# Discriminability Tests for Visualization Effectiveness and Scalability

Rafael Veras and Christopher Collins

**Abstract**—The scalability of a particular visualization approach is limited by the ability for people to discern differences between plots made with different datasets. Ideally, when the data changes, the visualization changes in perceptible ways. This relation breaks down when there is a mismatch between the encoding and the character of the dataset being viewed. Unfortunately, visualizations are often designed and evaluated without fully exploring how they will respond to a wide variety of datasets. We explore the use of an image similarity measure, the Multi-Scale Structural Similarity Index (MS-SSIM), for testing the discriminability of a data visualization across a variety of datasets. MS-SSIM is able to capture the similarity of two visualizations across multiple scales, including low level granular changes and high level patterns. Significant data changes that are not captured by the MS-SSIM indicate visualizations of low discriminability and effectiveness. The measure's utility is demonstrated with two empirical studies. In the first, we compare human similarity judgments and MS-SSIM scores for a collection of scatterplots. In the second, we compute the discriminability values for a set of basic visualizations and compare them with empirical measurements of effectiveness. In both cases, the analyses show that the computational measure is able to approximate empirical results. Our approach can be used to rank competing encodings on their discriminability and to aid in selecting visualizations for a particular type of data distribution.

**Index Terms**—Scalability, Discriminability, Simulation, Perception

## 1 Introduction

One measure of a visualization's effectiveness is whether data changes result in equivalent and perceptible visual changes. That is, can a viewer discriminate between visualizations of different data? Discriminability can be limited by both data scale (e.g., too much data results in overplotting such that views look the same) and by the perceptual scalability of the technique (e.g., small data changes result in changes to hue, position, or other encodings which are too small for a viewer to detect). Whether a visualization is robustly discriminable across the range of possible datasets is challenging to know at design time and expensive to empirically evaluate. In this work we introduce a process for using the Multi-Scale Structural Similarity Index (MS-SSIM), borrowed from image quality analysis, with simulated data variations, to estimate the discriminability of visual encodings. We evaluate our approach with a crowdsourced experiment comparing human similarity judgments to the MS-SSIM approach, and by using MS-SSIM to approximate previously reported empirical measures of effectiveness.

One predominant process for visualization design and validation is described by Munzner's nested model [30]. This model is structured as nested steps for visualization design and methods for validating each step: a) domain problem and data characterization; b) operation and data type abstraction; c) visual encoding and interaction design; d) algorithm design. At the first level, the designer "must learn about the tasks and the data of target users in some particular domain". Also known as elicitation of requirements, this phase borrows methods from human-centred design, such as ethnographic studies.

We argue that, in practice, this step is conflated into learning about the tasks of the users *in detriment* of the data. The very nested model is a victim of this reduction: the output of step (a) is a "set of questions asked about or actions carried out by the target users for some heterogeneous data". Note how the *characterization* of data is not present in the output. In the next level, operation and data type abstraction, the output is a description of operations and data types. Characterizing data is thus reduced to descriptions of data type. This gap gives rise to what we call *exemplary datasets*, a small collection of datasets taken as representative of the population and which the rest of the design process, including evaluation, is based upon. The outcome of the design process is commonly overfit to these few datasets.

The narrow scope of evaluation in the data axis threatens the validity of research claims and the robustness of visualization products. It affects any new encodings or techniques that are expected to be effective over a large range of data. In statistical terms, an exemplary dataset is only a single outcome of the random process that governs the data, and the more dimensions involved, the broader is the data universe. Thus, validity depends on the relation between the tested data and the possible data.

To strengthen validity, we can evaluate a visualization against many datasets, either real or produced by simulation. However, a large collection of test datasets across a large number of participants creates scalability problems for user studies. In this research we propose an automated way to evaluate visualization effectiveness and scalability against large data collections. We contribute a) a procedure for quantifying the discriminability of visual encodings based on a computational measure of image similarity; b) validation of this measure against empirical plot similarity data; and c) validation of the discriminability scores against empirical data on the effectiveness of visual encodings.

## 2 Related Work

In this section, we position our work in the taxonomy of quality measures proposed by Behrisch et al. [3], and use some of its categories to discuss how our discriminability measure relates to existing measures. We also review how the concept of discriminability has appeared in previous work.

### 2.1 Quality Measures

Most quality measures for visualization score the quality of a single view of a dataset. They are numerical functions whose arguments include an image or some visualization description, a dataset, and a task. In contrast, discriminability only makes sense if computed over a *family* of datasets. Our discriminability measure scores the quality of an *encoding* given an arbitrarily large data space. Despite this distinction in scope, many measures have been proposed that have features in common with our measure.

Behrisch et al. [3] showed that most quality measures in the visualization literature were not developed for evaluation purposes; instead, their purpose is to enable querying of visual patterns or automated recommendations. For example, assuming a user is looking to find clusters and there are too many possible views of a large dataset, the Hough Space measure for parallel coordinate plots [37] can be used to automatically find views where clusters are well-defined. Most measures are specific to task and/or encoding. There are measures for scatterplots [4, 45], node-link diagrams [11], line charts [35], and

• *The authors are with Ontario Tech University.*
*E-mail: rafael.verasguimaraes@uoit.ca; christopher.collins@uoit.ca*

*Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.*
*Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TVCG.2019.2934432*

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.
1077-2626 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

cartograms [1], to name a few; and for class separation [37], outlier detection [21], and change detection [38].

A few measures that are more general in scope could be used to evaluate new encodings. The information-theoretic measures of Chen and Jänicke [8], which include *Visualization Capacity* and *Visual Mapping Ratio*, attempt to quantify the intrinsic "power" of a visual encoding and its quality given a particular dataset. Similarly, the saliency measure proposed by Jänicke and Chen [20], evaluates the match between the saliency of a visualization against the saliency of the underlying data. Tufte proposed the *Data-ink Ratio*, a measure grounded on the principle of minimalism that penalizes visual embellishments [39]. In addition, several measures were proposed for quantifying crowding, occlusion, or overplotting [4, 5, 12].

Unfortunately, we have not observed these measures being applied to the evaluation of new encodings. One reason may be the fact that none has been empirically validated; that is, the connections between measure and effectiveness were not established.

### 2.2 Discriminability

Discriminability has been studied in visualization as a property of visual channels—how many distinguishable values they can be divided into [31]—and as a property of encodings—how to bind visual values to data values so as to ensure differences in data values can be perceived well [9, 10]. It has been extensively used in the design and evaluation of color encodings [36], including sequential [6, 25] and categorical color palettes [6, 13, 24], and as a criterion for texture design [18] and glyph design [47]. Rensink suggested that discriminability should be considered as one of the evaluation measures of a scientific framework for visualization [34].

The graphical inference framework [44], which proposes plot "lineups" as a method for evaluating the significance of visual discoveries, relies on encoding discriminability. Within this framework, Hofmann et al. [17] computed the "power" of competing visual encodings, which is the extent to which they enable the identification of the observed data in a line-up of plausible distractors. This power—which stems from statistical power—is, in essence, discriminability.

In vision science, discriminability is often measured with the computation of just noticeable differences (JND). The higher the JND, the lower the discriminability. The JND paradigm appears frequently in visualization research (e.g., [2, 36, 46]).

## 3 The Discriminability Criterion

At the core of the discriminability criterion is the premise that visualizations are visual embeddings of data [10]; as such, the notion of visualization quality is fundamentally tied to the preservation of structures in the data. An important consequence of the visual embedding model is that changes in data should always yield perceptible changes in the visual representation. This requirement is known as the *principle of unambiguous data depiction* [23] and, in theory, can be assessed with a discriminability test. Likewise, the principle of *visual-data correspondence*, which states that changes in data should yield changes of equivalent magnitude in the visual representation, can in theory be verified in terms of discriminability.

Low discriminability indicates that changes in data values result in low visual change, implying that viewers may have trouble decoding values accurately (due to ambiguity). It can indicate common visual mapping problems, such as low utilization of the space, narrow encoding range, overplotting, and high clutter. Discriminability is one of the most fundamental quality dimensions in visualization because a difficulty to decode values may affect other tasks, such as outlier detection, estimation of means, and visual comparison.

It is also closely linked to visualization scalability. With large datasets, most visualizations that do not rely on aggregation or sampling *saturate*. This saturation can be described as follows: a saturated plot causes a family of datasets to be mapped to identical or similar plots. Consider the common "hairball graph" example, where adding more nodes to a saturated graph does not change the appearance. A discriminability measure can determine the saturation levels of a visualization at increasing data scales. The curve formed by discriminability over data scale can be understood as a description of visual scalability.

Unlike measures that stem from vision science (e.g., clutter, saliency), which are applicable to all images (including natural images), discriminability is a quality of the visual mapping, a visualization-native property. In particular, the level of clutter of a plot may say something about its visual quality, but it says nothing about the visualization method as an instrument to observe data.

For the scope of this paper, we propose the following definition of discriminability:

**Discriminability** Given a collection of datasets, the average perceptual distance between the corresponding visualizations.

Thus, a discriminability test comprises a data scope, which is given by the collection of datasets, a visual encoding method, and a visualization similarity (or distance) measure. The similarity measure can be either an empirical function—judgments are collected through user studies—or a computational measure that approximates the empirical judgments. Because our motivation is to discover cheap and fast methods for evaluating new encodings, we will examine here the use of a computational measure of similarity.

Alternatively, discriminability could be defined in terms of the average data distance needed to produce a just noticeable difference in the visualization. Or, given a seed dataset and corresponding visualization, the effort needed to produce a second dataset (beyond a certain data distance) that yields an ambiguous visualization. If an intelligent agent is trained to generate such ambiguity-inducing dataset pairs, the effort could be measured in terms of model complexity. This ambiguity induction is conceptually the same procedure proposed by Matejka and Fitzmaurice [26] to generate wildly different scatterplots that have the same statistics.

## 4 Measuring Structural Similarity

In this section, we investigate in depth the possibility of a computational measure of similarity that can approximate human perceived similarity. With such a measure we could perform large-scale discriminability tests, involving not only variation in dataset size, but also in dataset distribution, entropy, etc.

The Structural Similarity Index (SSIM) was developed for quality assessment of compressed images [40]. Different than previous measures (e.g., mean squared error, and peak signal-to-noise ratio) that assumed that the perception of image quality depends on the visibility of errors, SSIM assumes that image quality depends on the preservation of structural information. As such, image quality can be quantified by a general measure of structural similarity between the original image and the compressed images. While the error-sensitivity paradigm tries to reproduce early-stage, low-level processing of the human visual system, such as thresholding informed by psychophysical experiments, the structural similarity paradigm tries to emulate the hypothesized *function* of the overall human visual system. This function consists in probing the structures of observed objects.

The SSIM is defined as the weighted product of luminance similarity, contrast similarity, and structural similarity.

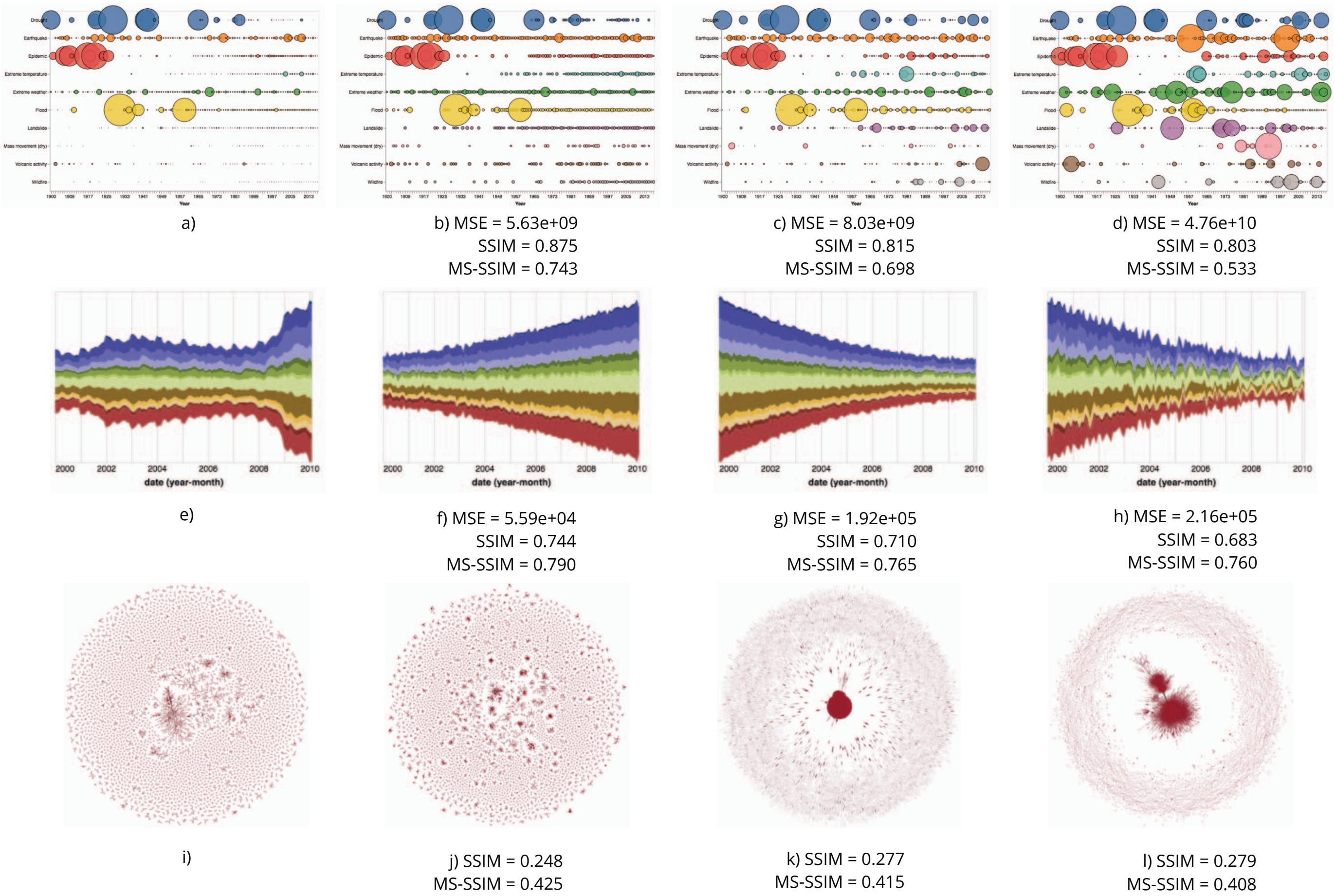
$$SSIM(x,y) = l(x,y)^{\alpha} c(x,y)^{\beta} s(x,y)^{\gamma} \quad (1)$$

where $x \in R^D$ and $y \in R^D$ are vectors (of the same size) containing the grayscale pixel intensities of each image. The SSIM calculation normalizes the images with respect to luminance in the contrast similarity calculation, and then normalizes the images with respect to contrast in the structural similarity step. This way, the similarity components are made independent. We can think of Equation 1 as a pipeline (from left to right) where a feature is subtracted after it has been the subject of a similarity assessment.

Luminance $\mu$ is the mean pixel intensity, and luminance similarity is defined as follows:

$$l(x,y) = \frac{2\mu_x\mu_y}{{\mu_x}^2 + {\mu_y}^2} \quad (2)$$

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.



a)

b) MSE = 5.63e+09
SSIM = 0.875
MS-SSIM = 0.743

c) MSE = 8.03e+09
SSIM = 0.815
MS-SSIM = 0.698

d) MSE = 4.76e+10
SSIM = 0.803
MS-SSIM = 0.533

e)

f) MSE = 5.59e+04
SSIM = 0.744
MS-SSIM = 0.790

g) MSE = 1.92e+05
SSIM = 0.710
MS-SSIM = 0.765

h) MSE = 2.16e+05
SSIM = 0.683
MS-SSIM = 0.760

i)

j) SSIM = 0.248
MS-SSIM = 0.425

k) SSIM = 0.277
MS-SSIM = 0.415

l) SSIM = 0.279
MS-SSIM = 0.408

Fig. 1: Data and image similarity measures: Mean-Squared Error (MSE), Structural Similarity Index (SSIM), and Multi-Scale SSIM (MS-SSIM). Leftmost images in each row are the references. Top: global deaths from natural disasters (Vega-lite gallery) and simulated perturbations. Middle: unemployment across industries (Vega-lite gallery) and simulated perturbations. Bottom: graphical models of passwords [48]. MSE is inversely proportional to similarity. MS-SSIM weights: [0.1, 0.1, 0.1, 0.2, 0.5].

where $x$ and $y$ are vector representations of the images. Contrast is estimated as the standard deviation of the pixel intensities. Note that the standard deviation $\sigma$ inherently subtracts the mean intensity (luminance) from the signal. Contrast similarity is defined analogously to luminance similarity:

$$c(x,y) = \frac{2\sigma_x\sigma_y}{{\sigma_x}^2 + {\sigma_y}^2} \tag{3}$$

Finally, the structural similarity function operates on the signal normalized by luminance and contrast: $(x-\mu_x)/\sigma_x$. Readers familiar with machine learning will recognize this operation as *standardization*, which yields a z-score. The structural similarity is the correlation (inner product) of these normalized vectors:

$$s(\sigma_x,\sigma_y) = \frac{1}{D-1}\sum_{i=1}^{D}\frac{(x_i-\mu_x)}{\sigma_x}\frac{(y_i-\mu_y)}{\sigma_y}. \tag{4}$$

The SSIM is then computed in a local fashion (per pixel) with a 3x3 Gaussian window. This yields a similarity map over the image. The overall image similarity measure, a scalar value, is the mean similarity of this map:

$$\text{Mean-SSIM}(X,Y) = \frac{1}{M}\sum_{j=1}^{M}\text{SSIM}(x_j,y_j) \tag{5}$$

where $M$ is the number of Gaussian windows, $X$ and $Y$ are the images, and $x_j$ and $y_j$ are the image patches defined by each of the $M$ windows. When zero-padding is used, $M = D$. Despite the parent-child relation, the acronym SSIM usually refers to Mean-SSIM, and the distinction is rarely in effect. In this paper, we follow this convention. When the context suggests SSIM is a scalar value, it refers to the Mean-SSIM.

The SSIM is symmetrical, bounded, and has a unique maximum. The index lies in the interval $[-1,1]$ and a comparison between two identical images will always yield 1.

### 4.1 Multiscale-SSIM

Recall that the SSIM was created to measure the encoding quality of natural images, which depends on the impact of imperfections introduced by the encoding. Clearly, the perception of quality depends on the viewing distance, given that some imperfections are only noticeable at close inspection. In general, we can say that the perception of quality and similarity depends on the *scale* of the image, which varies with viewing distance or image size. Recognizing the challenges of assessing image quality at a single scale, Wang et al. [42] proposed Multi-Scale SSIM. This technique is a straightforward extension of SSIM where the contrast and structural similarities are computed at $K$ image scales. The original image is subject to low-pass filtering and downsampling by a factor of 2 in each of $K-1$ steps.

$$MS\text{-}SSIM(X,Y) = l(x,y)^{\alpha}\prod_{i=1}^{K}c(x_i,y_i)^{\beta_i}s(x_i,y_i)^{\gamma_i} \tag{6}$$

The weights indexed by $i$ are adjusted according to the desired relative importance of the scales to the similarity judgment. For simplicity,

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

and following Wang et al. [42], we always set $\alpha = 1$, and $\beta = \gamma = w_i$ within each scale:

$$MS\text{-}SSIM(X,Y) = l(x,y)\prod_{i=1}^{K}\Big(c(x_i,y_i)s(x_i,y_i)\Big)^{w_i} \qquad (7)$$

Throughout this paper we will use vector notation to communicate the scale parameters; for instance, in the parameter array $W = [w_1, w_2, ..., w_n]$, $w_1$ is the weight on the finest, detailed view (largest image), while $w_n$ is the weight on the coarsest, distant view (smallest image).

## 4.2 Comparing SSIM and MS-SSIM

To begin assessing the utility of SSIM as a measure of visualization similarity we designed a small sanity test. We chose two visualizations from the Vega-lite visualization gallery [19], a bubble chart and a stream chart, and produced data perturbations of different magnitudes. Then we measured the similarity between the visualizations of the perturbed data and the original visualization. These visualizations have encodings of different nature: point and area. We added also a third set of visualizations, which consists of plots of graphical models of password lists [48]. They were chosen because they are dense representations that tend to form distinct shapes.

Figure 1 shows the mean squared errors (MSE) computed on the dataset pairs, and both MS-SSIM and SSIM computed on the corresponding visualization pairs. The MSE summarizes the differences in values from one dataset to the other. In this experiment, it represents the baseline or true dataset difference. Most charts of unaggregated data where clutter is not an issue should allow us to recover, with some effort, the MSE between two datasets by mapping the visual marks back to data values and computing the measure. In fact, there are tools designed with the specific purpose of extracting data from existing visualizations [15, 27].

The SSIM produced similarity rankings that reflect the MSE rankings in the bubble chart and stream chart cases: larger SSIM values should correspond to lower MSE values. In the dense graph case the true data similarity is unknown (as we only have the images), so we will resort to a qualitative assessment. It is rather clear that two of the plots (k and l) feature a dense central region that forms a solid red blob, while the other two plots (i and j) feature a more well-distributed pattern. The output of the SSIM comparisons indicates that this notion is not captured by the measure; the graph that is perceived as most similar (j) to the reference (i) received the lowest SSIM score.

It is plausible that the similarity of plots is judged at different scales depending on the kind of plot. For instance, dense graphs form distinct global shapes that override local similarity comparisons. Other visualizations, such as scatterplots, may or may not form global shapes. When a global shape is not formed, the similarity judgment is done at a lower level, by scanning the scene in search of differences, a process that is well captured by the windowed calculation of SSIM.

MS-SSIM is built on the premise that *viewing conditions* determine the right scale. We, instead, posit that at *identical* viewing conditions the scale in which similarity judgments varies with the chart type. We customized the weights as following, so as to give more importance to differences in coarser features: [0.1, 0.1, 0.1, 0.2, 0.5]. The resulting scores (Figure 1) reflect the correct similarity ordering of the dense graphs. As a bonus, the MS-SSIM scores also comply with the correct data MSE ranks for the stream charts and bubble charts.

## 4.3 Limitations

Fundamental limitations arise when the SSIM is applied to data plots. In natural images every pixel counts towards a similarity judgment, although some extensions of the SSIM recognize that some regions matter more than others and attempt to weigh their importance based on saliency [29], recognized objects [32], and information theoretic measures [41]. In data plots, this characteristic manifests adversely as a hypersensitivity to visual accessories, such as grids and labels. Figure 2 (a-d) displays scatterplots of the Iris dataset that feature a grid. Note how the SSIM values do not correspond to the visual similarity of the plots. Upon close inspection we see that the grids, which are not consistently positioned, contribute disproportionately to the measurement.

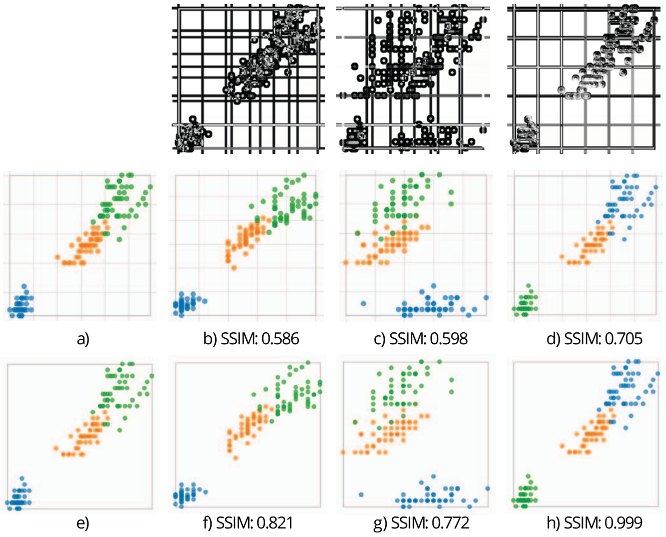

a)
b) SSIM: 0.586
c) SSIM: 0.598
d) SSIM: 0.705
e)
f) SSIM: 0.821
g) SSIM: 0.772
h) SSIM: 0.999

Fig. 2: The effect of grids on SSIM for scatterplots of the Iris dataset. (a-d): SSIM with grids. Top row: local SSIM values for these plots (brighter is more similar). (e-h): the same comparison without grids. Scores are relative to the leftmost plots in each row.

In the context of the proposed use of the measure, the discriminability tests, the tester has control over the production of the images, so the hypersensitivity problem can be completely disregarded if we assume that for testing purposes, the plots are generated without grids, labels, and other accessories. We also assume that the presence of these accessories would not change the result of a comparative evaluation. This is consistent with the findings of Bostock and Heer [16], who reported an effect of gridlines on effectiveness, but no interaction with chart type.

Better measurements are achieved by simply turning the grid off (Figure 2 (e-h)). However, this figure illustrates a more complicated limitation. The scatterplot labelled (d) is a clone of (a) that had the color mapping inverted (blue and green swapped); therefore, (d) should not be judged identical to (a), as the SSIM value implies. The SSIM operates on grayscale images and it is not capable of capturing changes in hue.

The color limitation does not affect color encodings of numerical, continuous data attributes which employ color schemes that vary luminance and saturation. It affects more strongly visualizations that use nearly equiluminant categorical color palettes. In the next section, we propose a modification to SSIM that addresses its “color blindness”.

## 4.4 SSIM on YUV Color Space

Since in the next sections we will be investigating the discriminability of a broad set of encodings with various color mappings, it is important to establish a more general use of SSIM that can accommodate both categorical and continuous color mappings.

Our goal is to introduce *some* sensitivity to color by using a color space where color components are represented independently from luminance. The YUV color space is well aligned with this goal, since it consists of a luminance component (Y), and two chrominance components (UV). Black and white images use only the Y component, so the original SSIM is equivalent to the computation on the Y channel.

We compute the SSIM on the YUV space by simply averaging the similarities computed in each color space component (Y, U, and V) independently. The computations on U and V can be interpreted as an assessment of the similarity existing in color structure. Thus, the proposed calculation consists in the *smoothing* of the original SSIM with values computed on color channels. Other color spaces that repre-

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

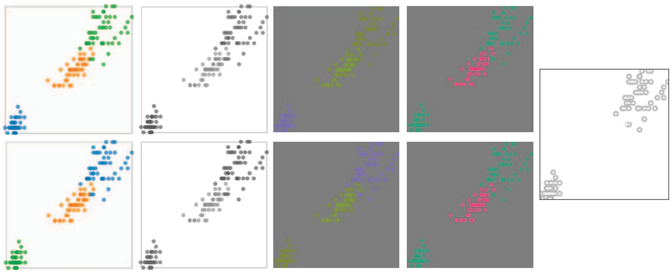


Fig. 3: SSIM applied on YUV image representations. Each row shows images in their original form, and decomposed into Y, U, and V channels of the YUV color space. On the right is the similarity map resulting from averaging the similarities computed on each channel independently. Note how the color difference in the original images appears in the final similarity map.

sent luminance or lightness independently are suitable, including the perceptually uniform ones; however, our calculations on individual components do not benefit from the perceptual uniformity property.

In the pathological example depicted in Figure 3, where two groups had their color swapped, this strategy is enough to prevent the visualizations from being scored identical. The standard SSIM similarity of 0.999 (Figure 2(h)) fell to 0.968 using SSIM on YUV space. SSIM on YUV preserves SSIM's characteristic of being driven by spatial structure. Additional research is needed to determine the best approach if capturing color change is of primary importance. Given that it shows improvement over basic SSIM, we chose to work with SSIM on YUV when color is involved.

## 5 Computing Plot Similarity

In this section we compare MS-SSIM judgments with empirical similarity judgments. Our goal is to test whether a parameterization of MS-SSIM is capable of approximating empirical judgments for a certain visualization type. A positive result in this validation should indicate that other parameterizations can help us approximate judgments for other visualization types, assuming that the judgments will vary mostly with respect to scale and the use of color. If instead we find that no parameter set can approximate well empirical judgments, that should prompt discussion about what factors are involved in similarity perception of data plots. This applies in particular to spatial encodings.

For this analysis we chose the data collected by Pandey et al. [33], which consists of human similarity judgments (13 participants) for a set of 247 single-color scatterplots. The similarity judgments were collected with a spatial arrangement interface in which scatterplot thumbnails are displayed in an "image carousel" and can be dragged and dropped into a large, initially empty, canvas. Participants were instructed to arrange the scatterplots into groups according to their similarity, explicitly mark the boundaries of each group, and finally, assign labels to them. Notably, they were told not to worry about within-group or between-group distances; that is, only group membership mattered.

Pandey et al. [33] calculated the consensus distances for each pair of plots as the complement of their probability of co-occurrence averaged across participants:

$$d_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\left(1 - \frac{c_{i,j}}{min(c_i, c_j)}\right)_k \tag{8}$$

where $N$ is the number of participants, $c_{i,j}$ is the number of clusters that contain both plots $i$ and $j$, and $c_i$ and $c_j$ are the number of clusters that contain the plots $i$ and $j$, respectively. Note that the interface allowed participants to assign plots to multiple groups. A hierarchical clustering of the plots based on the consensus perceptual distance matrix was calculated, and it is displayed in Figure 4a.

We compared MS-SSIM and empirical judgments using cluster quality measures, which are traditionally used to quantify the agreement between two independent label assignments on the same dataset. We selected the following measures, all of which assume the ground truth is known: adjusted mutual information (AMI), normalized mutual information (NMI), Rand index (RI), and adjusted Rand index (ARI). All measures except RI assign values close or equal to 0 to random clusterings and assign 1 to perfect clustering (relative to the ground truth). Change adjusted measures (AMI and ARI) do not exhibit a dependency between the number of clusters and the number of samples; such dependency could boost the score of random clusterings that have many groups.

Table 1: Cluster quality measures for clusterings of 247 scatter plots based on MS-SSIM. The quality measures are relative to the clustering based on human similarity judgments reported by Pandey et al. [33]. Each row corresponds to a parameter set ($w_1..w_5$). The parameters in the first row were obtained through gradient descent.

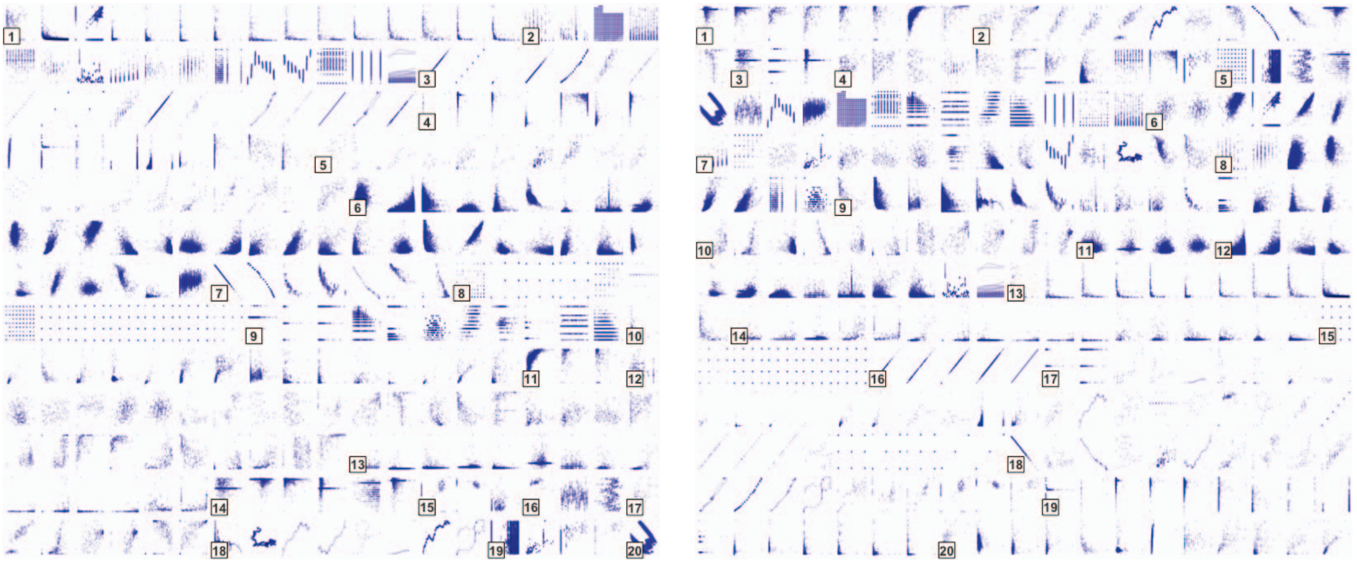
| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | **ARI** | **RI** | **AMI** | **NMI** |
|---|---|---|---|---|---|---|---|---|
| **0.32** | **0.73** | **0.82** | **1.00** | **1.00** | **0.20** | **0.90** | **0.35** | **0.51** |
| 0.10 | 0.10 | 0.10 | 0.30 | 0.40 | 0.16 | 0.86 | 0.30 | 0.46 |
| 0.10 | 0.20 | 0.20 | 0.20 | 0.30 | 0.13 | 0.83 | 0.25 | 0.42 |
| 0.10 | 0.15 | 0.15 | 0.30 | 0.30 | 0.10 | 0.81 | 0.22 | 0.40 |
| 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 0.13 | 0.81 | 0.24 | 0.42 |
| 0.40 | 0.20 | 0.20 | 0.10 | 0.10 | 0.13 | 0.81 | 0.26 | 0.44 |

In their experiments, Wang et al. [42] found optimal MS-SSIM weight parameters for natural images (0.04, 0.29, 0.30, 0.24, 0.13). In this section, MS-SSIM was set with five weight vectors manually chosen to represent different weight balancing strategies. As in the work of Wang et al., the vectors sum up to 1 and have components $< .5$. We compared these parameters with a weight vector tuned using gradient descent (the approach is described in detail in the next section). Table 1 presents the weight vectors ordered by importance on the finest scales. The clustering method was hierarchical under the Ward agglomeration strategy, with even-height tree cuts that yielded 20 clusters (the same number of clusters in the ground truth, although none of the quality measures requires an equal number of clusters).

The results (Table 1) show that the parameters found through gradient descent achieved the best fitness to the empirical clustering, as observed in all of the quality scores. The plot arrangement resulting from clustering with this best MS-SSIM parameter set is presented in Figure 4b, and the corresponding dendrogram in Figure 5. The fitted parameters and the plot arrangement comparison tell us much about the protocol used to collect the empirical measurements. First, the participants had only the chance of interacting with thumbnails, forcing them to make high-level perceptual judgments. This fact is expressed in the weights discovered with gradient descent, which clearly emphasize coarser judgments. Second, distances between and within clusters were not taken into account; as a consequence, the global structure in the empirical clusters is messy. The MS-SSIM clustering, in contrast, imposes a clear partition between dense and sparse plots (around cluster 13).

The cluster quality measures suggest a significant overlap between the clusterings, but far from full agreement. Some of the mismatch can be explained by the *cognitive interaction problem* [42], by which different user goals can result in very different judgments. Participants were not instructed to cluster plots based on *dataset similarity*. In a real-world scenario, analysts are making judgments about the data, with the visualization being a proxy. Some pairs of plots that bear some visual resemblance (in terms of shape) and are in the same empirical cluster, are unlikely to have been found similar if the question was about the underlying data. For instance, and have both a T-like shape, but represent very different relationships between the data variables. We can attribute much of the difference between the clusterings to this misalignment of goals. Empirical cluster #6, the one whose elements are spread the most across SSIM clusters, comprises elements with wildly distinct data patterns, but similar density (see Figure 4). Density-based agglomeration is still present in the SSIM clustering, but divided according to the position of the point-cloud. Likewise, empirical cluster #12 has plots with similar amount of "ink" but very different spatial

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.



(a) Empirical scatterplot clustering.

(b) MS-SSIM scatterplot clustering.

Fig. 4: Empirical and MS-SSIM clusterings of the scatterplots from the study of Pandey et al. [33]. MS-SSIM parameters were tuned to the empirical data via gradient descent. Plots are ordered from upper left to lower right in a wrapped sequence.

arrangements; it is also divided in several pieces in the MS-SSIM clusterings.

## 6 Tuning

This section explains in detail the MS-SSIM tuning procedure that we used to obtain the weights presented in the previous section. This procedure adjusts the scale weights so as to minimize the discrepancy between SSIM similarity and a set of empirical judgments. We used a vanilla stochastic numerical gradient descent algorithm, which, at each iteration, evaluates the gradient of the loss function with respect to the current parameters, then updates the parameters in the directions that reduce the loss.

Let's define a dataset of images $x_i \in R^D$, and a similarity function $s : R^D \times R^D \rightarrow R^1$. With the multi-scale SSIM, $s$ has the following form:

$$s(x_i, x_j) = \text{MS-SSIM}(x_i, x_j, W) \tag{9}$$

The above equation can be read as the similarity of $x_i$ and $x_j$ given the vector of weights $W$, which determines the importance of each scale to the overall similarity score, as seen in Section 4.1. Next, let's define a binary function that takes an image triplet $(x_i, x_j, x_k)$ and decides whether $x_i$ is more similar to $x_j$ than $x_i$ is to $x_k$:

$$f(x_i, x_j, x_k) = \mathbb{1}(s(x_i, x_j, W) \geq s(x_i, x_k, W)) \tag{10}$$

This equation embodies a triplet matching task and enables the definition of a loss function for comparison of SSIM scores with a ground truth. The ground truth data could be collected using triplet matching, triplet discrimination, spatial arrangement, or pairwise ratings on a Likert scale.

The loss function is defined as follows, where $f_{ijk}$ is an abbreviation for $f(x_i, x_j, x_k, W)$, the SSIM binary label, and $Y_{ijk}$ is the ground truth label:

$$L_{ijk}(W) = \sum_{f_{ijk} \neq Y_{ijk}} \Big( s(x_i, x_j, W) - s(x_i, x_k, W) \Big)^2 + R(W) \tag{11}$$

The loss defined in the equation above is composed of two terms, the data loss and the regularization loss. The data loss is simply the squared difference between the similarity scores when they are wrong. For instance, if $s(x_i, x_j, W) = 0.8$, $s(x_i, x_k, W) = 0.6$, and the ground truth is $s(x_i, x_j, W) < s(x_i, x_k, W)$, that is, $Y_{ijk} = 0$, then the loss is $(0.2)^2$. The regularization loss (or penalty) is a function of the weights and embeds our preference for weights in a certain range. In this case, the weights need to be between 0 and 1. The regularization loss has the following form

$$R(W) = \sum_{i=1}^{|W|} (W_i)^{\alpha-1} (1 - W_i)^{\alpha-1} \tag{12}$$

where $\alpha$ is a parameter that controls the steepness of the penalty as the values approach 0 or 1.

Table 2: Kim and Heer's experiment was divided into four tasks. $Q_1$ is a continuous variable.

| | |
|---|---|
| Read value | What is the $Q_1$ of the data point A? |
| Compare value | Which data point has more/less $Q_1$? |
| Find maximum | Which state has the data point with the highest $Q_1$? |
| Compare averages | Considering all data points for the State, which of the following two States has greater average $Q_1$? |

## 7 Discriminability and Effectiveness

The validation of the MS-SSIM against an empirical study of scatterplot similarity was useful for understanding the extent to which we can expect human similarity judgments to match MS-SSIM scores, but it did not shed light on the usefulness of discriminability as a quality criterion. We don't know if discriminability scores derived from similarities have any relationship to the effectiveness of visualizations. In this section, we seek to fill this gap.

There are a few empirical studies of the effectiveness of visualization encodings. We will base our investigation on the most recent of these studies, which has all materials publicly available [22]. As a plus, this study focused on the effect of data scale and distribution on performance, so it aligns with our interest in scalability. Kim and Heer [22] tested the effectiveness of twelve trivariate encodings, described here in the format $Q_1\_Q_2\_N$ where $Q_1$ and $Q_2$ are numerical, continuous variables, and $N$ is a categorical variable.

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

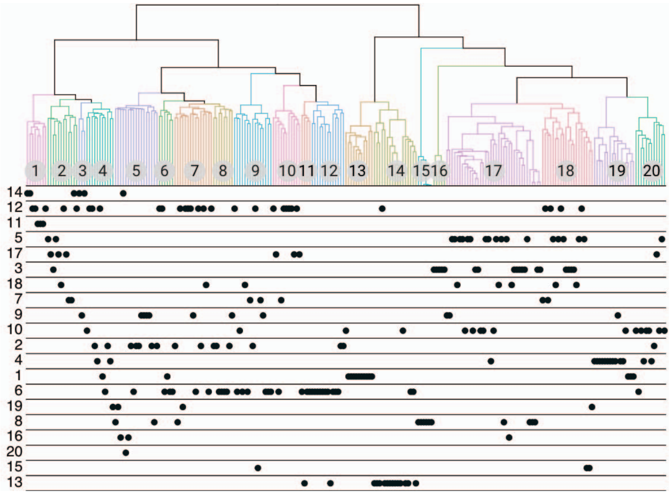


Fig. 5: A comparison of empirical and MS-SSIM clustering. The dendrogram represents the MS-SSIM clustering of scatterplots from Pandey et al. [33]. Each row in the bottom represents an empirical cluster, with each dot representing a plot. Dots are aligned with the dendrogram, allowing us to observe how the empirical clusters are disrupted by the dendrogram arrangement. If the clusterings were identical, all dots in each row would be adjacent. Rows are ordered according to leftmost match with dendrogram.

The data consists of 2016 U.S. monthly weather measurements, published as part of the Global Historical Climatology Network-Daily Database (GHCN) [28], and contains the categorical variables State and Month, and the following numerical variables: Maximum Temperature, Minimum Temperature, Average Wind Speed, Wind Direction, Strongest Gust Speed, Precipitation, Snowfall, and Snow Depth.

The stimuli of that experiment was produced by sampling from GHCN and it was divided into 24 experimental conditions that result from the crossing of the following factors: Cardinality (3, 10, 20), where cardinality is the number of categories $N$, #/Category (3, 30), $Entropy_{Q_1}$ (Low, High), and $Entropy_{Q_2}$ (Low, High). The specific variables $Q_1$ and $Q_2$ were not factors; thus, they vary randomly across stimuli. $N$ is always a derived variable resulting from the conflation of State and Month (as in TX-03), although in the stimuli it appears simply as State (participants were not exposed to Month).

Study participants were asked to perform tasks that involved questions about $Q_1$. The tasks were of the following types: Value tasks, further split into Read Value and Compare Value; and Summary tasks, further split into Find Maximum and Compare Averages. Table 2 lists the question templates for each task. Error rates and completion times were measured, and rankings of encodings were created based on the error rates.

The results of this experiment reveal that the effect of encoding on error rates depends on the task and on the various factors manipulated in the experiment; therefore, a different ranking of encodings is created within each task group and factor level. Furthermore, the differences in error rate and completion time for the encodings are not always statistically significant; for instance, in summary tasks involving datasets with three and ten categories, the ten best ranked encodings did not score significantly different error rates.

### 7.1 Measuring Discriminability

We conducted two benchmarks. The first is a *global* discriminability test, of the kind someone would run without a specific task in mind. It generates a variety of datasets then computes the average similarity across visualizations of these datasets for each encoding being considered. In essence, it measures the sensitivity of each encoding, or how much *overall* visual change we can expect of each encoding, on average. The link to effectiveness is in the assumption that the less sensitive an encoding, the harder it is to decode information: reading and comparing values is more difficult when the visual range is narrow.

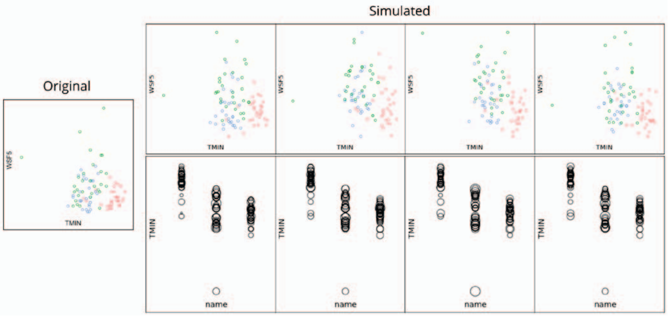


Fig. 6: Images generated for the global discriminability test. Left: Original plot used by Kim and Heer. Right: Plots depicting variations of the original data, resulting from sampling from statistical models fitted to Kim and Heer's data. Only the question variable $Q_1$ (WSF5 in this example) is simulated. The simulated data is depicted using position encoding (*y_x_color*), and size encoding (*size_y_x*) for $Q_1$.

The second experiment is task-specific. Kim and Heer's rankings for summary tasks (mean comparison and find maximum) are somewhat different than the rankings for value tasks. In the mean comparison tasks, participants are instructed to select the state with the highest mean out of only two options. It is safe to assume that in these tasks what matters is how easily people can segregate the values of the two states in question and compare their values. So we devised a scheme to test *local* discriminability in Benchmark 2.

### 7.2 Benchmark 1 - Global Discriminability

Kim and Heer's experiment is structured as follows: 8 different datasets were sampled from the GHCN records for each combination of factors *cardinality* × *#/category* × $entropy_{Q_1}$ × $entropy_{Q_2}$ × *encoding*. That is, within each condition, each encoding was tested with a different collection of datasets, all with similar characteristics (dictated by the experimental condition). The datasets vary randomly in $Q_1$, $Q_2$, and the specific data points and states that the questions center on, in order to avoid a combinatorial explosion of conditions. In the discriminability tests, we prioritized symmetry by testing all encodings within a given experimental condition on the *same* datasets. Furthermore, $Q_1$ and $Q_2$ were not varied randomly; instead, they were a factor in the experiment (between-encodings). These changes were made because the scale of the test is not a problem here, so we can test every possible cross between $Q_1$, $Q_2$, and the rest of the factors. In summary, we created 20 datasets by simulation for every combination of factors *cardinality* × *#/category* × $entropy_{Q_1}$ × $entropy_{Q_2}$ × $Q_1$ × $Q_2$.

In order to simulate data that are similar to the data used by Kim and Heer [22], we sampled values from generalized linear models (GLMs) fitted to the GHCN data. The simulation consisted in randomly drawing a dataset that matched the given experimental condition, then replacing its $Q_1$ values by values sampled from the model. The replacement

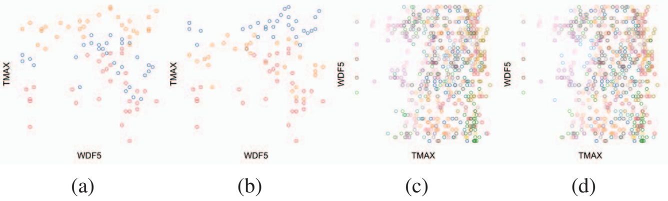


Fig. 7: Pairs of colored scatterplots (*y_x_color*) with y values swapped between two categories. a) and b) have 3 categories in total, while c) and d) have 30 categories. These pairs (a,b) and (c,d) are used to measure the visual discriminability of two categories (other categories fixed) along one variable.

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

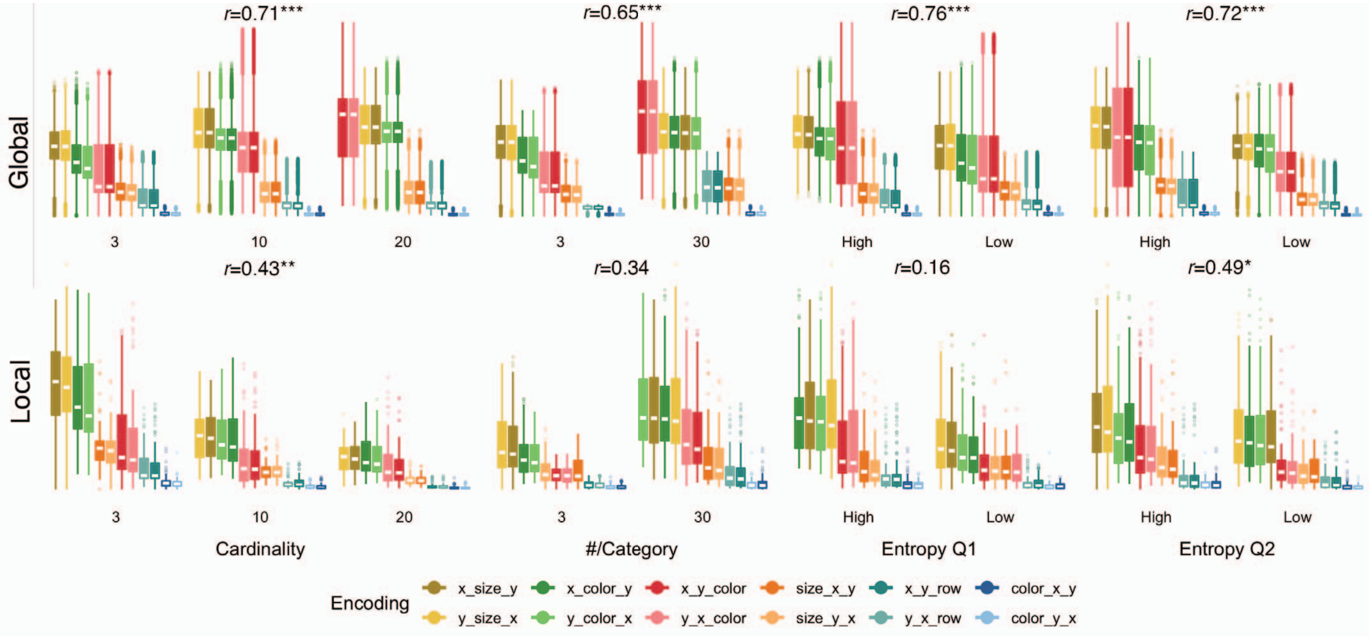


Fig. 8: Global and Local discriminability scores computed with MS-SSIM ($W = [1,1,1,1,1]$), aggregated by data attributes. For each dimension, we computed the Pearson correlation coefficient against the empirical accuracy measured by Kim and Heer [22], using the mean values for each encoding. The significance codes correspond to the null hypothesis that correlation is 0. Significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

step was repeated 20 times. The GLMs were fitted as follows. Given a condition, all records in Kim and Heer's data that match $Q_1$ were collected. Then a GLM was fitted to these records with $Q_1$ as the response variable and State as the covariate. Since all datasets have low correlation, $Q_2$ was omitted from the model; thus, the GLMs simply learn one distribution for each state. Figure 6 shows a reference dataset and simulated datasets visualized with two different encodings.

For each encoding, pairwise similarity judgments were computed with the MS-SSIM on the YUV representations of the images. We used a uniform parameter vector ($W = [1,1,1,1,1]$), which embodies the "naive" hypothesis that all scales contribute evenly to similarity judgments. We assume that, *a priori*, the parameters fitted to the scatterplots in Section 5 are not applicable because the tasks and the presentation of the stimuli are quite different. We transformed these similarities into distances by computing $d(x,y) = (1 - \text{MS-SSIM}(x,y))/2$. Each tuple (*cardinality*, *#/category*, $entropy_{Q_1}$, $entropy_{Q_2}$, $Q_1$, $Q_2$, *encoding*) yields one discriminability score computed as the mean pairwise distance over 20 images. These scores are then aggregated to produce scores per factor level, used in the rankings of encodings. We compared the results of this experiment with the effectiveness data for value tasks.

### 7.3 Benchmark 2 - Local Discriminability

As mentioned earlier, we cannot expect a general test as the one presented in Benchmark 1 to explain accurately the effectiveness of a task that requires the comparison of two sections of a visualization, because that experiment evaluated global discriminability.

In order to test the discriminability of the visual representations of the two categories *within the context of the whole plot* we devised the following testing scheme. Given a plot, a subset of two categories, and the variable $Q_1$ subject to the mean comparison, a second plot is generated where the values of $Q_1$ are swapped between the two categories. The values for $Q_2$ in both categories remain fixed, as well as all data points in all other categories. The similarity is then computed on this pair of images, effectively measuring the visual similarity of the two groups of data points in the context of the rest of the data. Figure 7 illustrates the scheme.

This test did not employ statistical simulation. We modified the same datasets that served as stimuli in Kim and Heer's experiment, which had 2,304 mean comparison tasks. In our experiment, each of these datasets was modified once, resulting in 4,608 datasets. Discriminability was calculated as the average distance (as described in Benchmark 1) between source and modified datasets. We assigned the same weights to all scales ($W = [1,1,1,1,1]$). We compared the results of this experiment with the effectiveness data for summary tasks.

### 7.4 Results and Discussion

We used Pearson correlation to compare encoding discriminability scores with empirical effectiveness (task accuracy) within each study factor. The analysis of statistically significant differences within ranks is left for future work. The global discriminability experiment produced encoding discriminability scores that are highly correlated to empirical effectiveness in value tasks (Figure 8). The derived rankings in Figure 9 are nearly identical to the value task rankings of Kim and Heer, with spatial encodings exhibiting higher discriminability compared to encodings that rely on size and color, the exception being the spatial encoding *x_y_row*, which displays categories in different axes. The full details of the benchmark scores are provided as supplementary materials.

The local discriminability experiment produced encoding discriminability scores that are mildly correlated to empirical effectiveness in summary tasks (Figure 8). In particular, discriminability scores did not account for the radical drop in accuracy (relative to value tasks) of the encoding *x_y_color* (multi-class scatterplot). This drop is due to the difficulty humans have in separating colors in displays with many colors. Neither the sharp increase in the effectiveness of *size_x_y* was observed. But curiously, the local discriminability of these encodings did suffer changes (relative to global discriminability) in the same directions observed in the empirical data, even though the magnitude was not equivalent. The failure to account for the difficulty in separating colored groups in plots with many groups [14] is the main limitation of MS-SSIM as a measure of discriminability. Future efforts should concentrate on measures that account for complex perceptual effects.

The correspondence between the rankings suggest that the effectiveness of the encodings is, to a large extent, driven by encoding discriminability. The discriminability within six out of eight factors had correlation with effectiveness higher than .4. However, fine grained changes in Kim and Heer's rankings due to data entropy and scale were not matched by the discriminability rankings. This suggests that discriminability cannot fully explain the rankings. This is to be expected, since other factors are known to influence people's judgments. Among these factors are saliency and distortions in the perception of brightness, contrast, length, and area (as described by Steven's law).

MS-SSIM can be interpreted as an inverse measure of the strength

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

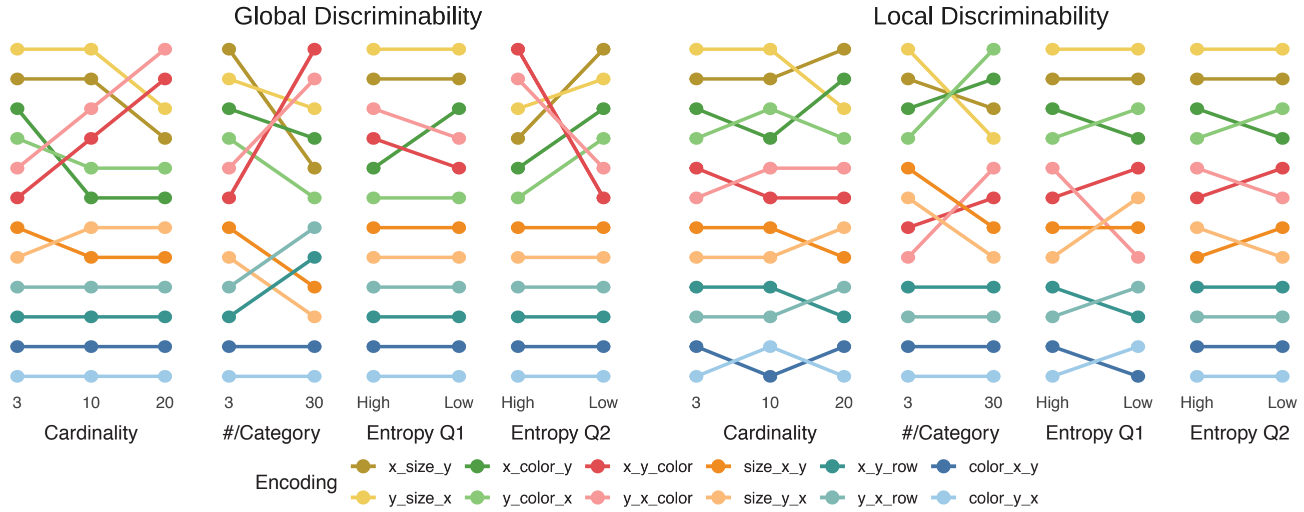


Fig. 9: Discriminability rankings of encodings (divided by data property) derived from discriminability scores.

of the visual difference generated by a visual encoding. The successive downsampling steps measure the preservation of the difference at increasing viewing or judgment distance. The utility of multiscale representation in visualization was first explored by Wattenberg and Fisher [43]. Moreover, the rankings of encodings suggest that the MS-SSIM seems to capture the notion of display space utilization, which Chen and Jänicke [7] define in information-theory terms. Encodings that tend to result in the least amount of blank space have higher discriminability.

## 8 Final Remarks

Computational measures of quality offer a scalable, low-cost, alternative to experiments with human participants. Discriminability, as a fundamental dimension of visualization quality, should be at the bottom of a stack of quality criteria for visual encoding evaluation. Future research should investigate what other criteria should compose this evaluation stack. Coupled with data simulation, discriminability tests enforce data characterization. Designers and researchers are required to document the data parameters and boundaries wherein proposed encodings are expected to produce high quality plots. This practice strengthens statements of the generality of research contributions and helps other researchers identify opportunities for new research.

Discriminability scores are tools to verify the principles of visual-data correspondence and unambiguous data depiction. In Kim and Heer's tasks there is little downside in perceiving a small data change as large. In other scenarios, proportionality is critical; for instance, a medical researcher examining effect size on a clinical experiment. In the future, discriminability scores may allow researchers to obtain an objective measure of the bias used in data communication. For instance, a scientific journal could create a standard maximum discriminability score for visualizations, in order to prevent exaggeration of effects. Our work also opens new directions for the study of ambiguity in visual encodings, which constitutes an overlooked source of uncertainty in visual information analysis.

Our exploration of the MS-SSIM brought to surface the role of perceptual scale on similarity judgments. The scale where a visualization is read—the level of detail considered—impacts people's perception of similarity. Currently, we know of no comprehensive studies that address this issue. Meanwhile, at least one visual analysis protocol [44] relies on accurate readings of similarity. Verifying the hypothesis that different encodings afford similarity judgments at different scales is a topic for future research. If this hypothesis is confirmed, MS-SSIM could help us discover and build a catalogue of these parameters for every chart type, in the same way that Steven's law has coefficients for stimulus types and Fitts' law has parameters that vary with device. This would require the design of a solid study protocol to collect data, and the fitting could be done as in Section 5. Informed by the appropriate weights, engineers could test the discriminability of a visualization on a dataset collection, and designers could test new designs by using weights fitted to visualizations that approximate the new design.

Our modifications to MS-SSIM add only some sensitivity to color. The proposed measure still embodies Wang's hypothesis that image similarity depends on preservation of *spatial* structure. Interestingly, the correlation with Kim and Heer's data is high despite our measure's lack of a sophisticated handling of color. So it is plausible that color is indeed less important to visualization discriminability (as it is the case with perceived image quality, Wang's problem). The results are aligned with the consensus that the color channel allows the encoding of fewer distinct values [31], which explains at least the poor performance of color encodings in the "Read Values" task.

## 9 Conclusions

In this paper, we introduced a general method for computing visual encoding discriminability that requires only a collection of datasets and the corresponding rendered visualizations. While discriminability has been a quality criterion in visualization for a long time, it has been mainly confined to theoretical discussions. This work constitutes the first methodical application of the discriminability criterion to the evaluation of visualization encodings.

We examined the suitability of SSIM and MS-SSIM for scoring plot similarity, and revealed limitations related to over-sensitivity to visual accessories (e.g., grids) and failure to capture differences in hue. To overcome these limitations, we proposed modifications that achieved satisfactory results. We demonstrated that a parameterization of the MS-SSIM can be found via gradient descent that achieves significant overlap with empirical plot similarity judgments. Most importantly, we devised a method for calculating encoding discriminability using the MS-SSIM and established a link between discriminability and task accuracy for a collection of basic encodings. We found that discriminability correlates with accuracy, especially for tasks that involve reading values of individual data points.

In conclusion, our results suggest that the MS-SSIM is useful for approximating plot similarity, and that discriminability scores based on MS-SSIM are associated with effectiveness. We recommend these scores to be used in early stages of visual encoding evaluation, as a low-cost computational measure of quality prior to committing to costly user studies.

## Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and Fundação CAPES (9078-13-4/Ciência sem Fronteiras).

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.

## References

[1] M. J. Alam, S. G. Kobourov, and S. Veeramoni. Quantitative measures for cartogram generation techniques. *Computer Graphics Forum*, 34(3):351–360, 2015.

[2] L. Bartram and M. C. Stone. Whisper, don't scream: Grids and transparency. *IEEE Trans. Visualization and Computer Graphics*, 17(10):1444–1458, 2011.

[3] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, et al. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018.

[4] E. Bertini and G. Santucci. Quality metrics for 2D scatterplot graphics: Automatically reducing visual clutter. In *Int. Symp. on Smart Graphics*, pp. 77–89. Springer, 2004.

[5] R. Brath. Metrics for effective information visualization. In *Proc. of IEEE Symp. on Information Visualization*, pp. 108–111. IEEE, 1997.

[6] A. Brychtová and A. Çöltekin. The effect of spatial distance on the discriminability of colors in maps. *Cartography and Geographic Information Science*, 44(3):229–245, 2017.

[7] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *IEEE Computer Graphics and Applications*, 29(1):12–19, 2009.

[8] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1206–15, 2010.

[9] C. Demiralp, M. S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE Trans. Visualization and Computer Graphics*, 20(12):1933–1943, 2014.

[10] C. Demiralp, C. E. Scheidegger, G. L. Kindlmann, D. H. Laidlaw, and J. Heer. Visual embedding: A model for visualization. *IEEE Computer Graphics and Applications*, 34(1):10–15, 2014.

[11] C. Dunne, S. I. Ross, B. Shneiderman, and M. Martino. Readability metric feedback for aiding node-link visualization designers. *IBM Journal of Research and Development*, 59(2/3):14–1, 2015.

[12] G. Ellis and A. Dix. The plot, the clutter, the sampling and its lens: occlusion measures for automatic clutter reduction. In *Proc. of the Working Conf. on Advanced Visual Interfaces*, pp. 266–269. ACM, 2006.

[13] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Trans. Visualization and Computer Graphics*, 23(1):521–530, 2017.

[14] S. Haroz and D. Whitney. How Capacity Limits of Attention Influence Information Visualization Effectiveness. *IEEE Trans. Visualization and Computer Graphics*, 18(12):2402–2410, dec 2012.

[15] J. Harper and M. Agrawala. Deconstructing and restyling D3 visualizations. In *Proc. of the ACM Symp. on User Interface Software and Technology*, pp. 253–262. ACM, 2014.

[16] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 203–212. ACM, 2010.

[17] H. Hofmann, L. Follett, M. Majumder, and D. Cook. Graphical tests for power comparison of competing designs. *IEEE Trans. Visualization and Computer Graphics*, 18(12):2441–2448, 2012.

[18] D. Holten, J. J. V. Wijk, and J.-B. Martens. A perceptually based spectral model for isotropic textures. *ACM Trans. Applied Perception (TAP)*, 3(4):376–398, 2006.

[19] Interactive Data Lab. Vega Lite example gallery, 2018. https://vega.github.io/vega-lite/examples/.

[20] H. Jänicke and M. Chen. A salience-based quality metric for visualization. *Computer Graphics Forum*, 29(3):1183–1192, 2010.

[21] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[22] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. *Computer Graphics Forum*, 37(3):157–167, 2018.

[23] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE Trans. Visualization and Computer Graphics*, 20(12):2181–2190, Dec. 2014.

[24] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum*, 32(3pt4):401–410, 2013.

[25] Y. Liu and J. Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2018.

[26] J. Matejka and G. Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 1290–1294. ACM, 2017.

[27] G. G. Méndez, M. A. Nacenta, and S. Vandenheste. iVoLVER: Interactive visual language for visualization extraction and reconstruction. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 4073–4085. ACM, 2016.

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

[29] A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):193–201, 2009.

[30] T. Munzner. A nested model for visualization design and validation. *IEEE Trans. Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009.

[31] T. Munzner. *Visualization Analysis & Design*. CRC Press, 2014.

[32] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In *Proc. IEEE Int. Conf. on Image Processing*, vol. 2, pp. II/169–II/172. IEEE, 2007.

[33] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 3659–3669. ACM, 2016.

[34] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of Human Centric Visualization*, pp. 147–175. Springer, 2014.

[35] G. Ryan, A. Mosca, R. Chang, and E. Wu. At a glance: Pixel approximate entropy as a measure of line chart complexity. *IEEE Trans. Visualization and Computer Graphics*, 25(1):872–881, 2019.

[36] D. A. Szafir. Modeling color difference for visualization design. *IEEE Trans. Visualization and Computer Graphics*, 24(1):392–401, 2018.

[37] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnork, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology*, pp. 59–66, 2009.

[38] Y. Tu and H. Shen. Visualizing changes of hierarchical data using treemaps. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1286–1293, Nov 2007.

[39] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd ed., 2001.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[41] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Processing*, 20(5):1185–1198, 2011.

[42] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. of the Asilomar Conf. on Signals, Systems Computers*, vol. 2, pp. 1398–1402 Vol.2, Nov 2003.

[43] M. Wattenberg and D. Fisher. Analyzing perceptual organization in information graphics. *Information Visualization*, 3(2):123–133, 2004.

[44] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for Infovis. *IEEE Trans. Visualization and Computer Graphics*, 16(6):973–9, Jan. 2010.

[45] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symp. on Information Visualization, 2005. INFOVIS 2005.*, pp. 157–164. IEEE, 2005.

[46] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Trans. Visualization and Computer Graphics*, 25(3):1474–1488, 2019.

[47] V. Yoghourdjian, T. Dwyer, K. Klein, K. Marriott, and M. Wybrow. Graph thumbnails: Identifying and comparing multiple graphs at a glance. *IEEE Trans. Visualization and Computer Graphics*, 24(12):3081–3095, 2018.

[48] Z. Zheng, H. Cheng, Z. Zhang, Y. Zhao, and P. Wang. An alternative method for understanding user-chosen passwords. *Security and Communication Networks*, 2018.

Authorized licensed use limited to: Linkoping University Library. Downloaded on November 27,2024 at 10:19:48 UTC from IEEE Xplore. Restrictions apply.