Eurographics Conference on Visualization (EuroVis) 2018
J. Heer, H. Leitte, and T. Ropinski
(Guest Editors)

Volume 37 (2018), Number 3

Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings

Younghoon Kim[1] and Jeffrey Heer[1]

[1]University of Washington

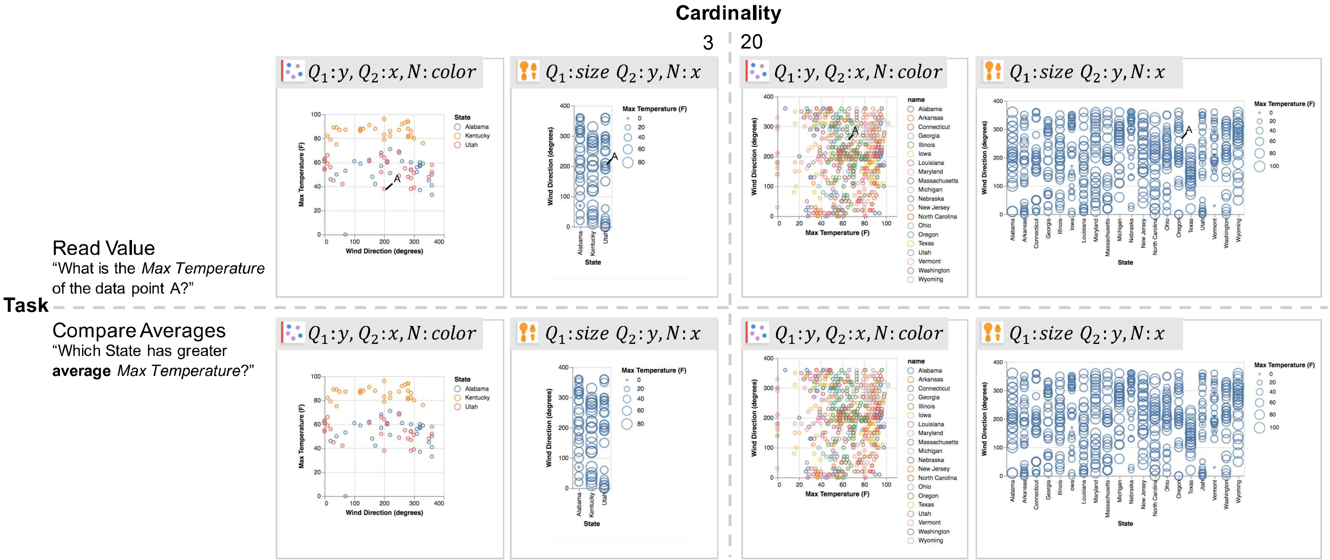


Figure 1: *A subset of experiment stimuli across tasks (Read Value and Compare Averages) and cardinalities (3, 20). Each stimulus is labeled with its mapping from data fields (Q_1, Q_2, N) to visual channels (x, y, color, size).*

Abstract

In addition to the choice of visual encodings, the effectiveness of a data visualization may vary with the analytical task being performed and the distribution of data values. To better assess these effects and create refined rankings of visual encodings, we conduct an experiment measuring subject performance across task types (e.g., comparing individual versus aggregate values) and data distributions (e.g., with varied cardinalities and entropies). We compare performance across 12 encoding specifications of trivariate data involving 1 categorical and 2 quantitative fields, including the use of x, y, color, size, and spatial subdivision (i.e., faceting). Our results extend existing models of encoding effectiveness and suggest improved approaches for automated design. For example, we find that colored scatterplots (with positionally-coded quantities and color-coded categories) perform well for comparing individual points, but perform poorly for summary tasks as the number of categories increases.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces]: User Interfaces—Evaluation

1. Introduction

Automated visualization design can promote effective encodings and facilitate rapid visual exploration [Mac86, Cas91, WMA*16, MHS07]. Traditional automatic visualization systems [Mac86, Cas91] use a greedy heuristic: given a list of importance-ordered data fields, they assign the highest priority field to the most effective visual channel according to pre-specified channel effectiveness rankings. They iteratively repeat this process for the next field and remaining encoding channels. More recent systems [MHS07, WMA*16] apply hand-tuned scores or constraints to specify more fine-grained preferences. However, their evaluation criteria remain limited, focused only on encoding channels and data types, with effectiveness criteria derived primarily in the context of individual value comparison tasks [Ber83, CM84].

Here, we extend prior work by additionally considering the effects of task (individual value vs. summary comparisons) and data distribution (including varying numbers of data points, cardinalities of categorical fields, and entropies of quantitative fields). To provide more detailed effectiveness criteria, we contribute a crowdsourced experiment with 1,920 subjects. We measure performance (completion time and error rate) across 12 encoding specifications of trivariate data involving 1 categorical field and 2 (primary and secondary) quantitative fields, including use of *x*, *y*, *color*, *size*, and *row* (faceting) channels. From our experimental results we offer contributions in three categories.

First, we measure the impact of task and data distribution on the effectiveness rankings of visual encodings. As one might expect, we find that colored scatterplots perform well for comparing individual values. On the other hand, they perform poorly for tasks involving aggregate judgments over a high-cardinality set of category values. We catalog such discrepancies to form nuanced effectiveness rankings parameterized by task and distribution.

Second, we identify interactions between visual channels within multidimensional visualizations. For example, *size* is often considered preferable to *color* for encoding quantitative fields [CM84]. However, we observe that encoding a secondary quantitative information using *size* can impede decoding of a primary quantitative field that is spatially-encoded (*x* or *y*).

Finally, we discuss how to apply our findings to improve automated visualization design systems. We describe the use of dynamic effectiveness criteria that are sensitive to task and data distributions, and incorporate interactions between channels. We conclude with a discussion of future research directions.

2. Related Work

Our study extends prior work on the effectiveness of visual encodings and automated visualization design, while drawing on research characterizing visual analysis tasks and data distributions.

2.1. Effectiveness of Visual Encodings

A rich body of work has examined empirical user performance with visual encodings [War12], validating and refining effectiveness rankings originally proposed by Bertin [Ber83]. Cleveland & McGill [CM84] conducted human-subjects experiments measuring decoding error across encoding channels, for example finding that position encodings outperform length encodings when comparing proportions. Heer & Bostock [HB10] later replicated and extended these results using crowdsourced participants. The effectiveness rankings of visual encodings by (nominal, ordinal, quantitative) data type produced by such studies in turn serve as a foundation for automatic visualization systems [Mac86, WMA*16].

In addition to studies of univariate encodings, researchers have examined interactions between visual encoding channels [GF70, War12]. For example, *integral* visual channels (*e.g., color* and *size*) may facilitate decoding when used to redundantly encode the same field or via interference impede decoding when visualizing different fields, whereas *separable* visual channels (*e.g., x* and *size*) may exhibit little to any cross-channel effects. Demiralp et al. [DBH14] evaluate methods for building empirical models (*perceptual kernels*) of these interactions, which can then be incorporated into automated design systems. Szafir [Sza17] focuses specifically on color encoding, contributing models to optimize color discrimination across varying mark sizes and shapes. Here, we focus on encodings of trivariate data in part to assess potential interference effects that might undermine the greedy approach used by classic visualization recommendation algorithms.

Much of the prior work focuses on the performance of reading and comparing values encoded by individual visual objects. A more recent trend is the study of tasks involving the perception of distributed (or *ensemble*) visual information [SHGF16]. Such tasks include perceiving summary values that are not explicitly encoded, such as averages [ACG14, GCNF13] or levels of correlation [HYFC14, KH16]. In this work, we investigate both *value tasks* (reading or comparing individual values) and *summary tasks* (finding a category containing the maximum value, comparing averages of two categories).

The work with aims most similar to our own is Sarikaya & Gleicher's [SG17] review of tasks, data characteristics, and design strategies for scatterplots. In that work, they evaluate the design space by surveying the literature. Our study has an overlapping scope but instead involves a controlled experiment. Our results provide quantitative evidence corroborating some of the issues raised by Sarikaya & Gleicher, including effects due to the number of data points, categorical cardinalities, and summary tasks.

2.2. Automated Visualization Design

Mackinlay's APT [Mac86] automatically recommends visualizations using *expressiveness* and *effectiveness* criteria informed by the work of Bertin [Ber83], Cleveland & McGill [CM84], and others. Given an importance-ordered list of data fields, APT outputs a ranked set of recommended visual encodings for a single plot. APT first employs a logical search procedure to enumerate candidate visualizations and prune those considered insufficiently expressive. It then ranks visualizations by effectiveness, using a greedy heuristic that privileges assigning the most important data fields to their most effective available channel.

Other projects, such as SAGE [RKMG94], consider an expanded set of data and visual encoding types. Casner's BOZ [Cas91] additionally takes a logical task specification as input and maps it to low-level perceptual tasks involving reading or comparison of individual values. In this work we include not only value tasks, but also summary tasks that involve comparison of aggregates.

Approaches such as APT might benefit from a more global optimization approach. For example, consider a situation involving three data fields where the "most important" fields are a primary quantitative field (Q_1) and a categorical field (N), with a secondary quantitative field (Q_2) considered less important. As positional (x, y) encodings are typically considered the most effective for all data types, a greedy system like APT would prioritize encoding Q_1, Q_2, and N using the y, *size*, and x channels, respectively. However, as *color* hue can be effective for categorical data with a bounded cardinality, encoding Q_1, Q_2, and N on y, x, and *color* might be preferable in some cases, whether for assessing relationships among Q_1 and N or for supporting additional assessments involving Q_2. Here we seek to evaluate such possibilities.

More recent visualization recommender systems [MHS07, WMA*16] support data exploration by generating visualizations for selected data fields. ShowMe [MHS07] uses heuristic rules to suggest feasible visual encodings, including faceted trellis plots, grouped into plot types with manual *a priori* rankings. To support open-ended data exploration, Voyager's Compass [WMA*16] recommends visualizations using hand-tuned scoring functions based on prior effectiveness rankings, but adapted to include considerations such as label legibility and space-efficiency. While neither system uses an importance ordering of fields, both systems use similar rankings as those underlying APT and involve subjective decisions that might be informed by more careful study.

2.3. Tasks & Data Characteristics

As a single comprehensive evaluation across all possible tasks is infeasible, we turned to prior work to identify visual analysis tasks on which to focus. Munzner's typology [Mun14] covers visualization tasks conceptually by characterizing *why* and *how* visualization tasks are conducted and *what* inputs and outputs the tasks have. The taxonomy of Amar et al. [AES05] categorizes low-level visualization analytic activities such as Retrieve Value, Find Extremum, *etc*. Sarikaya & Gleicher [SG17] enumerate scatterplot-specific tasks. In this work we primarily adopt Amar et al.'s taxonomy, but incorporate concerns from other classifications.

A number of researchers have sought data-driven characterizations for aiding visualization design. Wilkinson et al.'s scagnostics quantify distributional patterns within scatterplots [WAG05]. Pandey et al. [PKF*16] collect human judgments for a large set of scatterplots and identify perceptual features such as density, orientation, spread, regularity, and grouping. Sarikaya and Gleicher [SG17] further group data characteristics into several attributes for scatterplot design: class label (cardinality), number of points, number of dimensions, spatial nature, and bivariate relationship (*e.g.,* random, linear correlation, cluster). Similarly, our characterization of data distributions includes cardinality, number of points in each category, univariate entropy, clusteredness, and linear correlation between quantitative fields.

3. Experiment Design

The primary goal of our experiment is to measure user performance (in terms of both accuracy and response time) across encoding, task, and data conditions in order to form more nuanced effectiveness rankings. We compare the performance of 12 visual encoding schemes for trivariate data across 4 different tasks (2 value tasks, 2 summary tasks) and 24 data distribution conditions (with varied point counts, cardinalities, and entropies).

3.1. Datasets

We focus on trivariate data involving 1 categorical (N) and 2 quantitative fields (Q_1, Q_2). This combination is common within multidimensional visualization and provides a natural starting point for investigating interference among visual channels.

To generate our stimuli, we use data records sampled from 2016 U.S. daily weather measurements [MDV*12]. The data include fields for *State* and *Month*, as well as the eight quantitative measures *Maximum Temperature*, *Minimum Temperature*, *Average Wind Speed*, *Wind Direction*, *Strongest Gust Speed*, *Precipitation*, *Snowfall*, and *Snow Depth*. As described below, we sample the data into subsets matching desired data distribution characteristics. While one could sample from theoretical distributions, we found the diversity of distributions within the weather data amenable to our experimental needs, while providing the additional benefit of realistic data relatable to a popular audience.

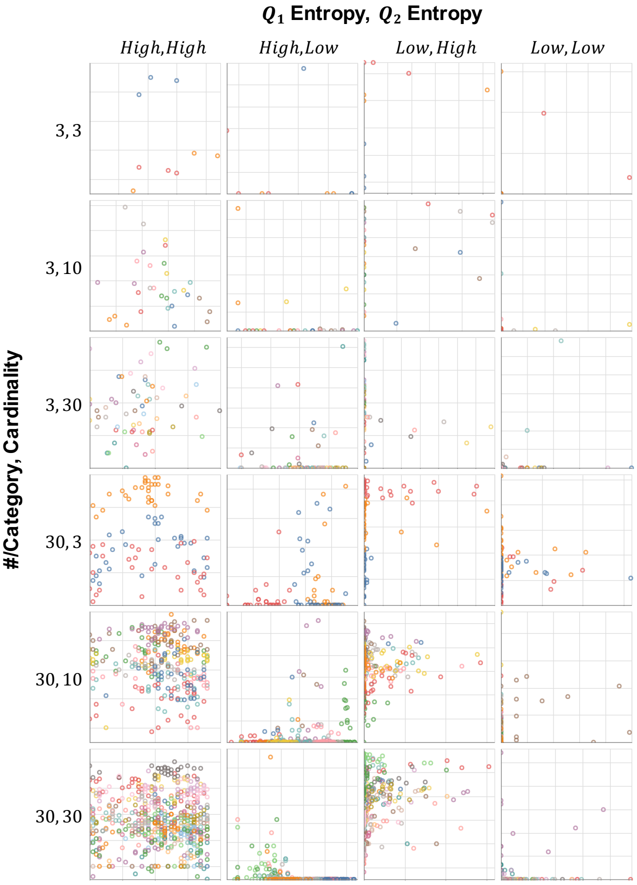


Figure 2: *Data distributions characterized according to four data attributes: number of records per category and cardinality (on rows), and normalized entropies for Q_1 and Q_2 (on columns). Each subplot encodes Q_1, Q_2, and N on x, y, and color).*

Akin to prior work [PKF*16, SG17], we characterize our data using the statistical attributes listed in Table 1: cardinality of N, number of records per category, normalized univariate entropies[†] for Q_1 and Q_2, Pearson's correlation between Q_1 and Q_2, and a measure of clusteredness computed as the sum of z-score distances for each category. We further discretize the continuous attributes. To ensure balance, we use quantiles to stratify entropy into $\{Low, Medium, High\}$ and clusteredness

[†] We calculate univariate entropy as $(-\sum_{i=1}^{20} p_i log(p_i))/log(20)$, where p_i is the fraction of data points in the i-th uniform bin.

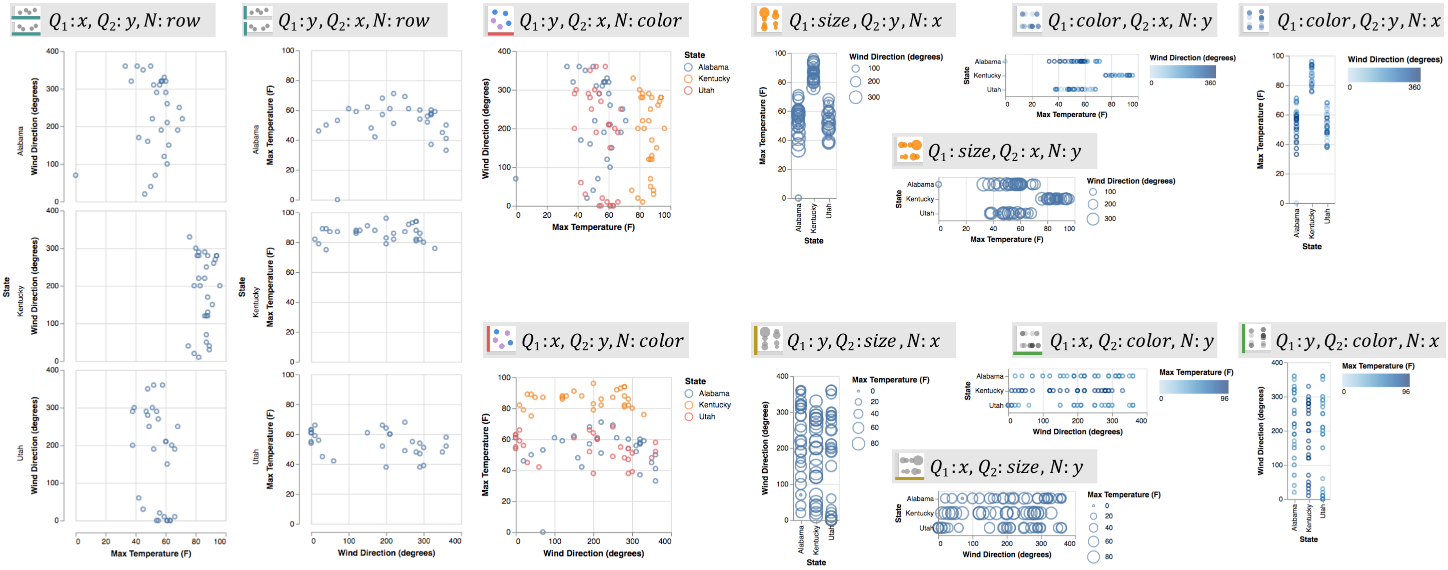


Figure 3: *For our experimental stimuli, we selected 12 visualization designs that varied the mapping between data fields and encoding channels. Each visualization design is labeled with the mapping used and includes an icon that denotes the type of visualization design.*

Attribute	All Conditions	Exp. Conditions
Cardinality	$\{3, 10, 20\}$	$\{3, 10, 20\}$
$\#/Category$	$\{3, 30\}$	$\{3, 30\}$
$Entropy_{Q_1}$	$\{Low, Medium, High\}$	$\{Low, High\}$
$Entropy_{Q_2}$	$\{Low, Medium, High\}$	$\{Low, High\}$
$Entropy_N$	$\{Low, Medium, High\}$	$High$
Correlation	$\{Weak, Mild, Strong\}$	$Weak$
Clusteredness	$\{Low, High\}$	$High(Low)$

Table 1: *Attributes used to characterize trivariate data distributions. All Conditions are stratified values of each attribute with representatives, quantiles, and specified intervals. For feasibility, we limit the Experimental Conditions to a subset of 24 conditions.*

into $\{Low, High\}$. For example, *High* entropy means the entropy value is in the top 33% of all entropy values. As absolute Pearson correlation values have a direct interpretation, we stratify those into the bins $\{Weak, Mild, Strong\}$ using the intervals $([0, 0.33), [0.33, 0.67), [0.67, 1])$.

As inclusion of all combinations of attribute strata leads to an untenably large combinatorial design, we focus on 24 experimental dataset conditions. We vary the cardinality, the entropies of the two quantitative fields ($Entropy_{Q_1}$, $Entropy_{Q_2}$), and the number of records in each category ($\#/Category$). We chose these attributes as they had the largest effect in pilot studies. For each dataset, we ensure an equal number of data points per categorical (N) value (i.e., maximal entropy). We also control for correlation and clusteredness by using the most frequent strata given the other attributes; this typically corresponds to *Weak* correlation and *High* clusteredness, except for conditions involving two *Low* quantitative entropies, in which case *Low* clusteredness is more prevalent.

Given these criteria, we first sample 3,000 trivariate datasets for each of our *Cardinality* and $\#/Category$ levels, for a total of 18,000 datasets. We use *State* as our categorical field N, and include records across weather stations and months. For Q_1 and Q_2, we randomly choose two of the eight quantitative measures listed above. Then, we then calculate entropies, correlation, and clusteredness to further subdivide the datasets. Finally, from each of our 24 experimental conditions we randomly sample 8 datasets to serve as replications. Figure 2 shows sampled datasets for each condition.

3.2. Visual Encodings

For visual encoding we use the Vega-Lite grammar [SMWH17]. Similar to popular commercial tools such as Tableau [MHS07], Vega-Lite plots are specified using a geometric mark type (*e.g., bar*, *point*, or *line*) and a set of encodings that map from data fields to visual enchoding channels such as *x*, *y*, *color*, *shape*, and *size*. In addition, *row* and *column* channels perform spatial subdivision (or faceting) to create trellis plots.

To limit the combinatorics of the experiment design, we apply a number of constraints. We use only *point* marks, as they accommodate all encoding channels considered. We also do not use all possible mappings between our three data fields (Q_1, Q_2, N) and the visual encoding channels. We omit the *column* channel (which partitions the data into horizontally ordered trellis plots), as it can require horizontal scrolling less common to web based interfaces. Faceting is still supported via the *row* channel. We exclude the *shape* channel as it has limited expressiveness (largely applicable to categorical data only) and prior work [Now97] has found it to be generally less effective than *color* for the same tasks. We require that both the *x* and *y* encoding channels be used, as using a single positional channel with two non-positional channels generally results in less expressive plots. Finally, we do not consider redundant encodings in which a field is mapped to more than one channel.

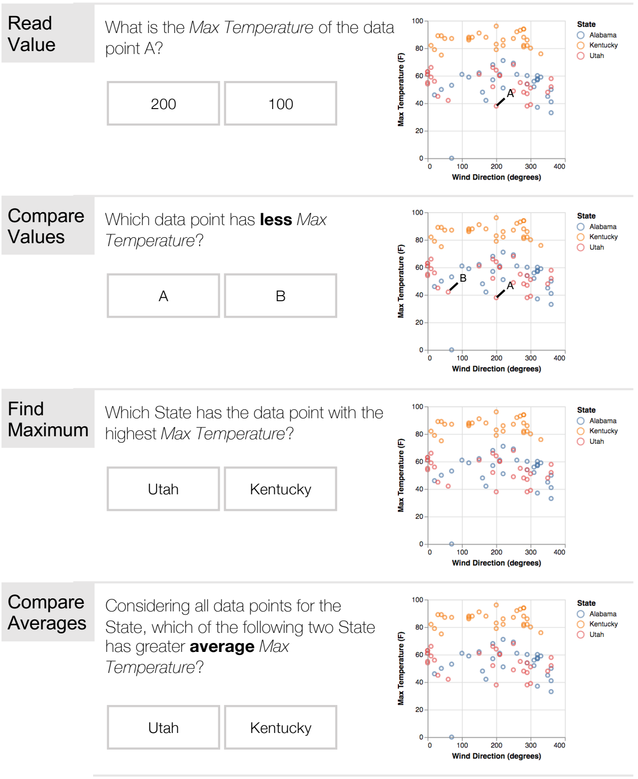
As shown in Figure 3, these considerations result in a total of 12 unique encoding specifications over the *x*, *y*, *color*, *size*, and *row* channels. For the *color* channel, we use the *Tableau-10* color scheme for categorical fields with cardinality 3 or 10 and *Tableau-20* for categorical fields with cardinality 20. For positional encodings, we use linear scales that include both the data range and zero for all quantitative fields except for temperatures, as degrees Fahrenheit is not a ratio measure. Throughout the paper we use both shorthand specifications (*e.g.,* Q_1:x, Q_2:y, N:*color*) and corresponding icons (●) to indicate visual encoding conditions. All of the icons are described with their corresponding encoding specifications in Figure 3.

3.3. Tasks

We begin with Amar et al.'s taxonomy of low-level analytic activities. From their Retrieve Value, Compute Derived Value, and Find Extremum tasks, we derive 3 concrete tasks: *Read Value*, *Find Maximum*, and *Compare Averages*. In addition, we include a *Compare Values* task, which is common to other task classifications [Mun14, SG17] and related to Amar et al.'s Filter task, which similarly requires logical comparison. Among other tasks identified by Amar et al., Determine Range is similar to Find Extremum and Sort is performed via interaction. We omit the remaining tasks (Data Distribution, Find Anomalies, Cluster, Correlate), which have more subjective definitions.

Our four tasks (illustrated in Figure 4) group into two categories: *value tasks* that require reading or comparing individual values and *summary tasks* that require identification or comparison of aggregate properties. We formulate the tasks as binary (two-alternative forced choice) questions pertaining to values of Q_1 and N. Based on pilot results we control question difficulty, as described below.

- *Read Value*: We visually annotate a randomly sampled data point with an arrow and label "A", and ask *"What is the Q_1 of the data point A?"*. Two values are provided as response options. The incorrect value is contained in the range of Q_1 (I_{Q_1}) and differs by exactly half the length of I_{Q_1} ($|I_{Q_1}|/2$).
- *Compare Values*: We visually annotate two sampled data points with arrows and labels "A" & "B". We ask "Which data point has more/less Q_1?", where "more" or "less" is decided randomly. When sampling the two points, half of the time A and B are sampled from the same category; for the other half, they are sampled from different categories. (Each participant answers four questions of each of these variants.) We also attempt to make the sampled Q_1 values differ by a half of $|I_{Q_1}|$, such that the average of the differences is 0.50 ($\sigma = 0.11$) of $|I_{Q_1}|$.
- *Find Maximum*: We ask "Which State has the data point with the highest Q_1?" We provide two categories as options: the one containing the highest Q_1 value (M) and the one having the highest Q_1 value closest to $M - |I_{Q_1}|/2$. The actual average difference is 0.51 ($\sigma = 0.21$) of $|I_{Q_1}|$. As the *Precipitation*, *Snowfall*, and *Snowdepth* fields are skewed with many zero values, we did not ask about minimum values.
- *Compare Averages*: We ask "Considering all data points for the State, which of the following two States has greater **average** Q_1?" We pick as options the two categories whose Q_1 average difference is closest to $0.3|I_{Q_1}|$. The resulting average difference is 0.25 ($\sigma = 0.11$) of $|I_{Q_1}|$.



Read Value

What is the *Max Temperature* of the data point A?

200 | 100

Compare Values

Which data point has **less** *Max Temperature*?

A | B

Find Maximum

Which State has the data point with the highest *Max Temperature*?

Utah | Kentucky

Compare Averages

Considering all data points for the State, which of the following two State has greater **average** *Max Temperature*?

Utah | Kentucky

Figure 4: *We compose four tasks as binary questions. The first two types ask subjects to read or compate individual values marked with arrow annotations. The last two aggregate tasks ask subjects to find maximum values and compare averages.*

3.4. Procedure

We employed a mixed design, using a within-subjects treatment for visual encodings and between-subjects treatments for tasks and data characteristics. We assigned each participant to one of the 24 data distributions and one of the 4 tasks, balanced across subjects. The subject then answered 8 questions for each of the 12 visual encodings. A different dataset was used for each question, though all came from the same distribution condition.

For each encoding, we first showed an example visualization using that encoding, encouraged the subject to examine it, and had them click a "Ready" button to begin. Next the subject was given an engagement check question that tests what information is presented in the channel encoding Q_1. The subject then completed 8 questions in the assigned task condition.

We randomized the presentation order of both the visual encodings and the questions. After completing this process for all 12 visual encodings (a total of 96 questions and 12 engagement checks), participants completed a short demographic survey.

3.5. Participants

We recruited a total of 1,943 participants on Amazon's Mechanical Turk. We limited participation to subjects located in the U.S. with a HIT approval rate $\geq$ 95%. People with self-reported color-

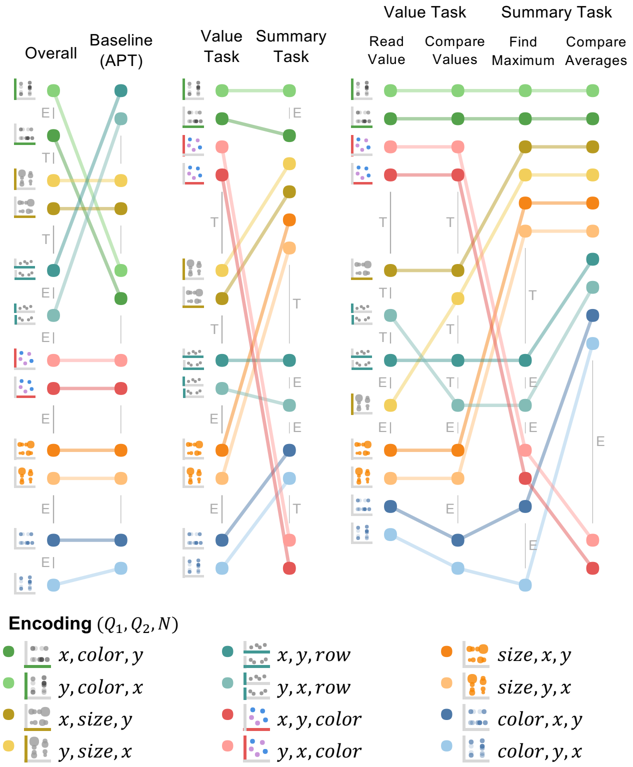


Figure 5: *Visual encoding effectiveness rankings. Higher is better; groupings indicate encodings of identical rank. The left side compares value tasks and summary tasks with a baseline ranking by APT [Mac86]. The right side compares performance for all tasks, averaged across data distribution conditions. Gray lines indicate sig. differences ($p < 0.05$) between groups in terms of error (E) or completion time (T).*

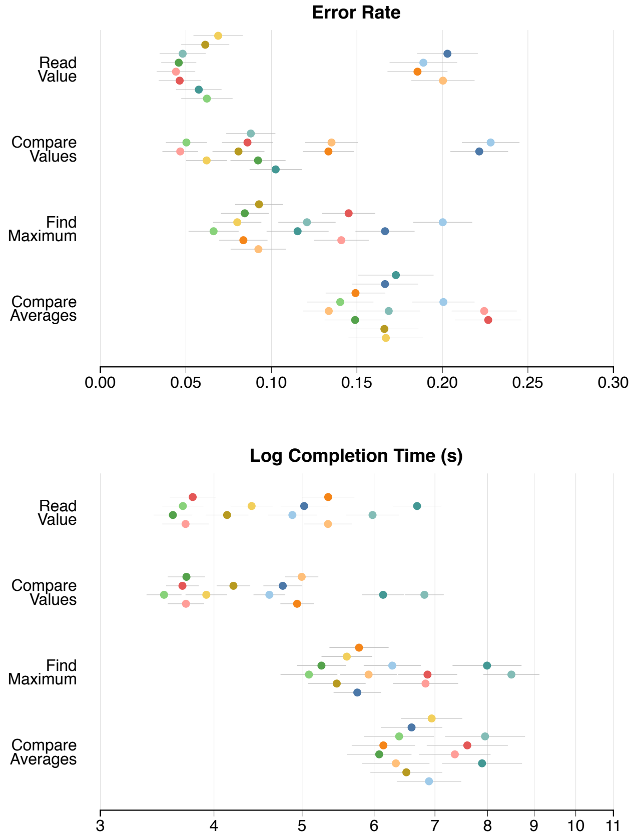


Figure 6: *Bootstrapped means and 95% confidence intervals for error rates and log completion times across tasks. For example, ▦ and ▦ exhibit reasonable error rates but longer completion times.*

blindness were screened out before they participated in the study. Each subject received $0.75 USD in compensation.

We filtered out 7 participants with error rates greater than 60% (higher than chance), and another 16 participants whose average task completion times were less than a half second (indicating they did not take sufficient time to examine the plot and answer a question). Upon removing a subject's data, we recruited a new subject in the same condition to ensure balance.

We ultimately analyzed responses from 1,920 subjects (60.21% female, 39.1% male, 0.63% other), evenly distributed with 20 subjects for each combination of task and data distribution ($4 \times 24 \times 20 = 1,920$). Among participants, 18% reported having a graduate degree, 51% a bachelors or associate degree, 24% some college coursework, and 7% a high school diploma.

4. Experiment Results

We analyze subject performance in both absolute and relative (ranked) terms. We assess error rate and (log transformed) response time across task and data characteristics. We focus first on main effects due to task, and then examine effects due to data characteristics within task groups. We do not examine interaction effects among different data characteristics (*e.g.,* between cardinality and entropies), leaving further analysis to future work.

We employ mixed effects models for statistical testing, using a logistic model for error and a linear model for log response time. We model *visual encoding*, *question order*, and *encoding order* as fixed effects and model *participants* as a random effect, with both a random intercept term and a random slope term for *visual encoding*. When comparing visual encoding performance across task and/or data conditions, we additionally include random intercept terms for *task* and *data distribution*. We perform post-hoc pairwise comparisons using Tukey's HSD test. To visualize effect sizes, we compute bootstrapped 95% confidence intervals of the mean by sampling participants with replacement.

Our primary results take the form of visual encoding effectiveness rankings across experiment conditions. We first rank encodings by error rate. We treat encodings as equivalent (same rank) if we do not observe a significant difference ($p < 0.05$) according to post-hoc pairwise comparison tests. We further separate (differentially rank) encodings with similar error rates if they exhibit a significant response time difference. Our data, analysis scripts, and full statistical results are included as supplemental material.

For comparison purposes, we use the expected ranks of visual encodings from the APT algorithm as a baseline for comparison.

We assume that the effectiveness rankings across visual channels are $x = y > size > color$ for quantitative fields and $x = y > row > color$ for the categorical field, as *row* uses a position encoding. We also assume an importance ordering of the fields as $Q_1 > N > Q_2$: all tasks involve Q_1 values, while the *Find Maximum* and *Compare Averages* tasks also require decoding of N values (categories).

Overall rankings by task category (*value* or *summary*) compared to APT are shown in Figures 5 & 7. Each circle represents a visual encoding, with similar hues indicating similar encodings with transposed x and y channels. Vertical position encodes rank (higher → lower rank) while spatial grouping indicates statistical ties with the same rank. The gray vertical lines with labels ('E', 'T') indicate significant differences ($p < 0.05$) between the two adjacent groups in terms of error or log completion time, respectively.

Figures 6 & 8 show bootstrapped means and 95% confidence intervals for error and log completion time. Due to differences between mixed effects models and bootstrapping, groupings in the CI plots may not precisely match those of the ranking diagrams.

4.1. Overall Effectiveness Ranking Relative to APT

Averaged across all tasks, our overall effectiveness ranking aligns well with APT's ranking (Figure 5, left side). In agreement with prior work, position encodings for Q_1 fare the best in rankings, while *size* and *color* encodings of the primary quantities lead to higher error rates. The close alignment is perhaps surprising given APT's greedy approach and its use of univariate effectiveness criteria based only on value comparison tasks, but bodes well for existing recommendation systems. However, we also observe deviations from APT that we discuss further below: in our results, faceted charts perform more slowly than others, and encodings with Q_2:*color* outperform Q_2:*size* for tasks focused on decoding Q_1.

4.1.1. Faceted charts (*row*) require more time.

A major discrepancy between the APT baseline and our experimental results concerns the ranks of faceted charts (▦, ▦). Though these charts are deemed "best" by APT's rules due their use of position encodings for all fields, our experimental results lead to a notably lower ranking. Despite reasonable accuracy, faceted views exhibit significantly longer completion times (seen in Figure 6). The increased time likely stems from the need to compare across multiple charts and scroll the display to view offscreen elements for all types of tasks. The scrolling occurred for reasonable monitor sizes and resolutions when the cardinality was 10 or 30.

4.1.2. Secondary quantities may interfere with decoding.

The top-performing encodings in our results use position encodings for Q_1 and N, while mapping Q_2 to *color*. APT, on the other hand, prefers mapping Q_2 to *size*, which is known to be a more effective encoding for quantities in univariate conditions. This discrepancy stems from our use of tasks focused on Q_1. Here, the use of *size* serves as a distractor that complicates decoding of Q_1, an example of interference between encoding channels. In the next section, we also see that this performance difference arises only in *value* tasks. These results indicate trade-offs between optimizing for the "most important" fields and supporting a broader range of comparisons (e.g., decoding Q_2 values). Interestingly, APT actually achieves the latter despite its greedy, importance-prioritized approach.

4.2. Effects of Tasks

The observed effectiveness ranks across different tasks show significant variation. For example, position channels (x, y) tend to convey Q_1 well, but for *summary* tasks a *size* encoding also performs well. Encodings that map Q_2 to the *size* channel perform more slowly in *value* tasks (interfering with positional decoding of Q_1), yet rank among the best encodings for *summary* tasks.

4.2.1. Position (x, y) conveys the primary quantities well.

As expected, visual encodings using position channels (x, y) for Q_1 (▦, ▦, ▦, ▦) are ranked higher on average. Notable exceptions involve the faceted charts (▦, ▦) discussed above and colored scatterplots (▦, ▦), which performed poorly in *summary tasks* (see also the discussion of cardinality below).

4.2.2. *Size* encoding performs well for *summary tasks*.

For *summary tasks*, dot plots with Q_1:*size* (▦, ▦) outperformed other encodings using position channels (▦, ▦, ▦, ▦), even though *size* is known to be less effective than position channels for value comparison tasks. This result suggests that *size* supports effective ensemble coding [SHGF16].

4.2.3. *Color* encoding performs well for *Compare Averages*.

Encodings using the *color* channel for Q_1 (▦, ▦) were placed in the top ranking group for *Compare Averages*. However, we do not observe a significant benefit for Q_1:*color* (▦, ▦) over position encodings of Q_1 (▦, ▦), as might be predicted from prior work examining perception of average values [ACG14].

4.2.4. *Size* & *color* exhibit asymmetric effects for Q_1 vs. Q_2.

For *value tasks*, charts with Q_2:*size* (▦, ▦) required more time than charts with Q_2:*color* (▦, ▦) (at least 1.13 times, $\forall p < 0.01$). This result suggests that marks with varied sizes may interfere with decoding of Q_1 values encoded on position channels (x,y). On the other hand, charts with Q_1:*size* (▦, ▦) performed better than charts using Q_1:*color* (▦, ▦) for all tasks (at least 6.3% lower error rate, $\forall p < 0.001$). These two observations reveal an asymmetry among *size* and *color* encodings for quantitative fields. For Q_1, *size* results in more accurate decoding of Q_1 than *color*. For Q_2, *size* more strongly interferes with decoding of Q_1 than *color*.

4.2.5. Faceted charts exhibit asymmetric performance.

We also observe asymmetric performance between the use of x and y channels within faceted charts (▦, ▦). ▦ exhibits faster completion times than ▦ for *Read Value* (0.9 times, $p < 0.01$), likely because ▦ requires participants to scroll down to the bottom of the display to see the axis for Q_1 values. ▦ performed faster than ▦ in *Compare Values* (0.92 times faster, $p < 0.01$) and was more accurate for *Find Maximum* (2.5% better, $p < 0.01$). Q_1 values on the y-axis appear harder to compare across vertically ordered subplots: unlike the x-axis, the y-axis lookups are not aligned across charts.

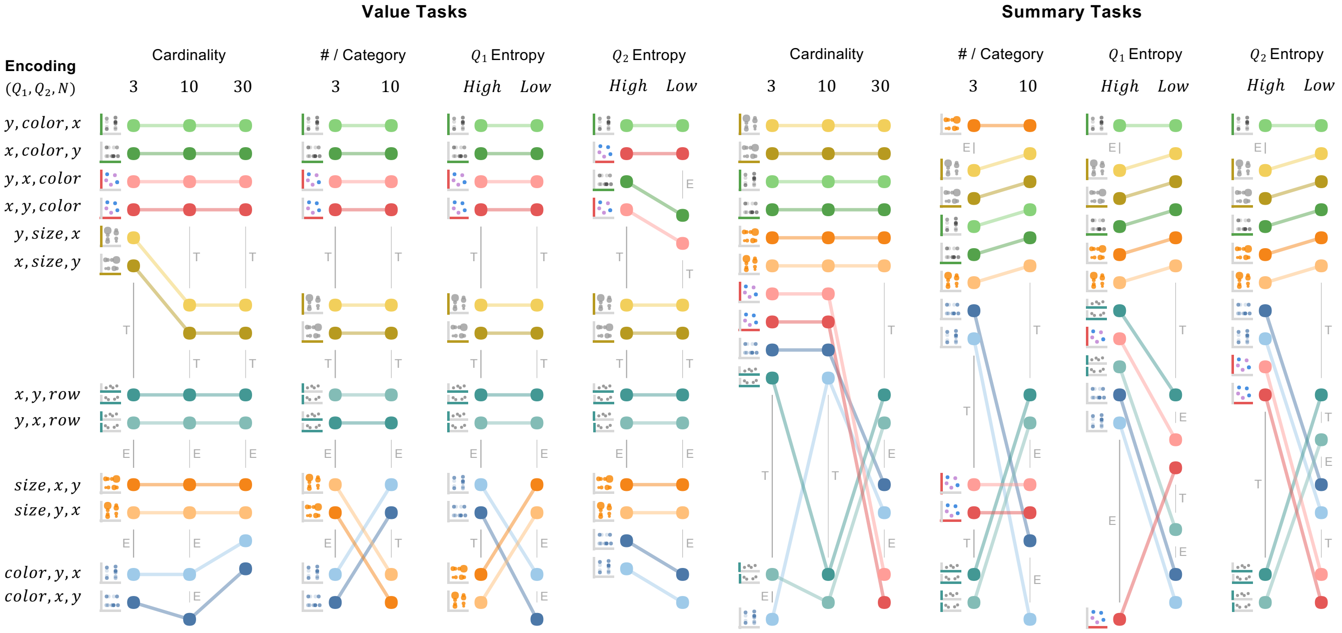


Figure 7: *Encoding effectiveness rankings by data distribution conditions, subdivided by task groups. Data characteristics exhibit different effects for value tasks and summary tasks. For example, increasing cardinality degrades* ⁘ *and* ⁘ *for summary tasks, but not value tasks.*

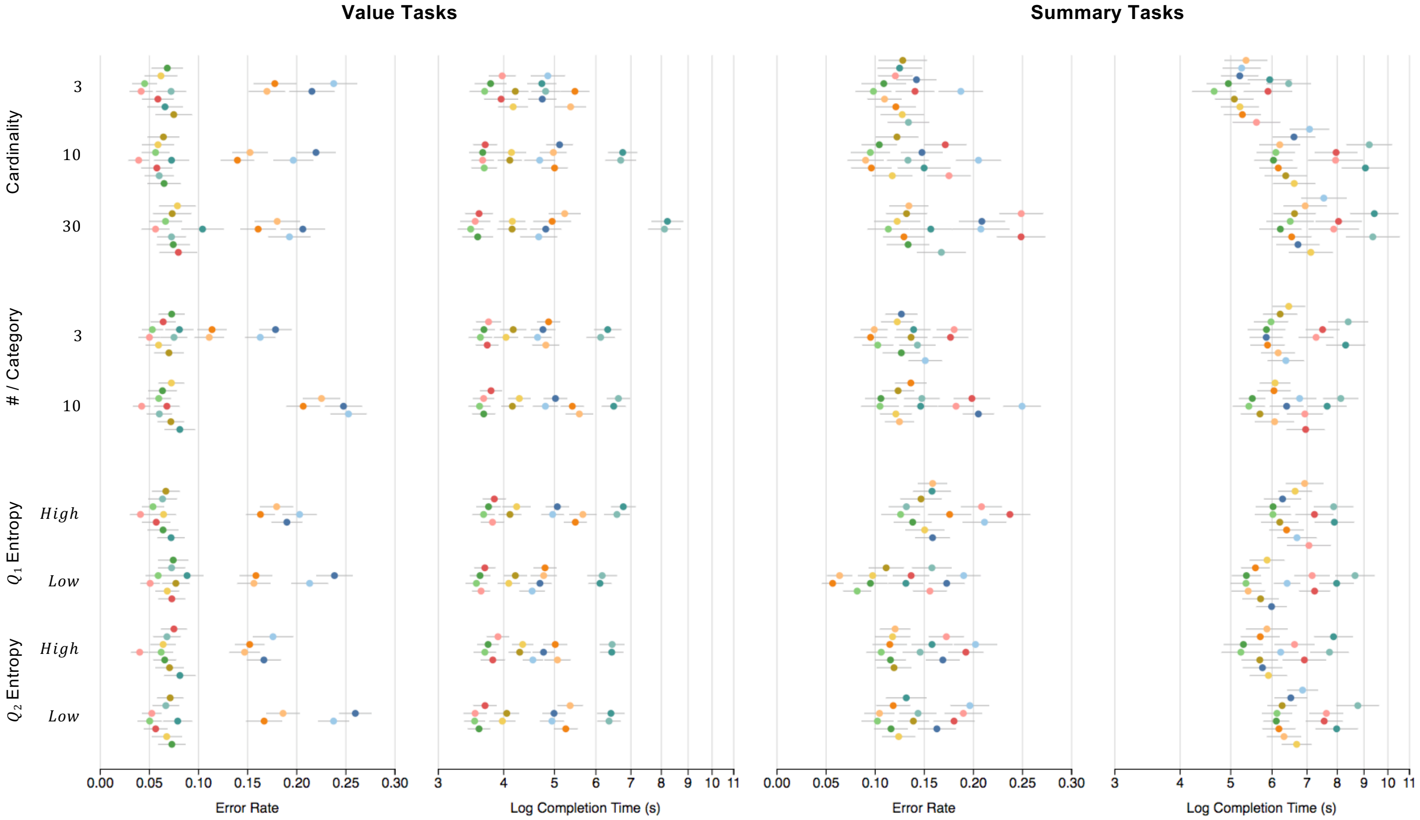


Figure 8: *Bootstrapped means and 95% confidence intervals for error rates and log completion times across data characteristics, again showing differing effects by task group. For example, as #/Category increases, the performance of* ⁘ *and* ⁘ *degrades for value tasks, but not for summary tasks.*

4.3. Effects of Data Characteristics

To analyze the effects of data characteristics, we separately assess performance across *value* and *summary* tasks. For *value* tasks, the effectiveness ranks are largely stable in terms of error rates across data characteristics. The exceptions concern changes among poorly-ranked encodings (Q_1:*color* and Q_1:*size*) in response to records per category and entropy. For *summary* tasks, the top-ranked visual encodings persist across data characteristics. However, colored scatterplots, faceted charts, and charts with Q_1:*color* exhibit significant variation. These findings are illustrated in Figure 7 and Figure 8.

4.3.1. Increased congestion / occlusion degrades performance.

For *summary tasks*, charts with Q_1:*color* (⁙, ⁙) and *N*:*color* (⁙, ⁙) exhibit higher error rates as cardinality increases, the number of records per category increases, or Q_2 entropy decreases. For *value tasks*, we also observe that ⁙, ⁙ exhibit more errors as Q_2 entropy decreases. These degradations correspond to increased visual congestion. For example, with ⁙ and ⁙, data points of different categories increasingly occlude each other as the point count rises, styming perception. When entropy decreases, data points cluster more tightly, again raising the odds of occlusion.

4.3.2. Trade-off between response time and error rate.

For *summary tasks*, the ranks of faceted charts (⁙, ⁙) can oscillate due to time-error tradeoffs. As cardinality increases, so does the number of subplots, requiring more comparisons and thus longer response times (though with no decrease in accuracy). However, in the high cardinality (20) condition, overplotting causes higher error rates for other encodings (⁙, ⁙, ⁙, ⁙), at which point the faceted charts reclaim an intermediate ranking. A similar dynamic occurs with changes to the number of records per category.

4.4. Unexplained Observations

We also observed significant differences for which we do not yet have plausible causal explanations. These include asymmetric results among encoding pairs with transposed *x* and *y* channels, and the effects of Q_1 entropy on encodings with Q_1:*color* and Q_1:*size*. More work is needed to explain or discount these observations.

4.4.1. Asymmetric performances between *x* and *y*

Across task and data characteristics, there are instances of asymmetric performance between visual encoding pairs with transposed *x* and *y* channels. Moreover, some of these asymmetries are also inconsistent, with the preferred encoding varying across conditions. For example, ⁙ outperforms ⁙ for *value tasks* when the cardinality is 10 (3.7% point, $p = 0.0435$), but the opposite holds for *summary tasks* when the cardinality is 3 (4.1% point, $p < 0.01$).

4.4.2. Primary quantity entropy interacts with *color* and *size*.

Encodings with Q_1:*color* and Q_1:*size* (⁙, ⁙, ⁙, ⁙) exhibit rank changes as Q_1 entropy decreases. Across tasks, the ranks of ⁙ and ⁙ degrade when Q_1 entropy is low, yet the ranks of ⁙, ⁙ relatively improve. This result suggests potential effects of entropy on the effectiveness of retinal channels such as *color* and *size*.

5. Application to Visualization Systems

Our experiment results suggest approaches for tailoring automatic visualization design systems by adding consideration of interactions between visual channels, tasks, and data distributions.

5.1. Incorporating Interactions Between Channels

Classic greedy approaches to automated design, such as APT [Mac86], prioritize visual encoding channels based on effectiveness rankings for univariate decoding tasks, overlooking potential interference effects among encoding channels. Our experimental results complicate this picture, finding instances where encodings in a "distractor" channel (*e.g.,* Q_2:*size*) degrade decoding performance of other channels. One challenge for visualization recommender systems is choosing how to weight cross-channel effects. Should one prioritize reading of "more important" fields, or attempt to balance performance across fields, even if this results in sub-optimal performance for individual fields? In practice, the answer may greatly depend on context.

In any case, greedy approaches are ill-suited for responding to these concerns. A rigid effectiveness order of channels, such as (*position* $>$ *size* $>$ *color*) or (*position* $>$ *color* $>$ *size*) is insufficiently expressive. The former ordering produces ⁙,⁙ > ⁙,⁙ and the latter produces ⁙,⁙ > ⁙,⁙, both of which are suboptimal for our studied *value tasks*. Another example of interaction arises across positional channels (*x*, *y*) within faceted charts (*N*:*row*), leading to asymmetric preferences for *x* and *y* across *Read Value*, *Compare Value*, and *Find Maximum* tasks. If we assume that an *N*:*column* mapping will exhibit similar (but transposed) results to *N*:*row*, any rigid order between *x* and *y* (*e.g.,* $x = y$, $x > y$, or $y < x$) can not address this dynamic preference (with *row*, $x > y$, but with *column*, $x < y$ for *Compare Value*).

In this light of these observations, recent systems [MHS07, WMA*16] that employ a set of heuristics or hand-tuned scores to evaluate *multivariate* encodings provide a more promising route. These approaches can handle cross-channel interactions with additional scoring terms or decision rules. For example, one might introduce a term which, if Q_1 is encoded on a spatial channel, promotes *color* channel for Q_2 instead of *size* for *value tasks*. Similarly, a system could recommend better faceted charts by introducing terms promoting either *x* or *y* based on the faceting channel (*row* or *column*).

5.2. Adapting Recommendations based on Task

Though APT's rankings were largely derived in the context of individual value comparison tasks [Mac86,CM84], they align well with our empirical rankings averaged across task types. This bodes well for the various visualization systems in use today that recommend charts using a similar logic. However, if a system has additional knowledge about the tasks users are performing, it might improve recommendations by using task-specific effectiveness rankings. For example, for *summary tasks* a system might include recommendations for charts that encode a primary quantitative field using *size* (⁙, ⁙), but not for *value tasks*.

By deriving effectiveness ranks for multiple tasks, visualization systems could recommend visualizations that balance multiple user needs. Consider the case of a user conducting an open-ended data exploration with limited prior knowledge about the specific dataset. If two quantitative fields are both subject to *value tasks*, colored scatterplots (,) would be preferable to dot plots (, , ,) as the scatterplots convey both quantitative fields well, but the dot plots prioritize a single quantitative field. Following a similar logic for *summary tasks*, sized dot plots (, , ,), might be preferable to colored dot plots (, , ,).

5.3. Adapting Recommendations based on Data Distribution

Our experiment results also suggest ways to dynamically adapt effectiveness rankings according to data characteristics. Most of the effects we observed in response to changing data characteristics (cardinality, #/category, entropy) correspond to changes in visual congestion and/or overplotting. It seems a particularly valuable addition to automatic visualization design systems would be to assess the level of overplotting across encoding choices and adjust the ranking accordingly. As one simple rule of thumb, we found that high cardinality rapidly degrades otherwise effective plots for *summary tasks* (*i.e.*, colored scatterplots (,)).

6. Conclusion & Future Work

We conducted an experiment measuring subject performance (time, error) with 12 visual encodings of trivariate data involving 1 categorical and 2 quantitative fields. We compared performance across 4 task types (*Read Value*, *Compare Values*, *Find Maximum*, and *Compare Averages*) and 24 data distributions characterized by univariate entropies of the two quantitative fields, cardinalities, and numbers of records per category. Our results extend existing models of encoding effectiveness and suggest improved approaches for automated design by considering interactions between channels, elementary task types, and data characteristics.

Still, a great deal of future work remains. First, the bounds of our experimental scope should be extended and refined. One might investigate more encodings with different marks and layouts such as bar, line, or pie charts as well as geographic maps. Or, one might increase the complexity of visual encodings by involving more data fields to investigate interactions among more channels, including the use of redundant encodings. Investigating additional task types we excluded from our scope, such as Cluster, Find Anomalies, or tasks associating multiple variables, might prove promising. Recent prior work has examined Correlation tasks [HYFC14, KH16], and the results of such studies might be integrated with our own to provide further task coverage. Regarding data-driven considerations, we used entropy as one means of characterizing a data distribution, but our treatment falls well short of all distributions of potential interest. Investigating common families of probability distributions (*e.g.,* normal, log-normal, Poisson, and mixtures thereof), might also be an interesting avenue for further work.

We assessed performance using binary response questions with fixed difficulty. Through pilot studies, we calibrated the difficulty to ensure users could, on average, complete the tasks successfully while still revealing meaningful accuracy differences. Still, investigating varied task difficulties would be helpful to further understand the effectiveness of visual encodings. As each visual channel can have different just noticeable differences, some high performers might encounter ceiling effects under our fixed difficulty. While we focused on binary responses, an alternative approach might use continuous responses such as magnitude estimates.

In our study, we examined the effectiveness of visual encodings using three data fields (N, Q_1, Q_2). Future work might compare visual encodings with different numbers and types of data fields. For example, do two-field dot plots ( ) perform better (or worse) than three-field sized dot plots ( ) for comparing Q_1 averages? Moreover, one might compare different design strategies: encoding all interesting fields in a single chart vs. dividing into multiple charts. For example, to compare Q_1 and Q_2 averages, do two basic dot plots (,) perform better than a single sized dot plot ( ) or colored scatter plot ( )?

In terms of channel interactions and data characteristics, our study provides results that can be directly incorporated into visualization recommender systems. However, to apply our results more fully, systems might also form a model of the tasks users need to perform. In accordance with task frameworks [LTM17], future work might consider means to elicit or infer such models, for example by analyzing user actions or gaze patterns.

7. Acknowledgements

This work was supported by the Paul G. Allen Family Foundation and a Moore Foundation Data-Driven Discovery Investigator Award.

References

[ACG14] Albers D., Correll M., Gleicher M.: Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 551–560. 2, 7

[AES05] Amar R., Eagan J., Stasko J.: Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (2005), IEEE, pp. 111–117. 3

[Ber83] Bertin J.: Semiology of graphics: diagrams, networks, maps. 1, 2

[Cas91] Casner S. M.: Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics 10*, 2 (1991), 111–151. 1, 2

[CM84] Cleveland W. S., McGill R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association 79* (1984), 531–554. 1, 2, 9

[DBH14] Demiralp Ç., Bernstein M. S., Heer J.: Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1933–1942. 2

[GCNF13] Gleicher M., Correll M., Nothelfer C., Franconeri S.: Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2316–2325. 2

[GF70] Garner W. R., Felfoldy G. L.: Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology 1*, 3 (1970), 225–241. 2

[HB10] Heer J., Bostock M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 203–212. 2

[HYFC14] Harrison L., Yang F., Franconeri S., Chang R.: Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1943–1952. 2, 10

[KH16] Kay M., Heer J.: Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 469–478. 2, 10

[LTM17] Lam H., Tory M., Munzner T.: Bridging from goals to tasks with design study analysis reports. *IEEE Transactions on Visualization and Computer Graphics* (2017). 10

[Mac86] Mackinlay J.: Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics 5*, 2 (1986), 110–141. 1, 2, 6, 9

[MDV*12] Menne M. J., Durre I., Vose R. S., Gleason B. E., Houston T. G.: An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology 29*, 7 (2012), 897–910. 3

[MHS07] Mackinlay J., Hanrahan P., Stolte C.: Showme: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1137–1144. 1, 3, 4, 9

[Mun14] Munzner T.: *Visualization Analysis and Design*. CRC press, 2014. 3, 5

[Now97] Nowell L. T.: *Graphical encoding for information visualization: using icon color, shape, and size to convey nominal and quantitative data*. PhD thesis, Virginia Tech, 1997. 4

[PKF*16] Pandey A. V., Krause J., Felix C., Boy J., Bertini E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 3659–3669. 3

[RKMG94] Roth S. F., Kolojejchick J., Mattis J., Goldstein J.: Interactive graphic design using automatic presentation knowledge. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1994), ACM, pp. 112–117. 2

[SG17] Sarikaya A., Gleicher M.: Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics* (2017). 2, 3, 5

[SHGF16] Szafir D. A., Haroz S., Gleicher M., Franconeri S.: Four types of ensemble coding in data visualizations. *Journal of vision 16*, 5 (2016), 11–11. 2, 7

[SMWH17] Satyanarayan A., Moritz D., Wongsuphasawat K., Heer J.: Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 341–350. 4

[Sza17] Szafir D. A.: Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics* (2017). 2

[WAG05] Wilkinson L., Anand A., Grossman R. L.: Graph-theoretic scagnostics. 157–164. 3

[War12] Ware C.: *Information Visualization: Perception for Design*. Morgan Kaufmann, 2012. 2

[WMA*16] Wongsuphasawat K., Moritz D., Anand A., Mackinlay J., Howe B., Heer J.: Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 649–658. 1, 2, 3, 9