Towards a Customizable Framework for Evaluating Visualization Recommendations

Kelsey Fulton, Debjani Saha, Katherine Scola, and Leilani Battle

Abstract—Visualization recommendation systems such as Voyager [40], VizDeck [20], SeeDB [37], and Foresight [11] have become popular in recent years. These systems use a variety of metrics to recommended visualizations to end users (e.g., data scientists and analysts). However, due to high variability in the design of recommendation algorithms, evaluations of these systems are often done in isolation, and lack standardization, potentially leading to biased outcomes. In this paper, we present a new framework to evaluate the recommendations these systems produce in a standardized, usable and unbiased manner. Our framework produces a score from 0 - 100, thereby allowing the user to make a determination as to which system they believe provides results most relevant to their work. Our framework incorporates a range of metrics derived directly from the visualization literature. Every component of the framework can be configured to better match different analysis scenarios and user goals. We present results from evaluating our framework through both individual case studies and surveying visualization experts, demonstrating the efficacy of our framework in practice in five different visualization cases. We highlight lessons learned through the development and evaluation of our framework, and propose future research directions to further our goal of designing a robust, customizable, and easy-to-use metric for evaluating visualization recommendation systems.

1 Introduction

The analysis of large datasets has become a necessary process across a many disciplines where visualizations enable more efficient interpretation of this data. While users of visualization tools (e.g., researchers, analysts) often have research questions and hypotheses planned a priori, sometimes they are interested in *exploring* the data first for the purposes of hypothesis generation. A thorough approach to exploration involves a systematic search of the entire data set, considering all permutations of data attributes [27]. However, given the size of many modern data sets, this is a difficult task to accomplish unaided. In response, several recommendation systems have been developed to reduce user effort in exploration tasks by automatically suggesting visualizations of potential interest as the user explores.

These recommendation systems vary in their methodology for producing recommendations. For example, Voyager puts most of its emphasis on following Mackinlay's design principles [23] while covering a large breadth of the data. This system produces visualizations encompassing every combination of the attributes in the data set while maintaining expressivity and effectiveness [40]. On the other hand, the VizML system uses machine learning to make its recommendations by collecting information about the most frequently used visualization designs to predict which visualizations the user will want to see next for the current data set [16]. In contrast, SeeDB recommends visualizations based on different criteria: "interestingness." SeeDB's creators define interestingness as proportional to deviation in the data. To this end, SeeDB uses a deviation-based metric to produce visualizations that it thinks analysts and data scientists will find interesting [37].

While the variability in these recommendation algorithms allows for flexibility, it in turn makes comparing new and existing recommendation systems difficult for researchers and developers because there is no standardized evaluation process by which to assess them. For example, suppose Carol, a graduate student in visualization, has developed an algorithm called *NewRec* to help analysts save time in generating reports by suggesting common standard visualizations. Now she wants to compare *NewRec* to existing methods. One possibility is to conduct a user study to compare *NewRec* to existing systems such as SeeDB, VizML or Foresight. However, when she reviews the tasks and metrics of the reported user studies for these systems, she realizes that because these systems optimize different recommendation metrics, it is unclear how to compare them to *NewRec*. Specifically, Carol is unsure how to design an evaluation that would allow her to argue for which of these methods performs "the best" for the target analysis context. Ultimately, we need a standardized process for evaluating visualization recommendation systems.

However, in order to formulate a standardized—and eventually an automated—evaluation process, we first need a starting point for evaluating the individual recommendations themselves in a consistent way. In a review of current evaluation practices, we find that many recommendation systems evaluate their recommendations in isolation [11, 37] or use some metric that measures a benchmark that their system already optimizes for [20, 37, 40]. None of the current evaluation methods help researchers and developers gain a clear understanding of the trade-offs between different visualization recommendations, making it difficult to determine whether these recommendations are truly as effective as claimed. We need a consistent methodology for comparing one recommended visualization to another. This methodology could consider multiple recommendation objectives, such as adhering to known graphic design principles, which affect the interpretability of recommended data, but also the analysis context, such as an analyst's goal(s) in using the recommendations. Such a methodology could provide a strong foundation for more sophisticated evaluation frameworks for comparing collections of recommendations.

We aim to address this evaluation challenge by introducing a standardized, yet configurable, framework for evaluation of individual visualization recommendations. Our framework calculates a single score to measure the efficacy of visualizations produced by different visualization recommendation systems. Our framework draws primarily upon known measures used from existing recommendation systems [37, 40]. To demonstrate the flexibility of the framework, we also incorporate evaluation methods that have not yet been used to assess recommendation systems but are well-known in the visualization community [7, 35].

We have provided a software implementation of the framework, but manual input is still required to calculate scores. Some of the incorporated metrics are objective (e.g., evaluating visual encodings), while others may be more subjective (e.g., evaluating the presence of chart junk). However, our framework is data driven and designed such that by collecting feedback from more users, the framework's results become more objective and precise. Furthermore, we plan to automate the framework in the future to minimize required human intervention.

To demonstrate the use of our framework, we present two case studies where we calculate scores for two real-world examples of visualizations from the web. We also present results from a survey of $n = 15$ visualization experts evaluating five different visualizations. We find that experts use a variety of heuristics to evaluate visualizations for usefulness, clear and detailed labels, and chart junk. For useful-

ness, experts aimed to address questions like "does the visualization convey the data well?" to determine whether they would consider the visualization to be useful. To determine whether a visualization had clear and detailed labels, experts used proxies along the lines of "does the label accurately describe the data at hand?" or "are the labels easy to find and understand?" Lastly, when deciding if a visualization contained chart junk, experts evaluated visualizations by determining if the visualization contained any unnecessary elements.

We have made several observations that could prove valuable in the design of future recommendation algorithms and evaluations. In varying the assignment of weights in our framework to favor different criteria (e.g., favoring good encodings, graphical excellence, or "interestingness"), we see marked shifts in output scores. These results suggest that favoring a single criterion may lead to biased recommendations. As such, we argue that *recommendation systems should be evaluated along multiple diverse criteria*. Furthermore, we argue that it is critical to align the weights of any evaluation framework with the user's evaluation goals. With the myriad of weights within the framework, this task could seem daunting to users. However, we see in the literature that certain tasks occur frequently during visual analysis [8]. Using *preset configurations for common analysis scenarios could reduce the complexity of evaluation*. Automation could further reduce the burden of framework configuration and ultimately evaluation. Specifically, *automation could enable broader and more rigorous testing of evaluation techniques for visualization recommendation systems.* We have begun the automation process by providing code to calculate recommendation scores, available as part of the Supplemental Materials.

2 Related Work

2.1 Principles of Visualization Design and Evaluation

There has been much prior work in the space of determining how best to create effective visualizations. Bertin proposed theoretical principles for visually encoding data by deriving several different graphical objects and relationships between them. He defined several different encoding techniques for data [7]. This work was furthered by Mackinlay, who automated the ideas put forth by Bertin to create the APT system and a visual design language. Mackinlay posited that there should be two main criteria for this language: expressiveness and effectiveness. A graphical language is expressive if it shows all the data the user wants to see and only the data the user wants to see. A graphical language is effective if a user can interpret the graphical representation with optimal accuracy. Mackinlay organized the encoding channels put forth by Bertin from least to most effective as far as what users are able to perceive from the data [23]. Shneiderman extends Mackinlay's work by including data types that were not covered in APT such as multidimensional data, trees, and networks [32].

Tufte also proposed several guidelines to consider when designing effective visualizations. The focus of our study centers around two of Tufte's main ideas [35]: graphical excellence and graphical integrity. A visualization upholds graphical excellence if the visualization portrays the greatest number of ideas, in the shortest amount of time, with the least ink, in the smallest space. A visualization upholds graphical integrity if it is an accurate representation of the underlying data. From these main ideas, Tufte offers five design principles [35]:

1. Tell the truth about the data (reduce the lie factor)
2. Minimize the data-to-ink ratio
3. Provide clear and detailed labels and annotations
4. Show data variation and not design variation
5. Avoid chart junk

Our framework was designed to incorporate the design principles of Bertin, Mackinlay, Shneiderman, and Tufte. We use their ideas, not to create good visualizations, but to determine if visualizations created by recommendation systems adhere to good graphical design principles.

Other prior work pursues a more holistic evaluation. Through means of a literature review, Lam et al. collated seven principles to guide the evaluation of visualization systems in general [22]. We used their work to gather insight on how to best to go about the more specific task of evaluating visualization recommendation systems. Specifically, their work on evaluating visual data analysis and reasoning aided in the selection of criteria for use in our evaluation algorithm. While Lam et al.'s work focuses on evaluating visualization systems as a whole, our work extends this prior work in part, focusing specifically on recommendation systems.

2.2 Heuristic Evaluation of Visualizations

A range of heuristics have been developed to support general-purpose, but low-level evaluation of visualizations. Kim et al. discuss ways to evaluate the effectiveness of basic visualization designs across 12 encoding channels, for different low-level tasks and dataset characteristics [21]. Wall et al. propose the ICE-T methodology, which uses four categories of low-level heuristics to assess the value of various visualizations [38]. Sakert et al. evaluate the effectiveness of basic visualization designs for a range of specific analysis tasks [29]. The context in which a visualization is used is critical to tailoring an appropriate evaluation, a principle which we also incorporate into our framework.

These heuristics are related to our own research goals, and we incorporate several of these concepts within our framework. However, our focus is to integrate these concepts into a broader evaluation of visualizations that also considers the design goals and optimization objectives of existing visualization recommendation systems. For example, we utilize a similar approach to Wall et al. [38], but specifically for metrics that have been explored in visualization recommendation contexts, which includes one's understanding of the data as investigated by Wall et al., but also others such as whether a user actually finds the visualization interesting or useful for future analysis.

2.3 Visualization Recommendation Systems

Visualization recommendation systems utilize a wide array of evaluation metrics to make suggestions to the user. Wongsuphasawat et al. used Mackinlay's principles within their system, Voyager, to make recommendations to the user, prioritizing recommendations based on the breadth of data the visualization(s) covered [40]. Vartak et al. use an "interestingness" metric based upon deviation in the data to make recommendations to their users [37]. Jayachandran et al. determine what recommendations to make to users in their system, DICE, by predicting what queries the user will make and then uses sampling to rapidly answer these queries [18]. Battle et al. utilize a similar prediction approach to pre-fetch relevant data for exploration [3]. Demiralp et al. use a ranking scheme to decide which visualizations to recommend in their system, Foresight. They rank visualizations based on the most relevant attribute tuples related to the insights of the user [11]. Hu et al. use machine learning to make recommendations to their users in their system, VizML. VizML first learns the most used visualization designs from a large corpus of data sets and their associated visualizations. The system then makes recommendations based upon the data set input by the user [16]. All of these recommendation systems utilize completely different metrics to determine which visualization to recommend to the user, but provide a strong list of candidate measures to consider for our evaluation framework.

The Draco system by Moritz et al. enables users to generate relevant visualizations by formulating their desired design features (e.g., the design principles described above) as rules passed to a constraint solver [25]. Our proposed framework provides a convenient approach for assessing visualizations produced by Draco and other visualization recommendation systems.

Finally, there has been some work focusing specifically on evaluating recommendation systems. Vartak et al. detail key design decisions that should be considered when creating and making recommendation systems in the future. They offer guidelines along three axes: recommendation axes, recommendation criteria, and architectural considerations. They suggest considering data characteristics, semantics and domain knowledge, user preference and competencies, visual ease of understanding, and intended task or insight when selecting the recommendation axes. When selecting recommendation criteria, the authors suggest considering the relevance, surprise, non-obviousness, diversity, and coverage of the recommendation system [36]. While we do not

focus on all of the suggestions put forth by Vartak et al., we chose to incorporate the consideration of data characteristics, visual ease of understanding, and intended task or insight into our algorithm when determining recommendation axes. We incorporated relevance, surprise, and non-obviousness into our algorithm when determining recommendation criteria. While the authors make valuable suggestions, their advice has never been concretely realized in an evaluation framework. Our contribution is to fill the current gap in implementation of these suggestions in an actionable and robust manner.

2.3.1 Recommender Systems at Large

Note that many recommendation systems exist outside the context of data visualization, and are subject to similar pitfalls that exist in the visualization space: deciding which one is most relevant to a particular domain is not a simple task. Gunawardana and Shani address the diversity of recommendation algorithms, and provide insight in how best to set up an offline experiment to determine which of these algorithms is most suited to the user's purposes [12]. Despite the diversity in recommendation algorithms, many produce item rankings based on some measure of similarity to user queries. Hurley et al. argue against this trend in favor of items that may seem novel or unusual, i.e. diversifying the recommendations produced by a system, akin to SeeDB's measure of "interestingness" [17, 37].

Further, Gunawardana and Shani stress the importance of selection of the proper evaluation metric (error-, precision-, or utility-based), warning that the use of an unsuitable metric may lead to choosing an algorithm that is suboptimal for the given task [12]. Bellogin et al. survey two error- and three precision-based evaluation metrics and find that four our of the five give comparable results, whereas the final one reliably overestimates performance [6]. These techniques focus primarily on user behaviors (e.g., what people buy or click on), ignoring contextual knowledge (e.g., perception, analysis task, user goals, etc.). In contrast, our framework combines a range of metrics that take analysis context into account.

3 Design Goals

Our primary design goal was to prioritize both good graphical design and the utility of recommended visualizations. Given that many works mentioned the visualization rules of Bertin, Mackinlay, Tufte, and Schneidermen, we felt it was best to use their metrics in our evaluation algorithm to evaluate the graphical design of the visualization. Furthermore, since we are not simply evaluating visualizations but are evaluating the visualizations put forth by recommendation systems, we felt it was important to emphasize criteria necessary to a good recommendation: usefulness and interestingness. This broader goal was refined to create a set of concrete metrics, ultimately resulting in the inclusion of five main evaluation criteria in our framework:

C1 Effective use of encoding channels

C2 Adherence to Graphical Excellence and Integrity principles

C3 Visualization "Interestingness"

C4 Visualization "Usefulness"

C5 Configure-ability and Ease of Use

Furthermore, we wanted to emphasize the usability of this tool for our user base. We specifically created this tool to be used by researchers or developers who are looking to create a recommendation system. We see our algorithm as being a tool for creators of new systems to measure their recommendation system against existing systems.

Encoding Channels (C1): This criterion focuses on determining the effectiveness of the encoding channels used in a given visualization. It incorporates design principles put forth by Bertin [7] and Mackinlay [23] (see §2.1). On a high level, this includes ensuring the encoding channels and mark types for each visualization are best suited to the particular type of data being displayed. It is essential that each dimension of data encoded in the visualization be displayed in a way that is easy for humans to perceive, and that this is done in a way that is effective for that data type. The use of suboptimal encoding channels and mark types can make a visualization unreadable or misleading, making this an important criterion for inclusion in our framework. Inclusion of these encoding principles is also supported by the literature and past research, as Voyager explicitly attempts to include Mackinlay's principles within its design [40].

Graphical Excellence and Graphical Integrity (C2): Tufte's principles provide one approach to assessing the usability of a graphic based on its design. This criterion considers how well a given visualization upholds Tufte's graphical excellence and graphical integrity principles [35]. We consider adherence to graphical excellence and integrity because it provides a useful measure of whether a visualization is: (a) hard to interpret or perceive, and (b) misleading to the user. Tufte proposes five main principles (see §2.1), which prioritize ensuring that the visualization is visually pleasing and that the design of the visualization does not take away from understanding the underlying data. For the purposes of this work, we chose to focus on two of these principles: "clear and detailed labels and annotations" and "avoiding chart junk". Our rationale for selecting these two principles is discussed in §4.4.

Visualization "Interestingness" (C3): The third criterion involves the evaluation of "interestingness," i.e. whether or not a given visualization is "interesting" to the user. **C3** was inspired by SeeDB's deviation-based utility metric, which tries to produce interesting visualizations by prioritizing the display of data that deviates from some reference data set [37]. However, a deviation-based metric has too narrow of a view of "interestingness" to be effective in a general-purpose framework (discussed further in §4.5). Nonetheless, we desired to include some measure of interestingess in our algorithm, because it is crucial to the effectiveness of a recommendation system. If a platform recommends visualizations that are uninteresting to the user, then the system will fail to capture the user's attention. We define a visualization to be interesting if it relays something novel about the data to the user; specifically, whether the system produces a visualization the user may not have necessarily thought of themselves. To this end, we derived a metric for our framework to assess the novelty of a visualization based on how likely the user was to to come up with the visualization on their own (without the aid of a recommendation system).

Visualization "Usefulness" (C4): Our fourth criterion was the evaluation of usefulness, i.e. whether or not the visualizations being produced by a system are useful to the user. For the purposes of this project, we define usefulness as a measure of how helpful the recommended visualization is for answering the user's research question(s). This criterion complements **C3** since it is possible for a visualization to be interesting without being useful, or vice versa. Without incorporating both criteria, our framework could potentially award high scores to visualizations that were unhelpful to the user's end goals. Thus, it was important to us that we include usefulness in our formula.

Configurability and Ease of Use (C5): No single metric or formula will perfectly match any and all visual analysis goals. Rather than complicating the process by creating yet another metric by which to compare systems, we take a broader view that considers how existing metrics can be combined to provide a more holistic measure. In the next section, we provide recommendations on how existing metrics can be folded into a single score, and how weights can be assigned to different parts of the scoring mechanism (in this case, a formula) to address different user needs and goals. By providing a single score, we aim to make it easier for users to assess whether a recommendation system is providing effective visualizations for the given context.

4 Methods

In this section, we describe the scoring mechanism for our framework which incorporates multiple criteria into a single formula for scoring. The formula uses the design considerations discussed in §3 to compute a score from 0 to 100. This range was chosen because it is simple for users to understand scores that are formatted like percentages, allowing for very easy-to-understand evaluation of relevant systems. We also note that this formula is flexible, allowing for easy re-weighting of any and all components to adjust the scores to better match user goals and needs, addressing design criterion **C5**.

4.1 Evaluation Formula

First we present the full formula and provide a high level breakdown of its constituent expressions:

$$S = w_{c1} \cdot \frac{1}{n}\Big[\sum_{i=1}^{n_q}\sum_{q=1}^{11} w_q p_q + \sum_{j=1}^{n_o}\sum_{o=1}^{11} w_o p_o + \sum_{k=1}^{n_c}\sum_{c=1}^{11} w_c p_c\Big] + \quad (1)$$

$$w_{c2} \cdot [w_1 p_1 + w_2 p_2] + \quad (2)$$

$$w_{c3} \cdot p_3 + \quad (3)$$

$$w_{c4} \cdot p_4 \quad (4)$$

Parts (1) and (2) evaluate the graphical design of the visualization, whereas parts (3) and (4) evaluate the utility of the visualization as a means by which to convey some message about the data. The suggested (relative) weighting of each criterion is discussed in §4.2. Part (1) pertains to effectiveness of encoding channels (design criterion **C1**), and is discussed in §4.3. Part (2) pertains to adherence to graphical excellence and integrity principles (design criterion **C2**), discussed in §4.4. Part (3) pertains to the interestingness of the visualization (design criterion **C3**), discussed in §4.5. Part (4) pertains to the usefulness of the visualization (design criterion **C4**), discussed in §4.6.

4.2 Relative Weighting of Criteria

As mentioned in §3, there are four main evaluation criteria. However, not all criteria should necessarily be weighted the same in each evaluation scenario. In our example with Carol, the focus is on producing recommendations that will be useful for generating reports. In this case, Carol will likely care more about well-chosen encodings (**C1**), graphical excellence (**C2**), and the usefulness of the resulting visualizations (**C4**), and care less about interestingness (**C3**). As such, Carol will want the weights of the framework to reflect these goals. This necessitates the ability to assign weights to each of the criteria in a context-specific manner. Here, we present three different weighting schemes, or *presets*, implemented in our framework, described below (see Table 1).

Preset	w_{c1}	w_{c2}	w_{c3}	w_{c4}
Basic (default)	25	25	25	25
Encodings	70	10	10	10
Interestingness	10	10	70	10

Table 1. Suggested weighting schemes, or *presets*.

Basic Preset: The simplest configuration is to assign weights uniformly. We use this configuration as a default for the framework.

Encodings Preset: In this preset, we consider a scenario where the scoring is focused primarily on encoding choices, an important factor in other systems [23, 40]. This configuration gives encodings seven times the weight of any other criteria ($w_{c1} = 70$).

Interestingness Preset: This preset is meant to reflect our original inspiration from the SeeDB system [36, 37], where interestingness is the most important factor and thus weighted most heavily ($w_{c3} = 70$).

We emphasize that these presets should only be considered helpful suggestions for using our framework, to be modulated as one's goals and requirements are further refined. We stress that all framework weights are *fully customizable*: the weights are easily accessed and modified through JSON-formatted configuration files in our code.

4.3 Encoding Channels (C1)

The quality of a visualization can be greatly affected by the encoding channels used to render each attribute of the data. Inappropriate encodings may drastically reduce the ability of a visualization to convey meaningful information about the data it represents. We ground our ranking and relevant weighting of encoding channels in the well-established principles of Bertin and Mackinlay, introduced in [7] and [23], respectively.

Because the number of data attributes represented in a visualization may vary, we opted to compute the average score over all encodings present in the visualization. We begin by adding the scores for all data attributes being encoded. The expression

$$\sum_{i=1}^{n_q}\sum_{q=1}^{11} w_q p_q + \sum_{j=1}^{n_o}\sum_{o=1}^{11} w_o p_o + \sum_{k=1}^{n_c}\sum_{c=1}^{11} w_c p_c$$

in the final formula accomplishes this by splitting the encodings into three different sums representing well-recognized data types: quantitative data, ordinal data, and nominal (or categorical) data, respectively. The n_q, n_o, and n_c terms represent the number of encodings present for each respective data type. Then, for each data attribute encoded by the given data category, we sum over all 11 encoding channels considered by our framework[1]. The p_q, p_o, and p_c terms are binary parameters that take a value of one if the given encoding channel is used for that data attribute, and zero otherwise. The w_q, w_o, and w_c terms are the weights assigned to these encoding channels, which are discussed in §4.3.1, §4.3.2, and §4.3.3.

After summing all the possible encodings, we finally divide by n, the total number of attributes represented, to compute an average ranging from 0 to 1 (note that $n = n_q + n_o + n_c$). This value is then multiplied by w_{c1} to obtain the final encoding channels score.

We incorporated the graphical design principles of Bertin and Mackinlay in our assignment of suggested initial weights and scoring functions for encoding channels. The weights are currently linearly distributed within the range from 0 to 1, and are based primarily on existing rankings [7, 23]. Factors we took into consideration when devising scoring functions included but were not limited to:

- Does the effectiveness of the channel decrease with an increase in the number of unique values (or categories) in the data?
- Does the channel provide too little or too much granularity for the data type in question?

All components of the encodings score can be configured in our framework code by adjusting the associated JSON-formatted configuration file and other framework parameters (e.g., image width).

4.3.1 Quantitative Data

For quantitative data, we assign constant weights to each encoding channel (see Table 2), which we base on existing perceptual rankings. For example, position, being the best channel to encode quantitative data, earns the maximum weight of 1, while texture and shape, being incongruent (and arguably ineffective) channels to encode quantitative data, earn scores 0.1 and 0.0, respectively.

Quantitative Encoding Channel	**Weight** (w_q)
Position	1.0
Length	0.9
Angle	0.8
Slope	0.7
Area	0.6
Volume	0.5
Density	0.4
Color Saturation	0.3
Color Hue	0.2
Texture	0.1
Shape	0.0

Table 2. Encoding channel rankings, and corresponding weights for quantitative data.

4.3.2 Ordinal Data

For ordinal data, we also assign weights to each encoding channel based on prior research as to which channels are most effective for ordinal data (see Table 3).

We recognize that many channels become less effective as the cardinality of the given data attribute (n_{cat}), or the number of unique values within this attribute, increases. Note that we assume that n_{cat} takes a

[1] Note that although they appear in Mackinlay's original paper, we omit the "connection" and "containment" channels, as our initial focus is on evaluating visualizations of relational rather than graph data.

Ordinal Encoding Channel	Weight (w_o)
Position	1.0*min[1, 1/log$_p$(n_{cat})]
Density	0.9*min[1, 1/log$_5$(n_{cat})]
Color Saturation	0.8*min[1, 1/log$_5$(n_{cat})]
Color Hue	0.7*min[1, 1/log$_5$(n_{cat})]
Texture	0.6*min[1, 1/log$_5$(n_{cat})]
Shape	0.5*min[1, 1/log$_5$(n_{cat})]
Length	0.4*min[1, 1/log$_p$(n_{cat})]
Angle	0.3*min[1, 1/log$_5$(n_{cat})]
Slope	0.2*min[1, 1/log$_5$(n_{cat})]
Area	0.1*min[1, 1/log$_p$(n_{cat})]
Volume	0.0*min[1, 1/log$_p$(n_{cat})]

Table 3. Encoding channel rankings and corresponding weights for ordinal data, where n_{cat} is the cardinality of the data attribute (i.e., total unique values observed), and p is pixel width/height of the visualization.

default value of 5, informed by prior visual perception work (e.g., for color encodings [13]). To account for this decay, we use an inverse log function to modulate the weights, based on Stevens' power law [34]. The base of the logarithm denotes how many categories we believe to be reasonable to distinguish between for that particular channel before comprehensibility is compromised. For example, because we believe it is reasonable for humans to distinguish five different lengths representing ordinal data at a glance, the base weight for shape (0.5) is multiplied by $\min[1, 1/\log_5(n_{cat})]$, meaning a penalty is incurred once the number of categories in the data being encoded surpasses five.

According to prior work, position is the best channel for encoding ordinal data (receiving a base weight of 1.0). However there is a limitation on the effectiveness of position based on the pixel width of the visualization in question, hence the base of the logarithm is p, i.e. pixel width. Since the weights are modulated by inverse log functions, they will approach zero as n_{cat} increases. Note that all bases and weights can easily be configured through JSON-formatted configuration files for our framework code.

4.3.3 Nominal Data

We use the same inverse log function scheme for nominal (or categorical) encoding channels as for ordinal encoding channels, since the effectiveness of the channels can decrease when too many categories are represented in the data (n_{cat}).

The full assignment of weights can be seen in Table 4. Position is the best channel for encoding for nominal data (receiving a base weight of 1), and is modulated using the same scheme as seen with ordinal data. Color hue is close behind, with a base weight of 0.9.

Nominal Encoding Channel	Weight (w_c)
Position	1.0*min[1, 1/log$_p$(n_{cat})]
Color Hue	0.9*min[1, 1/log$_5$(n_{cat})]
Texture	0.8*min[1, 1/log$_5$(n_{cat})]
Density	0.7*min[1, 1/log$_5$(n_{cat})]
Color Saturation	0.6*min[1, 1/log$_5$(n_{cat})]
Shape	0.5*min[1, 1/log$_5$(n_{cat})]
Length	0.4*min[1, 1/log$_p$(n_{cat})]
Angle	0.3*min[1, 1/log$_5$(n_{cat})]
Slope	0.2*min[1, 1/log$_5$(n_{cat})]
Area	0.1*min[1, 1/log$_p$(n_{cat})]
Volume	0.0*min[1, 1/log$_p$(n_{cat})]

Table 4. Encoding channel rankings and corresponding weights for nominal data, where n_{cat} is the cardinality of the data attribute, and p is pixel width/height of the visualization.

4.4 Graphical Excellence and Integrity Principles (C2)

Recall the main principles for graphical excellence and graphical integrity in visualization design, described in §2.

We currently evaluate two principles in our framework ("clear and detailed labels" and "avoiding chart junk") which complement our other framework components rather than duplicate them. For example, "showing data variation and not design variation" is largely dependant on the utilization of proper encoding channels and good design principles, both of which are already addressed with our existing criteria.

Of the principles included in our framework, we believe clear and detailed labels to be the more critical of the two. The reason for this is that clear and detailed labels are a necessary condition for effectively interpreting the data being presented. If labels are ambiguous or nonexistent, it becomes a challenge for users to glean anything from a visualization, making it a bad recommendation.

Tufte points out that avoiding chart junk is important for making interpretation of visualizations a fast and efficient process, but chart junk generally does not completely impede one's ability to interpret a visualization. However in contrast, poor or missing labels *can* prevent users from interpreting a visualization. For this reason, we suggest that w_1, the weight for clear and detailed labels, should be set to 0.7, and w_2, the weight for avoiding chart junk, to 0.3. Of course, there may be situations where one may need prioritize reduction of chart junk (or other criteria). Our framework is designed to allow for easy re-weighting of any component to accommodate these situations (e.g., by updating the corresponding configuration file in our code).

Evaluating Clear and Detailed Labels: When evaluating the effectiveness of labels in a visualization, we consider five possible categories of labels: title, subtitle, x-axis label, y-axis label, or data (per-mark) labels. We consider each of them separately, because it is possible to have effective labels in some categories while having ineffective labels in others, necessitating some degree of granularity.

We score labels by asking users (e.g. analysts, data visualization experts) to rate whether each of the five label types are clear and detailed using a three-point Likert scale: yes (assigned a score of 0.2), somewhat (0.1), or no (0.0). The median is taken for each label category across all users, and these five medians are then summed to yield the value for p_1 (with a maximum of 1). Note that if a label is correctly absent (such as axes labels on a pie chart), users are to select "yes" = 0.2. Then p_1 is multiplied by the corresponding weight w_1 to return the final score for clear and detailed labels.

Evaluating Chart Junk: To evaluate chart junk, we ask users (e.g. analysts) to identify the degree of chart junk present in a visualization, also on a three-point Likert scale: excessively (assigned a score of 0.0), somewhat (0.5), and none (1.0). The median across all users is taken, and this value is assigned to parameter p_2. Then p_2 is multiplied by the weight w_2, yielding the final score for avoiding chart junk.

Combining Labels and Chart Junk Scores: Once we have a final weighted score for evaluating labels and a final weighted score for evaluating chart junk, we add these two scores together. This sum returns a value between 0 and 1, depending on how well the visualization adheres to the aforementioned principles. Lastly, we multiply this value by the overall criterion weight of w_{c2}, discussed in §4.2, to calculate the overall graphical excellence/integrity score of the visualization.

4.5 Interestingness (C3)

The interestingness criterion was inspired by SeeDB's deviation-based interestingness metric [37], however their exact approach raises some concerns voiced in recent work [43]. In particular, such an approach may encourage users to p-hack. Moreover, large deviation is not necessarily what the user is interested in, so the use of such a metric precludes the recommendation of visualizations that may show little deviation, but still communicate something novel about the data.

Instead, we define a visualization as interesting if the attributes (and encodings) are non-obvious. Most data sets will have a set of "default" visualizations that users will immediately think of constructing on their own, but encouraging users to explore non-obvious visualizations may ultimately lead to a deeper understanding of the data and more nuanced hypotheses or findings. This approach is also considered in other recommendation systems, such as Voyager, which encourages users to interact with new, unexplored attributes [40].

We score our interestingness criterion by asking users (e.g., analysts, visualization experts) whether they would have come up with a particular visualization for a given data set on their own. If most users would create the visualization unprompted, this suggests that it is an

obvious and therefore uninteresting visualization; if most users would not consider the visualization until it is recommended to them, then it may be an interesting one to recommend. In particular, a user rates how likely they would be to come up with a given visualization using a four-point Likert scale: extremely unlikely (clearly novel), somewhat unlikely, somewhat likely, extremely likely (clearly not novel). We propose two methods for converting this feedback into a usable score.

Method 1: Median In this method, we assign numerical scores to the Likert ratings as follows: extremely unlikely = 1.0, somewhat unlikely = 0.75, somewhat likely = 0.25, and extremely likely = 0. We then calculate p_3 as the median score across all users (see §4.1).

Method 2: Proportion In this method, we collapse each Likert scale rating into a binary score: "extremely unlikely" and "somewhat unlikely" map to "unlikely", and "extremely likely" and "somewhat likely" map to "likely". To calculate p_3, we subtract from 1 the proportion of users responding with "likely." Thus, if no one would have thought of the given visualization unprompted, $p_3 = 1$, and if everyone would have thought of the visualization, $p_3 = 0$ (see §4.1).

Method choice is flexible, and can be adjusted depending on user preference. Either way, to obtain the final interestingness score, the value of p_3 is multiplied by the weight of w_{c3}, as discussed in §4.2.

4.6 Usefulness (C4)

Our usefulness criterion was designed to complement our interestingness criterion, because the interestingness criterion is unable to capture whether the recommended visualization is actually useful to the user. For example, it may be the case that a visualization is non-obvious because it is not very helpful to begin with. This component of the evaluation framework was implemented by assigning a score based on how many users (e.g., analysts) either bookmark the visualization for later use, or rate it pertinent to the research question(s) they are trying to answer. In particular, a user rates how likely they would be to save a visualization for future reuse on the following four-point Likert scale: extremely likely (clearly useful), somewhat likely, somewhat unlikely, extremely unlikely (clearly not useful). We propose similar methods for scoring feedback as our interestingness criterion.

Method 1: Median We assign numerical values between 0 and 1 for each Likert rating: extremely likely = 1.0, somewhat likely = 0.75, somewhat unlikely = 0.25, and extremely unlikely = 0. We then calculate p_4 as the median score across all users (see §4.1).

Method 2: Proportion here, we also collapse the Likert scale to a binary ranking. We then calculate p_4 by subtracting from 1 the proportion of users responding with "unlikely" (see §4.1).

For consistency, we suggest using the same scoring method for both interestingness and usefulness. To obtain the final usefulness score, the value of p_4 is multiplied by the weight w_{c4} (see §4.2).

4.7 Assessing Interestingness and Usefulness

To test our interestingness and usefulness criteria, we formulated several example research questions, and utilized two evaluation strategies: 1) polling the current research team on interestingness and usefulness, discussed in §5; and 2) seeking feedback from outside experts, discussed in §6. In the future, the usefulness criterion could be tracked automatically in systems that have a bookmarking feature, or experts could be polled in a larger user study on a wider range of visualization designs. In this way, collaborative filtering (CF) techniques could leveraged to aid in the visualization recommendation process.

5 Evaluation 1: Initial Case Studies

To test our framework, we selected five visualizations to evaluate using our metric. To gain an understanding how the formula works in practice, We provide a step-by-step evaluation of two of these examples. Note that we use the basic preset (weighting scheme) described in §4.2.

5.1 Visualization Example 1

Consider the scenario of a company that sells fruit juices, and we are interested in identifying the juices that brought in the most revenue. Suppose that some recommendation system suggested the visualization

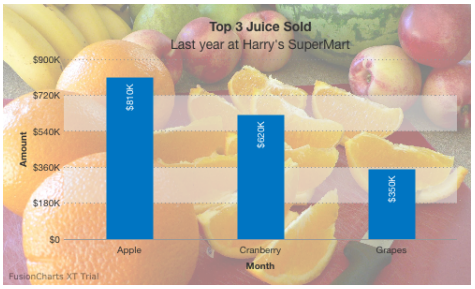

Fig. 1. A sample visualization from an existing visualization corpus [4, 5].

presented in Fig. 1 for this analysis. Together, three of the authors evaluated this visualization using each of the four criteria.

Encodings Score: This chart encodes two data attributes: one quantitative encoded by length (revenue), and one nominal encoded by position (type of juice), resulting in a final encoding channels score of:

$$\frac{1}{2}\Big[\sum_{i=1}^{1}\sum_{q=1}^{11} w_q p_q + \sum_{j=1}^{0}\sum_{o=1}^{11} w_o p_o + \sum_{k=1}^{1}\sum_{c=1}^{11} w_c p_c\Big]$$
$$= \Big[\frac{0.9+1.0}{2}\Big] = \mathbf{0.95}$$

Graphical Excellence and Integrity Score: Three authors evaluated this visualization on graphical excellence/integrity. All three agreed that both the title and data labels were clear and detailed. Thus these two categories receive the maximum of 0.2. All three authors found the subtitle to be somewhat vague ("Last year" is ambiguous), along with the y-axis label ("Amount" is ambiguous), yielding a score of 0.1 for both categories. All three authors rated the x-axis label as not being clear and detailed, assigning it the minimum score of 0. Summed together, this yields: $p_1 = 0.2+0.1+0+0.1+0.2 = \mathbf{0.6}$.

All three authors concurred that this visualization includes chart junk. Though distracting, it does not prevent the user from understanding the message the visualization aims to convey, resulting in a unanimous score of $p_2 = \mathbf{0.5}$. This visualization receives an overall graphical excellence/integrity score of: $0.7 \cdot 0.6 + 0.3 \cdot 0.5 = \mathbf{0.57}$

Interestingness and Usefulness Scores (Using Method 1): One author thought they were extremely unlikely to have come up with this visualization on their own (they found it interesting), while the other two thought they were somewhat likely to have done so (they found it somewhat uninteresting). These results dictate a value of $p_3 = median[0.25, 0.25, 1.0] = \mathbf{0.25}$

When considering usefulness, one author said they were somewhat likely to save this visualization for future use (they found it somewhat useful), while the other two said there were extremely unlikely to do so (they found it unuseful). This yields a value of $p_4 = median[0.0, 0.0, 0.75] = \mathbf{0.0}$.

Interestingness and Usefulness Scores (Using Method 2): Applying our proportion-based method to the above scores leads to an interestingness score of $p_3 = 1 - \frac{2}{3} = \mathbf{0.33}$, and a usefulness score of $p_4 = 1 - \frac{2}{3} = \mathbf{0.33}$.

Final Score: If this bar chart were suggested by a visualization recommendation system, it would be awarded a final score of:

$$25 \cdot 0.95 + 25 \cdot 0.57 + 25 \cdot 0.25 + 25 \cdot 0.0$$
$$= 23.75 + 14.25 + 6.25 + 0.0 = \mathbf{44.25}.$$

if using Method 1 for **C3** and **C4**, or:

$$25 \cdot 0.95 + 25 \cdot 0.57 + 25 \cdot 0.33 + 25 \cdot 0.33$$
$$= 23.75 + 14.25 + 8.33 + 8.33 = \mathbf{54.67}.$$

if using Method 2. Both scores line up with our intuition for the overall value of this visualization as a potential recommendation. While the data is presented faithfully, the confusion caused by the abuse of Graphical Excellence/Integrity principles lowers the final score significantly.

The final scores above reflect the use of the basic (default) preset described in §4.2. These scores were also recalculated using the other presets (encodings and interestingness, see Table 5). For completenesss, both methods for scoring interestingness and usefulness (median and proportion) are addressed. The encodings preset yielded the highest scores by a large margin, which is expected given the optimal use of encoding channels to represent the data. The interestingness preset yielded the lowest score. With the exception of the encodings preset, the scores generated using methods 1 and 2 were about 10 points apart.

Preset	Method 1 Score	Method 2 Score
Basic	44.25	54.67
Encodings	74.7	78.87
Interestingness	32.7	41.87

Table 5. The distribution of scores for Fig. 1 using the presets in §4.2.

5.2 Visualization Example 2

We now repeat our scoring process for visualizing the popularity of baby names starting with"Ki" over time. Suppose a recommendation system recommends the visualization in Fig. 2 for this analysis. All four authors evaluated this visualization.

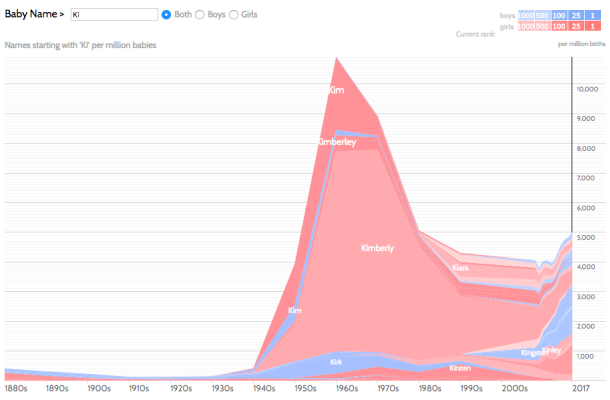

Fig. 2. A visualization showing the popularity of baby names starting with "Ki" over time, taken from [2].

Encodings Score: This visualization encodes four data attributes: one is quantitative data encoded by position (birth year), one is quantitative data encoded by area (number of babies), one is quantitative data encoded by color saturation (popularity of the name), and one is nominal data encoded by two distinct color hues (gender associated with name). Using our formula, we arrive at a final encodings score of:

$$\frac{1}{4}\left[\sum_{i=1}^{3}\sum_{q=1}^{11} w_q p_q + \sum_{j=1}^{0}\sum_{o=1}^{11} w_o p_o + \sum_{k=1}^{1}\sum_{c=1}^{11} w_c p_c\right]$$
$$= \left[\frac{1.0+0.6+0.3+0.9}{4}\right] = \mathbf{0.70}$$

Graphical Excellence and Integrity Score: All authors agreed that both the title and y-axis labels were clear and detailed (both receive the maximum of 0.2). There is no subtitle in the visualization, but no subtitle is needed (also receiving 0.2). One author found the x-axis to be unclear (0.0), two found it somewhat clear (0.1), and one found it to be clear and detailed (0.2), producing a median of 0.1. Two authors found the data labels to be somewhat clear (0.1); the other two found them to be clear and detailed (0.2), yielding a median value of 0.15. The final result is:

$$p_1 = 0.2+0.2+0.1+0.2+0.15 = \mathbf{0.85}$$

All authors agreed this visualization includes no chart junk, resulting in a unanimous score of $p_2 = \mathbf{1.0}$, and a final score of:

$$0.7 \cdot 0.85 + 0.3 \cdot 1.0 = \mathbf{0.895}$$

Interestingness and Usefulness Scores (Method 1): One author thought they were extremely likely to have come up with the visualization on their own (i.e., found it uninteresting), while the others would have been somewhat unlikely to create this visualization on their own (i.e., found it somewhat interesting). These results dictate a value of $p_3 = median[0.0, 0.75, 0.75, 0.75] = \mathbf{0.75}$.

When evaluating usefulness, three authors said they were extremely likely to save the visualization for later use (i.e., found it useful), while the other said they were somewhat unlikely to save it for later use (i.e., somewhat unuseful). This produces a value of $p_4 = median[0.25, 1.0, 1.0, 1.0] = \mathbf{1.0}$

Interestingness and Usefulness Scores (Method 2): Proportion scoring produces an interestingness score of $p_3 = 1 - \frac{1}{4} = \mathbf{0.75}$, and a usefulness score of $p_4 = 1 - \frac{1}{4} = \mathbf{0.75}$.

Final Score: If this visualization were suggested by a recommendation system, it would be awarded a final score of:

$$25 \cdot 0.70 + 25 \cdot 0.895 + 25 \cdot 0.75 + 25 \cdot 1.0$$
$$= 17.5 + 22.38 + 18.75 + 25.0 = \mathbf{83.63}.$$

if using Method 1 for **C3** and **C4**, or:

$$25 \cdot 0.70 + 25 \cdot 0.895 + 25 \cdot 0.75 + 25 \cdot 0.75$$
$$= 17.5 + 22.38 + 18.75 + 18.75 = \mathbf{77.38}.$$

if using Method 2. Both scores support our intuition for the value of this visualization as a recommendation, especially relative to the case study in §5.1. While it is not perfect (e.g., data labels were hard to read, etc.), this visualization is overall an effective way to communicate the data to the user. This is readily apparent from the scores for **C2**, **C3**, and **C4**. Final scores generated using all presets are described in Table 6. We see that the encodings preset resulted in the lowest scores, but also note that the distribution of scores across presets is narrower than seen in §5.1. This is due to the fact that all four criteria scores were somewhat evenly distributed, unlike in the previous case study: encodings were well executed, but the visualization suffered greatly in **C2**, **C3**, and **C4**. These discrepancies are amplified by the weights assigned to the criteria in different presets.

Preset	Method 1 Score	Method 2 Score
Basic	83.63	77.38
Encodings	75.45	72.95
Interestingness	78.45	75.95

Table 6. The distribution of scores for Fig. 2 using the presets in §4.2.

5.3 Results

These examples demonstrate that our formula can score recommendations in a way that is not only intuitive to potential users, but also robust to a variety of visualization schemes and user analysis goals. One lesson learned from this exercise, is that the interpretation of user feedback on interestingness and usefulness can result in markedly different scores. In both cases, we see clear differences in the results of the two scoring methods (see Tables 5 and 6), though the method that produced the higher score varied by case. These results may suggest that calculating multiple scores from the same user feedback could be beneficial, providing different perspectives on the design of a recommended visualization. However, it is unclear from these case studies whether the two methods may be suited to specific contexts. We investigate this question further in our user study, described in §6.

6 Evaluation 2: Surveying Experts

To gain a better understanding of the behavior of our framework, we conducted a user study to gather expert feedback on criteria **C2 - C4** (see §3). This was accomplished by administering an online survey asking visualization experts to evaluate five different visualizations: our two initial case studies, two well-known online visualization designs, and one classic and well-known visualization design. [2]

[2]This study was approved by our institution's Institutional Review Board.

	Method 1 (Median Scoring)					Method 2 (Percentage Scoring)				
	C1	**C2**	**C3**	**C4**	**Final**	**C1**	**C2**	**C3**	**C4**	**Final**
Pie Chart [4,5]	18.75	23.25	**6.25**	18.75	67	18.75	23.25	**8.33**	15	65.33
Baby Names [2]	17.5	19.5	**6.25**	18.75	62	17.5	19.5	**10**	13.33	60.33
Fruit Juice [4,5]	23.75	8.75	18.75	**6.25**	57.5	23.75	8.75	15	**5**	52.5
Cholera Map [33]	24.17	**16.25**	18.75	18.75	77.92	24.17	16.25	**15**	**15**	70.42
Facebook IPO [1]	17.5	17.75	18.75	**6.25**	60.25	17.5	17.75	13.33	**10**	58.58
Baby Names (Case Study) [2]	**17.5**	22.38	18.75	25	83.63	**17.5**	22.38	18.75	18.75	77.38
Fruit Juice (Case Study) [4,5]	23.75	14.25	6.25	**0.0**	44.25	23.75	14.25	**8.33**	**8.33**	54.67

Table 7. Framework scores using the Basic preset for every visualization evaluated in our case studies (§5) and user study (§6). A copy of each image is provided in the supplemental materials. The weakest category for each visualization is highlighted in bold. We see that interestingness (**C3**) and usefulness (**C4**) generally have the weakest scores. We also find that in most cases, proportion scoring (Method 2) leads to lower scores.

6.1 User Study

6.1.1 Recruitment

We sought study participants with prior experience in visualization and/or data analytics. To this end, subjects were recruited by emailing the researchers' contacts in the visualization and analytics communities and public mailing lists. Potential respondents were first directed to an initial screening survey that assessed their expertise in the fields of data visualization and analytics. Based on their responses, participants were contacted at the email address they provided at the end of the screening to complete the actual study survey. Fifteen people participated in the study survey; participants' expertise ranged from visualization faculty and graduate students to industry analysts and practitioners.

6.1.2 Survey Design

The survey was implemented in Qualtrix, and designed to gather feedback on three criteria: **C2, C3,** and **C4**. To assess **C2**, we first asked participants whether they considered each of the five label categories (title, subtitle, x-axis, y-axis, and data labels) to be clear and detailed using the three-point Likert scale discussed in §4.4. These Likert ratings were processed as described in §4.4 to obtain the value of p_1. To complete our assessment of **C2**, we then asked participants to determine the degree of chart junk present in a visualization, using the three-point Likert scale also discussed in §4.4. These Likert ratings were processed as described in §4.4 to obtain the value of p_2.

To assess **C3**, we asked respondents how likely they would be to have come up with a visualization on their own, using the four-point Likert scale described in §4.5. To assess **C4**, we asked respondents how likely they would be to save a visualization for future reuse, using the four-point Likert scale described in §4.6. For **C3** and **C4**, both proposed methods (median and proportion) were used to obtain their respective values of p_3 and p_4 (as described in §4.5 and §4.6, respectively).

The evaluations described above were completed for each of five data visualizations by a total of $n = 15$ respondents.

6.2 Results

In Table 7, we present the criteria and final scores for each of the five visualizations, calculated using the same process described in §5.1 and §5.2. Due to space limitations, we only present results using the basic (default) preset. Note that all scores for all presets are calculated in our framework code examples included in the supplemental materials.

6.2.1 Differences in Case Study and User Study Results

Figure 3 shows the spread of ratings provided by participants for the Baby Names visualization. The results for all visualizations are provided in the supplemental materials. For many evaluation criteria, we see relatively strong agreement across participants, represented as a vast majority of participants providing the same rating (e.g., all participants rating chart junk as "Somewhat" for Baby Names).

For interestingness (**C3**) and usefulness (**C4**), the inclusion of more users in completing these evaluations reduces the gaps between final scores. For example, in the case study of the Baby Bames visualization in §5.2, the difference between Methods 1 and 2 was over 6 points, whereas in the user study this difference dropped to less than 2 points. In the user study results, we also find that Method 1 generally produces higher scores, which differs from our case study results. These results

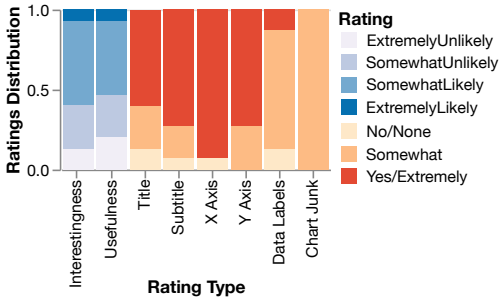

Fig. 3. Spread of responses ($n = 15$) from our study for the Baby Names visualization (see Fig. 2).

suggest that collecting feedback from a large number of experts can help to reduce noise in our scoring techniques.

However, we also notice a significant difference in the case study scores and the study scores, where scores appear to move towards an average of 65. In fact, averaging the five scores for each visualization from the study leads to the score of 64.93 (for Method 1), and 4 out of 5 visualizations score in the 60's or lower. Though only five visualizations were evaluated, these results fall into a consistent scoring range of about 55 to 65 points, using the basic preset. These results may suggest that routine visualizations tend to fall within this scoring range, however more data is required to confirm this hypothesis.

6.2.2 The Influence of Multiple Criteria on Scoring

Our results emphasize the importance of considering multiple factors when evaluating visualizations for recommendation. Well-chosen encodings (**C1**) alone do not make for an overall effective recommendation, as seen with the Fruit Juice visualization, which achieves the second best encoding score of all the visualizations. Similarly, strong adherence to Graphical Excellence and Integrity (**C2**) also fails to signal a strong visualization recommendation on its own, as demonstrated by the Pie Chart visualization. Again, neither interestingness nor usefulness alone signal a strong recommmendation, as seen in most visualizations in Table 7. These results suggest that prior evaluations clearly have blind spots, due to their emphasis on only one or two of these criteria, such as SeeDB's primary focus on interestingness [37]. As such, the corresponding presets should be used with care.

Our scores become particularly interesting when considering the Cholera Map visualization. This visualization is the only one to achieve both relatively high interestingness *and* usefulness scores, and relatively strong scores all around. These findings seem to suggest that our framework could provide a strong signal for good overall recommendations, where we define good as: visualizations that are not only well-designed (**C1** and **C2**), but also show the user a design that they may not have tried on their own (**C3**), and furthermore show something the user actually wants to investigate further (**C4**).

6.2.3 Interestingness, Usefulness, and Graphical Excellence

In Table 7, we highlight the lowest scoring criterion for each visualization, both for our case study and user study results. In general, we find that the weakest scores tend to occur for interestingness (**C3**) and usefulness (**C4**). In fact, out of all five visualizations evaluated, none of them

have their weakest score occur in the encodings criterion (**C1**). We also find that people seem to rarely find these visualizations both interesting *and* useful. These results apply to both "bad" visualizations like the juice visualization and more well-known visualizations, such as the Baby Names visualization (originally presented by Wattenberg [39]).

These results appear to be consistent with how many visualizations (and thus recommendation systems) are optimized for effectiveness: they tend to focus more on well-established but low-level metrics for graphic design (e.g., [23–25], and framework criterion **C1**), and less on higher-level evaluation criteria relevant to data analysis tasks, such as interestingness [37], usefulness and Graphical Excellence. Our results suggest that exploring how to measure and ultimately support these higher-level criteria could lead to more effective recommendations.

Our study can provide insight in this direction. For example, many participants based their usefulness ratings on how quickly and easily they could interpret the data from the visualization, suggesting a relationship between usefulness (**C3**) and Graphical Excellence (**C2**). In the Facebook IPO example, many participants voiced concern over not knowing what color and circle size encoded, making them question their understanding of the data, and generally more hesitant assign higher usefulness scores. One participant summarized the issue well: "I don't understand what the size and colour are supposed to encode... That confusion makes me unsure if I would pursue this design." Another commented that "it is pretty, but unclear." Thus, even when good encoding choices are made (e.g., using position to encode year and public offering, using redundancy to highlight relationships), if these encodings are not described, e.g. through clear labels and legends, users are less likely to find these visualizations useful. These results support existing evaluation heuristics that emphasize accurate interpretation of the data presented [38]. Similarly, when evaluating whether labels within a visualization were "clear and detailed," participants emphasized whether the labels: were easy to find, were clear, and accurately described the data. These results may suggest that clarity and accuracy of the titles, labels and annotations in conveying visualization design decisions are core metrics in evaluating usefulness.

7 Discussion

Based on our results, we offer recommendations for the creation and evaluation of visualization recommendation systems moving forward, discuss current limitations, and highlight possible areas of future work.

7.1 Diversity in Evaluation Criteria

Though many recommendation criteria exist, visualization recommendation systems are rarely evaluated on more than one criterion. Moreover during testing, researchers often emphasize the criteria they optimize for, potentially resulting in biased assessments. Furthermore, recommendation systems are evaluated using many different methods and metrics, making it difficult to directly and systematically compare different techniques. Based on the work presented here, we recommend that evaluations of visualization recommendations comprise a diverse but consistent set of criteria, enabling a fairer comparison of different techniques. Our framework provides a starting point for integrating multiple evaluation criteria into a single, easy-to-interpret score.

Insight 1*: Recommendation systems should be evaluated along multiple diverse criteria. Further, these criteria should be consistent.*

7.2 Appropriate Weighting

While selecting a diverse subset of criteria is critical, considering the relative weight of each criterion can be equally important. We argue that not every criterion is created equal for the purposes of evaluating visualization recommendation systems, and some criteria should be weighted more heavily than others in certain scenarios. Inappropriate weighting could allow for poorly designed systems to receive much higher ratings than they practically should in a given context.

Based on our evaluation and research, we recommend that future evaluation metrics be designed such that the relative weights given to the various evaluation criteria are assigned appropriately for the given context or task. This will help ensure that a system is evaluated fairly on the criteria included in the framework. While assigning weights to these various criteria is arguably a subjective process at times, this procedure could be made more systematic by consulting analysis and visualization experts for input on the weighting distribution. Additionally, making more presets for the framework based on existing evaluations may further enable users of the framework to quickly evaluate recommendation systems for well-known and common analysis contexts.

Insight 2*: The relative weights assigned to evaluation criteria should match the given analysis context. Preset configurations for common analysis scenarios could reduce the complexity of evaluation.*

7.3 Limitations and Future Work

One limitation lies in the subjectivity of some of our criteria. First, the selection of criteria could itself be seen as a subjective process. This dilemma is unavoidable, but we sought to ameliorate the issue by selecting criteria well-known within the visualization community. We plan to incorporate more criteria in our framework, through review of more evaluation methods, as well as through feedback from the community. For example, we could take into account whether a system is potentially using *p*-hacking [43] when recommending visualizations and deduct from the total score accordingly.

Second, the weights assigned to various criteria could also be subjective. While these weights were chosen through a principled assessment of existing work, and we attempted to reduce subjectivity by providing three different presets, our weighting system could still be biased. This evaluation space would have to be explored further in order to test the rigor, robustness, and accuracy of our weighting system, perhaps through additional user studies, or polling experts for feedback on the scores produced by our framework. As such, we see our framework as a useful *starting point* for evaluating recommendation systems.

An important limitation to consider our focus on static visualizations. Interaction is considered a core aspect of information visualization [15, 28]. However, interactions make it extremely difficult to predict what a user ultimately will glean from a visualization. As such, we follow the work of others and use static visualizations as our starting point [24, 25]. In the future, we plan to consider a wider variety of visualization designs in our evaluation, including interactions [15].

A final limitation to consider is that we do not have a fully automated implementation: input must still be provided directly from users. As such, our framework is challenging to scale to large visualization collections. We plan to investigate opportunities for automation, such as applying machine learning techniques to construct proxies for some criteria, or using collaborative filtering techniques to infer interestingness and/or usefulness. We could also potentially alter the framework to take features of the data itself into consideration, such as dimensionality (and cardinality) of the data being rendered, and incorporate these parameters into our formula. Ideally, the end goal is to develop software to complete the entire evaluation without requiring manual input. In general, automation enables the testing of many more configurations and possibilities in terms of recommendation design, influencing all other aspects of our future work, which leads to our final insight:

Insight 3: *Automation could enable broader and more rigorous testing of evaluation techniques for visualization recommendation systems.*

8 Conclusion

We have developed an evaluation framework that can be used to score individual visualizations produced by recommendation systems. In the past, such systems were evaluated in isolation, lacking standardization and comparisons between systems. To ensure that it is broadly applicable to multiple recommendation tools, our framework incorporates scores for a range of criteria derived directly from the visualization literature, such as the graphical design principles utilized within a recommended visualization, and the utility of this visualization in recommendation contexts. As we further develop this framework, it can be used to advance the state of research in this area by providing researchers with a way to systematically evaluate the output of different visualization recommendation systems and thereby empirically compare different tools. While this remains a work in progress, our proposed framework provides a first step towards developing a systematic approach to evaluating visualization recommendation systems.

REFERENCES

[1] The facebook offering: How it compares. https://archive.nytimes.com/screenshots/www.nytimes.com/interactive/2012/05/17/business/dealbook/how-the-facebook-offering-compares.jpg, 2012 (accessed March 18, 2019).

[2] Baby name voyager: Names starting with 'ki' per million babies. http://www.babynamewizard.com/voyager#prefix=ki&sw=both&exact=false, 2019 (accessed March 18, 2019).

[3] L. Battle, R. Chang, and M. Stonebraker. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*, pages 1363–1375. ACM, 2016.

[4] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker. Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 594. ACM, 2018.

[5] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker. Beagle project: Info and dataset. http://www.cs.umd.edu/~leilani/beagle.html, 2018 (accessed March 18, 2019).

[6] A. Bellogin, P. Castells, and I. Cantador. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 333–336, New York, NY, USA, 2011. ACM.

[7] J. Bertin. Semiology of graphics, univ. of wisconsin, 1983. *Translation of Semiologie graphique*, 1967.

[8] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[9] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[10] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[11] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Recommending visual insights. *Proceedings of the VLDB Endowment*, 10(12):1937–1940, 2017.

[12] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.*, 10:2935–2962, Dec. 2009.

[13] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96*, pages 263–270. IEEE, 1996.

[14] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 203–212. ACM, 2010.

[15] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30, 2012.

[16] K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo. Vizml: A machine learning approach to visualization recommendation. *arXiv preprint arXiv:1808.04819*, 2018.

[17] N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, Mar. 2011.

[18] P. Jayachandran, K. Tunga, N. Kamat, and A. Nandi. Combining user interaction, speculative query execution and sampling in the dice system. *Proceedings of the VLDB Endowment*, 7(13):1697–1700, 2014.

[19] L. Jiang, P. Rahman, and A. Nandi. Evaluating interactive data systems: Workloads, metrics, and guidelines. In *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data*, pages 1637–1644. ACM, 2018.

[20] A. Key, B. Howe, D. Perry, and C. Aragon. Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 681–684. ACM, 2012.

[21] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, volume 37, pages 157–167. Wiley Online Library, 2018.

[22] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2012.

[23] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[24] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics*, 13(6):1137–1144, 2007.

[25] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2019.

[26] T. Munzner. *Visualization analysis and design*. AK Peters/CRC Press, 2014.

[27] A. Perer and B. Shneiderman. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 109–118. ACM, 2008.

[28] W. A. Pike, J. Stasko, R. Chang, and T. A. O'connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009.

[29] B. Saket, A. Endert, and C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE transactions on visualization and computer graphics*, 2018.

[30] B. Saket, D. Moritz, H. Lin, V. Dibia, C. Demiralp, and J. Heer. Beyond heuristics: Learning visualization design. *arXiv preprint arXiv:1807.06641*, 2018.

[31] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics*, 23(1):341–350, 2017.

[32] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[33] J. Snow. Broad street cholera map. https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Snow-cholera-map-1.jpg, 2019 (accessed March 18, 2019).

[34] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957.

[35] E. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 2001.

[36] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran. Towards visualization recommendation systems. *ACM SIGMOD Record*, 45(4):34–39, 2017.

[37] M. Vartak, S. Rahman, S. Madden, A. Parameswaran, and N. Polyzotis. See db: efficient data-driven visualization recommendations to support visual analytics. *Proceedings of the VLDB Endowment*, 8(13):2182–2193, 2015.

[38] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):491–500, 2019.

[39] M. Wattenberg. Baby names, visualization, and social data analysis. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 1–7. IEEE, 2005.

[40] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization & Computer Graphics*, (1):1–1, 2016.

[41] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[42] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[43] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 479. ACM, 2018.