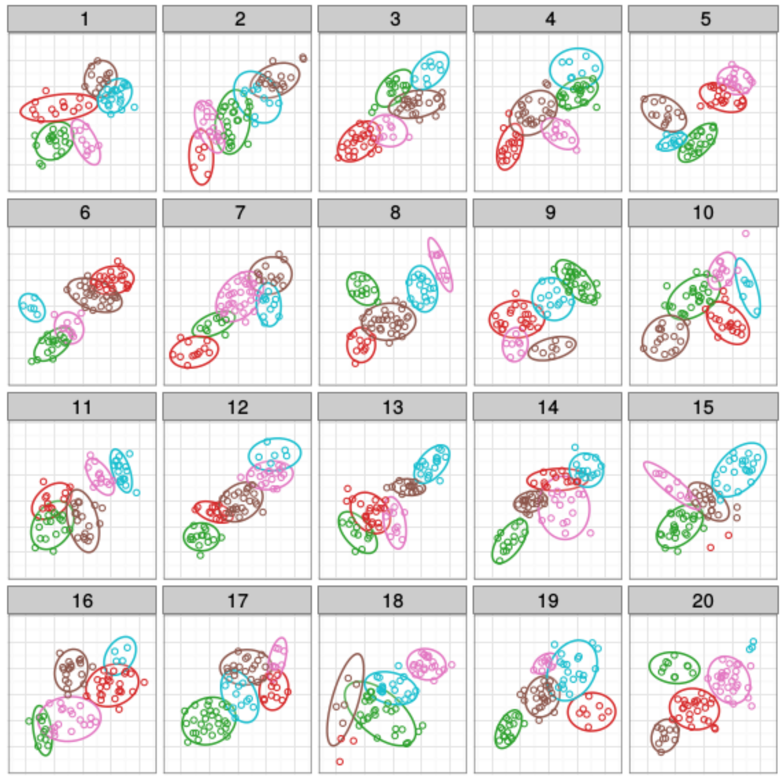
Null plot data is from a data-generating method consistent with the null hypothesis

The `nullabor` package helps with null data creation

Which plots are the most different?

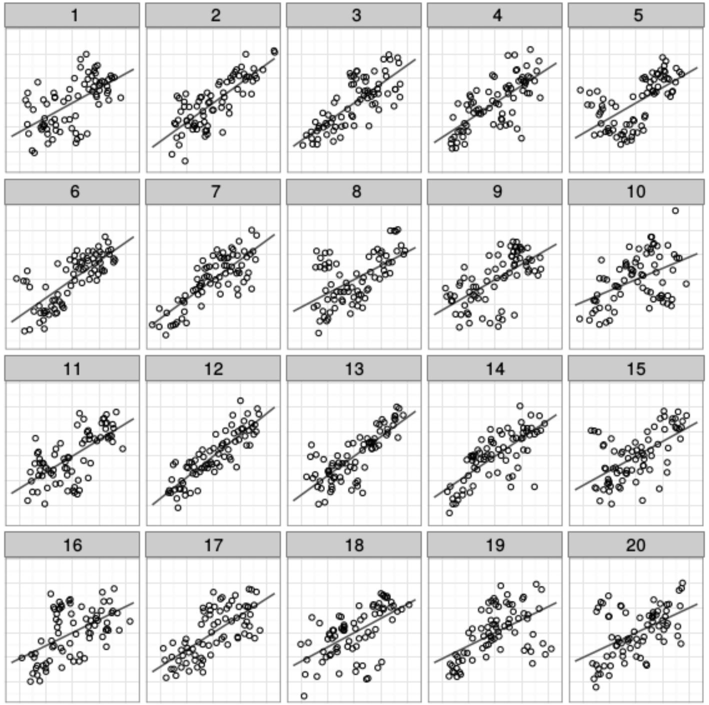


Which plots are the most different?



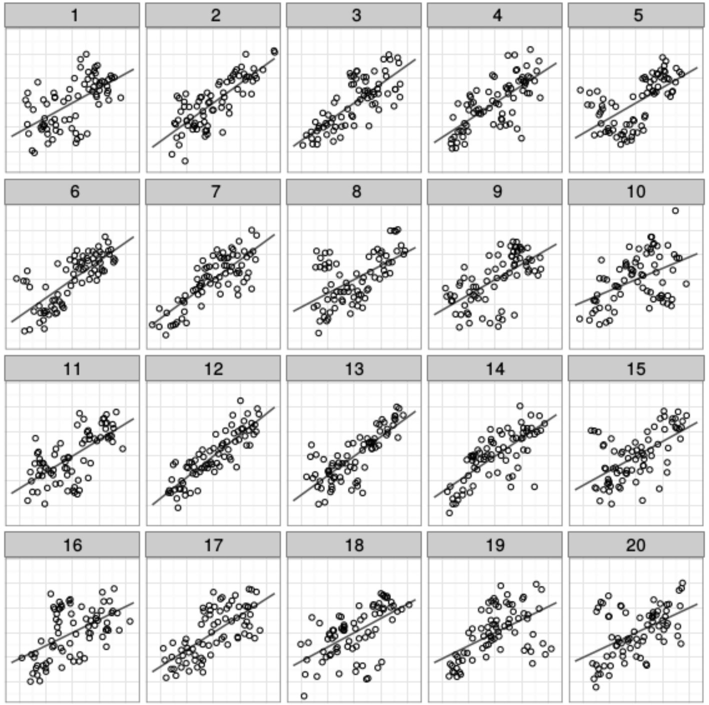
Which plots are the most different?

- 22 Evaluations



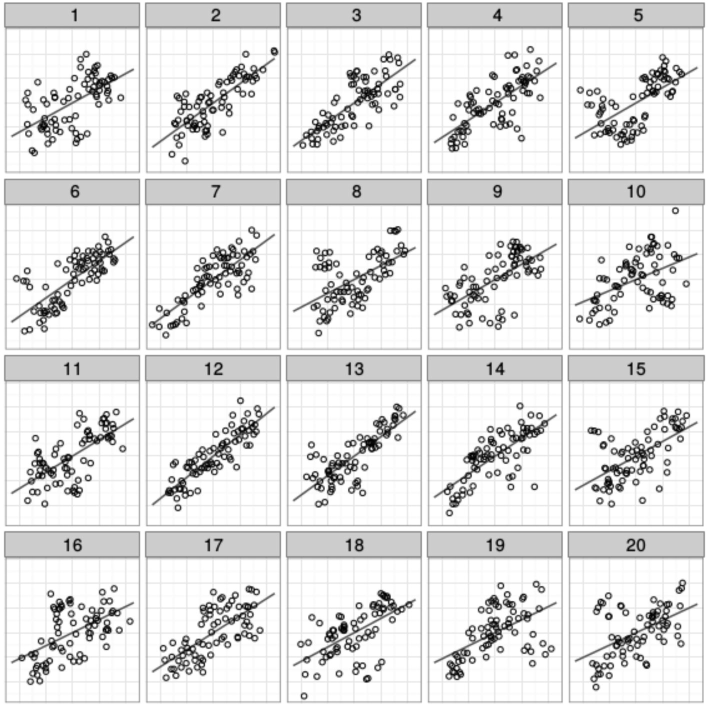
Which plots are the most different?

- 22 Evaluations
- Plot 12: 59.1%



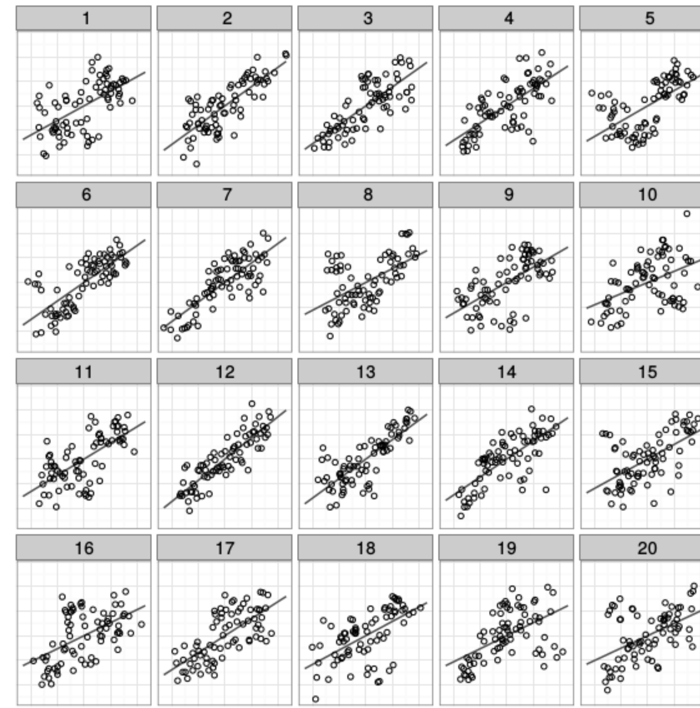
Which plots are the most different?

- 22 Evaluations
- Plot 12: 59.1%
- Plot 5: 9.1%



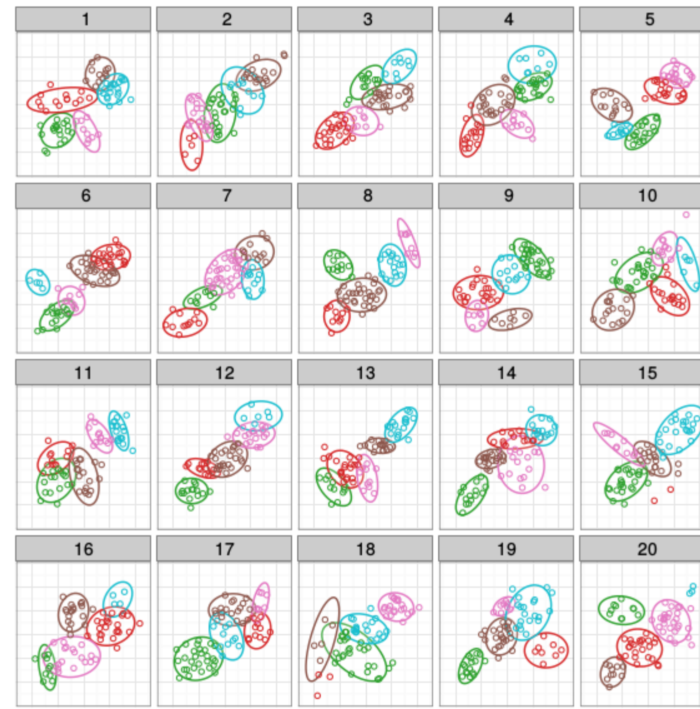
Which plots are the most different?

- 22 Evaluations
- Plot 12: 59.1%
- Plot 5: 9.1%
- Other: 31.7%



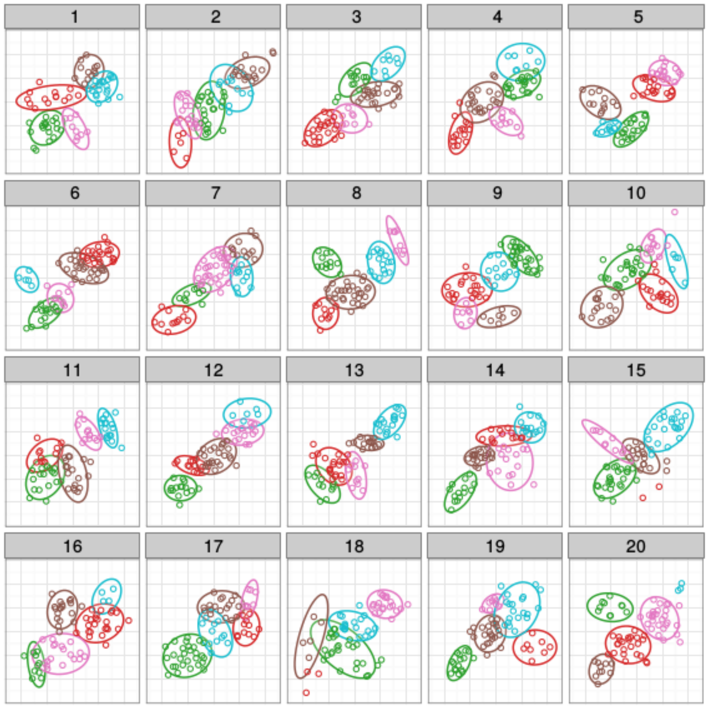
Which plots are the most different?

- 31 Evaluations



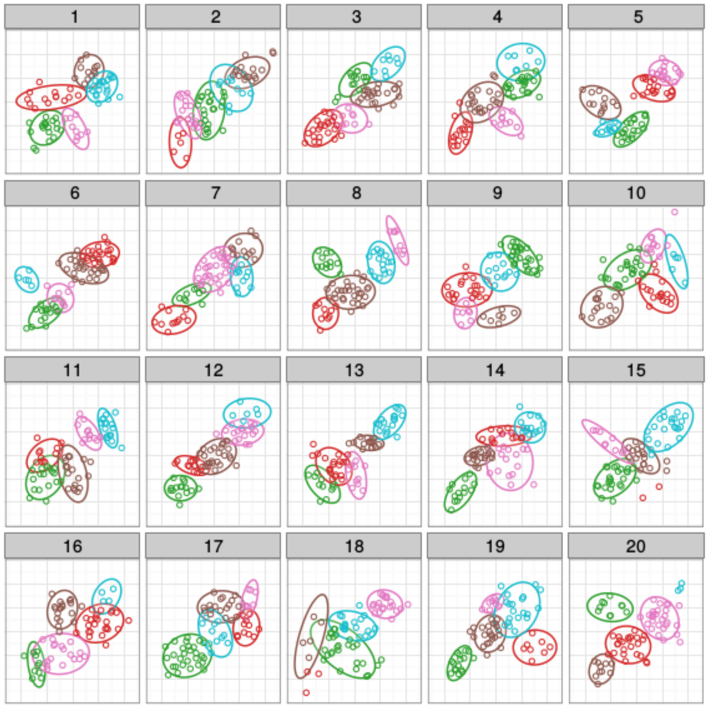
Which plots are the most different?

- 31 Evaluations
- Plot 12: 9.7%



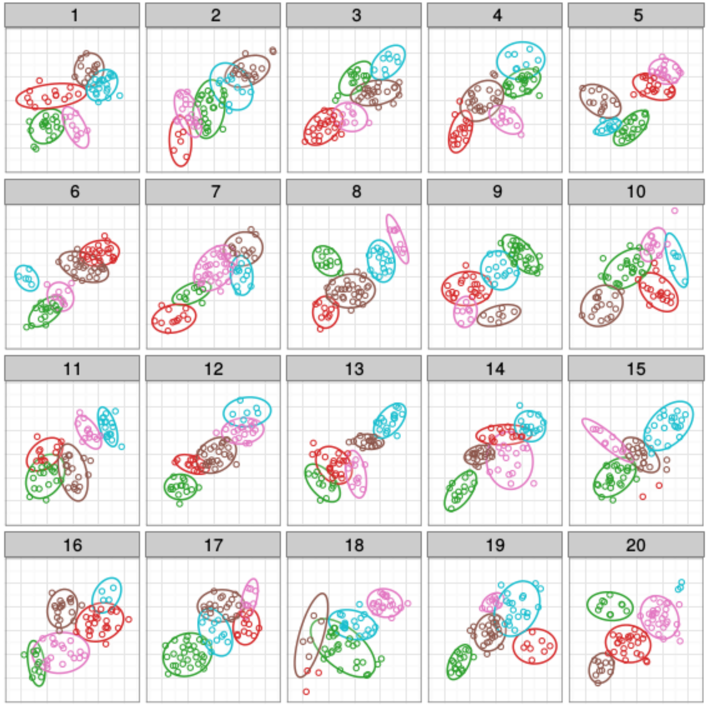
Which plots are the most different?

- 31 Evaluations
- Plot 12: 9.7%
- Plot 5: 29.0%



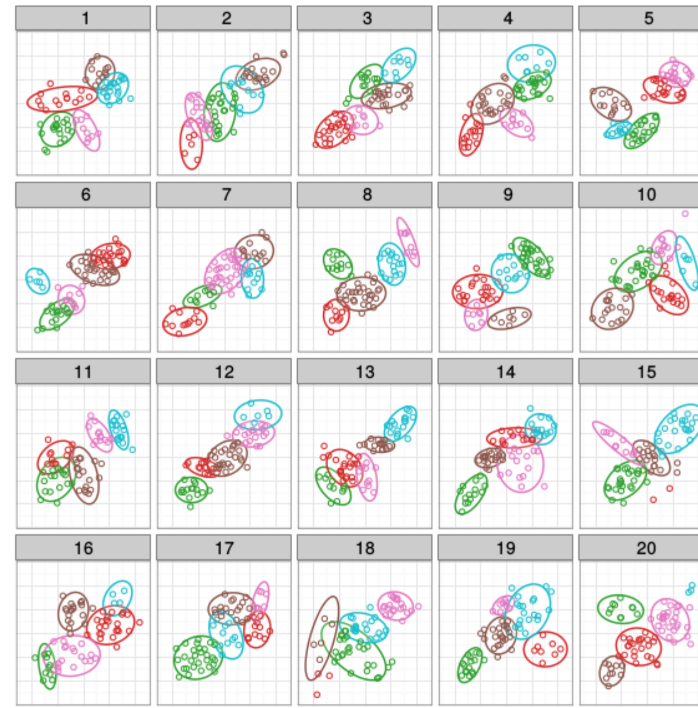
Which plots are the most different?

- 31 Evaluations
- Plot 12: 9.7%
- Plot 5: 29.0%
- Plot 18: 32.3%



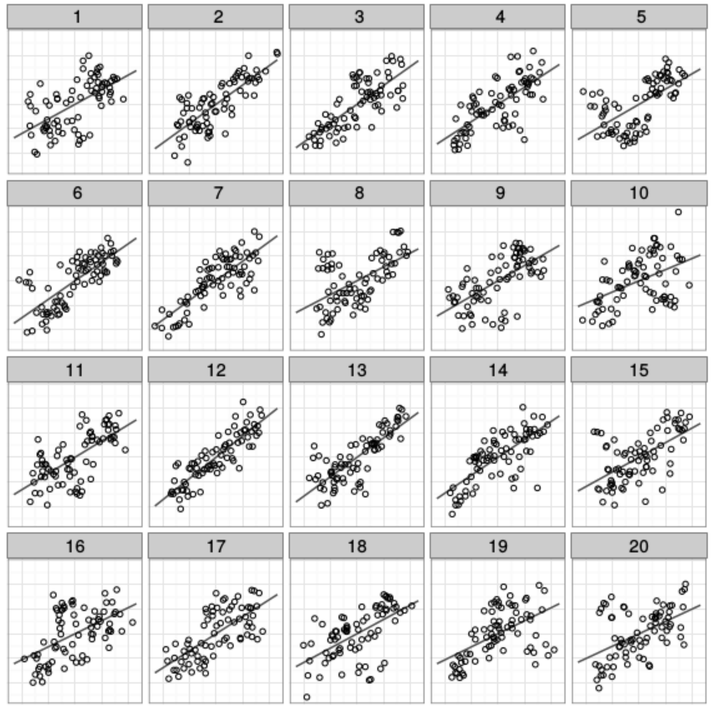
Which plots are the most different?

- 31 Evaluations
- Plot 12: 9.7%
- Plot 5: 29.0%
- Plot 18: 32.3%
- Other: 29.1%



Two-Target Lineups

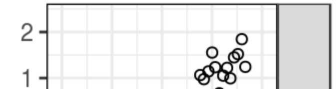
- Modify lineup protocol for tests of competing hypotheses H_1 and H_2
- H_1 and H_2 target plots
- 18 null plots generated using a mixture model consistent with H_0



Data Generating Mechanism

- Generate data from a linear model M_T (trend)
- Generate data from a k cluster model M_C
- Generate null data from a mixture model M_0
  - n_c observations from M_C
  - $n_t = N - n_c$ observations from M_T

Linear Model

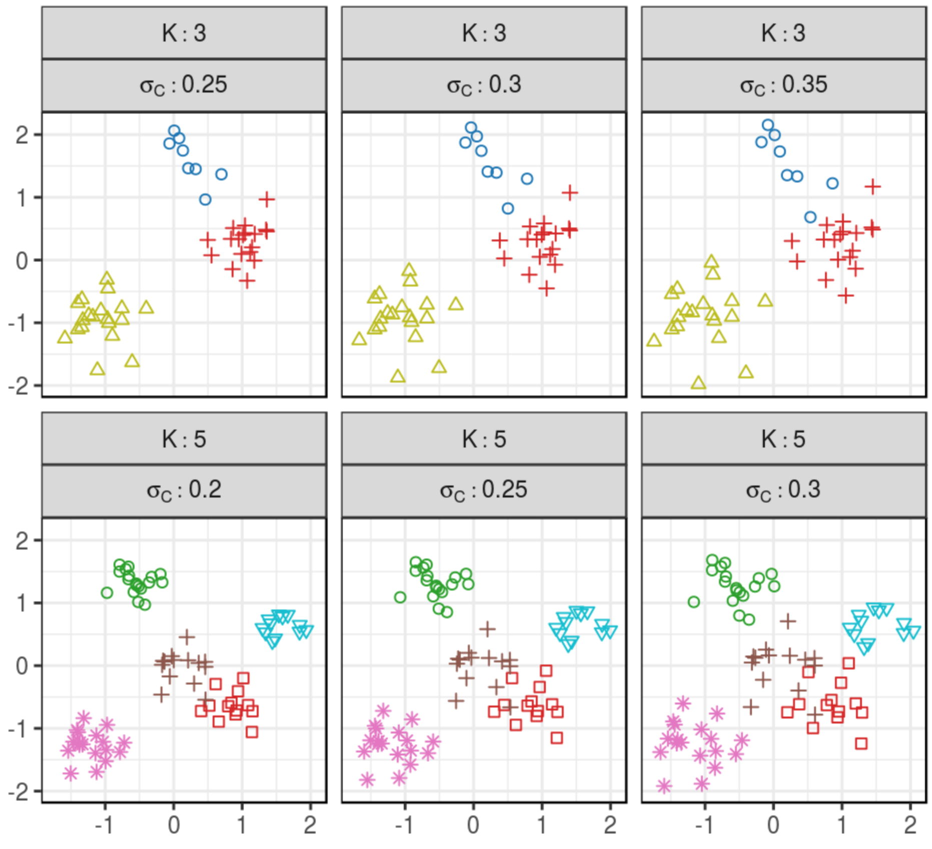


Cluster Model

Parameters: K clusters, σ_C cluster variability

1. Generate K cluster centers c^x, c^y on a $K \times K$ grid such that $cor(c^x, c^y) \in [.25, .75]$
2. Center and standardize c^x, c^y
3. Determine cluster size $g_1, ..., g_K \sim Multinomial(K, p)$
4. Generate points around cluster centers: $(x_i, y_i) = (c^x_{g_i}, c^y_{g_i}) + (e^x_i, e^y_i)$ where $e_i \sim N(0, \sigma_c^2)$
5. Center and scale x_i, y_i

Cluster Model

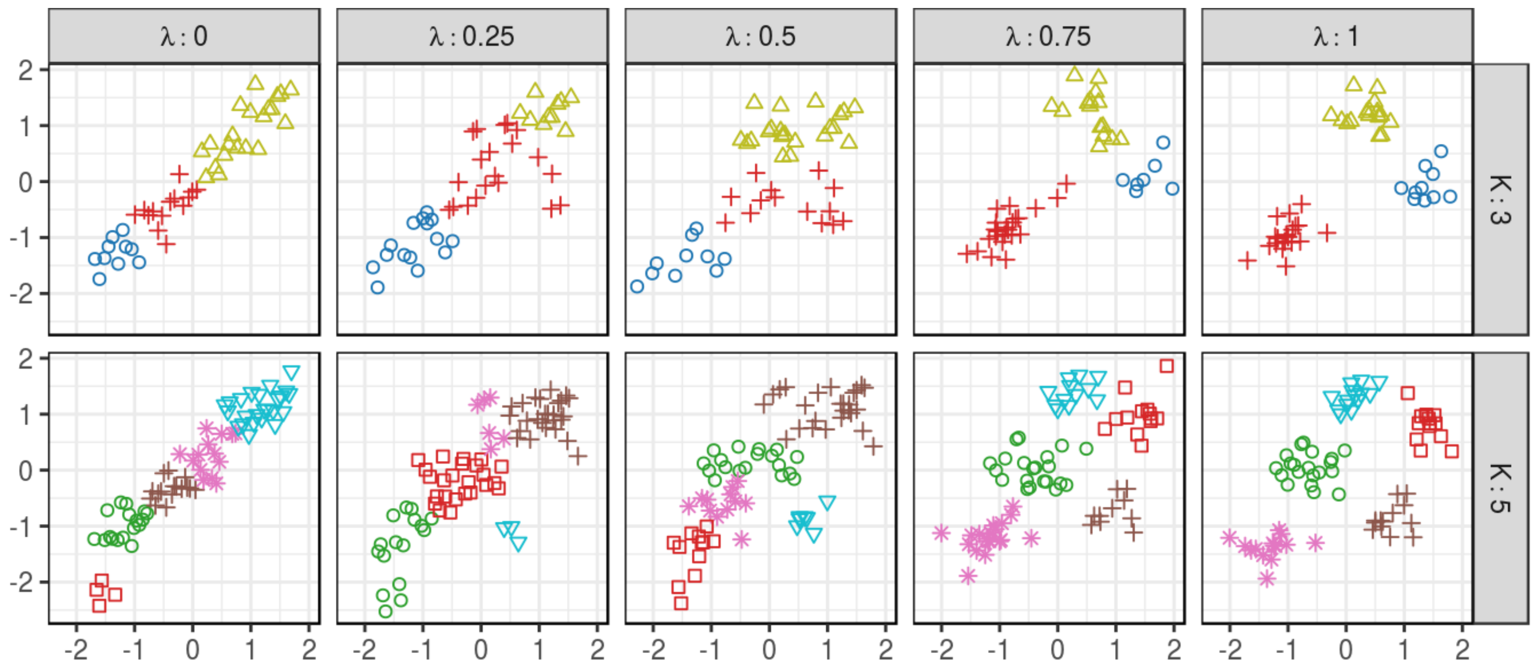


Mixture Model

- n_c points from M_C, where $n_c \sim Binomial(N, \lambda)$
- $N - n_c = n_T$ points from M_T

Groups created by k-means clustering

Mixture Model



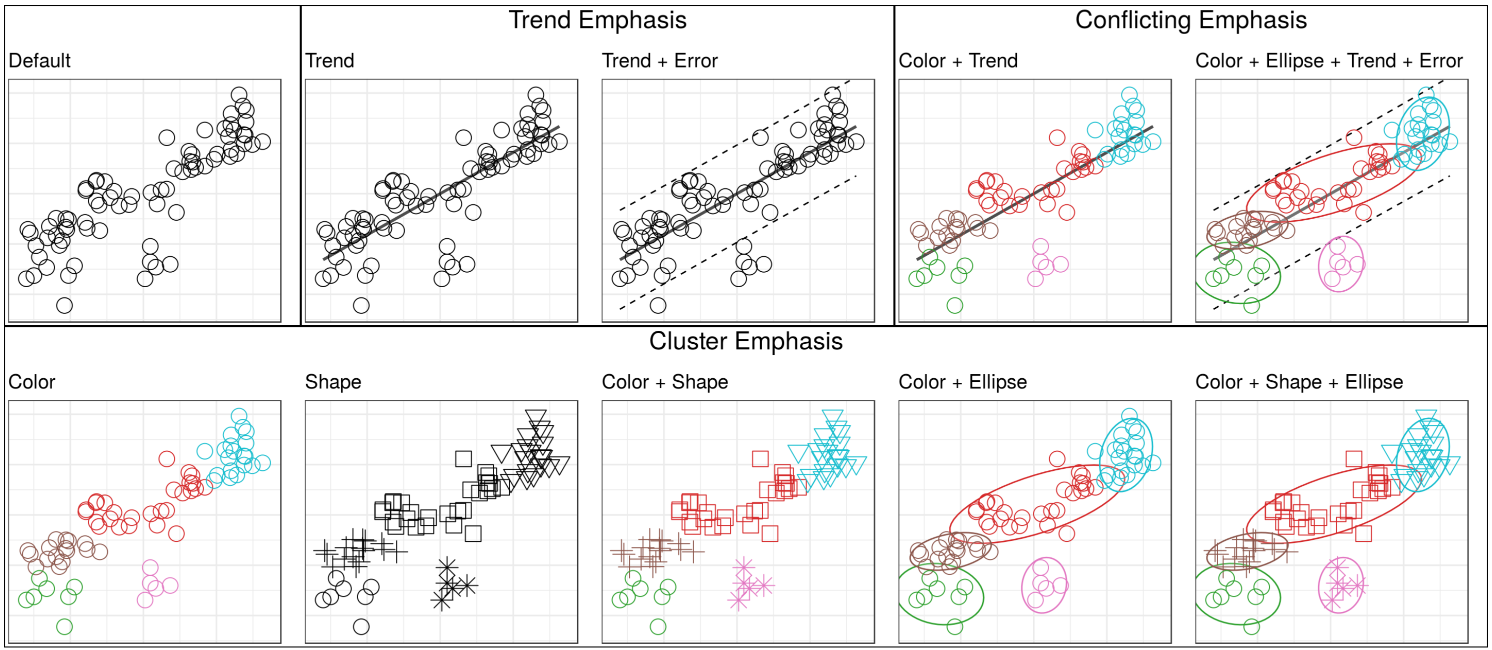
Experimental Design - Data Parameters

- $K = 3, 5$
- $N = 15K$
- $\sigma_T = 0.25, 0.35, 0.45$
- $\sigma_C = \begin{array}{l} 0.25, 0.30, 0.35 (K = 3) \\ 0.20, 0.25, 0.30 (K = 5) \end{array}$
- $\lambda = 0.5$

18 combinations of plot parameters ($2K \times 3\sigma_T \times 3\sigma_C$)

3 replicates of each parameter set; 54 total lineup data sets

Experimental Design - Plot Aesthetics

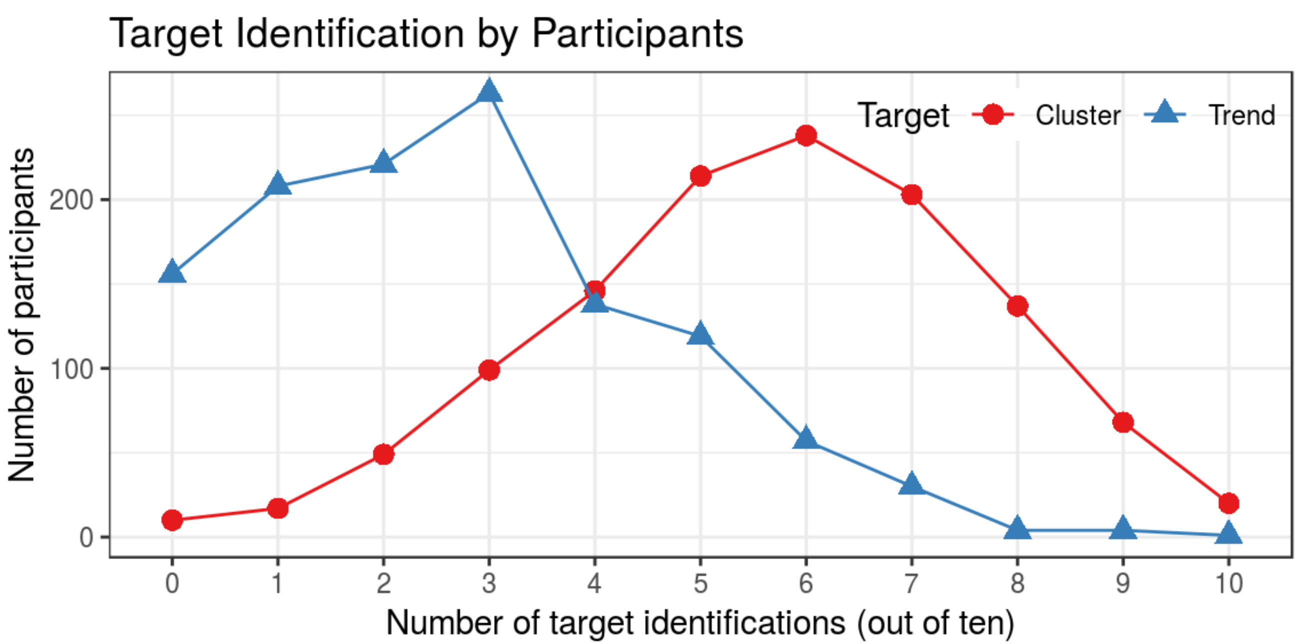


10 Aesthetics × 54 data sets = 540 plots

Experimental Design

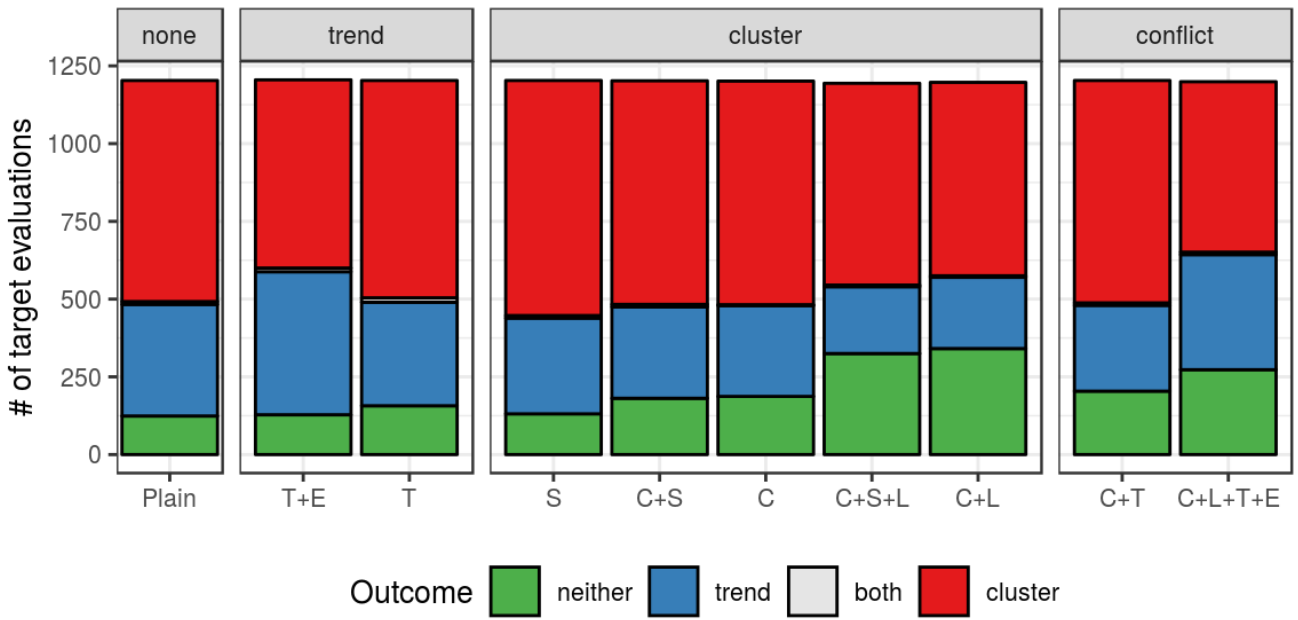
- 1201 participants from Mechanical Turk
- Each participant evaluates 10 plots (12010 evaluations)
  - Each $\sigma_C \times \sigma_T$ value with one replicate, randomized across K values
  - All 10 aesthetic types
- Participants select the plot or plots which are most different
  - Provide a short explanation
  - Rate confidence level

Results



Most participants identified a mix of cluster and trend targets

Results


none
trend
cluster
conflict
# of target evaluations
0
250
500
750
1000
1250
Plain
T+E
T
S
C+S
C
C+S+L
C+L
C+T
C+L+T+E
Outcome
neither
trend
both
cluster

Faceoff Model

- Examine trials in which participants identified at least one target (9959)
- Compare P(select cluster target) to P(select trend target)

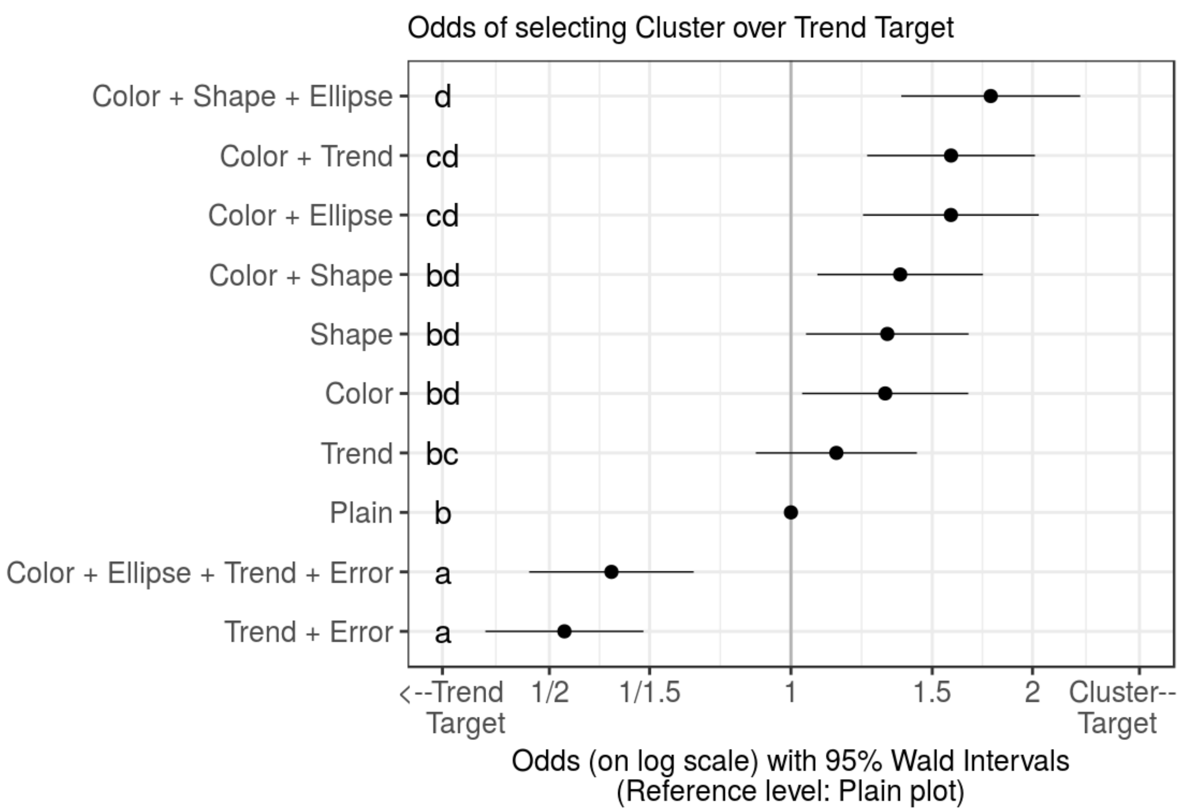
$$C_{ijk} := \left\{ \begin{array}{c} \text{Participant } k \text{ selects the cluster target} \\ \text{for dataset } j \text{ with aesthetic } i \end{array} \right\}$$

Faceoff Model

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

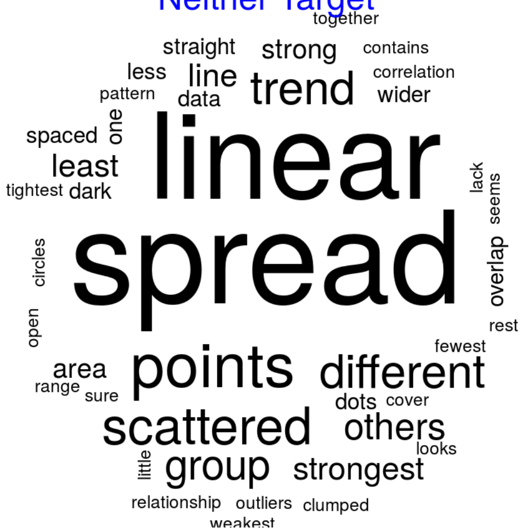
- α: vector of fixed effects describing the effect of data parameters σ_C, σ_T, K
- β: vector of fixed effects describing the effect of aesthetics $1 \leq i \leq 10$
- γ_j: random effect of dataset, $\gamma_j \sim N(0, \sigma^2_{\text{data}})$
- η_k: random effect of participant $\eta_k \sim N(0, \sigma^2_{\text{participant}})$
- ϵ_{ijk}: error associated with single evaluation of plot ij by participant k, $\epsilon_{ijk} \sim N(0, \sigma^2_e)$

Faceoff Model

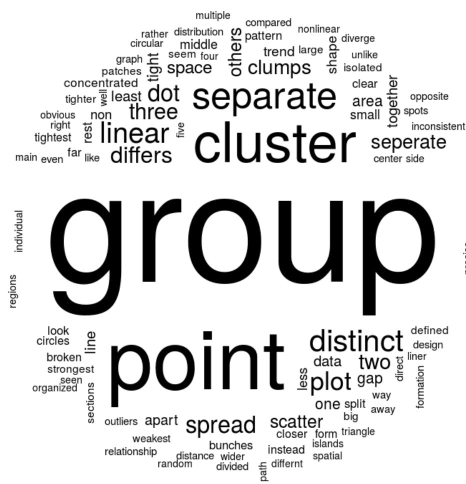


Participant Reasoning: Plain plots

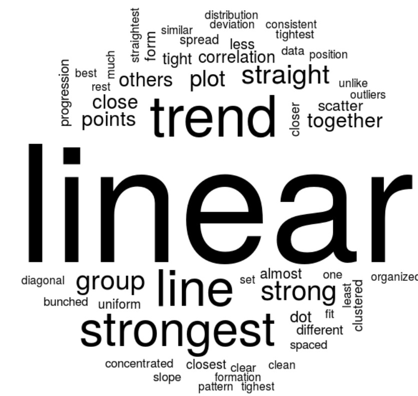
Neither Target



Cluster Target

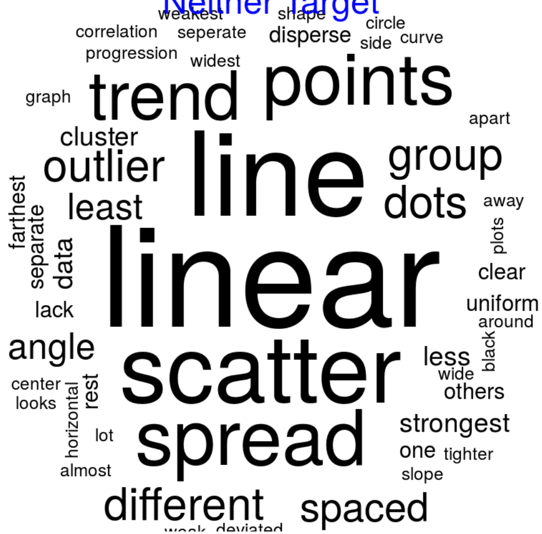


Trend Target

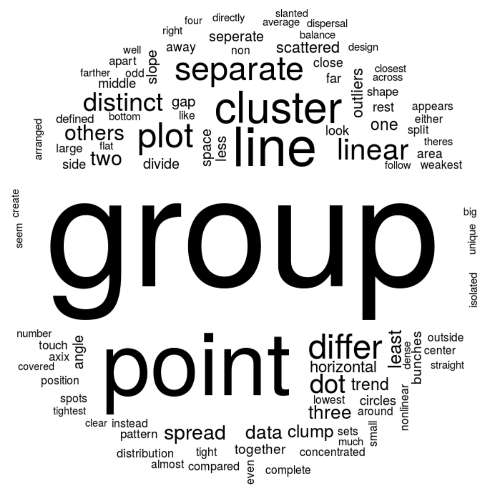


Participant Reasoning: Trend plots

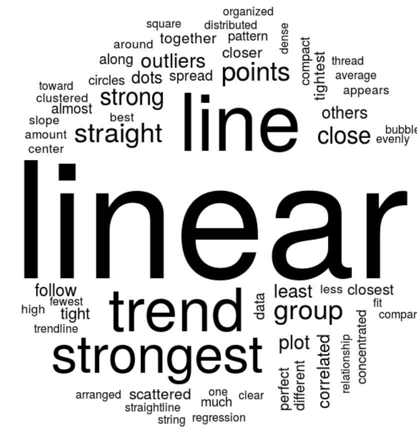
Neither Target



Cluster Target

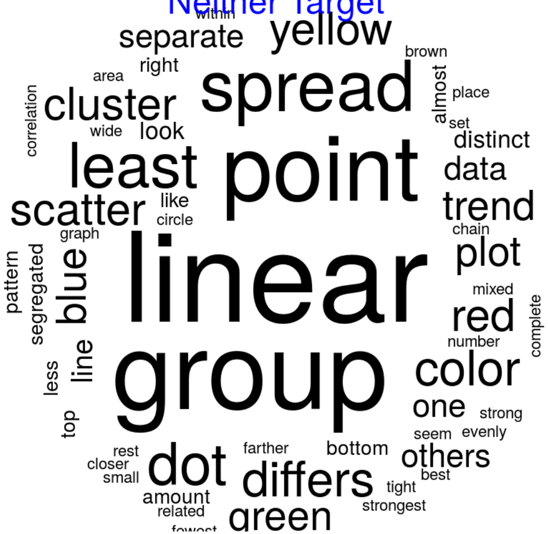


Trend Target

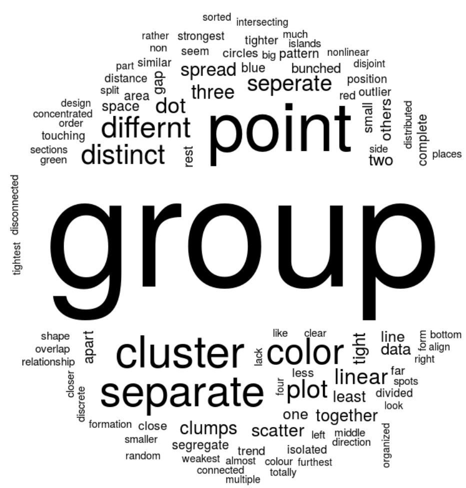


Participant Reasoning: Color plots

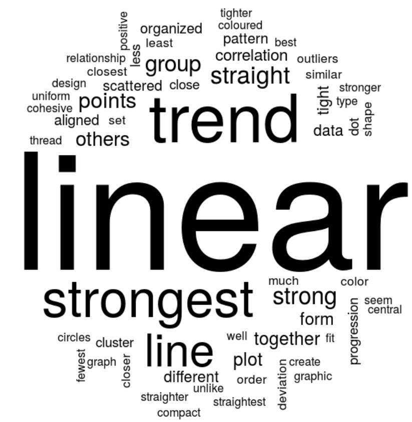
Neither Target



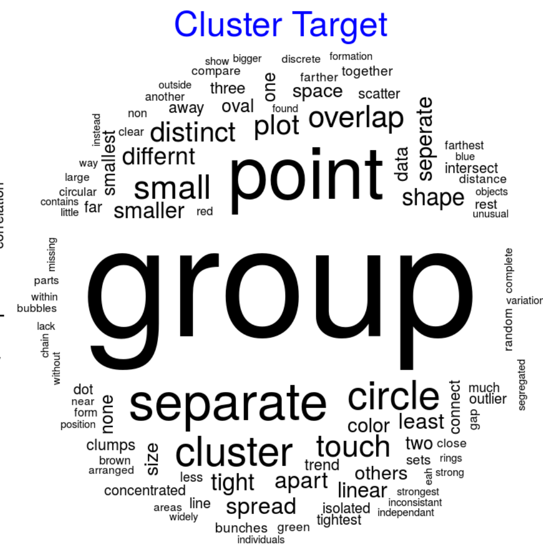
Cluster Target

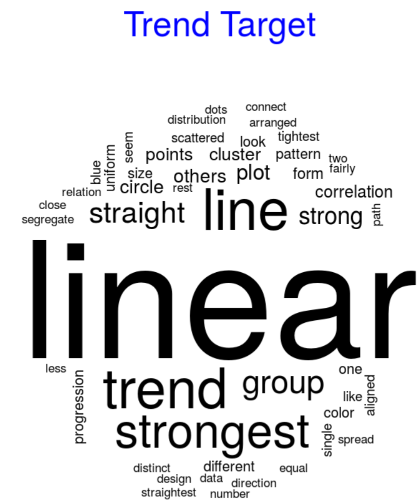


Trend Target

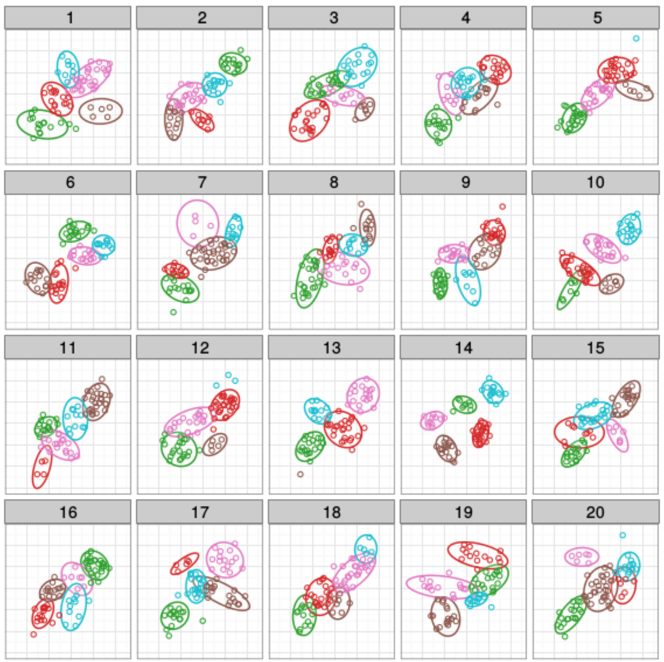


Participant Reasoning: Color + Ellipse plots





Participant Reasoning



Conclusion

- Plot aesthetics matter
  - non-additive effects
  - what do you want to emphasize?
- Multiple encoding is useful -

  “show the data” in a way that makes it easy to understand

Conclusion

- Error bands and cluster ellipses highlight important features in the data: outliers, group size inequality, variability, clustering
- Null data-generating models are hard!

  The brain runs 100s of visual "tests" and designing for all of them simultaneously is impossible