Are We There Yet? A Systematic Review of Visual Perception Knowledge for Visualization Recommendation Systems

Zehua Zeng and Leilani Battle

Abstract—Visualization recommendation systems must select appropriate visual encodings, yet few findings from visual perception are typically applied within these systems. Knowledge bases provide one way to inscribe perception guidelines, but how do we systematically translate the perception literature into a structured format? We present a literature review across 59 papers that study how to rank effective visualizations based on user performance in various visual analysis tasks. We contribute a comprehensive schema to collate existing theoretical and experimental knowledge and summarize study outcomes at three levels: between encodings, within chart types, and between chart types. We demonstrate how the resulting survey dataset can be utilized to inform automated encoding decisions with three representative visualization recommendation systems. Based on our findings, we highlight new challenges and opportunities for the community in collating visualization design knowledge for a range of visualization recommendation scenarios.

Index Terms—Literature Review, Human Perception, Visualization Design

1 INTRODUCTION

Visualization design can be a complex process [16, 17]; multiple competing factors must be weighed [59, 63], such as reconciling the size and type of data being visualized [48, 60, 61], the user’s analysis task [45, 77], and even individual user characteristics [112]. Visualization recommendation systems [26, 59, 61, 97, 105, 106] have been developed to reduce user effort in visual analysis and exploration tasks by partially or fully automating the visualization design process. These systems can be divided into two types: data query or visual encoding recommendation systems [104, 108]. Data query recommendation systems suggest new data queries to explore (e.g., [26, 97]), while visual encoding ones suggest how to visualize a given data query (e.g., [60, 61]). Some systems do both (e.g., [105, 106]). However, regardless of type, all of these systems must still select appropriate visual encodings in order to display their recommendations, making visual perception a critical factor for any visualization recommendation system.

To select effective encoding channels, many visualization recommendation systems leverage existing theoretical [60, 61, 105, 106] and/or empirical [48, 64, 77] guidelines on human visual perception. However, the scope of guidelines utilized is limited. For example, we find that these systems generally utilize guidelines from at most three visual perception studies. Furthermore, we observe a steady stream of changes and contradictions in encoding guidelines as the community’s knowledge in visual perception continues to evolve [50]. Thus, if a visualization recommendation system only uses one or even a few papers to guide its encoding strategy, it runs the risk of making ineffective decisions, which could lead users to misinterpret the data.

Recent work attempts to but falls short of closing the gap between visualization design guidelines and visualization recommendation systems. For example, existing surveys either summarize existing visual perception papers [101] or summarize the behavior of existing visualization recommendation systems [84, 110] and ignore the connection between the two. Recent work proposes a new technique to model theoretical and empirical visualization design knowledge as constraints and recommends visualizations based on these constraints [64]. However, users still need to translate existing visual perception guidelines into corresponding datasets and/or constraints to inform recommendation decisions. Alternative frameworks suffer from similar limitations (e.g., [86, 104]). To the best of our knowledge, no current work integrates visual perception papers as a whole to contribute holistic guidelines for generating optimal encodings within visualization recommendation systems.

In this paper, we present a literature review of 59 papers that compare visualization designs in terms of visual perception and human performance under different analysis tasks. We systematically document the visualization designs studied in each paper and other factors influencing how designs are compared, contributing a dataset that can be ingested directly by recommendation systems. Then, we summarize study outcomes at three levels—between encodings, within chart types, and between chart types—to synthesize concrete recommendation rules for generating effective visualization designs for specific data characteristics and analysis tasks. We demonstrate how to use our results to improve existing visualization recommendation systems with three representative systems: Foresight [26], Voyager [105, 106] and Draco [64]. Finally, we discuss significant challenges towards building a thorough knowledge base for informing visualization recommendation systems, such as contradictory results and missing visualization design pairings in the literature. Without “enough” knowledge, the visualization community may fall short in generating recommendations that provide measurable improvements in user performance [108].

All of our data are available online: https://osf.io/twj3s/

2 RELATED WORK

In this section, we discuss existing work in visualization recommendation systems and also the design principles behind those systems.

2.1 Visualization Recommendation Systems

We provide a summary of visualization recommendation systems here and refer to existing surveys for more details [108, 110]. Existing visualization recommendation systems can be divided into two main categories according to their strategies to rank visualization designs: rule-based or machine learning-based [41, 108]. Rule-based systems utilize either existing theoretical principles in visual perception (e.g., [105, 106]) or propose new metrics to rank visualization designs (e.g., [26, 47, 97]). For example, Wongsuphasawat et al. [105, 106] use Mackinlay’s principles [60] to make recommendations, prioritizing recommendations based on the breadth of data covered within the visualizations. Vartak et al. [97] use an “interestingness” metric based on deviation in the data to identify visualizations of potential interest. Both Key et al. [47] and Demiralp et al. [26] apply statistical features of the dataset into their systems for guiding exploratory visual analysis.

Machine learning-based systems design and train models based on (often large) visualization design corpora. For example, Hu et al. [47] trained a deep learning model using millions of Plotly visualizations and
recommended visualization designs for new datasets using the trained model. In a similar spirit, Luo et al. [59] implemented a visualization recommendation system by combining deep learning techniques with hand-written rules. Moritz et al. [64] introduced the Draco system, which enables users to generate relevant visualizations by formulating design requirements as rules passed to a constraint solver. One of the Draco applications, Draco-Learn, was implemented with a training model which learns effectiveness criteria from two prior empirical studies [45, 77].

Based on prior surveys of visualization recommendation systems [108, 110], we can see that most visualization recommendation algorithms rely on either theoretical principles or empirical results in visual perception. However, these systems only utilize a limited amount of research work as the design guidelines, which introduces the risk of suggesting ineffective visual encodings. To improve the performance of existing visualization recommendation systems, we aim to summarize study outcomes from the visual perception literature as a whole and share them as an actionable dataset.

2.2 Principles of Visualization Design

Many works investigate how to best design effective visualizations. Theory works such as Bertin’s visual encoding principles [8], and Mackinlay’s APT work [60] have been highly influential in information visualization research. Cleveland & McGill [19] organized the encoding channels put forth by Bertin from least to most effective in terms of quantitative data and validated this ranking in part through visual perception studies. Mackinlay [60] later extended the ranking to include ordinal and nominal data in the APT system. Shneiderman [85]’s task taxonomy then broadened Mackinlay’s work by including data types that were not covered in APT, such as multidimensional data, trees, and networks. The design principles proposed by Bertin, Cleveland & McGill, Mackinlay, and Shneiderman inform the structure of our framework, which focuses on organizing comparison among not only different visual encodings but also various visualization types.

Numerous later experiments build on these foundational theoretical works. For example, the experimental results of Cleveland & McGill were replicated and validated by Heer & Bostock [56] through crowdsourcing of visual perception experiments. Talbot et al. [73] also designed four follow-up experiments on the perception of bar charts to further explore and explain Cleveland & McGill’s results. Their main goal was to understand how different bar chart designs impact analysis task performance. Kim et al. [48] discuss ways to evaluate the effectiveness of twelve 3-encoding visualization designs for different low-level tasks and dataset characteristics. Kosara [51] finds that pie charts may be perceived differently than initially hypothesized by Cleveland and McGill. Saket et al. [77] evaluate the effectiveness of basic visualization types for a specific set of analysis tasks.

All these theoretical and empirical works can (and probably should) inform the design of visualization recommendation systems, yet their influence is still limited [78]. Rather than assessing visual perception from a structural and implementation perspective, existing surveys primarily summarize visual perception research to educate non-specialists [100, 101]. We believe this paper is the first to systematically synthesize visual perception literature into actionable data and guidelines for visualization recommendation systems. Furthermore, our work demonstrates how visual perception knowledge that has generally been overlooked in visualization recommendation systems can be used to improve their performance.

3 Methodology

Our goal is to enhance the ability of visualization recommendation systems to reason intelligently about the effectiveness [60] of various visualization designs across analysis tasks and datasets. To achieve this, we first need to understand the space of visualization designs and visual comparisons that are most relevant to visualization recommendation systems. In this section, we formally define the visualization design space that we focus on in this paper. Then, we describe our method and rationale for collecting and filtering relevant theory and experiment papers in visual perception.

3.1 Visualization Design Space

First, we need to strategically define the visualization design space in which we believe a single recommendation system (or algorithm) can be effective. On the one hand, it is impractical to derive a single visualization recommendation system to cover all possible visualizations. On the other hand, it is equally impractical to expect visualization users to learn a completely different system for every conceivable visualization use case. Here, we first establish the boundaries of the visualization design space and then discuss how we specify visualization designs in the space with detailed components.

3.1.1 Establishing Design Space Boundaries

Our boundaries are informed by existing literature on (1) visualization design spaces, which formally define the range of visualization designs that could be recommended; and (2) visual perception studies, which can be used to identify a subset of designs that can be fairly compared in terms of user performance. We summarize our findings as the following constraints on the visualization design space.

B1. Exclude 3D visualizations. As found in previous work, users often have difficulty in perceiving information within 3D visualizations [72]. Moreover, in many cases, multiple linked 2D views prove to be more effective than a single 3D visualization of the same data [96]. Thus, we exclude 3D visualizations from our design space.

B2. Exclude network graph visualizations. As discussed in previous work [53], graph analysis tasks are generally considered separate from tabular data in visualization research and should likely be studied separately. In this paper, we focus on existing literature that studies visualizing tabular data. Thus, we exclude graph visualizations, like trees, treemaps, networks, radar charts, chord diagrams, etc.

B3. Focus on static visualization designs. Although animations and transitions can improve a user’s perception of an underlying dataset [67], many if not most visualizations are still designed without any animations or transitions. Given a lack of data in the literature evaluating the animation and transition design spaces, we do not include these design elements within our visualization design space. Similarly, the design space of interactions is still an under-explored area in visualization, and enumeration of this space has only recently become viable [61]. In this case, the lack of data and theoretical principles is already evident and does not require an in-depth literature review. As a result, we exclude animations and interactions from our analysis. We plan to revisit this gap in our review as more data becomes available.

3.1.2 Specifying Visualization Designs

After establishing the design space boundaries, we then discuss how to specify individual visualization designs to be compared. Many factors could impact the effectiveness of visualization, such as the distribution of the targeted data, as well as the encoding channels, the mark type, and the scale used in the visualization. With the guidance from existing visualization grammars [81, 102], we use six components to specify each visualization design observed:

- **Data Types**: quantitative, nominal, or ordinal.
- **Data Characteristics**: cardinality, entropy, and correlation.
- **Data Transformations**: aggregation or bin.
- **Encoding Channels**: position (X/latitude, Y/longitude), length, angle, area, texture, shape, color saturation, color hue, orientation, column, row.
- **Mark Types**: point, line, area-circle, area-rect, area-arc, area-other, text, geoshape, box-plot.
- **Scales**: linear, log, nominal, or ordinal.

We utilize data types, characteristics, and transformations to describe the data while encoding channels, mark types, and scales to specify the visualization design itself. There exist more measurements for data characteristics like scagnostics [103]. However, scagnostics are mainly used for one chart type—scatterplot, which is already covered by a recent survey [80]. In this paper, we focus on comparing different visualization designs instead of emphasizing one or two specific chart types; thus, we select the three most commonly used measurements for data characteristics—cardinality, entropy, and correlation.
When selecting encoding channels for our analysis, we start with the encoding channels discussed in the ranking of perceptual tasks proposed by Cleveland & McGill [19] and later extended by Mackinlay [60]. We remove the connection and containment channels, because they are mainly used for graph visualizations which we exclude from the analysis (Sect. 3.1.1). We also find that orientation has been discussed frequently in the literature (e.g., [83]), and is similar to the direction channel proposed by Cleveland & McGill [19] and the slope channel mentioned by Mackinlay [60]; thus we combine them into orientation channel. We split the position channel into positionX and positionY, since there are 2 directions of position in the 2D Cartesian plane, which could impact a user’s perception of these values. We also add column and row encodings for faceting charts, bringing the number of encoding channels to 12 (see Fig. 1). To save space, we use an icon to represent each channel. — is positionX, • is positionY, ⬠ is length, ⬠ is angle, ⬠ is area, ⬠ is texture, ⬠ is shape, ⬠ is color saturation, ⬠ is color hue, ⬠ is orientation, ⬠ is column, and ⬠ is row.

3.2 Paper Collection and Filtering

To initially find relevant papers for our literature review, we checked all papers in well-known visualization-related conferences and journals (e.g., IEEE TVCG, ACM SIGCHI, EuroVIS) in the last ten years, in which we searched for the keywords “encoding”, “perception”, “effectiveness”, “evaluate” in the titles, abstracts, and keywords. We also reviewed the references for each paper found through colleagues or online searches; any relevant papers were also included in our review. In total, we found 132 candidate papers for our literature review.

We then excluded papers that fall outside the boundaries of the visualization design space described in Sect. 3.1.1. For example, we excluded papers that only evaluate 3D visualizations, graph visualizations, or animated visualizations. Given our focus on providing guidelines for visualization recommendation systems, we use the following filters to guide our paper selection process:

F1. Focus on human perception and task performance. An essential facet of visualization recommendation systems is encoding selection, which directly impacts a user’s ability to perceive the underlying information [49,108]. Even if a visualization system suggests certain data attributes to explore, these findings will be inaccessible to the user if the data is presented incorrectly. Thus, we focus on results that speak to a user’s ability to perceive different visual encodings and differences in user performance across tasks and visualization designs.

F2. Focus on evaluation with standard displays. Although some existing work has researched the effect of display size on visual perception or task performance [5,24,36], and some are building new systems to better support different display sizes [61,118,40], the vast majority of existing visualization evaluations are still conducted in regular displays (e.g., computer screens). Thus, we focus on reviewing the literature in visualization evaluation and comparison with standard desktop and laptop displays.

F3. Focus on evaluation with visualizations that can be generated by automatic processes. Recent work [23] combines natural language analysis techniques with visualization synthesis to automatically generate infographics; however, it remains challenging for recommendation systems to understand the semantic meaning of the dataset and then select corresponding encodings or visualizations. Thus, we exclude papers only evaluating visualizations that usually require certain amount of manual generation, like visual embellishments [10,44,87,89], and semantically color assignments [55,83], etc.

F4. Compare different visualization designs. In order for algorithms to select the most relevant visualization design for a given dataset, they must be able to compare and ultimately rank the effectiveness of different designs [108]. To determine which designs should be preferred by these algorithms, we need experimental results that compare different visualization designs or theoretical rules and guidelines to prune irrelevant designs. To this end, we include any paper in our review that compares the user’s ability to effectively perceive and reason about information encoded using different visualization designs (at least one of the six components from Sect. 3.1.1 are different).

This filtering step excluded 83 of the 132 candidate papers, leaving 59 papers for our analysis.

4 Systematically Recording Perceptual Results

In this section, we present a schema to record extracted visualization rankings. We use this schema to generate a data record for each of the 59 papers in our literature review, contributing a JSON dataset of visual perception results that could be imported directly into visualization recommendation systems. Our schema also enables a fine-grained analysis of how much visual perception knowledge exists to inform encoding choices within these systems. Our schema has four components:

Category: either theory, experiment, or hybrid. Theory papers tend to emphasize theoretical principles. Experiment papers focus on experiments. Hybrid papers present theoretical hypotheses and experiments conducted to test (at least some of) the proposed hypotheses.

Designs: a list of all the visualization designs tested by each paper in our review, each specified using our visualization space design parameters from Sect. 3.1.2.

Tasks: a set of visual analysis tasks used in the existing literature for guiding the evaluation of visualization designs. Previous research has indicated that the effectiveness of visualization depends on the data attributes to be visualized [79] and the task to be performed [4]; thus, we also include tasks in our schema. We use prior work [4,48] as guidelines and develop a hierarchical task taxonomy.

Rankings: ordered lists representing the rankings for the proposed or tested visualization designs. Results are separated into theoretical rankings and experimental results.

4.1 Visualization Designs

Listing 1: Example of a covered visualization design, where a bar chart overlaid on a line chart, as shown in Fig. 2.

```json
1"E-4": { 2 "layers": [ 3 {"encodings": [ 4 {"data-type": "quantitative", 5 "data-chars": {}, 6 "data-trans": {}, 7 "channel": "positionY", 8 "scale": "linear"}, 9 {"data-type": "ordinal", 10 "data-chars": {}, 11 "data-trans": {}, 12 "channel": "positionX", 13 "scale": "ordinal"}], 14 "mark": "line"}, 15 {"encodings": [ 16 {"data-type": "quantitative", 17 "data-chars": {}, 18 "data-trans": {"aggregate": "mean"}, 19 "channel": "length", 20 "scale": "linear"}, 21 {"data-type": "ordinal", 22 "data-chars": {}, 23 "data-trans": {}, 24 "channel": "positionX", 25 "scale": "ordinal"}], 26 "mark": "area-rect"}]}
```
5 Integrating Current Rankings & Guidelines

In this section, we review existing visual perception theories and experiments that could be used to guide visualization recommendation systems. We synthesize existing performance rankings across different visualization designs and summarize the impact of data characteristics and tasks on these rankings. We generate tables summarizing our findings, which system designers can use to specify encoding rules for visualization recommendation systems, e.g., as queries [104] or constraints [64]. We have three research goals for this work: (1) summarize how to rank individual encodings according to their expressiveness and effectiveness; (2) summarize how to rank variations on a single chart type to enhance its design; and (3) summarize rankings for comparing different chart types, to identify the best performing visualization designs for specific data characteristics or task types.

### 5.1 Encoding Channels

We first cluster research papers by the encoding channels they cover.

#### 5.1.1 Literature Coverage

As shown in Table 2, all twelve encoding channels are covered by existing literature. To analyze encoding coverage, we break down visualization designs into encoding sets. For example, a paper studying scatterplots covers both the \( \mathbb{w} \) and \( \mathbb{l} \) encodings. As another example, a paper that studies grouped bar charts covers three encodings: \( \mathbb{w}, \mathbb{d}, \) and \( \mathbb{l} \). We can see that the \( \mathbb{w} \) and \( \mathbb{l} \) encodings are well-covered (49/59, 83.05%), \( \mathbb{d} \) (29/59, 49.15%), \( \mathbb{d} \) (28/59, 47.46%), and \( \mathbb{d} \) (23/59, 38.98%) are the most discussed encodings, while other encodings such as \( \mathbb{d} \) (9/59, 15.25%), \( \mathbb{d} \) (7/59, 11.86%), \( \mathbb{d} \) (6/59, 10.17%) and \( \mathbb{d} \) (2/59, 3.39%) are less mentioned.

#### 5.1.2 Study Outcomes

In Table 3, we summarize findings of encoding perception organized by Mackinlay’s principles of expressiveness and effectiveness [60]. A visualization design is considered expressive if it shows all and only the data the user wants to see and effective if a user can accurately interpret...
Table 1: A taxonomy of visual analysis tasks based on the task categorizations proposed by Amar et al. [4] and Kim and Heer [48].

| Tasks          | Descriptions                                                                 | Relevant Work |
|----------------|-----------------------------------------------------------------------------|---------------|
| Retrieve Value | Identify values of the specified attributes                                 | [1, 15, 21, 24, 40, 57, 62, 73, 77, 80, 98] |
| Filter         | Find data points satisfying the specified conditions                         | [11, 67, 68, 77, 80, 97, 98] |
| Sort           | Compare a set of data points by the specified ordinal metric                 | [14, 19, 21, 36, 47, 52, 56, 62, 72, 80, 90, 93, 98, 159] |
| Cluster        | Detect clusters of similar attribute values                                 | [12, 25, 30, 34, 37, 47, 50, 60, 99] |
| Correlate      | Determine or estimate the correlation within the specified attributes        | [11, 14, 17, 22, 35, 47, 50, 62, 73, 77, 80] |
| Aggregate      | Compute or compare the aggregate value of the specified attributes           | [11, 14, 29, 29, 33, 39, 43, 47, 54, 62, 63, 69, 71, 73, 77, 80] |
| Find Extremum  | Find data points with an extreme value of the specified attribute            | [2, 3, 11, 13, 45, 48, 70, 72, 77, 91, 98] |
| Determine Range| Find the span of values of the specified attribute                           | [23, 34, 42, 47, 51] |
| Characterize Distribution | Identify the distribution of given attributes                        | [2, 3, 4, 5, 6, 14, 33, 44, 63, 77, 80] |
| Find Anomalies | Identify anomalies within the data set                                       | [2, 3, 4, 5, 6, 14, 33, 44, 63, 77, 80] |

Table 2: Literature coverage for the 12 encoding channels. The papers are grouped by their category: theory, experiment, and hybrid.

| Relevant Work |
|---------------|
| Q | N | O |
| ![](image1) | ![](image2) | ![](image3) |
| ![](image4) | ![](image5) | ![](image6) |
| ![](image7) | ![](image8) | ![](image9) |
| ![](image10) | ![](image11) | ![](image12) |
| ![](image13) | ![](image14) | ![](image15) |
| ![](image16) | ![](image17) | ![](image18) |
| ![](image19) | ![](image20) | ![](image21) |
| ![](image22) | ![](image23) | ![](image24) |
| ![](image25) | ![](image26) | ![](image27) |

Table 3: Encoding guidelines summarized from existing theoretical literature. Q means quantitative, N means nominal, and O means ordinal data. An encoding channel is recommended (✓), can partially support (○) or should not be used (×) for the corresponding data type.

| Expressiveness | Effectiveness |
|----------------|--------------|
| Q | N | O |
| ![](image28) | ![](image29) | ![](image30) |
| ![](image31) | ![](image32) | ![](image33) |
| ![](image34) | ![](image35) | ![](image36) |
| ![](image37) | ![](image38) | ![](image39) |
| ![](image40) | ![](image41) | ![](image42) |
| ![](image43) | ![](image44) | ![](image45) |
| ![](image46) | ![](image47) | ![](image48) |

Table 4: Ranking for encodings representing quantitative data from existing experimental literature, group by task type and metric. () differentiates the same encoding but with different mark types.

| Task          | Metric | Rank | Ref. |
|---------------|--------|------|------|
| sort          | accuracy | ![](image49) | ![](image50) |
| retrieve value | accuracy | ![](image51) | ![](image52) |
| retrieve value | bias | ![](image53) | ![](image54) |

The graphical representation. We cluster research papers regarding data types to learn which encoding works better for a specific data type.

Quantitative. As shown in Table 3, we can see that existing theoretical principles [60] do not recommend using □ and □ for quantitative data since they usually cannot be perceived to be ordered. However, empirical results from Chung et al. [18] show that □ and □ can be orderable; in particular, marks with countable differences (e.g., the number of spikes or lines) can be perceived as ordered (see Fig. 3).

In terms of effectiveness, Cleveland & McGill [19] propose a ranking for encodings representing quantitative data, and Mackinlay [60] extends this ranking to include nominal and ordinal data (see Table 3). Cleveland & McGill test part of the ranking with follow-up experiments. Their results show that □ encoding outperforms □ and □ encodings on sort tasks. Heer and Bostock [56] replicate these experiments but also add □ encoding in the test and adjust the experiments to make results among tested encodings comparable. Their results are similar to Cleveland & McGill’s. McColeman et al. [62] re-examine these encoding rankings with a different task and find they do not hold. Furthermore, they find that other factors—such as cardinality—have more influence on task performance than the encoding choice (see Table 4).

On the one hand, theoretical works [12, [13, 60] suggest that the full-color spectrum is not ordered, but part of □ still can be used for quantitative data. □, in comparison, is preferable to represent quantitative data. On the other hand, experiments are conducted to evaluate the human performance of perceiving □ and □ conveying quantitative data with various visual analysis tasks (see Table 5). Although Liu & Heer [57] and Reda et al. [73, 74] arrive at a similar finding which is that participants can discriminate more minor gradient variations with multi-hue colormaps (□ + □) than with single-hue ones (□ only), they draw different conclusions for the rainbow colormap (□). Liu & Heer [57] suggest that the rainbow colormap is not intuitive and performs the worst for ordering colors and should be jettisoned; however, Reda et al. do not discard the rainbow colormap. They recommend using rainbow or other multi-hue colormaps for value tasks at high spatial frequency. Schloss et al. [82], on the other hand, investigate how the different colormaps would be affected by background color. They find that when colormaps vary less in opacity, human perception is unaffected by the background; however, the role of the background increases when apparent variation in opacity increases.

Nominal. Mackinlay’s expressiveness rules—the only relevant theory work observed—do not recommend using □ and □ for nominal data since they would probably be perceived to be ordered. Empirically-focused papers [25, 90] aim to learn how discriminable different encod-
Table 5: Ranking for colormaps representing quantitative data, group by task type and metric (only top 3 colormaps are shown). > means the left performs better than the right, while ≥ means the same order but with some uncertainty.

| Task      | Metric     | Rank | Ref. |
|-----------|------------|------|------|
| sort      | accuracy   |      |      |
|           | time       |      |      |
| aggregate | JND        |      |      |

Fig. 4: Shape, color hue and area palettes proposed by Demiralp et al. [25]: bottom palettes are re-ordered to maximize perceptual distance.

Ordinal. As for expressiveness, existing theory work does not recommend hh to represent ordinal data and recommends hh instead [59]. We find one empirical work evaluating hh and hh for conveying ordinal data [31]. They confirm that hh (red) performs better than hh (blue) in both accuracy and time for small tasks.

Summary. Theoretical work helps eliminate bad encoding choices for a specific data type, while experimental work examines some hypotheses on how visual encodings perform in practical scenarios. Compared to quantitative and nominal data, evaluations for using different encodings conveying ordinal data have so far received little research attention. More work is needed to determine which encoding performs better. Existing results show that human performance depends on the encoding choice as well as task type and data characteristics; also, encodings exhibit asymmetric effects on each other. In addition, state-of-art palettes [25,27,32,99] are proposed based on empirical results or mathematical formulation.

These findings can (but not fully) inform the encoding choices of visualization recommendation systems, such as picking an effective colormap for conveying quantitative data under a specific task and optimizing the perceptual distance of each categorical class. In Sect. 6.1.1 we demonstrate how we can use existing encoding design knowledge to improve Voyager’s [105,106] recommendations.

5.2 Chart Types

Although research findings at the individual encoding level can help systems avoid obvious pitfalls (e.g., choosing a poor color scheme), they fail to clarify interference effects [48] between encodings. To address the complexities of combining encodings into full visualization designs, we re-cluster research papers by the chart types they cover. We again summarize observed coverage and study outcomes from the literature towards achieving our two remaining goals: (1) comparing different variants of a single chart type, and (2) comparing different chart types to inform selecting better visualization designs.

Table 6: Literature coverage for different chart types. The papers are grouped by their category, theory, experiment, and hybrid.

| Task              | Relevant Work                                      |
|-------------------|---------------------------------------------------|
| Scatterplot       | 60, 63, 99, 17, 22, 28, 30, 38, 20 44, 46          |
| Bar Chart         | 60, 69, 70, 92, 93, 24, 25, 35, 38, 20 44, 46      |
| Line Chart        | 22, 35, 38, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |
| Area Chart        | 35, 38, 32, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |
| Pie/Donut Chart   | 22, 35, 38, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |
| Heatmap           | 22, 35, 38, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |
| Parallel Coordinates | 22, 35, 38, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |
| Geomap/Cartogram  | 22, 35, 38, 30, 60, 70, 92, 93, 24, 25, 35, 38, 46 |

5.2.1 Literature Coverage

We summarize the chart type coverage by existing literature in Table 6. Here we use higher-level visualization types to cluster papers. For example, we lump bubble charts into scatterplots and any variants of area charts like stream graphs into area charts. We group pie charts and donut charts into one category, as well as geomaps and cartograms. We can see that bar charts (26/59, 44.07%), scatterplots (24/59, 40.68%), and line charts (15/59, 25.42%) are the top 3 studied chart types, while only a few research papers cover other chart types like area charts (5/59, 8.47%), heatmaps (4/59, 6.78%), and parallel coordinates (4/59, 6.78%).

This result is pretty consistent with Beagle [7], based on which bar charts and line charts are the most popular visualization types among all SVG-based visualizations mined from web. On the other hand, we also observe the dearth of theoretical work in the space (see Table 6); thus, we focus on summarizing the study outcomes from empirical work in Sect. 5.2.2 and Sect. 5.2.3.

5.2.2 Within Chart Type Comparison

Scatterplots. Besides the evaluation of the class separability perception [90,99] (mentioned in Sect. 5.1.2), scatterplots are also examined with different data characteristics [28,32,48], and marks [28,39,48,56]. We summarize the findings from these research papers in Table 7. Kim & Heer [48] suggest that in general using + performs better in summary tasks and ◦ performs better in aggregate tasks when representing quantitative data; however, the performance exhibits significant variance across different data characteristics (entropy and cardinality). Gleicher et al. [28] found that higher cardinality (more numbers of points) leads to marginally better performance in aggregate tasks; on the other hand, using redundant encodings (like using the combination of + and ◦ for a same nominal attribute) would not influence the task performance significantly. Gramazio et al. [33] suggest that using larger marks can reduce participants’ response time; however, Hong et al. [39] find that larger and darker marks lead to more bias. Liu et al. [58] studied if the mark orientation would affect task performance and found that the mark orientation that is consistent with the trend of the scatterplots can reduce errors in estimate trend tasks.
Bar Charts. Unlike scatterplots, bar charts are often studied with different variants [15, 45, 67, 91, 93] and arrangements [93, 109]. Srinivasan et al. [91] and Nothifer & Franconeri [67] evaluate different bar chart variants for comparing changing data. They both find that visualizing data differences yields better performance and suggest using charts with difference overlays since only visualizing deltas would lose the context of base values. For visualizing disproportionate values, Karduni et al. [45] propose using Du Bios wrapped bar charts. They find that wrapped bar charts lead to higher accuracy but sometimes at the cost of more time needed than basic bar charts. Other experiments focus more on the perception of bar charts. Talbot et al. [93] explore variations of bar charts originally studied by Cleveland & McGill [19] and find that shorter bars are more difficult for sort tasks. Zhao et al. [109], on the other hand, investigate whether neighborhood would influence the perception of bars with sort tasks. The results show that neighborhood does have an effect, but the effect size is small; other factors like data characteristics have dominated effects. Godau et al. [29] find consistent underestimations with bar charts, which are not affected by the height of bars; moreover, the bias persists even adding a numerical scale or outliers. Caja et al. [15] recently find that bars with wide aspect ratios are overestimated, bars with tall ratios are underestimated, and bars with square ratios show no systematic bias.

Line Charts. We only find one paper that studies line charts solely. Aigner et al. [1] evaluate three types of line charts (juxtaposition on linear scale, superimposition on log scale, and indexing) with various tasks. They find that using indexing generally yields higher accuracy and user preference than the two other types; the advantages with completion time are less clear, although some benefits are shown.

Small Multiples. Both Ondov et al. [69] and Jardine et al. [42] study how different arrangements of small multiples would affect the task performance. In their experiments, five arrangements (stacked, adjacent, mirrored, overlaid, animated) are tested with three chart types—bar charts (with find extremum, correlate, determine range, aggregate tasks), line and donut charts (with find extremum task only). The results suggest that it is unlikely to discover an easy guideline that specifies the best arrangement or encoding for a given task.

Summary. Our summary tables provide not only a performance overview of the relevant literature but also design guidelines that could be encoded within visualization recommendation systems. For example, using a wide bar chart to represent quantitative data might improve the performance of scatterplots in aggregate tasks, and using redundant encodings might not provide any additional benefit. In Sect. 6.1.3, we demonstrate how these guidelines can be translated into constraints towards informing Draco’s [64] design decisions.

However, we note that the literature tends to focus more on studying variants of scatterplots and bar charts than other chart types. Although the line chart is one of the most popular chart types used on the web [7], not many evaluations of variants exist; thus, we urge more experiments examining different variations of line charts. Moreover, we also find that existing study results differ based on individual factors like data characteristics, tasks, experiment setups, and participants.

5.2.3 Between Chart Type Comparison

Time Series Data. We observe that visualization time series data is often discussed in existing literature [2, 20, 30, 43, 70, 71]. Correll et al. [20] perform an empirical experiment of a aggregate task for time series data; four display conditions were tested: ordered/permutated line chart and ordered/permutated colorfield chart. The results suggest that colorfield charts outperform in accuracy across all difficulty levels. Albers et al. [2] extend the work of Correll et al. in a follow-on experiment by testing more visualization designs with more tasks. The results confirm that different designs support different tasks; position-based charts outperform in some tasks (find extremum, determine range) while color-based charts perform better in others (aggregate, find anomalies). Instead of testing encoding performance (position vs. color), Javed et al. [43] explore user performance for find extremum, determine range, sort tasks for different line graph techniques (shared-space vs. split-space). They find no significant difference between these two techniques in terms of accuracy; however, shared-space techniques are faster in find extremum while split-space ones are faster in sort tasks.

Systematic Bias. Bias evaluation has attracted more attention recently [15, 21, 29, 107]. Godau et al. [29] test whether there is a bias in the central tendency perceived in bar charts, and they find that the mean value is systematically underestimated in bar charts (but not in scatterplots). Their other experiments also confirm that the underestimation of the average persists with varying bar heights or adding outliers. However, Xiong et al. [107] reach a completely different conclusion. They conducted three empirical studies to investigate position perception bias with visualizations containing a single bar/line, multiple bars/lines, and one line with one set of bars. In contrast to the results of Godau et al., they find that the perceived average position was significantly biased in both single line charts and single bar charts. Line positions are consistently underestimated, while bar positions are overestimated.

In the experiments involving multiple data series (multiple lines and/or bars), they also observe an effect of “perceptual pull”, where the average position estimate for each series was “pulled” toward the other. Aiming to explain this contradiction, recent research by Caja et al. [15] finds that the systematic bias in bars is related to the aspect ratio of bars. No systematic bias is shown with square bars, while wide bars are overestimated and tall bars are underestimated. We summarize the study outcomes in Table 8.

| Underestimate | Overestimate | Perception Pull | No Bias | Ref. |
|---------------|--------------|-----------------|---------|-----|
| ![Underestimate](image1) | ![Overestimate](image2) | ![Perception Pull](image3) | ![No Bias](image4) | 29 |
| ![Underestimate](image5) | ![Overestimate](image6) | ![Perception Pull](image7) | ![No Bias](image8) | 107 |

Scatterplots vs. Parallel Coordinates. Several current works focus on comparing scatterplots with parallel coordinates [19, 36, 52, 75, 98]. Li et al. [54] focus on studying correlate task performance between scatterplots and parallel coordinates and find that the degree of correlation between attributes is underestimated in parallel coordinates, suggesting that scatterplots are better options for correlate tasks. Kanjanabose et al. [44] also perform experiments comparing scatterplots and parallel coordinates, but focus on other visual analysis tasks include retrieve value, cluster, find anomalies and determine range. The results suggest that parallel coordinates outperform in accuracy across cluster, find anomalies and determine range tasks, and in completion time with retrieve value and determine range tasks.

Bar Charts vs. Pie Charts. We find five papers [10, 45, 52, 75, 98] comparing bar charts and pie charts. Cleveland & McGill [19] propose an order of encoding channels based on graphical perception but also test parts of this theory through experiments. They use bar charts to assess position and length encodings and pie charts for angle encoding. Heer & Bostock [36] replicate these experiments but also adjust the experiments to make results between length and angle encodings comparable. They conduct the experiments with sort tasks, and both of the results suggest that bar charts perform better than pie charts in terms of the accuracy metric. Later on, Waldner et al. [98] also report that radial charts perform less accurately, efficiently, and preferably than bar charts in many analytical tasks. Kosara [52] also has similar findings (bar > pie) with find extremum tasks. However, by comparing the performance of pie charts and bar charts with multiple variants, Redmond [75] finds that pie charts perform more accurately with retrieve value tasks.

Multi-Chart Comparisons. Some experiments involve a large range of chart types [22, 35, 66, 77, 78, 88]. Saket et al. [77] conduct an
Table 9: Visualization types recommended by empirical work for each visual analysis task. Multiple visualizations recommended for the same task might not be comparable.

| Task          | Designs                                      | Task          | Designs                                      |
|---------------|----------------------------------------------|---------------|----------------------------------------------|
| correlate     | 156, 104, 77                                 | distribution  | 77                                          |
| cluster       | 77, 44                                       | filter        | 77, 44                                      |
| retrieve value| 75, 44                                       | determine range| 44                                          |
| find anomalies| 77, 44                                       | find extremum | 53, 77                                      |
| sort          | 19, 36, 77                                   | estimate trend| 22                                          |

Fig. 5: An example of improving Voyager’s recommendations using our synthesized guidelines (specifically, using [25, 90]).

6.1.2 Foresight [26]

The Foresight system suggests data attributes based on their “insight” scores and presents the recommendation with either a bar chart, a box plot, or a scatterplot. As we can see from Fig. 6, Foresight uses different chart types to visualize different “insight classes”. Bar charts are used for distributions (heavy tails), while box plots are used for outliers and scatterplots for correlations. However, the results from experiments by Saket et al. [77] suggest that scatterplots perform the best for finding anomalies, while line charts for correlation, and bar charts for distribution tasks. Thus, replacing existing chart types with the best ones (see Table 9) based on empirical results for corresponding tasks might help users gain more insights more effectively with Foresight.

6.1.3 Draco [64]

The Draco system models visualization design guidelines as hard or soft constraints. Draco first excludes the visualizations that violate experimental design guidelines from consideration.
the hard constraints and then searches for the most preferred visualizations using soft constraints. Although Draco already applies some of the existing visualization design knowledge in its applications (Draco-API [60], Draco-CQL [60,61], and Draco-Learn [48,77]), the number of utilized research papers is limited. In this paper, we collect visual perceptual results from 59 existing literature works. First, our synthesized dataset of perceptual results (see Sect. 4) can be used as an input corpus for Draco-Learn. By learning visualization rankings from a large number of research papers (instead of two papers), Draco-Learn could potentially support more chart types and produce more effective recommendations for observed data characteristics and task types. Second, our synthesized guidelines in Sect. 5 could be translated into hard or soft constraints to further enhance Draco’s recommendations. For example, we translate two visualization design rules from Sect. 5.2.2 (e.g., preferring color encodings for the aggregate task) into Draco soft constraints in Listing 3.

Listing 3: Examples of translating our synthesized guidelines into Draco soft constraints [63].

%Prefer to use fewer encodings with fields
soft(encoding_field,E) :- encoding(E), field(E,_).
%Prefer to use color for aggregate task
soft((aggregate_color,E) :- task(aggregate), channel(E, color).

6.2 Challenges & Open Research Areas

Here we identify challenges toward a comprehensive ranking of visualization designs and discuss directions for future research based on our analysis.

6.2.1 Gaps in Visualization Comparison Coverage

Gaps in theoretical work. Although all twelve encodings are ranked (or pruned) according to theoretical hypotheses (see Table 2), only a few visualization types are discussed in theory work (see Table 6). In other words, interference effects among visual encodings are rarely theoretically studied. Solutions: Deriving theoretical principles from existing experiment results for multi-encoding designs to infer how well different encodings will work together and whether the performance of encoding combinations still follow the same rankings would be an impactful area. New theories can also help to prune the design space to identify gaps that truly warrant new experiments.

Gaps in experimental work. On the one hand, we observe much fewer experiments evaluating how effectively each encoding could convey ordinal data (only two encodings are tested: II, III). Solutions: We urge more empirical work to conclude which encoding to pick under different task scenarios.

On the other hand, existing evaluations mainly focus on specific encodings (e.g., II, III, IV, V, VI) or charts (scatterplots, and bar charts), while other chart types are either only compared with one or two other charts, or never studied with different variants (see Sect. 5.2). Solutions: Evaluating the performance of previously ignored encodings (e.g., X, Y) and chart types (e.g., area charts, heatmap) under different analysis tasks would contribute more “ground truth” evidence to further validate our approaches to automating the visualization design process.

6.2.2 Inconsistencies and Conflicts in the Literature

Between theories and experiments. We observe that theoretical hypotheses might not necessarily be “correct” in a practical sense. For example, as previously mentioned, five attribute-encoding pairs ((Q, X/2), (O, Y), (N, W/M)) are considered inexpensive based on Mackinlay’s work [60] (see Table 3), however, a more recent evaluation [18] shows different results. Mackinlay suggests that X and Y are not relevant to quantitative data, but according to the results from Chung et al.’s experiments, both X and Y encodings perform better in accuracy than II conveying quantitative data with estimate trend and find extremum tasks. Solutions: Refining core theory work in light of recent experimental results could further enhance the performance of visualization recommendation systems.

Between different experiments. Even when experiments were similar, we may find contradictory results. Even though Godau et al. [29] and Xiong et al. [107] both conducted experiments to test human bias in perceiving average position for length (bar charts) and position encodings (scatterplot or line charts), they have completely different results (as shown in Table 5). Godau et al. only find underestimation in bar charts but no bias for point positions (scatterplots). However, Xiong et al. find significant bias in both bar charts and line charts, where line positions are underestimated while bar positions are overestimated. In another example, Harrison et al. [33] find Weber’s law to be a convincing model for how people perceive data correlations; however, in a re-analysis of the same data, Kay and Heer [46] find Weber’s Law not to be a good fit. It is natural in science to improve upon existing results and theories; however, there is currently no easy way to identify and track these discrepancies within the literature and translate them into concrete improvements to visualization recommendation systems. Solutions: Redesigning experiments to test visualizations with controversial results, conducting more comprehensive comparisons between more nuanced design decisions, and involving more metrics in our assessments could lead to more precise visualization design rankings for recommendation systems.

Summary. Given the (multiple) discrepancies we have observed, we argue that the findings of both visualization theory and experimental research should be treated as hypotheses until subsequent experiments converge on a consistent set of results. Furthermore, we argue that replication experiments should be held in high regard within the visualization community regardless of whether their findings reinforce or challenge our current assumptions, since either way, they are the only way to validate our understanding of how people perceive and use visualizations. We need them to ensure that visualization recommendation algorithms are built upon a solid foundation of theoretical and empirical findings; we should reward them accordingly.

6.3 Limitations & Future Work

Our literature review contributes a detailed record of how different visualization designs are compared and ranked in 59 different theoretical and experimental papers. This record not only specifies all researched visualization designs, but also keeps track of the ranking of their performance (accuracy, bias, AND time, user-preference) under different task scenarios. A next step to extend this work could be to apply the findings to develop better encoding strategies within visualization recommendation systems, such as adapting the recommendation strategy based on the user’s current analysis task.

Given our initial goal is to understand how different visualization designs (specially different encodings) are ranked in the existing theoretical and experimental work, our schema only records {data types, data characteristics, data transformations, encoding channels, mark types, scales} for each covered design (details in Listing 1). For more granular design decisions, we add notes to specify them. For example, to record Talbot et al.’s experiments testing bar charts [73], we add notes to specify each variant of bar chart tested, such as whether two bars are aligned or separated, whether distractors are added, the indicator location, etc. However, it is hard to parse these notes automatically. We see our schema as a starting point for collating existing encoding design knowledge, and encourage the visualization community to extend this schema to support more nuanced visualization designs.

Informed by existing work on visualization design spaces and visual perception studies, we excluded (1) 3D visualizations, (2) graph visualizations, and (3) visualizations with animations or interactions from our defined visualization space. When more theoretical and experimental findings become available in the literature, expanding our work to include these excluded designs would be interesting.

We also note that by focusing on visual perception, we are unable to account for other factors that may influence the overall effectiveness of a visualization design, such as visual aesthetics [95], intuition, and metaphors [111], as well as user background and preferences [112]. Developing a broader framework encompassing both visual perception and these other factors would be exciting future work.
