# Learning to Automate Chart Layout Configurations Using Crowdsourced Paired Comparison

Aoyu Wu
Hong Kong University of Science and Technology
awuac@connect.ust.hk

Liwenhan Xie
Hong Kong University of Science and Technology
liwenhan.xie@connect.ust.hk

Bongshin Lee
Microsoft Research
bongshin@microsoft.com

Yun Wang
Microsoft Research Asia
wangyun@microsoft.com

Weiwei Cui
Microsoft Research Asia
Weiwei.Cui@microsoft.com

Huamin Qu
Hong Kong University of Science and Technology
huamin@cse.ust.hk

## ABSTRACT

We contribute a method to automate parameter configurations for chart layouts by learning from human preferences. Existing charting tools usually determine the layout parameters using predefined heuristics, producing sub-optimal layouts. People can repeatedly adjust multiple parameters (*e.g.*, chart size, gap) to achieve visually appealing layouts. However, this trial-and-error process is unsystematic and time-consuming, without a guarantee of improvement. To address this issue, we develop Layout Quality Quantifier ($LQ^2$), a machine learning model that learns to score chart layouts from paired crowdsourcing data. Combined with optimization techniques, $LQ^2$ recommends layout parameters that improve the charts' layout quality. We apply $LQ^2$ on bar charts and conduct user studies to evaluate its effectiveness by examining the quality of layouts it produces. Results show that $LQ^2$ can generate more visually appealing layouts than both laypeople and baselines. This work demonstrates the feasibility and usages of quantifying human preferences and aesthetics for chart layouts.

## CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods**; Visualization toolkits.

## KEYWORDS

Machine Learning, Visualization, Crowdsourced, Visual Design, Image Quality Assessment

**ACM Reference Format:**
Aoyu Wu, Liwenhan Xie, Bongshin Lee, Yun Wang, Weiwei Cui, and Huamin Qu. 2021. Learning to Automate Chart Layout Configurations Using Crowdsourced Paired Comparison. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI '21).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*CHI '21, May 08–13, 2021, Online Virtual*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Data visualizations have been ubiquitous in everyday life, such as social media, magazines, and websites. They are widely used by the general public to express complex data in an intuitive, concise, and visually appealing manner. However, creating effective and elegant visualizations is a challenging task even for professions [52]. Individuals usually need to engage in a time-consuming process to craft designs that clearly convey information and insights, while satisfying the aesthetic goals. As such, there have been huge efforts from both industry and research communities to aid the design process by automated approaches.

Existing approaches have predominantly focused on studying and optimizing performance metrics for data analytics concerning usability and utility. For example, commercial software such as Excel automatically recommends chart types based on selected data. Besides, much recent research proposes automated visualization systems that retain data integrity [28], highlight interesting data facts [37], and recommend effective visual encodings [14, 29, 47]. Nevertheless, those systems utilize pre-defined heuristics to generate visual styles, which could be sub-optimal (Figure 1). This paradigm results in a quasi-automated process where individuals need to manually adjust the visual style of the automatically generated charts (*e.g.*, [63, 70]). However, performing manual adjustments can be unsystematic and difficult, especially for lay users without design backgrounds [73]. Users might be unaware of guidance or find it tedious to adjust multiple parameters simultaneously. To address this problem, we aim to propose a systematic data-driven approach that recommends parameter configurations by learning from crowdsourcing human preference data. Particularly, we study layouts because it is a fundamental element of chart design [61]. We focus on bar charts which are one of the most common chart types [2].

Despite the increasing acknowledgement of the importance of visual styles in charts [9, 24, 33, 46, 54], little work has attempted to understand and quantify layout qualities through large-scale user studies. Research in graphical design has provided various layout metrics such as alignment and segmentation [49], but they are not readily applicable to charts that are data-driven and yield different visual perception [7, 66]. There is a lack of empirical studies to understand metrics for chart layouts. This is challenging due to the subjective nature of layout qualities, which requires a large number of participants to score charts. The scores, however, might

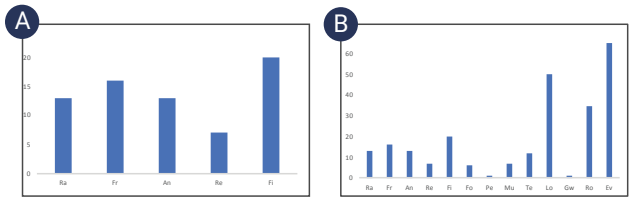



**Figure 1: Visualization tools such as Microsoft Excel utilize a default heuristic to generate layouts: (A) the chart with five bars; (B) increasing the number of bars results into a chart with the same size but different bandwidths. Those layouts have room for improvements through manual refinement.**

not be precise since participants might be hesitant and feel difficult to give an accurate score [55, 64]. Besides, the scoring scales can be inconsistent among participants [21]. Those limitations constrain the reliability of utilizing the scores as benchmarks for machine learning models. To that end, our approach is inspired by the successful applications of pair-wise ranking for assessing natural image qualities [35]. We propose a two-alternative forced-choice experiment [17], asking participants to select a better chart layout between two candidates. This data acquisition method allows us to obtain more precise and consistent results [6, 55].

We propose a novel approach, called Layout Quality Quantifier ($LQ^2$), for learning to score and rank the chart layout configurations from human preference through crowdsourced pair-comparison experiments. $LQ^2$ utilizes neural networks to predict the score of an individual chart by taking comparison pairs as training data. $LQ^2$ predicts the pair-wise ranking with the accuracy of 78%, showing that it could reasonably learn human preference for layout configurations. We further interpret the trained model by investigating the impact of layout parameters on human preference, thereby summarizing rule-of-thumb for layout configurations in bar charts. Finally, quantitative user studies demonstrate that $LQ^2$ could recommend more visually appealing layouts than manual results by laypeople and default styles in Excel and Vega-Lite [57]. Overall, our work demonstrates the possibility of quantifying human aesthetics for charts. We open source all our code and experimental material[1]. In summary, our contributions are as follows:

- A novel approach for quantifying human preference for chart layouts through crowdsourced paired comparison
- A machine-learning method, $LQ^2$, for ranking and scoring layout configurations in bar charts
- A set of qualitative and quantitative evaluations as well as two user studies that demonstrate the effectiveness and usefulness of $LQ^2$

## 2 RELATED WORK

Our work is related to aesthetics for visualizations, automated visualization designs, as well as data collection and training for visualization research.

[1]https://github.com/shellywhen/LQ2

### 2.1 Aesthetics for Visualizations

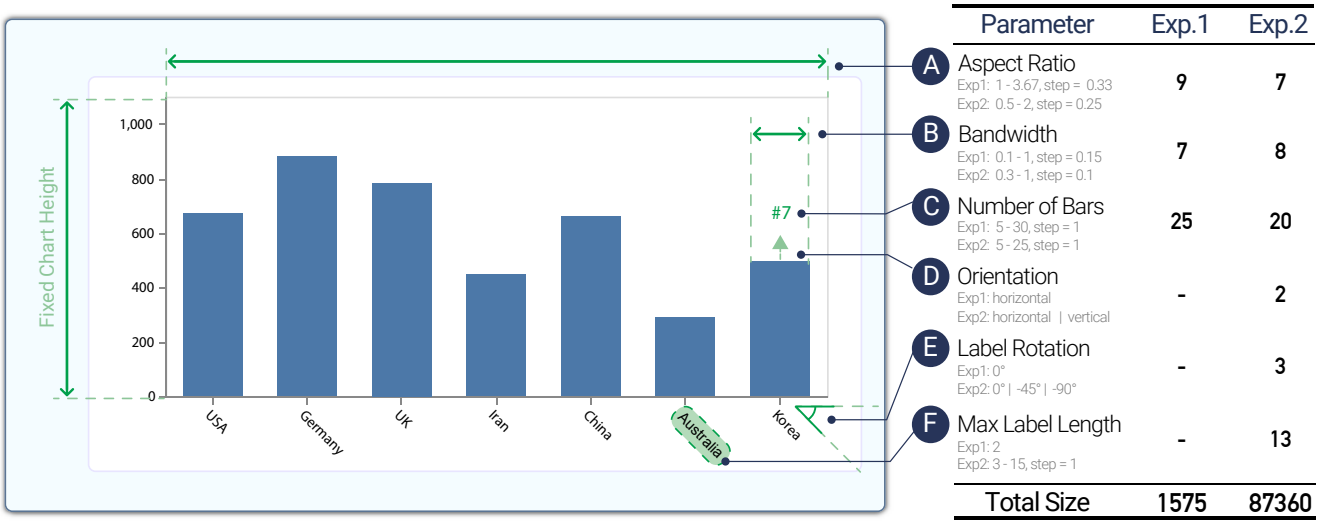
In a broader sense, our work is related to the aesthetic qualities of data visualizations. In the book *Information is Beautiful*, McCandless [45] lists aesthetics as one of the four criteria for a good visualization. However, aesthetics were traditionally considered as an add-on feature that was typically implemented at the very end of the design process. Already 13 years ago, Cawthon and Moere [9] argued for increased recognition for visualization aesthetics, by demonstrating the relationships between aesthetics and usability in data visualizations. Since then, many empirical studies have shown that the aesthetics of data visualizations could contribute to various factors such as first impressions [24], memorability [5], emotional engagement [33], and task performances [54].

Nevertheless, little work has studied what makes a data visualization visually appealing. Moere *et al.* [46] demonstrated that the visual styles could lead to different comments regarding aesthetics. Quispel *et al.* [53] found that laypeople were attracted to designs they perceived as familiar and easy to use. However, they investigate aesthetics as a qualitative reflection of personal judgment rather than a quantifiable and comparable entity. Human preferences for aesthetics in the context of charts are still not methodically quantifiable from a data-driven perspective, and seem underrepresented in large-scale empirical studies. We address this gap by proposing a systematic machine learning approach for ranking and scoring layout qualities from crowdsourcing experiment data. Besides, we propose our approach for interpreting the trained model, whereby summarizing speculative hypotheses that warrant future empirical research to confirm.

### 2.2 Automated Visualization Design

Recently, there have been growing interests in applying machine learning methods for automated visualization designs. Researchers have proposed many systems [14, 29, 44, 47] that recommend visualizations based on data structures and characteristics. Those systems focus on deciding the effective chart type, visual encoding, and data transformation. In addition to effectiveness, much research has been devoted to optimizing visualizations from other aspects. For example, VisuaLint [28] addresses the data integrity by surfacing chart construction errors such as truncated axes. DataShot [70] and Calliope [59] focus on generating visualizations with interesting data-related facts from tabular data. Dziban [41] attempts to balance automated suggestions with user intent by preserving similarities with anchored visualizations. Different from them, our work studies to improve the aesthetic quality of chart layouts.

Researchers have also recently proposed many approaches for improving the visualization layout. Several work automatically extracts reusable layout templates from visualizations [11, 12] and infographics [43]. Nevertheless, they do not propose metrics for extracted layouts. Other systems optimize the layout according to various metrics such as mobile-friendliness [71], similarities with user-input layouts [63] and graph features [23, 69]. However, their metrics are not derived from empirical studies and therefore might not reflect the overall perceived quality [3]. Therefore, our work studies how to quantify and optimize layout qualities from crowdsourcing experiments.




| Parameter | Exp.1 | Exp.2 |
|---|---|---|
| A Aspect Ratio<br>Exp1: 1 - 3.67, step = 0.33<br>Exp2: 0.5 - 2, step = 0.25 | 9 | 7 |
| B Bandwidth<br>Exp1: 0.1 - 1, step = 0.15<br>Exp2: 0.3 - 1, step = 0.1 | 7 | 8 |
| C Number of Bars<br>Exp1: 5 - 30, step = 1<br>Exp2: 5 - 25, step = 1 | 25 | 20 |
| D Orientation<br>Exp1: horizontal<br>Exp2: horizontal \| vertical | - | 2 |
| E Label Rotation<br>Exp1: 0°<br>Exp2: 0° \| -45° \| -90° | - | 3 |
| F Max Label Length<br>Exp1: 2<br>Exp2: 3 - 15, step = 1 | - | 13 |
| Total Size | 1575 | 87360 |

**Figure 2: The layout parameters in Experiment 1 and 2. The table displays the sampled values of each parameter. Experiment 1 concerns 3 parameters with 1,575 possible combinations, while Experiment 2 includes 6 parameters with 87,360 combinations.**

## 2.3 Data Collection and Training for Visualization Research

Recent years have witnessed a growing recognition of data collection to facilitate research in machine learning for visualizations. Many efforts have been made to collect real datasets of charts from websites [2, 29, 50] and scientific literature [10, 39]. Those datasets include annotations or original data as ground-truth labels and assume that those charts share the same quality. Therefore, another line of research conducts crowdsourcing user studies to obtain quality metrics such as task completion time and accuracy [56] as well as attentions [34], which can be measured objectively by devices. However, it is much more challenging to generate a reliable dataset for subjective metrics, as participants might not share a reliable and consistent scoring scale [21, 55, 64]. To that end, Saket *et al.* [56] extended their experiment by asking participants to rank five different visualization types in the order of preference and found a positive correlation between user preference and task accuracy. However, the collected data is for statistical analysis instead of machine learning tasks.

To generate dataset for machine learning, Luo *et al.* [44] proposes a pair-wise comparison approach, *i.e.*, to ask participants to choose which chart is better from two candidates, which yields more precise results. They subsequently compute the overall order from pair-wise comparison, and choose the top-ranked ones as training data. However, their approach is limited for two reasons. First, it is inefficient as they only obtained 2,520/30,892 good/bad charts after 285,236 comparisons. Second, they formulate the problem as a classification task, neglecting the subtle differences among charts. To that end, we propose $LQ^2$ which predicts a numerical score of a single chart through regression neural networks, while directly taking paired comparisons as the training input. $LQ^2$ is built on similar learning frameworks for image assessment [68], but integrates a parameter module and two sampling strategies to learn the visualization-specified features.

# 3 OVERVIEW

Data visualizations represent data with graphical elements according to visual specifications. Specifications can be classified into two types: *visual encodings* that map data to visual properties (*e.g.*, color, position, size) of graphical elements, and *visual styles* that specify the remaining visual properties irrespective of data (*e.g.*, label rotation, bar bandwidth). Our work aims to automate the parameter configurations for the latter, *i.e.*, visual styles, which are largely neglected in existing automated charting tools. Concretely, we focus on the layout properties in bar charts, since this is one of the first work that attempts to rank and recommend parameter specifications of visual styles leveraging machine-learning approaches. In this section, we describe our experiments, the design considerations, and the problem formulation.

## 3.1 Experiments

Different from visual encodings that are typically described as discrete mappings, visual styles usually have greater cardinality and continuous values that increase their complexity. To keep this study complexity manageable, we select two concrete yet underexplored experiments (Figure 2).

Our first experiment considers three basic layout-related parameters, *i.e.*, the number of bars, the aspect ratio of the chart, and the bandwidth. This is because we observe that existing charting tools determine those values by predefined heuristics. For instance, Microsoft Excel fixes the aspect ratio and computes the bandwidth according to the number of bars (Figure 1). In this paper, we argue and demonstrate that such default heuristics could result in suboptimal layouts. Individuals, therefore, need to repeatedly adjust several parameters to achieve visually appealing layouts. However, such manual adjustments are unsystematic and time-consuming, without a guarantee of improvement. Therefore, we study how to automatically configure those parameters by learning from human preferences.

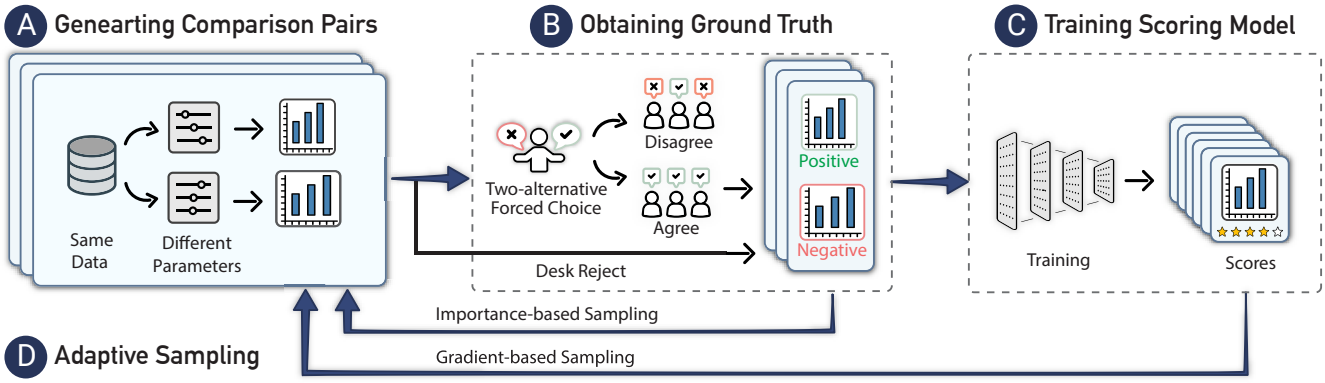



**Figure 3: The data collection process in an iterative manner: (A) Generating paired charts with same data and different layout parameters; (B) Labelling the training data through crowdsourcing experiments; (C) Training the scoring models; (D) Utilizing two offline adaptive sampling strategies to increase the representativeness of the training dataset.**

The second experiment is extended with another three parameters, namely, the chart's orientation, the max length and rotation degree of axis tick labels. This experiment is motivated by the practical needs for responsive visualization design, *i.e.*, how to adjust the chart layout to fit into different sizes. This is a challenging task since chart creators need to manually examine and edit layouts for multiple chart sizes [27]. Therefore, we add those three parameters that are often subject to adjustments in responsive visualization designs. This experiment extends existing automated responsive visualization approaches [71] by considering the aspect ratio and allowing rotating chart orientations.

## 3.2 Design Consideration

The problems above can be summarized as optimization problems, *i.e.*, to find values of visual styles that maximize the layout quality. To guide the design of our solution, we summarize two primary considerations:

*C1: To quantify and score layout qualities.* One of the primary challenges in optimization problems is to define the objective function. Hence, our primary goal is to learn a loss function that maps values of layout parameters onto a numerical score intuitively representing the layout qualities. This loss function can be subsequently used for mathematical optimization.

*C2: To learn the overall quality from human feedback.* Since judgments of layout qualities involve a wide range of factors, previous work in graphic designs [40, 49] usually utilizes human-crafted metrics (*e.g.*, symmetry) to measure layout qualities. Their methods face challenges in the context of data visualizations since few human-crafted metrics are available for chart layouts. Besides, it is difficult to weigh different metrics to reflect the overall quality perceived by users. Therefore, we aim to measure the overall quality by learning from human feedback, and conduct post hoc analysis to summarize rule-of-thumb from the trained model.

## 3.3 Problem Abstraction and Formulation

Guided by the design considerations, our main task is to develop a machine-learning model that learns to predict a layout quality score of the given parameters from human feedback data. We formulate this task as a learning-to-rank problem [42], which purposes to acquire a global ranking from partial orders. The ground truth of partial orders is harvested from experimental data on human preference. Specifically, we conduct a paired comparison experiment, asking participants to choose their preferred layout from two candidates. The results from paired comparisons contain partial orders, which constitute the training data.

# 4 DATA COLLECTION

We describe our process of constructing the training dataset containing ranked pairs of chart layout configurations. As shown in Figure 3, the process is iterative and contains four steps. In this section, we describe the step A, B, and D in Figure 3 in detail. Step C will be introduced in section 5.

## 4.1 Generating Comparison Pairs

Our first step is to create paired charts for crowdsourcing comparison. We decide to synthesize charts since it is difficult to harvest real-world chart pairs that are fairly comparable, that is, they are controlled to represent the same data. Charts are created using Vega-Lite [58], which allows specifying the aforementioned parameters in a declarative manner. For each pair, we choose data from two popular real-world datasets, namely the Car dataset[2] and the Baseball dataset[3], and randomly select entries according to the number of bars. In Experiment 1, we replace the tick labels with meaningless two-character tokens. In Experiment 2, the tick labels are truncated according to the parameter of label lengths.

The remaining parameters are generated with different values within a chart pair, including the aspect ratio, chart orientation, bandwidth, and rotation of axis labels. It is expensive to conduct controlled experiments for each parameter, since those parameters may not be interdependent. Therefore, we choose to randomize all parameters, intuitively intending to obtain a wide variety of chart configurations. However, exhaustive enumeration of possible

[2]https://vega.github.io/vega-datasets/data/cars.json
[3]https://github.com/vincentarelbundock/Rdatasets/blob/e38552ac3cb40a532941b09d7332b03d19409919/doc/plyr/baseball.html

values and combinations of parameters is infeasible due to their continuous distributions. Thus, we decide to randomly sample from uniform distributed values (*e.g.*, the bandwidths range from 0.1 to 1.0 with a step of 0.15). We choose a relatively large step in order to make the differences notable. While parameters are sampled randomly, we make the sampled values evenly distributed to avoid the data imbalance problem.

Figure 2 shows the sample values and the possible combinations of parameter values in our experiments. We update the sampling values in Experiment 2 according to our findings in Experiment 1. For instance, we truncate the maximal aspect ratio to 2, since we observe that larger aspect ratios are less favored. It should be noted that the resulting design space is still considerably large that poses challenges in solving the optimization problem.

## 4.2 Obtaining Ground Truth

We harvest ground truths of ranked pairs of charts through a two-step process.

*Desk Reject.* First, we "desk reject" charts that violate a set of predefined rules and label them as negative in a pair. We discard a chart pair if both charts violate rules. Specifically, two rules are included in Experiment 2: the axis labels should not overlap with each other, and the axis label should not rotate in a horizontal bar chart. This approach allows us to train an ML model that learns human-crafted rules during the training time and therefore better reflect the overall quality. An alternative approach in visualization recommendation systems is to utilize rules as hard constraints in the optimization phase [47], which, however, poses challenges in solving the optimization problem with the increasing number of rules. This might be undesirable since it could prolong the execution time and therefore degrade the usability.

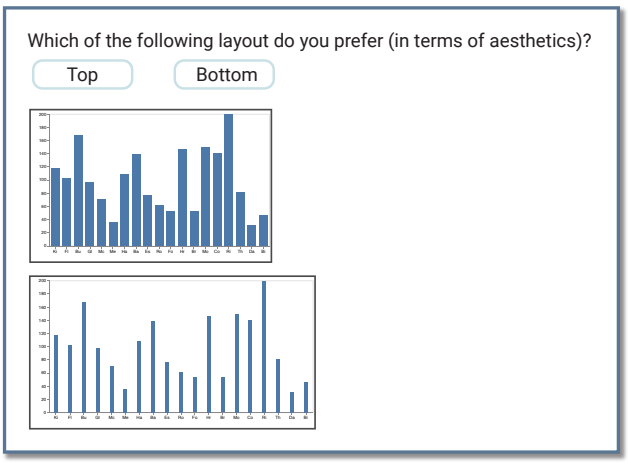



**Figure 4: Illustration of the MTurk interface for crowdsourcing experiments: we propose a two-alternative forced-choice design that makes it easier for participants to evaluate the relative quality of paired charts than scoring a single chart.**

*Crowdsouring Experiments.* Second, we conduct crowdsourcing experiments on Amazon Mechanical Turk (MTurk) to obtain experimental data from human preference. Figure 4 illustrates the settings of the MTurk experiment. We propose a two-alternative forced-choice (2AFC) experiment, asking participants to choose "which of the following two layouts do you prefer, in terms of aesthetics?". Two charts are placed vertically within a viewport since it is easy to compare by moving eyes between side-by-side views [48]. We choose a forced-choice method in an attempt to capture the subtle differences [75]. Each MTurk HIT consists of 10 comparison tasks, and each task (paired comparison) is assigned to 3 participants. For quality control, we randomly duplicate one comparison task within a HIT and swap the order of paired charts. We keep HITs where participants offer consistent answers for duplicated tasks.

We measure the inter-observer reliability by the joint probability of agreement. It is observed that three participants make the same choices in 45.6% of pairs for two experiments. This observation probability is much higher than that of the agreement by chance, *i.e.*, 25%, showing that human preference exhibits a fair degree of agreement on layout qualities. This fair agreement can have several reasons. First, the differences between the two charts in a pair might be small and therefore cause uncertainties, since the layout parameters are generated randomly. Second, individual participants have different preferences. Third, it might be because of the noises of MTurk experiments.

We select paired comparisons with full agreements among participants as the training data to reduce noises [75]. Each pair consists of two charts, denoted $\langle \mathcal{I}^+, \mathcal{I}^- \rangle$, where $\mathcal{I}^+$ is preferred over $\mathcal{I}^-$.

## 4.3 Adaptive Sampling

It is crucial to employ an effective pair sampling strategy to select the most important pairs for rank learning [68]. Our uniform sampling and random pairing strategy in subsection 4.1 is sub-optimal, since we are interested in finding the most "optimal" chart configurations. Therefore, we propose two offline adaptive sampling strategies to improve the qualities and representativeness of the comparison pairs. The term "offline" here is referred in the context of machine learning, that is, we re-sample comparison pairs when the initial training phase has finished.

*Importance-based Sampling.* We are interested in finding important pairs that allow us to determine the "best" chart configurations. Therefore, we borrow the idea of elimination tournaments, intuitively conducting a second round of comparisons among previous winners. However, this is not readily applicable since our sample size is much smaller than the huge number of possible parameter combinations. Thus, we propose an important-based sampling scheme, which intends to increases the probability of sampling important charts with "good" parameter values.

Suppose each chart $\mathcal{I}$ is configured by a set of parameters $p_i \in \mathcal{P}$, where the possible values of $p_i$ are $v_i^j \in \mathcal{V}_i$. Let $w_i^j$ denote that number of times that the chart $\mathcal{I}$ whose configuration contains $v_i^j$ has won in paired comparisons in Figure 3(B). We update the probability of sampling the value:

$$P(v_i^j) = \frac{\min\{w_i^j, T\}}{\sum_j \min\{w_i^j, T\}}, \tag{1}$$

where $T$ is a parameter responding the exploration-exploitation trade-off by avoiding empty probabilities.

*Gradient-based Sampling.* Having a large step size in uniformly sampling might cause the model to overlook a maximal. To address this problem, we use a gradient-based sampling method to sample important parameter values with a smaller step size. As gradients are computed on a differentiable function, we refer to our scoring model trained in Figure 3(C). This scoring model learns a regression function $f(\cdot)$ that maps the parameter vector $p = \{p_1, p_2..., p_i\}$ to a numerical score. We compute the locations where the gradient of $f$ along with $p$, $\nabla f(p)$, is smaller than a given threshold. We sample parameter values within those locations with a smaller step-size, *i.e.*, 1/3 of the original step-size.

In both experiments, we conduct each of the following adaptive sampling strategies once, and merge the resulting dataset. This procedure results in 1,177 pairs in Experiment 1 and 1,333 pairs in Experiment 2. Overall, our data collection process involves 416 unique MTurk participants.

## 5 METHOD

With the obtained pairs $\mathcal{D} = \{\langle \mathcal{I}^+, \mathcal{I}^- \rangle\}$, LQ$^2$ aims to quantify the aesthetic scores of a given chart. Specifically, we formulate the problem as a regression problem, that is, to output a numerical score $\mathcal{S}$ for an input chart $\mathcal{I}$. Our goal is to learn a regression function $f(\cdot)$ that predicts a higher score for the preferred chart in a pair:

$$\mathcal{S}^+ > \mathcal{S}^-, \forall \langle \mathcal{I}^+, \mathcal{I}^- \rangle \in \mathcal{D} \tag{2}$$

where $\mathcal{S} = f(\mathcal{I})$.

*Model Architecture.* LQ$^2$ adopts a Siamese neural network structure, *i.e.*, to work in tandem on two different inputs with the same weights to compute comparable output [13]. As shown in Figure 5, it consists of two identical scoring networks, and the loss function is defined on the combined output of scoring networks. The scoring network takes the parameter values as input and outputs a numerical score. We employ fully-connected neural networks (NN), which have proven effective in handling features describing design choices (*i.e.*, parameters) in VizML [29]. Our NN contains 6 hidden layers, each consisting of different numbers of neurons with

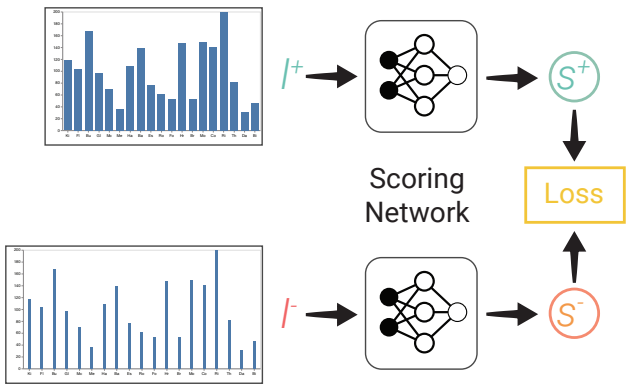



**Figure 5: LQ$^2$ utilizes a Siamese neural network structure to work in tandem on a pair to compute comparable output.**

ReLU activation functions and dropout layers. We perform min-max normalization on the parameter values so that each parameter contributes approximately proportionately to the results.

We also tried to take the graphical features (*i.e.*, images) as the training input with off-the-shelf Convolutional Neural Networks (CNNs) models. However, this method did not bring about remarkable performances despite the expensive training time. Our findings conform to earlier work [20, 22] that CNNs might not readily capture human perception in visualizations. Thus, we utilize the parameter as the input, which serves as a compact and learning representation that reduces the computational costs.

*Loss Function.* We adopt the Pairwise Ranking Loss as the loss function, which explicitly exploits relative rankings of chart pairs [35]:

$$\mathcal{L}(\mathcal{S}^+, \mathcal{S}^-) = max(0, \mathcal{S}^+ - \mathcal{S}^- + m), \tag{3}$$

where $m$ is a specified margin hyper-parameter. This loss imposes a ranking constraint by penalizing mistakes for assigning a lower score to the preferred chart.

*Implementation and Training.* We implement LQ$^2$ with Pytorch. During training, we split the data by a ratio of 8:2 with the purpose of training and validation. We tune several hyper-parameters by diagnosing the learning curves so that the plots of training and validation data converge to a good point of stability and have a small gap. The model is trained with the Adadelta optimizer for 200 epochs. The learning rate is 1, and subsequently is reduced by half per 30 epochs. We found that only the margin hyper-parameter $m$ had a significant impact on the training performance, while weight decay, optimizer, and dropouts had small effects.

## 6 EXPERIMENT

To evaluate the effectiveness of our method, we conduct experiments with baseline approaches and perform qualitative analyses with the trained scoring network across different layout parameters.

### 6.1 Performance

We compare the performance of our model with several baseline approaches. For experiment reproducibility, we adapt a Monte Carlo Cross-Validation strategy [72], that is, to randomly split the data into training data and testing data with an 80-20 ratio, run the experiment, and repeat the above process ten times.

*Model Baseline.* Our problem is formed as a learning-to-rank problem. Therefore, we consider the Ranking Support Vector Machine (**RankSVM**) [31] as the baseline approach, which is a well-established method for computing the overall ranking based on pairwise preference. Similar to Draco [47], we use a linear SVM model with hinge loss.

*Scoring Baseline.* We also compare our learned scoring network with existing human-crafted metrics for layout qualities in graphical design [49]. We select four metrics that are mostly applicable to the context of charts, including **White Space**, **Scale**, **Unity**, and **Balance**. Those metrics are implemented according to instructions in the supplemental material. We discard Alignments and Overlapping whose value does not vary among our charts. Besides, Emphasis and Flow are not considered since they are mainly concerned with key text or graphics, which are not well defined in

**Table 1: Comparison of the performances between our method and baseline approaches in terms of the prediction accuracy (%) via Monte-Carlo Cross-Validation for 10 runs with an 80-20 training-testing split ratio.**

| | Ours | RankSVM | White Space | Scale | Unity | Balance | All |
|---|---|---|---|---|---|---|---|
| **Exp. 1** (N = 1,177) | 76.60 | 70.83 | 57.28 | 56.26 | 52.00 | 56.08 | 60.81 |
| **Exp. 2** (N = 1,333) | 78.27 | 64.48 | 58.24 | 61.72 | 56.21 | 63.18 | 68.73 |

**Table 2: The Pearson correlation between predicted scores and each layout parameter in Experiment 1 and 2.**

| | Number of Bars | Aspect Ratio | Bandwidth | Max Label Length | Label Rotation | Orientation |
|---|---|---|---|---|---|---|
| **Exp. 1** | -0.38 | 0.20 | 0.27 | - | - | - |
| **Exp. 2** | -0.09 | 0.37 | -0.05 | -0.09 | -0.43 | 0.04 |

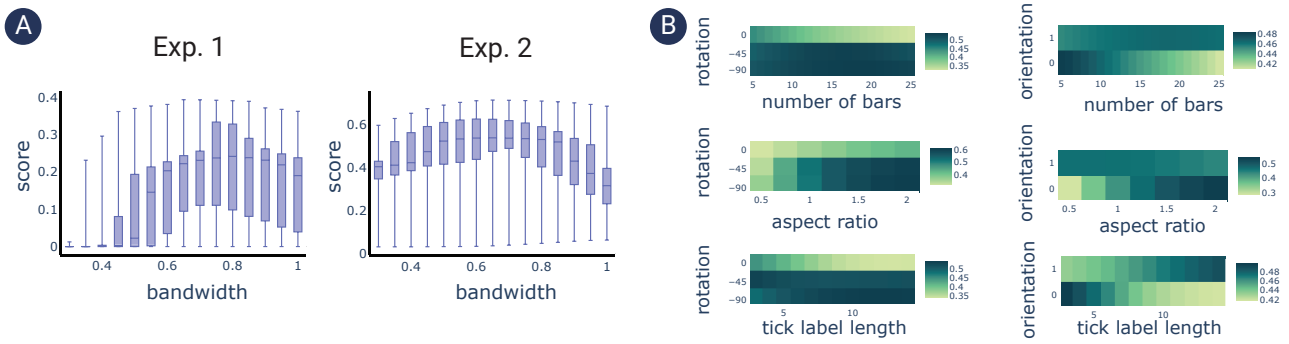



**Figure 6: Visualizing the predicted scores with (A) a single parameter by box-plots and (B) multiple parameters by heat-maps.**

charts. We also combine those metrics (**All**). Each metric consists of several features, which are fed into RankSVM to learn their weights.

*Result.* Table 1 shows the results of the performances. In both experiments, our model outperforms the baseline RankSVM approaches. In particular, RankSVM performs much poorer in Experiment 2, showing that the relations between the impacts of each parameter on predicted scores tend to be non-linear. All scoring baselines cannot achieve desired performances, suggesting that those hand-crafted features for layout qualities in graphic design cannot be readily applicable to charts.

## 6.2 Interpreting Models

To understand the impact of layout parameters on the perceived layout quality, we conduct the quantitative and qualitative analyses with the trained scoring model. Those analyses help relate our work with prior knowledge about chart layout designs, inform design guidelines, and provide qualitative support for our methods. Specifically, we calculate the predicted layout quality score of different combinations of parameters.

We study the relationships between the predicted score and each parameter by computing the correlations and visualizing distributions. Our findings are summarized in the following text. Those findings should be interpreted carefully since they are derived from the black-box ML models. Thus, they should not be considered as guidelines, but instead speculative hypotheses that warrant future empirical research to confirm.

*6.2.1 Quantitative Parameter Analysis.* Table 2 shows the Pearson correlations between the predicted scores and each parameter. We first note the negative impacts of the number of bars in both experiments, showing that it is more challenging to find good layouts with more bars. The aspect ratio contributes positively to the overall score, suggest that landscape layouts might be superior to portraits.

The impacts of bandwidths differ between our two experiments. Figure 6(A) visualizes the predicted scores versus bandwidths using box-plots. In Experiment 1, the average scores are relatively higher when the bandwidth is between 0.6 and 0.95, with a subtle peak at 0.8. However, in Experiment 2 where horizontal bar charts are introduced, the "optimal" interval become between 0.5 to 0.85, followed by a sharp drop after 0.9. Based on those observations, we form hypotheses for future studies to confirm: first, as a rule-of-thumb, the optimal bandwidth in vertical bar charts is 0.8; second, the optimal bandwidth in horizontal bar charts is less than that in vertical bar charts. Our second hypothesis conforms to existing rule-of-thumb that suggests a bandwidth between 0.57 to 0.67 in horizontal bar chart [19].

The label rotation has a moderate negative correlation (-0.43) with the score. Figure 6(B) presents the combined effects of the label rotation with other parameters on the predicted score. It is observed that non-rotation (zero-degree) is acceptable when the number of bars is small, when the aspect ratio is large, and when axis labels are short. This is because axis labels are less likely to overlap with each other under those conditions. Otherwise, the axis

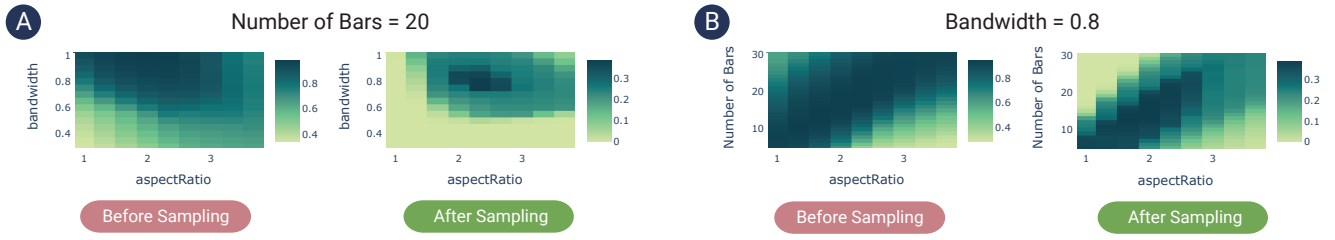



**Figure 7: Heatmaps showing the predicted scores with different parameter combinations in Experiment 1. Our adaptive sampling strategies allow us to obtain fine-tuned results.**

labels need to rotate to avoid overlapping. In general, rotations by 45 degrees yield higher scores than rotations by 90 degrees.

We observe no correlation (0.04) between the orientation and the scores, showing that both horizontal and vertical bar charts have own advantages. As shown in Figure 6(B), vertical bar charts achieve much fewer scores when the number of bars is large, when the aspect ratio is small than 1, and when the length of axis labels are large. On the other hand, the scores of horizontal bar charts are less sensitive to the number of bars and the aspect ratio, while horizontal charts seem strongly useful when axis labels are lengthy.

We also note that the score distributions are different between two experiments (Figure 6(B)). The predicted scores in Experiment 1 vary between 0 to 0.39, while the range in Experiment 2 is 0.03 to 0.72. This might be due to the existence of comparison "deadlocks" in Experiment 1, *e.g.*, $\mathcal{I}_1 > \mathcal{I}_2, \mathcal{I}_2 > \mathcal{I}_3, \mathcal{I}_3 > \mathcal{I}_1$. This left optimization constraints in Equation 3, *i.e.*, $\mathcal{S}_1 - \mathcal{S}_2 > m, \mathcal{S}_2 - \mathcal{S}_3 > m, \mathcal{S}_3 - \mathcal{S}_1 > m$, without feasible solutions. In Experiment 1, we observe 21 three-node, 23 four-node, 10 five-node, and 32 six-node circles, and the value of $m$ is 0.12.

*6.2.2 Qualitative Ablation Analysis.* We investigate the effectiveness of adaptive sampling strategies by conducting ablation analysis. Specifically, we train two models on the dataset before and after the adaptive sampling process in Experiment 1, denoted **BS** and **AS**. Two datasets are down-sampled to ensure the same data size. Figure 7 presents two qualitative examples showing the learned relationships between the predicted scores and different parameters. It is observed that the **BS** model yields a small region of light colors (low scores) and a majority of dark colors (high scores). That said, it learns to reject bad conditions but could not further differentiate conditions scored "borderline and above". On the contrary, the **AS** model is able to identify a small region of parameters that yield higher scores, which accords with our goal to optimize layout parameters. Besides, it identifies difficult conditions. For instance, Figure 7(B) (right) suggests that the optimal aspect ratio increases with the number of bars before it reaches 3. This implies the difficulties in finding good layout parameters for charts with an aspect ratio over 3, which are uncommon and less favored.

## 7 APPLICATION

To demonstrate the usefulness of $LQ^2$, we present a novel application, *i.e.*, automatic optimization of layouts. Existing charting tools typically generate layouts by predefined heuristics, which requires tedious manual adjustments. It would therefore be useful to automate this process by recommending layout parameters that improve the quality. To that end, we propose an automatic optimization approach and conduct two user studies.

### 7.1 Method

We present two user studies in line with our experiments. In User Study 1, we propose a common real-world scenario - presentations, where individuals usually wish to create charts to convey data insights in an aesthetically pleasing manner. The task is to adjust the aspect ratio and the bandwidth given the data. User Study 2 concerns adaptive visualization designs, where a maximal width is posed as a hard constraint and the task is to adjust four parameters including the aspect ratio, the bandwidth, the orientation, and the label rotation. We create 50 and 80 design cases for two studies, respectively, each case encoding randomly chosen data. We compare our results (**Ours**) with those generated by laypeople (**Human**) and default parameters (**Default**), and random values (**Random**).

*Our Approach.* Our optimization approach aims to find parameter values that maximize the layout quality score predicted by $LQ^2$. For that purpose, we adopt a brute-force method that enumerates combinations of values and selects the one with the highest predicted score. We choose brute-force methods since the maximal enumeration size is 87,360 which computers could operate within seconds. Advanced optimization techniques are desired to cope with the expanding parameter space by avoiding enumeration [1].

*Human Baseline.* To obtain human baselines, we run an experiment on MTurk. Participants are instructed to "adjust the parameters until you are mostly satisfied with the layout" with a What You See Is What You Get (WYSIWYG) editor implemented with Vega-Lite. Akin to standard charting tools, participants are provided with a slider and an input box for adjusting continuous values, and a radio group for editing discrete parameters. We record their editing history and the time used from the first editing to final submission. Each participant is assigned to one and only one design task.

*Default Baseline.* In User Study 1, we choose Microsoft Excel as the default baseline. In order to keep the comparison fair, we remove components that do not exist under other conditions, which include the chart title and y-axis gridlines. Besides, the bars are filled with the default color of Vega-Lite. In User Study 2, we compare our method against the Responsive Bar Chart feature in Vega-Lite[4].

[4]https://vega.github.io/vega-lite/examples/bar_size_responsive.html

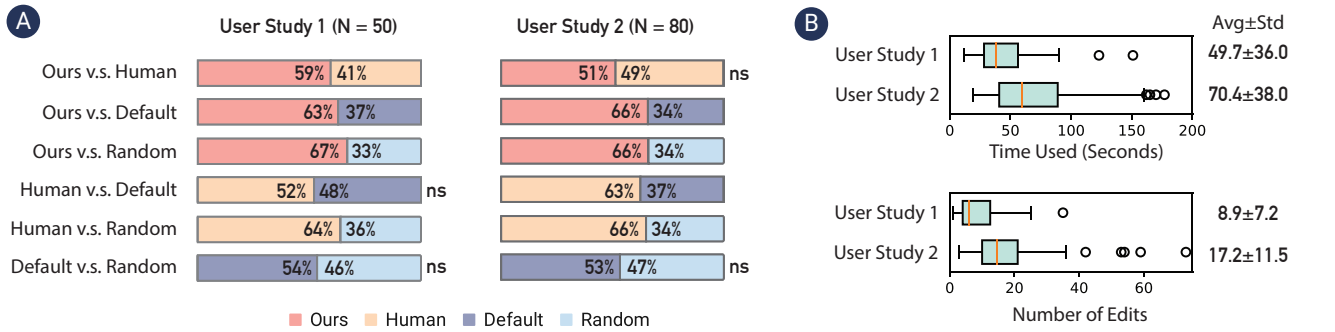



**Figure 8: Results of the user study: (A) displays the results of group-wise comparison among four groups in terms of percentages of favored votes. An "ns" denotes no statistical significance via Wilcoxon signed-rank tests; (B) presents two box-plots visualizing the time used and the number of edits by laypeople in configuring the chart layout.**

*Random Baseline.* The random baseline takes random parameters, which are sampled from values observed in the training data to make the comparison fair.

## 7.2 Evaluation and Results

We run another MTurk experiment, asking participants to compare the results among the above four groups. Similar to the labeling process, we conduct pair-wise comparisons between every two groups. Each between-group comparison includes 50 paired charts in User Study 1 and 80 in User Study 2. Each paired chart is evaluated in a two-alternative force decision (2AFC) paradigm by 10 participants. There is one duplicate pair out of per 10 for quality control.

Figure 8(A) summarizes the results of group-wise comparisons in terms of the percentages of preferred votes in the 2AFC procedures. In User Study 1 (US1), our method outperforms Human, Default, and Random ($p < 0.05$, Wilcoxon signed-rank tests). It is also noted that while Human has a higher preference than Default, the difference is not statistically significant. Similarly, Default has a small yet not significant superiority over Random. However, Human performs significantly better than Random. This said, laypeople could only achieve a relatively small improvement in the layout quality, although they spent notable efforts, *i.e.*, 49.7 seconds and 8.9 adjustments on average (Figure 8(B)).

In User Study 2 (US2), both Ours and Human outperform Default and Random ($p < 0.05$), while Default and Random are evenly matched. This shows while Vega-Lite enables automatic adaptive visualization, the generated layouts are sub-optimal. Our method only achieves compatible results with Human, which might due to three reasons. First, participants have spent more efforts, *i.e.*, 70.4 seconds and 17.2 adjustments on average (Figure 8(B)). It is therefore expected that participants could achieve better results. Second, US2 presents a much more challenging task than US1, since the total number of possible parameter combinations is 1,575 in US1 and 87,360 in US2 (Figure 2). However, the training data size is only 1,333 in US2, which is far from fully representing the whole design space and therefore could not find the "optimal" solution all the time. Still, our results are positive as our method has achieved human-level performances by leveraging a small training data, showing the effectiveness of our sampling strategies. Future work could further extend our work by augmenting the training data. Finally, our sample size (80) is relatively small, considering the variety of the parameter space and the random nature in task generation. In the future, we plan to conduct larger-scale user studies to better understand the scalability.

In summary, our results show that the default heuristics for generating layouts in existing charting tools could result in sub-optimal results. To improve the layout quality, laypeople need to engage in a time-consuming process to adjust the parameters over and again. Our automatic approach could achieve at least human-level performance via small-sample learning, while removing the heavy costs of manual adjustments.

# 8 DISCUSSION AND CONCLUSION

We reflect on the implications and future work of our research.

## 8.1 Implication

*Do not trust the defaults.* Charting tools and libraries provide default settings for user-configurable parameters. Default settings are proven to introduce a default effect that people would blindly trust and stick with them [15]. However, default settings are designed to be reasonable under most cases, *i.e.*, to prevent stupid mistakes. Thus, they are just acceptable but not good for all. We provide empirical evidence that the default layout parameters for bar charts in Excel and Vega-Lite are sub-optimal, which can be significantly improved by manual or automatic fine-tuning. Those results support the needs of increasing recognition for utilizing default values prudently.

*Augmenting empirical studies with a machine learning approach.* Our experiments can be considered as empirical studies aiming to identify the "best" combinations of variables. This is challenging due to the vast design space, *i.e.*, 87,360 possible combinations, making exhaustive enumeration and controlled studies infeasible. In response, we propose a ML approach that learns to rank the combinations from small samples (1,333 pairs), yielding notable results. More importantly, we formulate hypothesises of optimal variables via interpreting the ML model. Future work could verify those hypothesises by conducting controlled experiments.

*Quantifying visualizations with subjective metrics.* Recent years have witnessed a growing research interest in quantifying and benchmarking visualizations for machine learning (e.g. [30, 56]). Those work has predominately focused on objective metrics such as accuracy and effectiveness. Subjective metrics, however, are relatively neglected, while they are considered more challenging to measure. Our work extends this line of research by benchmarking charts with subjective metrics, *i.e.*, human preference over layouts, through crowdsourcing experiments. We describe our procedures and strategies for quantifying subjective metrics, hoping to inform future research to measure and improve visualizations from more diverse perspectives, *e.g.*, understandability [60].

*Improving aesthetic qualities of visualizations from a data-driven perspective.* A good visualization consists of four necessary elements: information, story, goal, and beauty [45]. In a broader sense, our work addresses the beauty, that is, the aesthetic quality. We propose a data-driven method to learn human preference for layouts, which outperforms hand-crafted layout metrics. Our results demonstrate the promising research possibilities of understanding and improving the aesthetic quality via data-driven machine learning approaches. This research direction is supported by real-life practical needs, *i.e.*, existing charting tools generate sub-optimal results, while laypeople tend to rely on default values [15] or need to engage in a time-consuming process to tune the results until they are satisfied with the result (Figure 8(B)). These needs call for an increasing recognition for understanding what makes a chart visually appealing and proposing more advanced automatic methods for improving the aesthetic quality.

## 8.2 Critical Reflection

We assess the quality of charts by asking participants "which do you prefer?". Compared with scoring a single chart, this paired comparison method is easier for participants and yields more precise and consistent results. As such, we see potentials of adopting it for various purposes in visualization research. To better inform future research, we discuss our critical reflections on this method.

*Combating decision paralysis.* Decision-makings are not always easy, especially when the differences between two charts are small. It could cause analysis paralysis where individuals overthink the situation that makes decision-making "paralyzed" [38]. Subsequently, individuals tend to choose an arbitrary decision hesitantly [64]. As shown in Figure 8(B), some participants spent much more efforts on editing the parameters than the average, showing that they seemed subject to analysis paralysis. To alleviate this problem, future research should propose more effective sampling strategies that avoid over-subtle differences between paired charts. Besides, we might borrow the idea of agile methodologies in software engineering to overcome the anti-pattern of decision paralysis [8]. One promising approach in the context of empirical research is to set time limits for viewing visualizations and making decisions (*e.g.*, [24]).

*"Evils" can attract.* Psychological studies reveal the physical attractiveness stereotype that people tend to assume "what is beautiful is good" [16]. In the context of data visualizations, this is exemplified by chartjunk [18, 65], where laypeople are attracted by elements that are visually appealing but usually at the expense of effectiveness. This contributes to the paradigm in charting tools that "compromise" with such human preference. For example, Google Sheets supports 3D pie charts despite their criticism by the visualization research community. Google Sheets also offers a Smooth Line Chart that improves the aesthetics but compromises the integrity of the underlying numbers. Future research should be aware of this trade-off when designing the experiment settings.

*Incorporating crowdsourced opinions with expert knowledge.* Visualization researchers have increasingly leveraged crowdsourcing experiments for the sake of scalability and diversity. However, crowdsourcing experiments face challenges such as reduced control in the assessment of participants' capability that might harm the validity [4]. Besides, we observe disputes in crowdsourced opinions. To that end, we envision that expert knowledge could help increase validity, resolve disputes, and reduce costs. For instance, one might select expert-generated charts as positive and randomly-generated charts as negative in pair [51]. However, it is worthy noting that expert judgement could clash with crowdsourced opinions that warrants deeper investigation [36].

## 8.3 Limitation and Future Work

*Balancing human preference and perceptional effectiveness.* Our work takes only the first step in improving the visual quality of data visualizations via a data-driven approach that learns from human preference. In particular, we study six layout parameters in bar charts. We do not conduct comprehension experiments to evaluate their effects on perceptional effectiveness, because the effect size of layouts on perceptions is typically small in standard bar charts [62, 74]. How to balance human preference and perceptional effectiveness is a clear next step for future work. This is critical because layouts have proven to impose more influences in some other charts (*e.g.*, [25, 26]). A key challenge here is that human preference and perception should be measured conjunctively in order to obtain the training data for machine learning approaches.

*Moving towards an more adaptive approach.* Although we propose two sampling strategies that enable learning from small data sets (Small Sample Learning), our model in Experiment 2 only achieves human-level performance. This presents a significant challenge as the size of the design space grows exponentially with the increasing number of parameters. Future work should propose advanced sampling approaches to improve effectiveness. Recent research in online adaptive sampling [32] that automatically updates the sampling strategy during training is a promising method to address this problem. An interesting research problem would be how to dynamically adjust the sampling probabilities during crowdsourcing experiments. Moreover, we see research opportunities in leveraging the authoring provenance (*e.g.*, the editing histories) to augment the training data and develop a ML model that adaptive recommends design suggestions based on the current configuration.

*Understanding the representations and models for visualization research.* In a broader sense, it remains an open challenge to choose the feature representations and machine learning models for visualizations. Similar with Draco [47] and VizML [29], $LQ^2$ is trained on the parameter features that are compact and computationally

inexpensive, which, however, might not generalize to unobserved parameter values (*e.g.*, more than 30 bars) and different chart types or require labour-intensive feature engineering. Although graphical features (*i.e.*, bitmaps) might embrace generalisability, recent studies [20, 22] suggest that CNNs, the most common model for analyzing visual imagery [67], seem not currently capable of processing visualization images. This underscores the research needs to explore advanced ML models, *e.g.*, VAE [20]. Furthermore, $LQ^2$ does not include the underlying data distributions and non-layout parameters (*e.g.*, colors) in the training representations, which could influence the perceived aesthetic qualities. To that end, future research should study how to choose and fuse multiple representations including the underlying data, parameters, and graphics.

*Debating "what is beautiful is good".* Finally, we propose a research agenda towards more understanding of the roles of aesthetic qualities in data visualizations. This is critical since nowadays more and more people are able to create visualizations, so does their exposure to the greater masses. This phenomenon contributes to the increasingly popular pursuits of aesthetic qualities. We even see extreme cases where aesthetic concerns play a more crucial role than usability and even usefulness, *e.g.*, the Smooth Line Chart. How should the research community respond to this shifting boundary?

## ACKNOWLEDGMENTS

The research is partially supported by Hong Kong RGC GRF grant 16213317.

## REFERENCES

[1] Satyajith Amaran, Nikolaos V Sahinidis, Bikram Sharda, and Scott J Bury. 2016. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research* 240, 1 (2016), 351–380. https://doi.org/10.1007/s10479-015-2019-x

[2] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. 2018. Beagle: Automated extraction and interpretation of visualizations from the web. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–8. https://doi.org/10.1145/3173574.3174168

[3] Michael Behrisch, Michael Blumenschein, Nam Wook Kim, Lin Shao, Mennatallah El-Assady, Johannes Fuchs, Daniel Seebacher, Alexandra Diehl, Ulrik Brandes, Hanspeter Pfister, et al. 2018. Quality metrics for information visualization. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, Hoboken, NJ, USA, 625–662. https://doi.org/10.1111/cgf.13446

[4] Rita Borgo, Luana Micallef, Benjamin Bach, Fintan McGee, and Bongshin Lee. 2018. Information visualization evaluation using crowdsourcing. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, Hoboken, NJ, USA, 573–595. https://doi.org/10.1111/cgf.13444

[5] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

[6] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[7] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron Hertzmann. 2017. Learning visual importance for graphic designs and data visualizations. In *Proc. of the Annual Symposium on User Interface Software and Technology (UIST)*. ACM, NY, USA, 57–69. https://doi.org/10.1145/3126594.3126653

[8] Barbara A Carkenord. 2009. *Seven steps to mastering business analysis*. J. Ross Publishing, FL, USA.

[9] Nick Cawthon and Andrew Vande Moere. 2007. The effect of aesthetic on the usability of data visualization. In *Proc. of the International Conference Information Visualization (IV)*. IEEE, NY, USA, 637–648. https://doi.org/10.1109/IV.2007.147

[10] Xi Chen, Wei Zeng, Yanna Lin, Hayder Mahdi Al-maneea, Jonathan Roberts, and Remco Chang. 2020. Composition and Configuration Patterns in Multiple-View Visualizations. arXiv:2007.15407 [cs.HC]

[11] Zhutian Chen, Wai Tong, Qianwen Wang, Benjamin Bach, and Huamin Qu. 2020. Augmenting Static Visualizations with PapARVis Designer. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376436

[12] Zhutian Chen, Yun Wang, Qianwen Wang, Yong Wang, and Huamin Qu. 2019. Towards Automated Infographic Design: Deep Learning-based Auto-Extraction of Extensible Timeline. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 917–926. https://doi.org/10.1109/TVCG.2019.2934810

[13] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, NY, USA, 539–546. https://doi.org/10.1109/CVPR.2005.202

[14] Victor Dibia and Çağatay Demiralp. 2019. Data2Vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE Computer Graphics and Applications* 39, 5 (2019), 33–46. https://doi.org/10.1109/MCG.2019.2924636

[15] Isaac Dinner, Eric J Johnson, Daniel G Goldstein, and Kaiya Liu. 2011. Partitioning default effects: why people choose not to choose. *Journal of Experimental Psychology: Applied* 17, 4 (2011), 332. https://doi.org/10.1037/a0024354

[16] Karen Dion, Ellen Berscheid, and Elaine Walster. 1972. What is beautiful is good. *Journal of personality and social psychology* 24, 3 (1972), 285. https://doi.org/10.1037/h0033731

[17] Gustav Theodor Fechner. 1860. *Elemente der psychophysik*. Vol. 2. Breitkopf u. Härtel, Germany.

[18] Stephen Few and Perceptual Edge. 2011. The chartjunk debate. Retrieved August 20, 2020 from https://www.perceptualedge.com/articles/visual_business_intelligence/the_chartjunk_debate.pdf

[19] Stephen Few and Perceptual Edge. 2016. Bar Widths and the Spaces in Between. Retrieved August 20, 2020 from https://www.perceptualedge.com/articles/visual_business_intelligence/bar_widths.pdf

[20] Xin Fu, Yun Wang, Haoyu Dong, Weiwei Cui, and Haidong Zhang. 2019. Visualization Assessment: A Machine Learning Approach. In *Proc. of the IEEE Visualization Conference (VIS)*. IEEE, NY, USA, 126–130. https://doi.org/10.1109/VISUAL.2019.8933570

[21] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. 2015. Learning to rank for blind image quality assessment. *IEEE Transactions on Neural Networks and Learning Systems* 26, 10 (2015), 2275–2290. https://doi.org/10.1109/TIP.2017.2708503

[22] Daniel Haehn, James Tompkin, and Hanspeter Pfister. 2018. Evaluating 'graphical perception'with CNNs. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 641–650. https://doi.org/10.1109/TVCG.2018.2865138

[23] Hammad Haleem, Yong Wang, Abishek Puri, Sahil Wadhwa, and Huamin Qu. 2019. Evaluating the readability of force directed graph layouts: A deep learning approach. *Computer Graphics and Applications* 39, 4 (2019), 40–53. https://doi.org/10.1109/MCG.2018.2881501

[24] Lane Harrison, Katharina Reinecke, and Remco Chang. 2015. Infographic aesthetics: Designing for the first impression. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1187–1190. https://doi.org/10.1145/2702123.2702545

[25] Jeffrey Heer and Maneesh Agrawala. 2006. Multi-scale banking to 45 degrees. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 701–708. https://doi.org/10.1109/TVCG.2006.163

[26] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. 2009. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1303–1312. https://doi.org/10.1145/1518701.1518897

[27] Jane Hoffswell, Wilmot Li, and Zhicheng Liu. 2020. Techniques for Flexible Responsive Visualization Design. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376777

[28] Aspen K Hopkins, Michael Correll, and Arvind Satyanarayan. 2020. VisuaLint: Sketchy In Situ Annotations of Chart Construction Errors. *Computer Graphics Forum* 39, 3 (2020), 219–228. https://doi.org/10.1111/cgf.13975

[29] Kevin Hu, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. 2019. VizML: A Machine Learning Approach to Visualization Recommendation. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300358

[30] Kevin Hu, Snehalkumar'Neil'S Gaikwad, Madelon Hulsebos, Michiel A Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300892

[31] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, NY, USA, 133–142. https://doi.org/10.1145/775047.775067

[32] Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *Proc. of the International Conference on Machine Learning (ICML)*. PMLR, Georgia, USA, 2525–2534. http:

//proceedings.mlr.press/v80/katharopoulos18a.html
[33] Helen Kennedy and Rosemary Lucy Hill. 2018. The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology* 52, 4 (2018), 830–848. https://doi.org/10.1177/0038038516674675
[34] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–40. https://doi.org/10.1145/3131275
[35] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Proc. of the European Conference on Computer Vision*. Springer, London, UK, 662–679. https://doi.org/10.1007/978-3-319-46448-0_40
[36] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement?. In *Proc. of the International Conference on Research & Development in Information Retrieval (SIGIR)*. ACM, NY, USA, 805–814. https://doi.org/10.1145/3209978.3210033
[37] Chufan Lai, Zhixian Lin, Ruike Jiang, Yun Han, Can Liu, and Xiaoru Yuan. 2020. Automatic Annotation Synchronizing with Textual Description for Visualization. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376443
[38] Ann Langley. 1995. Between'paralysis by analysis' and'extinction by instinct'. *MIT Sloan Management Review* 36, 3 (1995), 63. https://doi.org/10.1016/0024-6301(95)94294-9
[39] Po-shen Lee, Jevin D West, and Bill Howe. 2017. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data* 4, 1 (2017), 117–129. https://doi.org/10.1109/TBDATA.2017.2689038
[40] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. 2020. Attribute-conditioned Layout GAN for Automatic Graphic Design. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.2999335 Early Access.
[41] Halden Lin, Dominik Moritz, and Jeffrey Heer. 2020. Dziban: Balancing Agency & Automation in Visualization Design via Anchored Recommendations. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376880
[42] Tie-Yan Liu. 2011. *Learning to rank for information retrieval*. Springer Science & Business Media, Berlin, Germany.
[43] Min Lu, Chufeng Wang, Joel Lanir, Nanxuan Zhao, Hanspeter Pfister, Daniel Cohen-Or, and Hui Huang. 2020. Exploring Visual Information Flows in Infographics. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376263
[44] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. Deepeye: Towards automatic data visualization. In *Proc. of the International Conference on Data Engineering (ICDE)*. IEEE, NY, USA, 101–112. https://doi.org/10.1109/ICDE.2018.00019
[45] David McCandless. 2009. *Information is Beautiful*. Collins, Scotland, UK.
[46] Andrew Vande Moere, Martin Tomitsch, Christoph Wimmer, Boesch Christoph, and Thomas Grechenig. 2012. Evaluating the effect of style in information visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2739–2748. https://doi.org/10.1109/TVCG.2012.221
[47] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing visualization design knowledge as constraints: Actionable and extensible models in Draco. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 438–448. https://doi.org/10.1109/TVCG.2018.2865240
[48] Tamara Munzner. 2014. *Visualization analysis and design*. CRC press, Boca Raton, FL, USA.
[49] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning layouts for single-page graphic designs. *IEEE Transactions on Visualization and Computer Graphics* 20, 8 (2014), 1200–1213. https://doi.org/10.1109/TVCG.2014.48
[50] Jorge Poco and Jeffrey Heer. 2017. Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images. In *Computer Graphics Forum*, Vol. 36. The Eurographics Association & John Wiley & Sons, Ltd., NJ, USA, 353–363. https://doi.org/10.1111/cgf.13193
[51] Chunyao Qian, Shizhao Sun, Weiwei Cui, Jian-Guang Lou, Haidong Zhang, and Dongmei Zhang. 2020. Retrieve-Then-Adapt: Example-based Automatic Generation for Proportion-related Infographics. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030448 Early Access.
[52] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *The VLDB Journal* 29 (2020), 93–117. https://doi.org/10.1007/s00778-019-00588-3
[53] Annemarie Quispel, Alfons Maes, and Joost Schilperoord. 2016. Graph and chart aesthetics for experts and laymen in design: The role of familiarity and perceived ease of use. *Information Visualization* 15, 3 (2016), 238–252. https://doi.org/10.1177/1473871615606478
[54] Irene Reppa and Siné McDougall. 2015. When the going gets tough the beautiful get going: aesthetic appeal facilitates task performance. *Psychonomic bulletin & review* 22, 5 (2015), 1243–1254. https://doi.org/10.3758/s13423-014-0794-z
[55] David M Rouse, Romuald Pépion, Patrick Le Callet, and Sheila S Hemami. 2010. Tradeoffs in subjective testing methods for image and video quality assessment. In *Human Vision and Electronic Imaging XV*, Vol. 7527. International Society for Optics and Photonics, CA, USA, 75270F. https://doi.org/10.1117/12.845389
[56] Bahador Saket, Alex Endert, and Çağatay Demiralp. 2018. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 7 (2018), 2505–2512. https://doi.org/10.1109/TVCG.2018.2829750
[57] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 341–350. https://doi.org/10.1109/TVCG.2016.2599030
[58] Arvind Satyanarayan, Ryan Russell, Jane Hoffswell, and Jeffrey Heer. 2015. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 659–668. https://doi.org/10.1109/TVCG.2015.2467091
[59] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2020. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030403 Early Access.
[60] Xinhuan Shu, Aoyu Wu, Junxiu Tang, Benjamin Bach, Yingcai Wu, and Huamin Qu. 2020. What Makes a Data-GIF Understandable? *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030396 Early Access.
[61] Tableau. 2020. Visual Best Practices. Retrieved Aug 20, 2020 from https://help.tableau.com/current/blueprint/en-us/bp_visual_best_practices.htm
[62] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320
[63] Tan Tang, Renzhong Li, Xinke Wu, Shuhan Liu, Johannes Knittel, Steffen Koch, Thomas Ertl, Lingyun Yu, Peiran Ren, and Yingcai Wu. 2020. PlotThread: Creating Expressive Storyline Visualizations using Reinforcement Learning. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030467 Early Access.
[64] Kristi Tsukida and Maya R Gupta. 2011. *How to analyze paired comparison data*. Technical Report. Washington University Seattle Dept of Electrical Engineering.
[65] Edward R Tufte. 2001. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT, USA.
[66] Sara Vaca. 2018. Difference between Graphic Design and Data Visualization. Retrieved Aug 20, 2020 from https://www.saravaca.com/project/graphicdesign-dataviz/
[67] Maria V Valueva, NN Nagornov, Pave A Lyakhov, Georgiy V Valuev, and Nikolay I Chervyakov. 2020. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation* 177 (2020), 232–243. https://doi.org/10.1016/j.matcom.2020.04.031
[68] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, NY, USA, 1386–1393. https://doi.org/10.1109/CVPR.2014.180
[69] Yong Wang, Zhihua Jin, Qianwen Wang, Weiwei Cui, Tengfei Ma, and Huamin Qu. 2019. Deepdrawing: A deep learning approach to graph drawing. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 676–686. https://doi.org/10.1109/TVCG.2019.2934798
[70] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2019. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 895–905. https://doi.org/10.1109/TVCG.2019.2934398
[71] Aoyu Wu, Wai Tong, Tim Dwyer, Bongshin Lee, Petra Isenberg, and Huamin Qu. 2020. MobileVisFixer: Tailoring Web Visualizations for Mobile Phones Leveraging an Explainable Reinforcement Learning Framework. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030423 Early Access.
[72] Qing-Song Xu and Yi-Zeng Liang. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* 56, 1 (2001), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2
[73] Andre Ye. 2020. *Easy Ways to Make Your Charts Look More Professional*. Retrieved August 20, 2020 from https://towardsdatascience.com/easy-ways-to-make-your-charts-look-more-professional-9b081655eae7
[74] Mingqian Zhao, Huamin Qu, and Michael Sedlmair. 2019. Neighborhood Perception in Bar Charts. In *Proc. of the Conference on Human Factors in Computing Systems (CHI)*. ACM, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300462
[75] Nanxuan Zhao, Ying Cao, and Rynson WH Lau. 2018. What characterizes personalities of graphic designs? *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–15. https://doi.org/10.1145/3197517.3201355