


Article

A Bounded Measure for Estimating the Benefit of Visualization (Part I): Theoretical Discourse and Conceptual Evaluation

Min Chen ^{1,*}  and Mateu Sbert ² 

¹ Oxford e-Research Centre (OeRC), Department of Engineering Science, University of Oxford, Oxford OX1 3QG, UK

² Department of Informàtica i Matemàtica Aplicada, University of Girona, 17071 Girona, Spain; mateu@ima.udg.edu

* Correspondence: min.chen@oerc.ox.ac.uk

Abstract: Information theory can be used to analyze the cost–benefit of visualization processes. However, the current measure of benefit contains an unbounded term that is neither easy to estimate nor intuitive to interpret. In this work, we propose to revise the existing cost–benefit measure by replacing the unbounded term with a bounded one. We examine a number of bounded measures that include the Jensen–Shannon divergence, its square root, and a new divergence measure formulated as part of this work. We describe the rationale for proposing a new divergence measure. In the first part of this paper, we focus on the conceptual analysis of the mathematical properties of these candidate measures. We use visualization to support the multi-criteria comparison, narrowing the search down to several options with better mathematical properties. The theoretical discourse and conceptual evaluation in this part provides the basis for further data-driven evaluation based on synthetic and experimental case studies that are reported in the second part of this paper.

Keywords: information theory; theory of visualization; cost–benefit analysis; divergence measure; benefit of visualization; human knowledge in visualization; abstraction; deformation; volume visualization; metro map



Citation: Chen, M.; Sbert, M. A Bounded Measure for Estimating the Benefit of Visualization (Part I): Theoretical Discourse and Conceptual Evaluation. *Entropy* **2022**, *24*, 228. <https://doi.org/10.3390/e24020228>

Academic Editor: Éloi Bossé

Received: 30 November 2021

Accepted: 27 January 2022

Published: 31 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

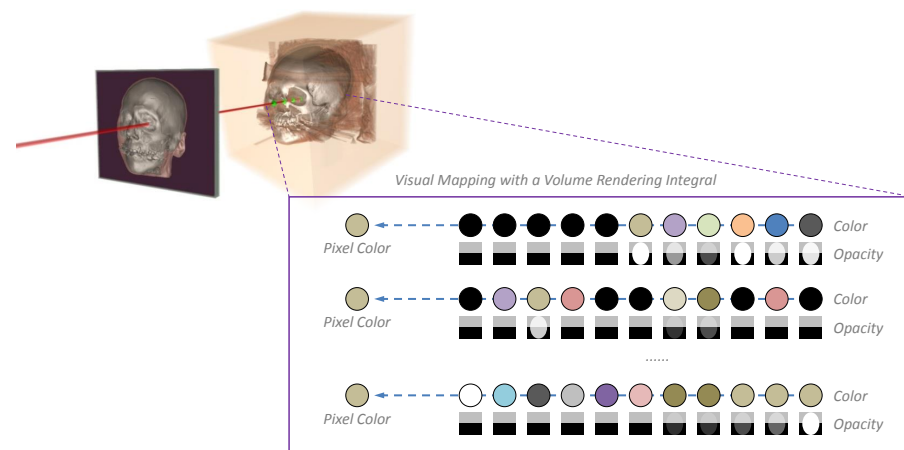


Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

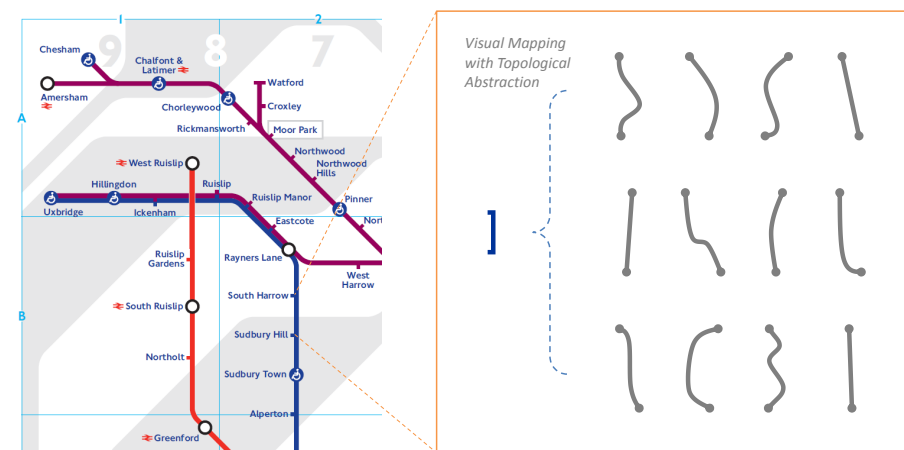
1. Introduction

To most of us, it seems rather intuitive that visualization should be accurate, different data values should be visually encoded differently, and visual distortion should be disallowed. However, when we closely examine most (if not all) visualization images, we can notice that inaccuracy is ubiquitous. The two examples in Figure 1 evidence the presence of such inaccuracy. In volume visualization, when a pixel is used to depict a set of voxels along a ray, many different sets of voxel values may result in the same pixel color. In a metro map, a variety of complex geographical paths may be distorted and depicted as a straight line. Since there is little doubt that volume visualization and metro maps are useful, some “inaccurate” visualization must be beneficial.

In terms of information theory, the types of inaccuracy featured in Figure 1 are different forms of information loss (or many-to-one mapping). Chen and Golan proposed an information-theoretic measure [1] for analyzing the cost–benefit of data intelligence workflows. It enables us to consider the positive impact of information loss (e.g., reducing the cost of storing, processing, displaying, perceiving, and reasoning about the information) as well as its negative impact (e.g., being misled by the information). The measure provides a concise explanation about the benefit of visualization because visualization and other data intelligence processes (e.g., statistics and algorithms) all typically cause information loss and visualization allows human users to reduce the negative impact of information loss effectively using their knowledge.



(a) mapping from different sets of voxel values to the same pixel color



(b) mapping from different geographical paths to the same line segment

Figure 1. Visual encoding typically features many-to-one mapping from data to visual representations, hence information loss. For example, (a) in volume visualization, the color of each pixel results from a complex process of combining a sequence of voxel values, and (b) in metro maps, different geographical paths are often represented using indistinguishable line segments. The significant amount of information loss in volume visualization and metro maps suggests that viewers not only can abide the information loss but also benefit from it. Measuring such benefits can lead to new advancements of visualization, in theory and practice.

The mathematical formula of the cost–benefit ratio features a term based on the Kullback–Leibler (KL) divergence [2] for measuring the potential distortion of a user or a group of users in reconstructing the information that may have been lost or distorted during a visualization process. The cost–benefit ratio instigates that a user with more knowledge about the source data and its visual representation is likely to suffer less distortion. While using the KL-divergence is mathematically intrinsic for measuring the potential distortion, its unboundedness property has some undesirable consequences. The simplest phenomenon of making a false representation (i.e., always displaying 1 when a binary value is 0 or always 0 when it is 1) happens to be a singularity condition of the KL-divergence. The amount of distortion measured by the KL-divergence often has many more bits than the entropy of the information space itself. This is not intuitive to interpret and hinders practical applications.

In this two-part paper, we propose to replace the KL-divergence with a bounded term. In the first part, we first confirm the boundedness is a necessary property. We then conduct multi-criteria decision analysis (MCDA) [3] to compare a number of bounded measures, which include the Jensen–Shannon (JS) divergence [4], its square root, and a new divergence measure \mathcal{D}_{new}^k (including its variations) formulated as part of this work. We

use visual analysis to aid the observation of the mathematical properties of these candidate measures, narrowing down from eight options to five. In the second part of this paper [5], we use synthetic and experimental case studies to instantiate values that may be returned by the five options. It also explores the relationship between measuring the benefit of visualization and measuring the viewers' knowledge used during visualization.

The search for the best way to measure the benefit of visualization will likely entail a long journey. The main aim of this work is to initiate this endeavor. The main technical contributions of this two-part paper include:

- Identifying a shortcoming of using the KL-divergence in the information-theoretic measure proposed by Chen and Golan [1] and evidencing the shortcoming using practical examples (Parts I and II);
- Presenting a theoretical discourse to justify the use of a bounded measure for finite alphabets (Part I);
- Proposing a new bounded divergence measure, while studying existing bounded divergence measures (Part I);
- Analyzing nine candidate measures using seven criteria reflecting desirable conceptual or mathematical properties, and narrowing the nine candidate measures to six measures (Part I);
- Conducting several case studies for collecting instances for evaluating the remaining six candidate measures (Part II);
- Demonstrating the uses of the cost–benefit measurement to estimate the benefit of visualization in practical scenarios and the human knowledge used in the visualization processes (Part II);
- Discovering a new conceptual criterion that a divergence measure is a summation of the entropic values of its components, which is useful in analyzing and visualizing empirical data (Part II);
- Offering a recommendation to revise the information-theoretic measure proposed by Chen and Golan [1] based on multi-criteria decision analysis (Parts I and II).

2. Related Work

Claude Shannon's landmark article in 1948 [6] signifies the birth of information theory. It has been underpinning the fields of data communication, compression, and encryption since. As a mathematical framework, information theory provides a collection of useful measures, many of which, such as Shannon entropy [6], cross entropy [7], mutual information [7], and Kullback–Leibler divergence [2] are widely used in applications of physics, biology, neurology, psychology, and computer science (e.g., visualization, computer graphics, computer vision, data mining, machine learning), and so on. In this work, we also consider Jensen-Shannon divergence [4] in detail.

Information theory has been used extensively in visualization [8]. It has enabled many applications in visualization, including scene and shape complexity analysis by Feixas et al. [9] and Rigau et al. [10], light source placement by Gumhold [11], view selection in mesh rendering by Vázquez et al. [12] and Feixas et al. [13], attribute selection by Ng and Martin [14], view selection in volume rendering by Bordoloi and Shen [15], and Takahashi and Takeshima [16], multi-resolution volume visualization by Wang and Shen [17], focus of attention in volume rendering by Viola et al. [18], feature highlighting by Jänicke and Scheuermann [19,20], and Wang et al. [21], transfer function design by Bruckner and Möller [22], and Ruiz et al. [23,24], multi-modal data fusion by Bramon et al. [25], isosurface evaluation by Wei et al. [26], measuring observation capacity by Bramon et al. [27], measuring information content by Biswas et al. [28], proving the correctness of “overview first, zoom, details-on-demand” by Chen and Jänicke [29] and Chen et al. [8], and confirming visual multiplexing by Chen et al. [30].

Ward first suggested that information theory might be an underpinning theory for visualization [31]. Chen and Jänicke [29] outlined an information-theoretic framework for visualization, and it was further enriched by Xu et al. [32] and Wang and Shen [33] in

the context of scientific visualization. Chen and Golan proposed an information-theoretic measure for analyzing the cost–benefit of visualization processes and visual analytics workflows [1]. It was used to frame an observation study showing that human developers usually entered a huge amount of knowledge into a machine learning model [34]. It motivated an empirical study confirming that knowledge could be detected and measured quantitatively via controlled experiments [35]. It was used to analyze the cost–benefit of different virtual reality applications [36]. It formed the basis of a systematic methodology for improving the cost–benefit of visual analytics workflows [37]. It survived qualitative falsification by using arguments in visualization [38]. It offered a theoretical explanation of “visual abstraction” [39]. It provided a theoretical basis to a design space that was structured according to different ways of “losing information” in origin–destination data visualization [40]. The work reported in this paper continues the path of theoretical developments in visualization [41], and is intended to improve the original cost–benefit formula [1], in order to make it a more intuitive and usable measurement in practical visualization applications.

The information-theoretic measure proposed by Chen and Golan [1] can be applied to a variety of processes for transforming some input data to some output data [42]. These include machine-centric processes (e.g., computing statistical measures, importance sampling, feature extraction, dimensionality reduction, etc.) and human-centric processes (e.g., data visualization, human–computer interaction, written communication, human cognition, etc.). In this work, we focus on processes of data visualization.

3. Overview and Motivation

A short introduction to information-theoretic cost–benefit analysis can be found in an arXiv report [43]. For self-containment, we provide a brief overview to accompany our description of the problem that motivated this work.

Visualization is useful in most data intelligence workflows, but the usefulness is not universally true because the effectiveness of visualization is usually data-, user-, and task-dependent. The cost–benefit ratio proposed by Chen and Golan [1] captures the essence of such dependency. Below is the qualitative expression of the measure:

$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\text{Alphabet Compression} - \text{Potential Distortion}}{\text{Cost}} \quad (1)$$

Consider the scenario of viewing some data through a particular visual representation. The term *Alphabet Compression* (AC) measures the amount of information loss due to visual abstraction [39] (or any transformation featuring many-to-one mappings). Since the visual representation is fixed in the scenario, AC is thus largely data-dependent. AC is a positive measure reflecting the fact that visual abstraction must be useful in many cases though it may result in information loss. This apparently counter-intuitive term is essential for asserting why visualization is useful. Note that the term also helps assert the usefulness of statistics, algorithms, and interaction since they all usually cause information loss [37].

The positive implication of the term AC is counterbalanced by the term *Potential Distortion*, while both being moderated by the term *Cost*. The term *Cost* encompasses all costs of the visualization process, including computational costs (e.g., visual mapping and rendering), cognitive costs (e.g., cognitive load), and consequential costs (e.g., impact of errors). As illustrated in Figure 2, increasing AC typically enables the reduction of cost (e.g., in terms of energy, or its approximation such as time or money).

The term *Potential Distortion* (PD) measures the informative divergence between viewing the data through visualization with information loss and reading the data without any information loss. The latter might be ideal but is usually at an unattainable cost except for values in a very small data space (i.e., in a small alphabet as discussed in [1]). As shown in Figure 2, increasing AC typically causes more PD. PD is data-dependent or user-dependent. Given the same data visualization with the same amount of information loss,

one can postulate that a user with more knowledge about the data or visual representation usually suffers less distortion. This postulation is a focus of this paper.

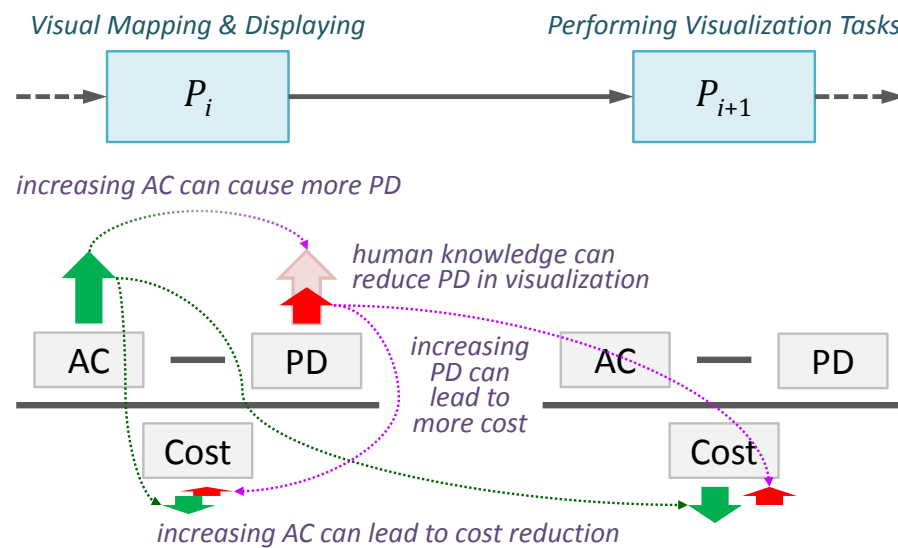






Figure 2. Each process in a data intelligence workflow can be characterized using three abstract measures: *alphabet compression* (AC), *potential distortion* (PD), and *cost*. They can be used to reason about the shortcomings in a workflow and identify possible solutions in abstraction [37]. For example, increasing data filtering in visualization (AC) may reduce the cost of P_i and P_{i+1} , especially when human knowledge can reduce perceptual errors (PD).

Consider the visual representation of a network of arteries in Figure 3. The image was generated from a volume dataset using the maximum intensity projection (MIP) method. While it is known that MIP cannot convey depth information well, it has been widely used for observing some classes of medical imaging data, such as arteries. The highlighted area in Figure 3 shows an apparently flat area, which is a distortion from the actuality of a tubular surface likely with some small wrinkles and bumps. The doctors who deal with such medical data are expected to have sufficient knowledge to reconstruct the reality adequately from the “distorted” visualization, while being able to focus on the more important task of making diagnostic decisions, e.g., about aneurysm.

As shown in some recent works, it is possible for visualization designers to estimate AC, PD, and Cost qualitatively [36,37] and quantitatively [34,35]. It is highly desirable to advance the scientific methods for quantitative estimation, towards the eventual realization of computer-assisted analysis and optimization in designing visual representations. This work focuses on one challenge of quantitative estimation, i.e., how to estimate the benefit of visualization to human users with different knowledge about the depicted data and visual encoding.

Building on the methods of observational estimation [34] and controlled experiment [35], one may reasonably anticipate a systematic method based on a short interview by asking potential viewers a few questions. For example, one may use the question in Figure 3 to estimate the knowledge of doctors, patients, and any other people who may view such a visualization. The question is intended to tease out two pieces of knowledge that may help reduce the potential distortion due to the “flat area” depiction. One piece is about the general knowledge that associates arteries with tube-like shapes. Another, which is more advanced, is about the surface texture of arteries and the limitations of the MIP method.

Question 5: The image on the right depicts a computed tomography dataset (arteries) that was rendered using a maximum intensity projection (MIP) algorithm. Consider the section of the image inside the red circle (also in the inset of a zoomed-in view). Which of the following illustrations would be the closest to the real surface of this part of the artery?

- (A)  *Curved, rather smooth*
- (B)  *Curved, with wrinkles and bumps*
- (C)  *Flat, rather smooth*
- (D)  *Flat, with wrinkles and bumps*

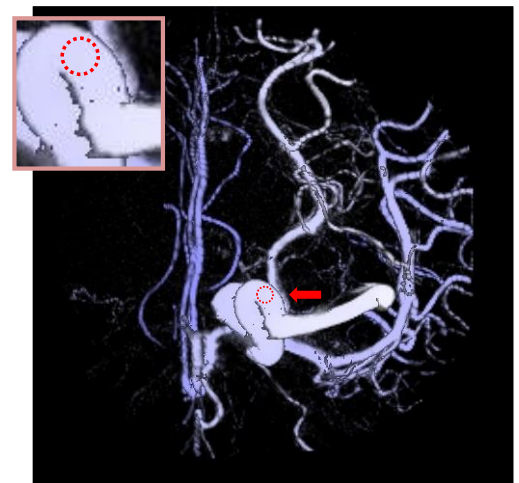


Figure 3. A volume dataset was rendered using the maximum intensity projection (MIP) method, which causes curved surfaces of arteries to appear rather flat. Posing a question about a “flat area” in the image can be used to tease out a viewer’s knowledge that is useful in a visualization process.

In the second part of this paper [5], the question in Figure 3 is one of eight questions used in a survey for collecting empirical data for evaluating the bounded measures considered in this paper. As this paper focuses on the theoretical discourse and conceptual evaluation, we use a highly abstracted version of this example to introduce the relevant information-theoretic notations and elaborate the problem statement addressed by this paper.

4. Mathematical Notations and Problem Statement

4.1. Mathematical Notation

Consider a simplified scenario in Figure 4, where three sequences of voxels are rendered using the MIP method, resulting in three pixel values on the left. Here the three sequence voxels exemplifies a volume with $N_x \times N_y \times N_z$ voxels (illustrated as $1 \times 3 \times 10$). Let each voxel value be an 8-bit unsigned integer. In information theory, the possible 256 values are referred to as an *alphabet*, denoted here as $\mathbb{D}_{\text{vxl}} = \{0, 1, 2, \dots, 255\}$. The 256 valid values $[0, 255]$ are its *letters*. The alphabet is associated with a *probability mass function* (PMF) $P(\mathbb{D}_{\text{vxl}})$. The Shannon entropy of this alphabet $\mathcal{H}(\mathbb{D}_{\text{vxl}})$ measures the average uncertainty and information of the voxel, and is defined as:

$$\mathcal{H}(\mathbb{D}_{\text{vxl}}) = - \sum_0^{255} p_i \log_2 p_i \quad \text{where } p_i \in [0, 1], \sum_0^{255} p_i = 1$$

where p_i indicates the probability for the voxel to have its value equal to $i \in [0, 255]$. When all 256 values are equally probable (i.e., $\forall i \in [0, 255], p_i = 1/256$), we have $\mathcal{H}(\mathbb{D}_{\text{vxl}}) = 8$ bits. In practice, an application usually deals with a specific type of volume data, the probability of different values may vary noticeably. For example, in medical imaging, a voxel at the boundary of a volume is more likely to have a value indicating an empty space.

The entire volume of $N_x \times N_y \times N_z$ voxels be defined as a composite alphabet \mathbb{D}_{vml} . Its letters are all valid combinations of voxel values. If the $N_x \times N_y \times N_z$ voxels are modelled as independent and identically distributed random variables, we have:

$$\mathcal{H}(\mathbb{D}_{\text{vml}}) = \sum_{k=1}^M \mathcal{H}(\mathbb{D}_{\text{vxl},k}) = 8 \times N_x \times N_y \times N_z \text{ bits}$$

where $M = N_x \times N_y \times N_z$. For the volume illustrated in Figure 4, $\mathcal{H}(\mathbb{D}_{\text{vml}})$ would be 240 bits. However, this is the maximum entropy of such a volume. In real world applications, it is very unlikely for the $N_x \times N_y \times N_z$ voxels to be independent and identically

distributed random variables. Although domain experts may not have acquired the ground truth PMF, by measuring a very large corpus of volume data, they have intuitive knowledge as to what may be possible or not. For example, doctors, who handle medical imaging data, do not expect to see a car, ship, aeroplane, or other “weird” objects in a volume dataset. This intuitive and imprecise knowledge about the PMF can explain the humans’ ability to decode visualization featuring some “short comings” such as various visual multiplexing phenomena (e.g., occlusion, displacement, and information omission) [30]. In the second part of this paper [5], we will explore means to measure such ability quantitatively.

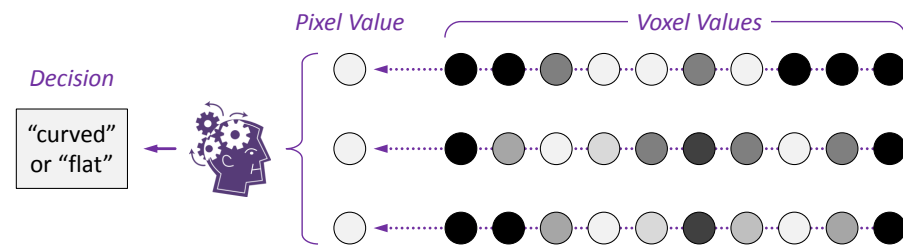


Figure 4. In this 2D illustration of a simplified scenario of volume visualization, three sequences of voxels are rendered using the MIP method. The volume on the right features a curved surface defined by those brightest voxels. By projecting the maximum voxel values to the pixels in the middle, the curvature information of the surface is lost. A viewer needs to determine if the surface in the volume is curved or flat, for which the viewer’s knowledge is critical.

Similarly, we can define an alphabet for a pixel \mathbb{R}_{pxl} and a composite alphabet for an image \mathbb{R}_{img} . For the example in Figure 4, we assume a simple MIP algorithm that selects the maximum voxel value along each ray, and assigns it as the corresponding pixel as an 8-bit monochromatic value. It is obvious that the potential variation of a pixel is much less than the potential combined variation of all voxels along a ray. Hence in terms of Shannon entropy, most likely $\mathcal{H}(\mathbb{D}_{\text{vlm}}) - \mathcal{H}(\mathbb{D}_{\text{img}}) \gg 0$, indicating significant information loss during the rendering process.

Given an analytical task to be performed through visualization, the analytical decision alphabet \mathbb{A} usually contains a small number of letters, such as $\{\text{contain artefact } X, \text{ no artefact } X\}$ or $\{\text{big, medium, small, tiny, none}\}$. The entropy of \mathbb{A} is usually much lower than that of \mathbb{D}_{img} , i.e., $\mathcal{H}(\mathbb{D}_{\text{img}}) - \mathcal{H}(\mathbb{A}) \gg 0$, indicating further information loss during perception and cognition. As discussed in Section 3, this is referred to as *alphabet compression* and is a general trend of all data intelligence workflows. The question is thus about how much the analytical decision was disadvantaged by the loss of information. This is referred to as *potential distortion*.

In the original quantitative formula proposed in [1], the potential distortion is measured using Kullback–Leibler divergence (or KL-divergence) [2]. Given an alphabet \mathbb{Z} with two PMFs P and Q , KL-divergence measures how much Q differs from the reference distribution P :

$$\mathcal{D}_{\text{KL}}(P(\mathbb{Z})||Q(\mathbb{Z})) = \sum_{i=1}^n p_i (\log_2 p_i - \log_2 q_i) = \sum_{i=1}^n p_i \log_2 \frac{p_i}{q_i} \quad (2)$$

where $n = \|\mathbb{Z}\|$ is the number of letters in the alphabet \mathbb{Z} , and p_i and q_i are the probability values associated with letter $z_i \in \mathbb{Z}$. \mathcal{D}_{KL} is also measured in *bit*. Because \mathcal{D}_{KL} is an unbounded measure regardless the maximum entropy of \mathbb{Z} , it is easy to relate, quantitatively, the value of potential distortion and that of alphabet compression. This leads to the problem to be addressed in this two-part paper.

Note: In this paper, to simplify the notations in different contexts, for an information-theoretic measure, we use an alphabet \mathbb{Z} and its PMF P interchangeably, e.g., $\mathcal{H}(P(\mathbb{Z})) = \mathcal{H}(P) = \mathcal{H}(\mathbb{Z})$. An arXiv report [43] provides a short introduction to the cost–benefit analysis and the relevant mathematical background of information theory, which some readers may find helpful.

4.2. Problem Statement

Recall our brief discussion about an analytical task that may be affected by the MIP image in Figure 3 in Section 3. Let us define the analytical task as binary options about whether the “flat area” is actually flat or curved. In other words, it is an alphabet $\mathbb{A} = \{curved, flat\}$. The likelihood of the two options is represented by a probability distribution or probability mass function (PMF) $P(\mathbb{A}) = \{1 - \epsilon, 0 + \epsilon\}$, where $0 < \epsilon < 1$. Since most arteries in the real world are of tubular shapes, one can imagine that a ground truth alphabet $\mathbb{A}_{G.T.}$ might have a PMF $P(\mathbb{A}_{G.T.})$ strongly in favor of the *curved* option. However, the visualization seems to suggest the opposite, implying a PMF $P(\mathbb{A}_{MIP})$ strongly in favor of the *flat* option. It is not difficult to interview some potential viewers, enquiring how they would answer the question. One may estimate a PMF $P(\mathbb{A}_{doctors})$ from doctors’ answers, and another $P(\mathbb{A}_{patients})$ from patients’ answers.

Table 1 shows two scenarios where different probability data is obtained. The values of PD are computed using the KL-divergence as proposed in [1]. In Scenario 1, without any knowledge, the visualization process would suffer 6.50 bits of potential distortion (PD). As doctors are not fooled by the “flat area” shown in the MIP visualization, their knowledge is worth 6.50 bits. Meanwhile, patients would suffer 1.12 bits of PD on average, their knowledge is worth $5.38 = 6.50 - 1.12$ bits.

Table 1. Imaginary scenarios where probability data is collected for estimating knowledge related to alphabet $\mathbb{A} = \{curved, flat\}$. The ground truth (G.T.) PMFs are defined with $\epsilon = 0.01$ and 0.0001 respectively. The potential distortion (as “→ value”) is computed using the KL-divergence.

	Scenario 1	Scenario 2
$Q(\mathbb{A}_{G.T.})$:	{0.99, 0.01}	{0.9999, 0.0001}
$P(\mathbb{A}_{MIP})$:	{0.01, 0.99} → 6.50	{0.0001, 0.9999} → 13.28
$P(\mathbb{A}_{doctors})$:	{0.99, 0.01} → 0.00	{0.99, 0.01} → 0.05
$P(\mathbb{A}_{patients})$:	{0.7, 0.3} → 1.12	{0.7, 0.3} → 3.11

In Scenario 2, the PMFs of $P(\mathbb{A}_{G.T.})$ and $P(\mathbb{A}_{MIP})$ depart further away, while $P(\mathbb{A}_{doctors})$ and $P(\mathbb{A}_{patients})$ remain the same. Although doctors and patients would suffer more PD, their knowledge is worth more than that in Scenario 1 (i.e., $13.28 - 0.05 = 13.23$ bits and $13.28 - 3.11 = 10.17$ bits respectively).

Similarly, the binary options about whether the “flat area” is actually smooth or not can be defined by an alphabet $\mathbb{A} = \{wrinkles-and-bumps, smooth\}$. Table 2 shows two scenarios about collected probability data. In these two scenarios, doctors exhibit much more knowledge than patients, indicating that the surface texture of arteries is a piece of specialized knowledge.

The above example demonstrates that using the KL-divergence to estimate PD can differentiate the knowledge variation between doctors and patients regarding the two pieces of knowledge that may reduce the distortion due to the “flat area”. When it is used in Equation (1) in a relative or qualitative context (e.g., [36,37]), the unboundedness of the KL-divergence does not pose an issue.

However, this does become an issue when the KL-divergence is used to measure PD in an absolute and quantitative context. From the two diverging PMFs $P(\mathbb{A}_{G.T.})$ and $P(\mathbb{A}_{MIP})$ in Table 1, or $P(\mathbb{B}_{G.T.})$ and $P(\mathbb{B}_{MIP})$ in Table 2, we can observe that the smaller ϵ is, the more divergent the two PMFs become and the higher value the PD has. Indeed, consider an arbitrary alphabet $\mathbb{Z} = \{z_1, z_2\}$, and two PMFs defined upon \mathbb{Z} : $P = [0 + \epsilon, 1 - \epsilon]$ and $Q = [1 - \epsilon, 0 + \epsilon]$. When $\epsilon \rightarrow 0$, we have the KL-divergence $\mathcal{D}_{KL}(Q||P) \rightarrow \infty$.

Meanwhile, the Shannon entropy of \mathbb{Z} , $\mathcal{H}(\mathbb{Z})$, has an upper bound of 1 bit. It is thus not intuitive or practical to relate the value of $\mathcal{D}_{KL}(Q||P)$ to that of $\mathcal{H}(\mathbb{Z})$. Many applications of information theory do not relate these two types of values explicitly. When reasoning such relations is required, the common approach is to impose a lower-bound threshold for ϵ (e.g., [35]). However, there is yet a consistent method for defining such a threshold

for various alphabets in different applications, while preventing a range of small or large values (i.e., $[0, \sigma)$ or $(1 - \sigma, 1]$) in a PMF is often inconvenient in practice. Indeed, for a binary alphabet with two arbitrary P and Q , in order to restrict its $\mathcal{D}_{\text{KL}}(P||Q) \leq 1$, one has to set $0.0658 \lesssim \sigma \lesssim 0.9342$, rendering some 13% of the probability range $[0, 1]$ unusable. In the following section, we discuss several approaches to defining a bounded measure for PD.

Table 2. Imaginary scenarios for estimating knowledge related to alphabet $\mathbb{B} = \{\textit>wrinkles-and-bumps, smooth}\}$. The ground truth (G.T.) PMFs are defined with $\epsilon = 0.1$ and 0.001 respectively. The potential distortion (as “ \rightarrow value”) is computed using the KL-divergence.

	Scenario 3	Scenario 4
$Q(\mathbb{B}_{\text{G.T.}}):$	$\{0.9, 0.1\}$	$\{0.999, 0.001\}$
$P(\mathbb{B}_{\text{MIP}}):$	$\{0.1, 0.9\} \rightarrow 2.54$	$\{0.001, 0.999\} \rightarrow 9.94$
$P(\mathbb{B}_{\text{doctors}}):$	$\{0.8, 0.2\} \rightarrow 0.06$	$\{0.8, 0.2\} \rightarrow 1.27$
$P(\mathbb{B}_{\text{patients}}):$	$\{0.1, 0.9\} \rightarrow 2.54$	$\{0.1, 0.9\} \rightarrow 8.50$

5. Bounded Measures for Potential Distortion (PD)

Let \mathbf{P}_i be a process in a data intelligence workflow, \mathbb{Z}_i be its input alphabet, and \mathbb{Z}_{i+1} be its output alphabet. \mathbf{P}_i can be a human-centric process (e.g., visualization and interaction) or a machine-centric process (e.g., statistics and algorithms). In the original proposal [1], the value of Benefit in Equation 1 is measured using:

$$\text{Benefit} = \text{AC} - \text{PD} = \mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1}) - \mathcal{D}_{\text{KL}}(\mathbb{Z}'_i||\mathbb{Z}_i) \quad (3)$$

where $\mathcal{H}()$ is the Shannon entropy of an alphabet and $\mathcal{D}_{\text{KL}}()$ is the KL-divergence of an alphabet from a reference alphabet. AC, which is $\mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1})$, defines the entropic difference between the input and output alphabets. Because the Shannon entropy of an alphabet with a finite number of letters is bounded, AC is also bounded. On the other hand, as discussed in the previous section PD (i.e., $\mathcal{D}_{\text{KL}}(\mathbb{Z}'_i||\mathbb{Z}_i)$) is unbounded. Although Equation (3) can be used for relative comparison, it is not quite intuitive in an absolute context, and it is difficult to imagine that the amount of informative distortion can be more than the maximum amount of information available.

Given a divergence or difference measure $\Delta(\alpha, \beta)$, the term *bound* may be used in two different contexts. (i) In the *general context*, the bounds of $\Delta(\alpha, \beta)$ are defined based on all possible α and β values in their generic variable domain (e.g., integer, real, or PMF). (ii) In a *specific or conditional context*, the bounds of $\Delta(\alpha, \beta)$ are defined based on possible α and β values subject to a specific condition. For example, if we have a specific condition $\alpha, \beta \in [-1, 1] \subset \mathbb{R}$, it is not unusual to expect a difference measure $\Delta(\alpha, \beta)$ to be bounded. However, as discussed in the previous section, the KL-divergence $\mathcal{D}_{\text{KL}}(P||Q)$, can still be unbounded even if we have a finite alphabet \mathbb{Z} and the Shannon entropy measures $\mathcal{H}(P(\mathbb{Z}))$ and $\mathcal{H}(Q(\mathbb{Z}))$ are bounded.

In this section, we present the unpublished work by Chen and Sbert [44], which reasons mathematically that for alphabets of a finite size, the KL-divergence used in Equation (3) should ideally be bounded. In their arXiv report, they also outlined a new divergence measure and compare it with a few other bounded measures. Building on the initial comparison by Chen and Sbert in [44], we use visualization in Section 6 to assist the multi-criteria analysis and selection of a bounded divergence measure to replace the KL-divergence used in Equation (3). In the second part of this paper [5], we will further examine the practical usability of a subset of bounded measures by evaluating them using synthetic and experimental data.

5.1. A Conceptual Proof of Boundedness

According to the mathematical definition of \mathcal{D}_{KL} in Equation (2), \mathcal{D}_{KL} is of course unbounded. We do not in anyway try to prove that this formula is bounded. We are inter-

ested in a scenario where an alphabet \mathbb{Z} is associated with two PMF, P and Q , which is very much the scenario of measuring the potential distortion in Equation (1). We ask a question: is it conceptually necessary for \mathcal{D}_{KL} to yield a unbounded value to describe the divergence between P and Q in this scenario despite that $\mathcal{H}(P)$ and $\mathcal{H}(Q)$ are both bounded?

We highlight the word “conceptually” because this relates to the concept about another information-theoretic measure, cross entropy, which is defined as:

$$\mathcal{H}_{\text{CE}}(P, Q) = \sum_{i=1}^n p_i \log_2 \frac{1}{q_i} = \sum_{i=1}^n p_i \log_2 \frac{p_i}{q_i} - \sum_{i=1}^n p_i \log_2 p_i = \mathcal{D}_{\text{KL}}(P||Q) + \mathcal{H}(P) \quad (4)$$

Conceptually, cross entropy measures the cost of a coding scheme. If a code (i.e., an alphabet \mathbb{Z}) has a true PMF P , the optimal coding scheme should require only $\mathcal{H}(P)$ bits according to Shannon’s source coding theorem [7]. However, if the code designer mistakes the PMF as Q , the resulting coding scheme will have $\mathcal{H}_{\text{CE}}(P, Q)$ bits. From Equation (4), we can observe that the inefficiency is described by the term $\mathcal{D}_{\text{KL}}(P||Q)$. Naturally, we can translate our aforementioned question to: should such inefficiency be bounded if there is a finite number of codewords (letters) in the code (alphabet).

Coding theory has been applied to visualization, e.g., for explaining the efficiency of logarithmic plots in displaying data of a family of skewed PMFs and the usefulness of redundancy in visual design [29]. Here, we focus on proving that $\mathcal{H}_{\text{CE}}(P, Q)$ is conceptually bounded.

Let \mathbb{Z} be an alphabet with a finite number of letters, $\{z_1, z_2, \dots, z_n\}$, and \mathbb{Z} is associated with a PMF, Q , such that:

$$\begin{aligned} q(z_n) &= \epsilon, \quad (\text{where } 0 < \epsilon < 2^{-(n-1)}), \\ q(z_{n-1}) &= (1 - \epsilon)2^{-(n-1)}, \\ q(z_{n-2}) &= (1 - \epsilon)2^{-(n-2)}, \\ &\dots \\ q(z_2) &= (1 - \epsilon)2^{-2}, \\ q(z_1) &= (1 - \epsilon)2^{-1} + (1 - \epsilon)2^{-(n-1)}. \end{aligned} \quad (5)$$

When we encode this alphabet using an entropy binary coding scheme [45], we can be assured to achieve an optimal code with the lowest average length for codewords. One example of such a code for the above probability is:

$$\begin{aligned} z_1 &: 0, & z_2 &: 10, & z_3 &: 110 \\ &\dots & & & & \\ z_{n-1} &: 111 \dots 10 & (\text{with } n - 2 \text{ “1”s and one “0”}) \\ z_n &: 111 \dots 11 & (\text{with } n - 1 \text{ “1”s and no “0”}) \end{aligned} \quad (6)$$

In this way, z_n , which has the smallest probability, will always be assigned a codeword with the maximal length of $n - 1$. Entropy coding is designed to minimize the average number of bits per letter when one transmits a “very long” sequence of letters in the alphabet over a communication channel. Here the phrase “very long” implies that the string exhibits the above PMF Q (Equation (5)).

Suppose that \mathbb{Z} is actually of PMF P , but is encoded as Equation (6) based on Q . The transmission of \mathbb{Z} using this code will have inefficiency. As mentioned above, the cost is measured by cross entropy $\mathcal{H}_{\text{CE}}(P, Q)$, and the inefficiency is measured by the term $\mathcal{D}_{\text{KL}}(P||Q)$ in Equation (4).

Clearly, the worst case is that the letter, z_n , which was encoded using $n - 1$ bits, turns out to be the most frequently used letter in P (instead of the least in Q). It is so frequent that all letters in the long string are of z_n . So the average codeword length per letter of this string is $n - 1$. The situation cannot be worse. Therefore, $n - 1$ is the upper bound of

the cross entropy. From Equation (4), we can also observe that $\mathcal{D}_{\text{KL}}(P||Q)$ must also be bounded since $\mathcal{H}_{\text{CE}}(P, Q)$ and $\mathcal{H}(P)$ are both bounded as long as \mathbb{Z} has a finite number of letters. Let \top_{CE} be the upper bound of $\mathcal{H}_{\text{CE}}(P, Q)$. The upper bound for $\mathcal{D}_{\text{KL}}(P||Q)$, \top_{KL} , is thus:

$$\mathcal{D}_{\text{KL}}(P||Q) = \mathcal{H}_{\text{CE}}(P, Q) - \mathcal{H}(P) \leq \top_{\text{CE}} - \min_{\forall P(\mathbb{Z})} (\mathcal{H}(P)) \tag{7}$$

There is a special case worth noting. In practice, it is common to assume that Q is a uniform distribution, i.e., $q_i = 1/n, \forall q_i \in Q$, typically because Q is unknown or varies frequently. Hence the assumption leads to a code with an average length equaling $\log_2 n$ (or in practice, the smallest integer $\geq \log_2 n$). Under this special (but rather common) condition, all letters in a very long string have codewords of the same length. The worst case is that all letters in the string turn out to be the same letter. Since there is no informative variation in the PMF P for this very long string, i.e., $\mathcal{H}(P) = 0$, in principle, the transmission of this string is unnecessary. The maximal amount of inefficiency is thus $\log_2 n$. This is indeed much lower than the upper bound $\top_{\text{CE}} = n - 1$, justifying the assumption or use of a uniform Q in many situations.

A more formal proof of the boundedness of $\mathcal{H}_{\text{CE}}(P, Q)$ and $\mathcal{D}_{\text{KL}}(P||Q)$ for an alphabet with a finite number of letters can be found in Appendix A with more detailed discussions. It is necessary to note again that the discourse in this section and Appendix A does not imply that the KL-divergence is incorrect. Firstly, the KL-divergence applies to both discrete probability distributions (PMFs) and continuous distributions. Secondly, the KL-divergence is one of the many divergence measures found in information theory, and a member of the huge collection of statistical distance or difference measures. There is no simply answer as to which measure is correct and incorrect or which is better. We therefore should not over-generalize the proof to undermine the general usefulness of the KL-divergence.

5.2. Existing Candidates of Bounded Measures

In practical applications, numerical approximation is commonly used to bound KL-divergence by setting a small value $0 < \epsilon < 0.5$ and adjusting probability values in a PMF to ensure all $\epsilon \leq p \leq 1 - \epsilon$. While numerical approximation may provide a bounded KL-divergence, it is not easy to determine the value of ϵ and it is difficult to ensure everyone to use the same ϵ for the same alphabet or comparable alphabets. For a small alphabet, ϵ has to be a fairly large value, reducing the probability range noticeably. For example, a binary alphabet has maximum Shannon entropy 1 bit. One would need to set an $\epsilon > 0.22$ in order bound any \mathcal{D}_{KL} for this alphabet within $[0, 1]$. It is therefore desirable to consider bounded measures that may be used in place of \mathcal{D}_{KL} .

Jensen-Shannon divergence is such a measure:

$$\begin{aligned} \mathcal{D}_{\text{JS}}(P||Q) &= \mathcal{D}_{\text{JS}}(Q||P) = \frac{1}{2} (\mathcal{D}_{\text{KL}}(P||M) + \mathcal{D}_{\text{KL}}(Q||M)) \\ &= \frac{1}{2} \sum_{i=1}^n \left(p_i \log_2 \frac{2p_i}{p_i + q_i} + q_i \log_2 \frac{2q_i}{p_i + q_i} \right) \end{aligned} \tag{8}$$

where P and Q are two PMFs associated with the same alphabet \mathbb{Z} and M is the average distribution of P and Q . Each letter $z_i \in \mathbb{Z}$ is associated with a probability value $p_i \in P$ and another $q_i \in Q$. With the base 2 logarithm as in Equation (8), $\mathcal{D}_{\text{JS}}(P||Q)$ is bounded by 0 and 1.

The square root of $\mathcal{D}_{\text{JS}}(P||Q)$, denoted as $\sqrt{\mathcal{D}_{\text{JS}}(P||Q)}$, is not only a bounded divergence measure, but also a distance metric [46,47]. It is thus interesting to include $\sqrt{\mathcal{D}_{\text{JS}}}$ as a candidate measure.

Another bounded measure is the conditional entropy $\mathcal{H}(P|Q)$:

$$\mathcal{H}(P|Q) = \mathcal{H}(P) - \mathcal{I}(P; Q) = \mathcal{H}(P) - \sum_{i=1}^n \sum_{j=1}^n r_{i,j} \log_2 \frac{r_{i,j}}{p_i q_j} \tag{9}$$

where $\mathcal{I}(P; Q)$ is the mutual information between P and Q and $r_{i,j}$ is the joint probability of the two conditions of $z_i, z_j \in \mathbb{Z}$ that are associated with P and Q . $\mathcal{H}(P|Q)$ is bounded by 0 and $\mathcal{H}(P)$. Because $\mathcal{I}(P; Q)$ measures the amount of shared information between P and Q (and therefore a kind of similarity), $\mathcal{H}(P|Q)$ thus increases if P and Q are less similar. Note that we use $\mathcal{H}(P|Q)$ and $\mathcal{I}(P; Q)$ here in the context that P and Q are associated with the same alphabet \mathbb{Z} , though the general definitions of $\mathcal{H}(P|Q)$ and $\mathcal{I}(P; Q)$ are more flexible.

The above two measures in Equations (8) and (9) consist of logarithmic scaling of probability values, in the same form of Shannon entropy. They are entropic measures. There are many other divergence measures in information theory, including many in the family of f -divergences [48]. However, many are also unbounded.

Meanwhile, entropic divergence measures belong to the broader family of statistical distances or difference measures. In this work, we considered a set of non-entropic measures in the form of Minkowski distances, which have the following general form:

$$D_M^k(P, Q) = \sqrt[k]{\sum_{i=1}^n |p_i - q_i|^k} \quad (k > 0) \tag{10}$$

where we use symbol D instead of \mathcal{D} because it is not entropic.

5.3. New Candidates of Bounded Measures

For each letter $z_i \in \mathbb{Z}$, $\mathcal{D}_{KL}(P||Q)$ measures the difference between its self-information $-\log_2(p_i)$ and $-\log_2(q_i)$ with respect to P and Q . Similarly, $\mathcal{D}_{JS}(P||Q)$ measures the difference of self-information with the involvement of an average distribution $(P + Q)/2$. Meanwhile, it will be interesting to consider the difference of two probability values, i.e., $|p_i - q_i|$, and the information content of the difference. This would lead to measuring $\log_2 |p_i - q_i|$, which is unfortunately an unbounded term in $[-\infty, 0]$.

Let $u = |p_i - q_i|$, the function $\log_2 u^k + 1$ (where $k > 0$) is an isomorphic transformation of $\log_2 u$. The former preserves all information of the latter, while offering a bounded measure in $[0, 1]$. Although $\log_2 u^k + 1$ and $\log_2 u$ are both monotonically increasing measures, they have different gradient functions, or visually, different shapes. We thus introduce a power parameter k to enable our investigation into different shapes. The introduction of k reflects the open-minded nature of this work. It follows the same generalization approach as Minkowski distances and α -divergences [49], avoiding a fixation on their special cases such as the Euclidean distance or \mathcal{D}_{KL} .

We first consider a commutative measure \mathcal{D}_{new}^k :

$$\mathcal{D}_{new}^k(P||Q) = \frac{1}{2} \sum_{i=1}^n (p_i + q_i) \log_2(|p_i - q_i|^k + 1) \tag{11}$$

where $k > 0$. Because $0 \leq |p_i - q_i|^k \leq 1$, we have

$$\frac{1}{2} \sum_{i=1}^n (p_i + q_i) \log_2(0 + 1) \leq \mathcal{D}_{new}^k(P||Q) \leq \frac{1}{2} \sum_{i=1}^n (p_i + q_i) \log_2(1 + 1)$$

Since $\log_2 1 = 0$, $\log_2 2 = 1$, $\sum p_i = 1$, $\sum q_i = 1$, $\mathcal{D}_{new}^k(P||Q)$ is thus bounded by 0 and 1. The formulation of $\mathcal{D}_{new}^k(P||Q)$ was derived from its non-commutative version:

$$\mathcal{D}_{ncm}^k(P||Q) = \sum_{i=1}^n p_i \log_2(|p_i - q_i|^k + 1) \tag{12}$$

which captures the non-commutative property of \mathcal{D}_{KL} . In this work, we focus on two options of \mathcal{D}_{new}^k and $\mathcal{D}_{ncm}^k(P||Q)$, i.e., when $k = 1$ and $k = 2$.

As \mathcal{D}_{JS} , \mathcal{D}_{new}^k , and \mathcal{D}_{ncm}^k are bounded by $[0, 1]$, if any of them is selected to replace \mathcal{D}_{KL} , Equation (3) can be rewritten as

$$\text{Benefit} = \mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1}) - \mathcal{H}_{\max}(\mathbb{Z}_i)\mathcal{D}(\mathbb{Z}'_i|\mathbb{Z}_i) \tag{13}$$

where \mathcal{H}_{\max} denotes maximum entropy, while \mathcal{D} is a placeholder for \mathcal{D}_{JS} , $\mathcal{D}_{\text{new}}^k$, or $\mathcal{D}_{\text{ncm}}^k$. Note that while $\mathcal{H}_{\max}(\mathbb{Z}_i)\mathcal{D}(\mathbb{Z}'_i|\mathbb{Z}_i)$ is bounded by $\mathcal{H}_{\max}(\mathbb{Z}_i)$, $\mathcal{H}_{\max}(\mathbb{Z}_i)$ can have any non-negative value and is calculated as $\log_2 \|\mathbb{Z}_i\|$, where $\|\mathbb{Z}_i\|$ is the number of letters in \mathbb{Z}_i .

We have considered the option of using $\mathcal{H}(\mathbb{Z}_i)$ instead of $\mathcal{H}_{\max}(\mathbb{Z}_i)$. However, this would lead to an undesirable paradox. Consider an alphabet $\mathbb{Z}_i = \{z_a, z_b\}$ with a PMF $P_i = \{p_a, 1 - p_a\}$. Consider a simple visual mapping that is supposed to encode the probability value p_a using the luminance of a monochrome shape with, $\text{luminance}(p_a) = p_a$, black = 0, and white = 1. Unfortunately, the accompanying legend displays incorrect labels as black for $p_a = 1$ and white for $p_a = 0$. The visualization results thus feature a ‘‘lie’’ distribution $P_i = \{1 - p_a, p_a\}$. An obvious paradoxical scenario is when $P_i = \{1, 0\}$, which has an entropy value $\mathcal{H}(\mathbb{Z}_i) = 0$. Although \mathcal{D}_{JS} , $\mathcal{D}_{\text{new}}^k$, and $\mathcal{D}_{\text{ncm}}^k$ would all return 1 as the maximum value of divergence for the visual mapping, the term $\mathcal{H}(\mathbb{Z}_i)\mathcal{D}(\mathbb{Z}'_i|\mathbb{Z}_i)$ would indicate that there would be no divergence. Hence $\mathcal{H}(\mathbb{Z}_i)$ cannot be used instead of $\mathcal{H}_{\max}(\mathbb{Z}_i)$.

6. Conceptual Evaluation of Bounded Measures

Given those bounded candidates in the previous section, we would like to select the most suitable measure to be used in Equation (13). In the history of measurement science [50], there have been an enormous amount of research effort devoted to inventing, evaluating, and selecting different candidate measures (e.g., metric vs. imperial measurement systems; temperature scales: Celsius, Fahrenheit, kelvin, Rankine, and Reaumur; and Seismic magnitude scales: Richter, Mercalli, moment magnitude, and many others). There is usually no ground truth as to which is correct, and the selection decision is rarely determined only by mathematical definitions or rules [51]. Similarly, there are numerous statistical distance and difference measures. selecting a measure in a certain application is often an informed decision based on multiple factors. Measuring the benefit of visualization and the related informative divergence in visualization processes is a new topic in the field of visualization. It is not unreasonable to expect that more research effort will be made in the coming years, decades, or unsurprisingly, centuries. The work presented in this two-part paper represents the early thought and early effort in this endeavor. In this work, we devised a set of criteria and conducted multi-criteria decision analysis (MCDA) [3] to evaluate the candidate measures described in the previous section.

Our criteria fall into two main categories. The first group of criteria reflect seven desirable conceptual or mathematical properties, as shown in Table 3. The second group of criteria reflect the assessments based on numerical instances constructed synthetically or obtained from experiments. This first part of the paper focuses conceptual evaluation based on the first group of criteria, while the second part focuses on empirical evaluation based on the second group of criteria [5].

Table 3. A summary of multi-criteria decision analysis in the first part of this paper. Each measure is scored against a conceptual criterion using an integer in [0, 5] with 5 being the best. The symbol ► indicates an interim conclusion after considering one or a few criteria. In the second part of the paper [5], we will discuss another five criteria.

Criteria	Importance	$0.3\mathcal{D}_{\text{KL}}$	\mathcal{D}_{JS}	$\sqrt{\mathcal{D}_{\text{JS}}}$	$\mathcal{H}(P Q)$	$\mathcal{D}_{\text{new}}^{k=1}$	$\mathcal{D}_{\text{new}}^{k=2}$	$\mathcal{D}_{\text{ncm}}^{k=1}$	$\mathcal{D}_{\text{ncm}}^{k=2}$	$\mathcal{D}_{\text{M}}^{k=2}$	$\mathcal{D}_{\text{M}}^{k=200}$
1. Boundedness ► $0.3\mathcal{D}_{\text{KL}}$ is eliminated but used below only for comparison. The other scores are carried forward.	critical	0	5	5	5	5	5	5	5	3	3
2. Number of PMFs	important	5	5	5	2	5	5	5	5	5	5
3. Entropic measures	important	5	5	5	5	5	5	5	5	1	1
4. Distance metric	helpful	2	3	5	2	4	3	2	2	5	5
5. Easy to understand	helpful	4	4	3	4	4	3	4	3	5	4
6. Curve shapes (Figure 5)	helpful	5	5	3	1	2	4	2	4	2	2
7. Curve shapes (Figure 6)	helpful	5	3	4	1	3	5	3	5	2	3
► Eliminate $\mathcal{H}(P Q)$, \mathcal{D}_{M}^2 , $\mathcal{D}_{\text{M}}^{200}$ based on criteria 1–7	sum:		30	30	20	28	30	26	29	23	23

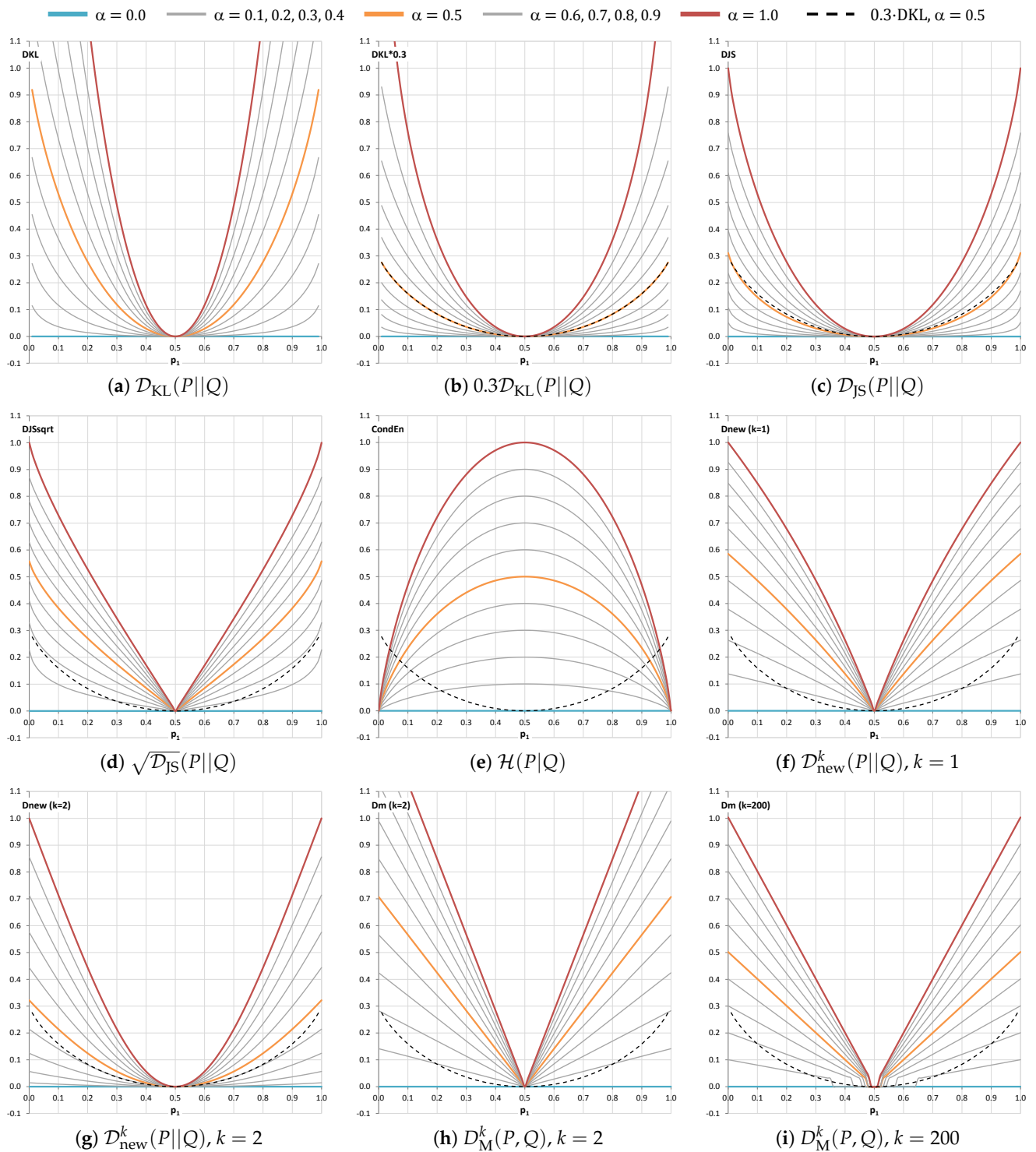


Figure 5. The different measurements of the divergence of two PMFs, $P = \{p_1, 1 - p_1\}$ and $Q = \{q_1, 1 - q_1\}$. The x-axis shows p_1 , varying from 0 to 1, while we set $q_1 = (1 - \alpha)p_1 + \alpha(1 - p_1)$, $\alpha \in [0, 1]$. When $\alpha = 1$, Q is most divergent away from P . The curve $0.3D_{KL}(\alpha = 0.5)$ is shown in a dashed black line, and is used as a benchmark for observing the corresponding curves (in orange) produced by the candidate measures in (c–i).

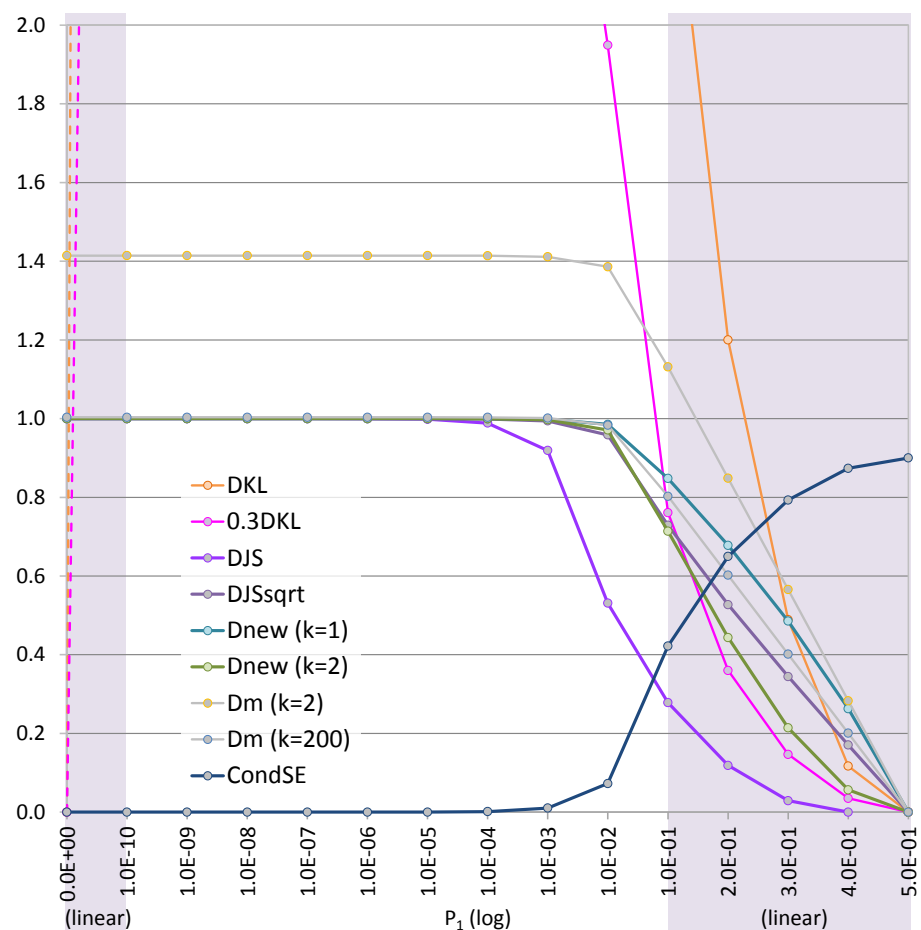


Figure 6. A visual comparison of the candidate measures in a range near zero. Similar to Figure 5, $P = \{p_1, 1 - p_1\}$ and $Q = \{q_1, 1 - q_1\}$, but only the curve $\alpha = 1$ is shown, i.e., $q_1 = 1 - p_1$. The line segments of \mathcal{D}_{KL} and $0.3\mathcal{D}_{KL}$ in the range $[0, 0.1^{10}]$ do not represent the actual curves. The ranges $[0, 0.1^{10}]$ and $[0.1, 0.5]$ are only for references to the nearby contexts as they do not use the same logarithmic scale as in $[0.1^{10}, 0.1]$.

For criteria 1, 6, and 7 in the first group, we use visualization plots to aid our analysis of the mathematical properties. Based on our analysis, we score each divergence measure against a criterion using ordinal values between 0 and 5 (0 unacceptable, 1 fall-short, 2 inadequate, 3 mediocre, 4 good, 5 best). We intentionally do not assign weights to these criteria. While we will offer our view as to the importance of different criteria, we encourage readers to apply their own judgement to weight these criteria. We hope that readers will reach the same conclusion as ours. We draw our conclusion about the conceptual evaluation in Section 7, where we also outline the need for data-driven empirical evaluation.

6.1. Criterion 1: Is It a Bounded Measure?

This is essential since the selected divergence measure is to be bounded. Otherwise we could just use the KL-divergence. Let us consider a simple alphabet $\mathbb{Z} = \{z_1, z_2\}$, which is associated with two PMFs, $P = \{p_1, 1 - p_1\}$ and $Q = \{q_1, 1 - q_1\}$. We set $q_1 = (1 - \alpha)p_1 + \alpha(1 - p_1), \alpha \in [0, 1]$, such that when $\alpha = 1$, Q is most divergent away from P . The entropy values of P and Q fall into the range of $[0, 1]$. Hence semantically, it is more intuitive to reason an unsigned value representing their divergence within the same range.

Figure 5 shows several measures by varying the values of p_1 in the range of $[0, 1]$. We can observe that \mathcal{D}_{KL} raises its values quickly above 1 when $\alpha = 1, p_1 \leq 0.22$. Its scaled version, $0.3\mathcal{D}_{KL}$, does not rise up as quick as \mathcal{D}_{KL} but raises above 1 when $\alpha = 1, p_1 \leq 0.18$.

In fact \mathcal{D}_{KL} and $0.3\mathcal{D}_{\text{KL}}$ are not only unbounded, they do not return valid values when $p_1 = 0$ or $p_1 = 1$. We therefore score them 0 for Criterion 1.

\mathcal{D}_{JS} , $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{H}(P|Q)$, $\mathcal{D}_{\text{new}}^k$, and $\mathcal{D}_{\text{ncm}}^k$ are all bounded by $[0, 1]$ and they can potentially be used in the rewritten formula Equation (13). We score them 5. Although \mathcal{D}_{M}^k is a bounded measure, its semantic interpretation is not ideal, because its upper bound depends on k and is always >1 . We thus score it 3. Although $0.3\mathcal{D}_{\text{KL}}$ is eliminated based on criterion 1, it is kept in Table 3 as a benchmark in analyzing criteria 2–7. Meanwhile, we carry all other scores forward to the next stage of analysis.

6.2. Criterion 2: How Many PMFs Does It Have as Dependent Variables

For criteria 2–7, we follow the base-criterion method [52] by considering \mathcal{D}_{KL} and $0.3\mathcal{D}_{\text{KL}}$ as the benchmark. Criterion 2 concerns the number of PMFs as the dependent (or input) variables of each measure. \mathcal{D}_{KL} and $0.3\mathcal{D}_{\text{KL}}$ depend on two PMFs, P and Q . All candidates of the bounded measures depend on two PMFs, except the conditional entropy $\mathcal{H}(P|Q)$ that depends on three. Because in most practical applications, it requires some effort to obtain a PMF, e.g., by observing an alphabet for a period. A joint probability distribution, which is required for calculating the mutual information term $\mathcal{I}(P; Q)$ in Equation (9), would need observation of both input and output alphabets of a process in a synchronized manner. The need for an extra PMF makes $\mathcal{H}(P|Q)$ much less favourable, and it is scored 2. All others are scored 5.

6.3. Criterion 3: Is It an Entropic Measure?

An entropic measure characteristically features a logarithmic transformation of some numerical compositions of probability values. The logarithmic transformation accentuates the change from a state of order to a state of disorder or vice versa. The probabilistic mean of such changes related to all letters in an alphabet pertains to the Shannon entropy. With base 2 logarithm, an entropic measure usually has or features the unit bit. As the AC term in Equation (3) and the original PD term (i.e., \mathcal{D}_{KL}) are measured in bits, we prefer to have an entropic divergence measure for the PD term so the “benefit” can be measured in bits. For this reason, \mathcal{D}_{M}^k is scored 1, and all others are given 5.

6.4. Criterion 4: Is It a Distance Measure?

When a measure is referred to as a divergence measure, it usually implies that it is not a distance metric. A true distance metric must have the following mathematical properties:

1. identity: $d(x, y) = 0 \iff x = y$,
2. symmetry: $d(x, y) = d(y, z)$,
3. triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$,
4. non-negativity: $d(x, y) \geq 0$.

The first three conditions are axioms of a metric system. Among the candidate measures, \mathcal{D}_{M}^k and $\sqrt{\mathcal{D}_{\text{JS}}}$ are metrics. They are scored 5. $0.3\mathcal{D}_{\text{KL}}$, $\mathcal{H}(P|Q)$, $\mathcal{D}_{\text{ncm}}^{k=1}$, and $\mathcal{D}_{\text{ncm}}^{k=2}$ satisfy only conditions 1 and 4. They are scored 2. \mathcal{D}_{JS} and $\mathcal{D}_{\text{new}}^{k=2}$ satisfy conditions 1, 2, 4, and they are scored 3. At the moment, we do not know if $\mathcal{D}_{\text{new}}^{k=1}$ is a metric or not. There is a mathematical proof to show that $\mathcal{D}_{\text{new}}^{k=1}$ is a metric for 2-letter alphabets [53], but a proof or disproof for n -letter alphabets is yet known. We thus give $\mathcal{D}_{\text{new}}^{k=1}$ a score 4.

6.5. Criterion 5: Is It Intuitive or Easy to Understand?

One reason that \mathcal{D}_{KL} is the most popular divergence measure is that it is easy to understand the meaning of its element function $f(p, q) = \log(p/q) = \log(p) - \log(q)$, where $p, q \in [0, 1]$ are two probability values. It is the difference of the logarithmic representations of p and q . As $0.3\mathcal{D}_{\text{KL}}$ introduces a global scaling transformation, it adds a barrier in understanding. We take one score away for such a barrier by giving $0.3\mathcal{D}_{\text{KL}}$ a score 4.

As shown in Equation (8), \mathcal{D}_{JS} introduces an intermediate value $m = (p + q)/2$. The element function splits into two parts $f(p, m)$ and $f(q, m)$. Such a transformation adds a

barrier in our appreciation of meaning of the measure. Note that one could take two points away as this transformation is rather complex. Nevertheless, for consistency, we take one point away per transformation. $\sqrt{\mathcal{D}_{JS}}$ introduces a square root as a global transformation, adding a further barrier in understanding. With \mathcal{D}_{KL} as the benchmark (score 5), we score \mathcal{D}_{JS} 4 and $\sqrt{\mathcal{D}_{JS}}$ 3 by counting the number of barriers in understanding.

Consider another element function $g(p, q) = \log |p - q|$, which is the logarithmic representation of the difference between p and q . It is as easy to understand as $f(p, q)$. $\mathcal{D}_{new}^{k=1}$ introduces a transformation as $\log(|p - q| + 1)$, while $\mathcal{D}_{new}^{k=2}$ introduces an additional one as $\log(|p - q|^2 + 1)$. Each transformation adds a new barrier in understanding. We therefore give $\mathcal{D}_{new}^{k=1}$ a score 4 and $\mathcal{D}_{new}^{k=2}$ a 3. Similarly we assign a score 4 to $\mathcal{D}_{ncm}^{k=1}$ and a 3 to $\mathcal{D}_{ncm}^{k=2}$.

For n -letter alphabet, D_M^k ($k = 2$) is the same as the n -dimensional Euclidean distance. We thus gives it a full score 5. As D_M^k ($k = 200$) is considered to have an extra barrier in understanding, we score it 4.

Finally, $\mathcal{H}(P|Q)$ is the composition of two commonly-used information-theoretic measures. We give it a score 4 by considering the composition as a transformation.

6.6. Criterion 6: Visual Analysis of Curve Shapes in the Range of (0, 1)

One may wish for a bounded measure to have a geometric behaviour similar to \mathcal{D}_{KL} since it is the most popular divergence measure. Since \mathcal{D}_{KL} rises up far too quickly as shown in Figure 5, we use $0.3\mathcal{D}_{KL}$ as a benchmark, though it is still unbounded. As Figure 5 plots the curves for $\alpha = 0.0, 0.1, \dots, 1.0$, we can visualize the “geometric shape” of each bounded measure, and compare it with that of $0.3\mathcal{D}_{KL}$.

From Figure 5, we can observe that \mathcal{D}_{JS} has almost a perfect match when $\alpha = 0.5$, while \mathcal{D}_{new}^k ($k = 2$) is also fairly close. They thus score 5 and 4 respectively in Table 3. Meanwhile, the lines of $\mathcal{H}(P|Q)$ curve in the opposite direction of $0.3\mathcal{D}_{KL}$. We score it 1. $\sqrt{\mathcal{D}_{JS}}$, \mathcal{D}_{new}^k ($k = 1$), and D_M^k ($k = 2, k = 200$) are of similar shapes. In terms of the direction of curvature, $\sqrt{\mathcal{D}_{JS}}$ correlates slightly better with $0.3\mathcal{D}_{KL}$ than D_M^k and \mathcal{D}_{new}^k ($k = 1$). We thus assign a score 3 to $\sqrt{\mathcal{D}_{JS}}$ and a score 2 to \mathcal{D}_{new}^k ($k = 1$) and D_M^k ($k = 2, k = 200$). For the PMFs P and Q concerned, \mathcal{D}_{ncm}^k has the same curves as \mathcal{D}_{new}^k . Hence \mathcal{D}_{ncm}^k has the same score as \mathcal{D}_{new}^k in Table 3.

6.7. Criterion 7: Visual Analysis of Curve Shapes in a Range near Zero, i.e., [0.1¹⁰, 0.1]

We now consider Figure 6, where the candidate measures are visualized in comparison with \mathcal{D}_{KL} and $0.3\mathcal{D}_{KL}$ in a range close to zero, i.e., [0.1¹⁰, 0.1]. The ranges [0, 0.1¹⁰] and [0.1, 0.5] are there only for references to the nearby contexts as they do not have the same logarithmic scale as that in the range [0.1¹⁰, 0.1]. We can observe that in [0.1¹⁰, 0.1] the curve of $0.3\mathcal{D}_{KL}$ rises as almost quickly as \mathcal{D}_{KL} . This confirms that simply scaling the KL-divergence is not an adequate solution. The curves of $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$ converge to their maximum value 1.0 earlier than that of \mathcal{D}_{JS} . The $\sqrt{\mathcal{D}_{JS}}$ curve appears between those of $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$. If the curve of $0.3\mathcal{D}_{KL}$ is used as a benchmark as in Figure 5, the curve of $\mathcal{D}_{new}^{k=2}$ is much closer to $0.3\mathcal{D}_{KL}$ than that of \mathcal{D}_{JS} . We thus score $\mathcal{D}_{new}^{k=2}$: 5, $\sqrt{\mathcal{D}_{JS}}$: 4, \mathcal{D}_{JS} : 3, $\mathcal{D}_{new}^{k=1}$: 3, D_M^k ($k = 200$): 3, D_M^k ($k = 200$): 2, and $\mathcal{H}(P|Q)$: 1. Same as Figure 5, \mathcal{D}_{ncm}^k has the same curves and thus the same score as \mathcal{D}_{new}^k .

The sums of the scores for criteria 1–7 indicate that $\mathcal{H}(P|Q)$ and D_M^k are much less favourable than \mathcal{D}_{JS} , $\sqrt{\mathcal{D}_{JS}}$, \mathcal{D}_{new}^k , and \mathcal{D}_{ncm}^k . Because these criteria have more holistic significance than the data-driven analysis in the second part of this paper [5], we can eliminate $\mathcal{H}(P|Q)$ and D_M^k for further consideration. Ordinal scores in MCDA are typically subjective. Nevertheless, in our analysis, ± 1 in those scores would not affect the elimination.

7. Discussions and Conclusions

In this paper, we have considered the need to improve the mathematical formulation of an information-theoretic measure for analyzing the cost–benefit of visualization as well as other processes in a data intelligence workflow [1]. The concern about the original

measure is its unbounded term based on the KL-divergence. As discussed in the early sections of this paper, although using the KL-divergence measure in [1] as part of the cost–benefit measure is a conventional or orthodox choice, its unboundedness leads to several issues in the potential applications of the cost–benefit measure to practical problems:

- It is not intuitive to interpret a set of values that would indicate that the amount of distortion in viewing a visualization that features some information loss, could be much more than the total amount of information contained in the visualization.
- It is difficult to specify some simple visualization phenomena. For example, before a viewer observes a variable x using visualization, the viewer incorrectly assumes that the variable is a constant (e.g., $x \equiv 10$, and probability $p(10) = 1$). The KL-divergence cannot measure the potential distortion of this phenomenon of bias because this is a singularity condition, unless one changes $p(10)$ by subtracting a small value $0 < \epsilon < 1$.
- If one tries to restrict the KL-divergence to return values within a bounded range, e.g., determined by the maximum entropy of the visualization space or the underlying data space, one could potentially lose a non-trivial portion of the probability range (e.g., 13% in the case of a binary alphabet).

To address these problems, we have proposed to replace the KL-divergence in the cost–benefit measure with a bounded measure. We have obtained a proof that the divergence used in the cost–benefit formula is conceptually bounded, as long as the input and output alphabets of a process have a finite number of letters.

We have considered a number of bounded measures to replace the unbounded term, including a new divergence measure $\mathcal{D}_{\text{new}}^k$ and its variant $\mathcal{D}_{\text{ncm}}^k$. We have conducted multi-criteria decision analysis to select the best measure among these candidates. In particular, we have used visualization to aid the observation of the mathematical properties of the candidate measures, assisting in the analysis of three criteria in considered in this paper.

From Table 3, we can observe the process of narrowing down from eight candidate measures to five measures. In particular, three candidate measures \mathcal{D}_{JS} , $\sqrt{\mathcal{D}_{\text{JS}}}$, and $\mathcal{D}_{\text{new}}^{k=2}$ received the same total scores. They are followed by $\mathcal{D}_{\text{ncm}}^{k=2}$ and $\mathcal{D}_{\text{new}}^{k=1}$. It is not easy to separate them. We therefore conducted two groups of case studies to collect empirical evidence for further evaluating these candidate measures.

In the history of measurement science [50], as shown in Figure 7, scientists encountered many similar dilemma in choosing different measures. For example, temperature measures Celsius, Fahrenheit, Réaumur, Rømer, and Delisle scales exhibit similar mathematical properties, their proposal and adoption were largely determined by practical instances:

- Rømer—0 degree: freezing brine, 7.5 degree: the freezing point of water, 60 degree: the boiling point of water;
- Fahrenheit (original)—0 degree: the freezing point of brine (a high-concentration solution of salt in water), 32 degree: ice water, 96 degree: average human body temperature;
- Fahrenheit (present)—32 degree: the freezing point of water, 212 degree: the boiling point of water;
- Réaumur—0 degree: the freezing point of water, 80 degree: the boiling point of water;
- Delisle—0 degree: the boiling point of water, −1 degree: the contraction of the mercury in hundred-thousandths.
- Celsius* (original)—0 degree: the boiling point of water, 100 degree: the freezing point of water;
- Celsius (1743–1954)—0 degree: the freezing point of water, 100 degree: the boiling point of water;
- Celsius (1954–2019)—redefined based on absolute zero and the triple point of VSMOW (specially prepared water);
- Celsius (2019–now)—redefined based on the Boltzmann constant.

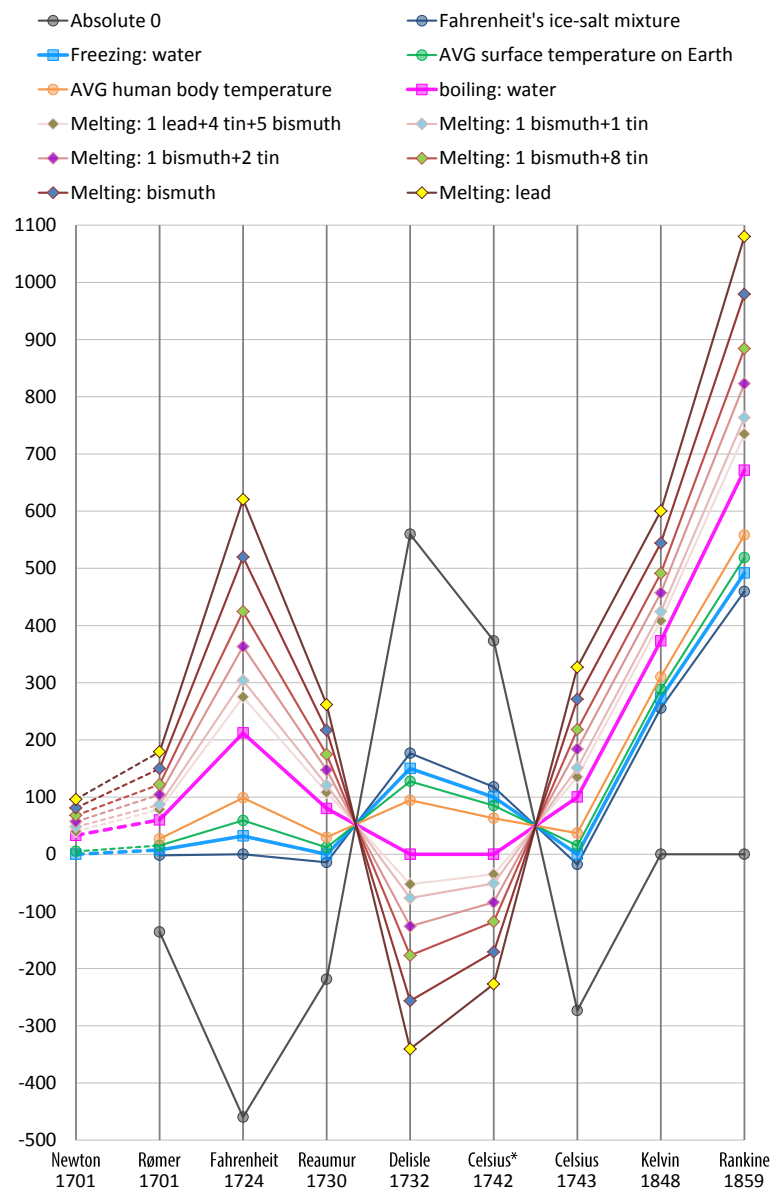


Figure 7. Some of the major temperature scales considered by scientists in the history. It took four decades from Isaac Newton’s instance-based proposal to arrive at the most-commonly used Celsius scale. It took another century to discover absolute zero as the lower bound.

Before the development of these scales, Newton proposed two temperature systems based on his observation of some 18 instance values [54]. The effort of these scientific pioneers suggested that observing how candidate measures relate to practical instances was part of the scientific processes for selecting different candidate measures.

The work reported in this paper outlines a conceptual notion that measuring the divergence that may be caused by a transformation from an input alphabet (e.g., data) to an output alphabet (e.g., visualization) is more intuitive if it has a lower bound 0 and an upper bound of the maximum entropy of the input alphabet (as enforced in Equation (13)). The more complex the input alphabet (i.e., the input information space), the wider the range of the potential divergence. As all candidate measures are bounded by [0, 1] and the maximum entropy of an alphabet is easy to calculate, we have addressed part of the problem where \mathcal{D}_{KL} has no upper bound regardless how complex the input alphabet is.

Building on the work presented in this paper, we carried out further investigation into a group of criteria based on observed instances in synthetic and experimental data. This data-driven evaluation is presented in the second part of this paper [5], where we aim to

narrow the remaining six candidate measures to one measure, and to revise the original cost–benefit ratio in [1] based on the combined conclusion derived from the conceptual evaluation (i.e., this work) and empirical evaluation. The empirical evidence collected in several case studies helps identify some additional strengths and weaknesses of the remaining six candidate measures. Based on conceptual and empirical criteria considered in both parts of the paper, we will offer a conclusion that the candidate measure $\mathcal{D}_{\text{new}}^k (k = 2)$ is ahead of $\sqrt{\mathcal{D}_{\text{JS}}}$, especially when we include an additional conceptual criterion discovered during the case studies. Readers can find the detailed description and analysis of these case studies in the second part of this paper [5].

Author Contributions: Conceptualization, M.C.; methodology, M.C. and M.S.; validation, M.C. and M.S.; formal analysis, M.C. and M.S.; investigation, M.C. and M.S.; writing—original draft preparation, M.C.; writing—review and editing, M.S.; visualization, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: M.C. would like to acknowledge that some collaborative meetings of this work were made possible by the Network of European Data Scientists (NeEDS), a Research and Innovation Staff Exchange (RISE) project under the Marie Skłodowska-Curie Program. M.S. has been supported in part by project PID2019-106426RB-C31 funded by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Alphabet Compression
JS	Jenson–Shannon
KL	Kullback–Leibler
MCDA	Multi-Criteria Decision Analysis
MIP	Maximum Intensity Projection
PD	Potential Distortion
PMF	Probability Mass Function

Appendix A. Conceptual Boundedness of $\mathcal{H}_{\text{CE}}(P, Q)$ and \mathcal{D}_{KL}

According to the mathematical definition of cross entropy:

$$\mathcal{H}(P, Q) = - \sum_{i=1}^n p_i \log_2 q_i = \sum_{i=1}^n p_i \log_2 \frac{1}{q_i} \quad (\text{A1})$$

$\mathcal{H}(P, Q)$ is of course unbounded. When $q_i \rightarrow 0$, we have $\log_2 \frac{1}{q_i} \rightarrow \infty$. As long as $p_i \neq 0$ and is independent of q_i , $\mathcal{H}(P, Q) \rightarrow \infty$. Hence the discussion in this appendix is not about a literal proof that $\mathcal{H}(P, Q)$ is unbounded when this mathematical formula is applied without any change. It is about that the concept of cross entropy implies that it should be bounded when n is a finite number.

Definition A1. Given an alphabet \mathbb{Z} with a true PMF P , cross-entropy $\mathcal{H}(P, Q)$ is the average number of bits required when encoding \mathbb{Z} with an alternative PMF Q .

This is a widely-accepted and used definition of cross-entropy in the literature of information theory. Firstly, with Shannon entropy $\mathcal{H}(P) = \sum_{i=0}^n p_i \log_2(1/p_i)$, the term $\log_2(1/p_i)$ is considered as the mathematically-supposed length of a codeword that is used

to encode letter $z_i \in \mathbb{Z}$ with a probability value $p_i \in P$. Note that if $p_i = 0$, z_i does not need to be encoded since it will never occur in communication. Here, a *codeword* is the digital representation of a letter in an alphabet. A *code* is a collection of the codewords for all letters in an alphabet. In communication and computer science, we usually use binary codes as digital representations for alphabets, such as ASCII code and variable-length codes.

When PMF Q is used for encoding \mathbb{Z} instead of PMF P , we can observe from the cross entropy formula in Equation (A1), $\log_2(1/q_i)$ is considered as the mathematically-supposed length of a codeword that is used to encode letter $z_i \in \mathbb{Z}$ with a probability value $q_i \in Q$. Because $\sum_{i=1}^n p_i = 1$, $\mathcal{H}(P, Q)$ is thus the weighted or probabilistic average length of the codewords for all letters in \mathbb{Z} , such that the weights are based on the actual PMF P and the codeword lengths are based on the supposed PMF Q .

When a letter $z_i \in \mathbb{Z}$ is given a probability value q_i , it is not necessary for z_i to be encoded using a codeword of length $\log_2(1/q_i)$ bits. More precisely, it is the nearest integer above or equal to it, i.e., $\lceil \log_2(1/q_i) \rceil$ bits, since a binary codeword cannot have fractional bits digitally. For example, consider a simple alphabet $\mathbb{Z} = \{z_1, z_2\}$. Regardless what PMF is associated with \mathbb{Z} , \mathbb{Z} can always be encoded with a 1-bit code, e.g., codeword 0 for z_1 and codeword 1 for z_2 , as long as neither of the two probability values in Q is zero, i.e., $q_1 \neq 0$ and $q_2 \neq 0$. (Note that if a codeword were to have a zero probability, we would not need to encode the codeword. It would not increase the number of bits required for coding.) However, if we had followed Equation (A1) literally, we would have created codes similar to the following examples:

- if $Q = \{\frac{1}{2}, \frac{1}{2}\}$, codeword 0 for z_1 and codeword 1 for z_2 ;
- if $Q = \{\frac{3}{4}, \frac{1}{4}\}$, codeword 0 for z_1 and codeword 10 for z_2 ;
- ...
- if $Q = \{\frac{63}{64}, \frac{1}{64}\}$, codeword 0 for z_1 and codeword 11111 for z_2 ;
- ...

As shown in Figure A1a, such a code is very wasteful. Hence, in practice, encoding \mathbb{Z} according to Equation (A1) literally is not desirable. Note that the discussion about encoding is normally conducted in conjunction with the Shannon entropy. Here, we use the cross entropy formula for our discussion to avoid a deviation from the flow of reasoning.

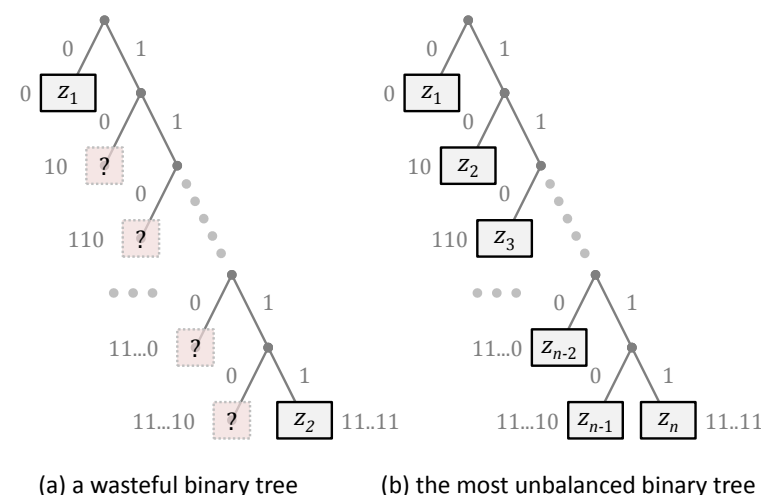


Figure A1. Two examples of binary codes illustrated as binary trees.

Let \mathbb{Z} be an alphabet with a finite number of letters, $\{z_1, z_2, \dots, z_n\}$, and \mathbb{Z} is associated with a PMF, Q , such that:

$$\begin{aligned} q(z_n) &= \epsilon, \quad (\text{where } 0 < \epsilon < 2^{-(n-1)}), \\ q(z_{n-1}) &= (1 - \epsilon)2^{-(n-1)}, \\ q(z_{n-2}) &= (1 - \epsilon)2^{-(n-2)}, \\ &\dots \\ q(z_2) &= (1 - \epsilon)2^{-2}, \\ q(z_1) &= (1 - \epsilon)2^{-1} + (1 - \epsilon)2^{-(n-1)}. \end{aligned} \tag{A2}$$

We can encode this alphabet using the Huffman encoding that is a practical binary coding scheme and adheres the principle to obtain a code with the Shannon entropy as the average length of codewords [45]. Entropy coding is designed to minimize the average number of bits per letter when one transmits a “very long” sequence of letters in the alphabet over a communication channel. Here the phrase “very long” implies that the string exhibits the above PMF Q (Equation (A2)). In other words, given an alphabet \mathbb{Z} and a PMF Q , the Huffman encoding algorithm creates an optimal code with the lowest average length of codewords when the code is used to transmit a “very long” sequence of letters in \mathbb{Z} . One example of such a code for the above PMF Q is:

$$\begin{aligned} z_1 &: 0, & z_2 &: 10, & z_3 &: 110 \\ &\dots & & & & \\ z_{n-1} &: 111 \dots 10 \quad (\text{with } n - 2 \text{ “1”s and one “0”}) \\ z_n &: 111 \dots 11 \quad (\text{with } n - 1 \text{ “1”s and no “0”}) \end{aligned} \tag{A3}$$

Figure A1b shows illustrates such a code using a binary tree. In this way, z_n , which has the smallest probability value, will always be assigned a codeword with the maximum length of $n - 1$.

Lemma A1. *Let \mathbb{Z} be an alphabet with n letters and \mathbb{Z} is associated with a PMF Q . If \mathbb{Z} is encoded using the aforementioned entropy coding, the maximum length of any codeword for $z_i \in \mathbb{Z}$ is always $\leq n - 1$.*

We can prove this lemma by confirming that when one creates a binary code for an n -letter alphabet \mathbb{Z} , the binary tree shown in Figure A1b is the worst unbalanced tree without any wasteful leaf nodes. Visually, we can observe that the two letters with the lowest values always share the lowest internal node as their parent node. The remaining $n - 2$ letters are to be hung on the rest binary subtree. Because the subtree is not allowed to waste leaf space, the $n - 2$ leaf nodes can be comfortably hung on the root and up to $n - 3$ internal node. A formal proof can be obtained using induction. For details, readers may find Golin’s lecture notes useful [55]. See also [7] for related mathematical theorems.

Theorem A1. *Let \mathbb{Z} be an alphabet with a finite number of letters and \mathbb{Z} is associated with two PMFs, P and Q . With the Huffman encoding, conceptually the cross entropy $\mathcal{H}(P, Q)$ should be bounded.*

Let n be the number of letters in \mathbb{Z} . According to Lemma A1, when \mathbb{Z} is encoded in conjunction with PMF Q using the Huffman encoding, the maximum codeword length is $\leq n - 1$. In other words, in the worst case scenario, there is letter $z_k \in \mathbb{Z}$ that has the lowest probability value q_k , i.e., $q_k \leq q_j \forall j = 1, 2, \dots, n$ and $j \neq k$. With the Huffman encoding, z_k will be encoded with the longest codeword of up to $n - 1$ bits.

According to Definition A1, there is a true PMF P . Let $L(z_i, q_i)$ be the codeword length of $z_i \in \mathbb{Z}$ determined by the Huffman encoding. We can write a conceptual cross entropy formula as:

$$\mathcal{H}(P, Q) = \sum_{i=1}^n p_i \cdot L(z_i, q_i) \leq \sum_{i=1}^n p_i \cdot L(z_k, q_k) \leq n - 1$$

where q_k is the lowest probability value in Q and z_k is encoded with a codeword of up to $n - 1$ bits (i.e., $L(z_k, q_k) \leq n - 1$). Hence conceptually $\mathcal{H}(P, Q)$ is bounded by $n - 1$ if the Huffman encoding is used. Since we can find a bounded solution for any n -letter alphabet with any PMF, the claim of unboundedness has been falsified.

Corollary A1. *Let \mathbb{Z} be an alphabet with a finite number of letters and \mathbb{Z} is associated with two PMFs, P and Q . With the Huffman encoding, conceptually the KL-divergence $\mathcal{D}_{\text{KL}}(P||Q)$ should be bounded.*

For an alphabet \mathbb{Z} with a finite number of letters, the Shannon entropy $\mathcal{H}(P)$ is bounded regardless any PMF P . The upper bound of $\mathcal{H}(P)$ is $\log_2 n$, where n is the number of letters in \mathbb{Z} . Since we have

$$\begin{aligned} \mathcal{H}(P, Q) &= \mathcal{H}(P) + \mathcal{D}_{\text{KL}}(P||Q) \\ \mathcal{D}_{\text{KL}}(P||Q) &= \mathcal{H}(P, Q) - \mathcal{H}(P) \end{aligned}$$

using Theorem A1, we can infer that with the Huffman encoding, conceptually $\mathcal{D}_{\text{KL}}(P||Q)$ is also bounded.

Further Discussion: The code created using Huffman encoding is also considered to be optimal for source coding (i.e., assuming without the need for error correction and detection). A formal proof can be found in [55].

Let \mathbb{Z} be an n -letter alphabet, and Q be its associated PMF. When we use the Shannon entropy to determine the length of each codeword mathematically, we have:

$$L(z_i, q_i) = \lceil \log_2 \frac{1}{q_i} \rceil, \quad z_i \in \mathbb{Z}, q_i \in Q$$

As we showed before, the length of a codeword can be infinitely long if $q_i \rightarrow 0$. Huffman encoding makes the length finite as long as n is finite. This difference between the mathematically-literal entropy encoding and Huffman encoding is important to our proof that conceptually $\mathcal{H}(P, Q)$ and $\mathcal{D}_{\text{KL}}(P||Q)$ are bounded.

However, we should not draw a conclusion that there is much difference between the communication efficiency gained based on the mathematically-literal entropy encoding and that gained using the Huffman encoding. In fact, in terms of the average length of codewords, they differ by less than one bit since both lie between $\mathcal{H}(Q)$ and $\mathcal{H}(Q) + 1$ [7], although their difference in terms of the maximum length of individual letters can be very different.

For example, if \mathbb{Z} is a two-letter alphabet, and its PMF Q is $\{0.999, 0.001\}$, the Huffman encoding results in a code with one bit for each letter, while the mathematically-literal entropy encoding results in 1 bit for $z_1 \in \mathbb{Z}$ and 10 bits for $z_2 \in \mathbb{Z}$. The probabilistic average length of the two codewords, which indicate the communication efficiency, is 1 bit for the Huffman encoding, and 1.009 bits for the mathematically-literal entropy encoding, while the entropy $\mathcal{H}(Q)$ is 0.0114 bits. As predicted, $0.0114 < 1 < 1.009 < 1.0114$.

Consider another example with a five-letter alphabet and $Q = \{0.45, 0.20, 0.15, 0.15, 0.05\}$. The mathematically-literal entropy encoding assigns five codewords with lengths of $\{2, 3, 3, 3, 5\}$, while the Huffman encoding assigns codewords with lengths of $\{1, 3, 3, 3, 3\}$. The probabilistic average length of the former is 2.65, while that of the Huffman encoding is 2.1, while the entropy $\mathcal{H}(Q)$ is 2.0999. As predicted, $2.0999 < 2.1 < 2.65 < 3.0999$.

References

1. Chen, M.; Golan, A. What May Visualization Processes Optimize? *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2619–2632. [[CrossRef](#)]
2. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
3. Ishizaka, A.; Nemery, P. *Multi-Criteria Decision Analysis: Methods and Software*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
4. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [[CrossRef](#)]
5. Chen, M.; Abdul-Rahman, A.; Silver, D.; Sbert, M. A Bounded Measure for Estimating the Benefit of Visualization (Part II): Case Studies and Empirical Evaluation. (earlier version: arXiv:2103.02502). *arXiv* **2022**, under review.
6. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
7. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
8. Chen, M.; Feixas, M.; Viola, I.; Bardera, A.; Shen, H.W.; Sbert, M. *Information Theory Tools for Visualization*; A K Peters: Natick, MA, USA, 2016.
9. Feixas, M.; del Acebo, E.; Bekaert, P.; Sbert, M. An Information Theory Framework for the Analysis of Scene Complexity. *Comput. Graph. Forum* **1999**, *18*, 95–106. [[CrossRef](#)]
10. Rigau, J.; Feixas, M.; Sbert, M. Shape Complexity Based on Mutual Information. In Proceedings of the IEEE Shape Modeling and Applications, Cambridge, MA, USA, 13–17 June 2005.
11. Gumhold, S. Maximum entropy light source placement. In Proceedings of the IEEE Visualization, Boston, MA, USA, 27 October–1 November 2002; pp. 275–282.
12. Vázquez, P.P.; Feixas, M.; Sbert, M.; Heidrich, W. Automatic View Selection Using Viewpoint Entropy and its Application to Image-Based Modelling. *Comput. Graph. Forum* **2004**, *22*, 689–700. [[CrossRef](#)]
13. Feixas, M.; Sbert, M.; González, F. A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Trans. Appl. Percept.* **2009**, *6*, 1–23. [[CrossRef](#)]
14. Ng, C.U.; Martin, G. Automatic selection of attributes by importance in relevance feedback visualisation. In Proceedings of the Information Visualisation, London, UK, 14–16 July 2004; pp. 588–595.
15. Bordoloi, U.; Shen, H.W. View selection for volume rendering. In Proceedings of the IEEE Visualization, Minneapolis, MN, USA, 23–28 October 2005; pp. 487–494.
16. Takahashi, S.; Takeshima, Y. A Feature-Driven Approach to Locating Optimal Viewpoints for Volume Visualization. In Proceedings of the IEEE Visualization, Minneapolis, MN, USA, 23–28 October 2005; pp. 495–502.
17. Wang, C.; Shen, H.W. LOD Map—A Visual Interface for Navigating Multiresolution Volume Visualization. *IEEE Trans. Vis. Comput. Graph.* **2005**, *12*, 1029–1036. [[CrossRef](#)]
18. Viola, I.; Feixas, M.; Sbert, M.; Gröller, M.E. Importance-Driven Focus of Attention. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 933–940. [[CrossRef](#)]
19. Jänicke, H.; Wiebel, A.; Scheuermann, G.; Kollmann, W. Multifield Visualization Using Local Statistical Complexity. *IEEE Trans. Vis. Comput. Graph.* **2007**, *13*, 1384–1391.
20. Jänicke, H.; Scheuermann, G. Visual Analysis of Flow Features Using Information Theory. *IEEE Comput. Graph. Appl.* **2010**, *30*, 40–49. [[CrossRef](#)] [[PubMed](#)]
21. Wang, C.; Yu, H.; Ma, K.L. Importance-Driven Time-Varying Data Visualization. *IEEE Trans. Vis. Comput. Graph.* **2008**, *14*, 1547–1554. [[CrossRef](#)] [[PubMed](#)]
22. Bruckner, S.; Möller, T. Isosurface similarity maps. *Comput. Graph. Forum* **2010**, *29*, 773–782. [[CrossRef](#)]
23. Ruiz, M.; Bardera, A.; Boada, I.; Viola, I.; Feixas, M.; Sbert, M. Automatic transfer functions based on informational divergence. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 1932–1941. [[CrossRef](#)]
24. Bramon, R.; Ruiz, M.; Bardera, A.; Boada, I.; Feixas, M.; Sbert, M. Information Theory-Based Automatic Multimodal Transfer Function Design. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 870–880. [[CrossRef](#)]
25. Bramon, R.; Boada, I.; Bardera, A.; Rodríguez, Q.; Feixas, M.; Puig, J.; Sbert, M. Multimodal Data Fusion based on Mutual Information. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 1574–1587. [[CrossRef](#)] [[PubMed](#)]
26. Wei, T.H.; Lee, T.Y.; Shen, H.W. Evaluating Isosurfaces with Level-set-based Information Maps. *Comput. Graph. Forum* **2013**, *32*, 1–10. [[CrossRef](#)]
27. Bramon, R.; Ruiz, M.; Bardera, A.; Boada, I.; Feixas, M.; Sbert, M. An Information-Theoretic Observation Channel for Volume Visualization. *Comput. Graph. Forum* **2013**, *32*, 411–420. [[CrossRef](#)]
28. Biswas, A.; Dutta, S.; Shen, H.W.; Woodring, J. An Information-Aware Framework for Exploring Multivariate Data Sets. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2683–2692. [[CrossRef](#)]
29. Chen, M.; Jänicke, H. An Information-theoretic Framework for Visualization. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1206–1215. [[CrossRef](#)]
30. Chen, M.; Walton, S.; Berger, K.; Thiyaalingam, J.; Duffy, B.; Fang, H.; Holloway, C.; Trefethen, A.E. Visual multiplexing. *Comput. Graph. Forum* **2014**, *33*, 241–250. [[CrossRef](#)]
31. Purchase, H.C.; Andrienko, N.; Jankun-Kelly, T.J.; Ward, M. Theoretical Foundations of Information Visualization. In *Information Visualization: Human-Centered Issues and Perspectives*; LNCS 4950; Springer: Berlin, Germany, 2008; pp. 46–64.
32. Xu, L.; Lee, T.Y.; Shen, H.W. An information-theoretic framework for flow visualization. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1216–1224. [[PubMed](#)]
33. Wang, C.; Shen, H.W. Information Theory in Scientific Visualization. *Entropy* **2011**, *13*, 254–273. [[CrossRef](#)]

34. Tam, G.K.L.; Kothari, V.; Chen, M. An analysis of machine- and human-analytics in classification. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 71–80. [[CrossRef](#)]
35. Kijmongkolchai, N.; Abdul-Rahman, A.; Chen, M. Empirically measuring soft knowledge in visualization. *Comput. Graph. Forum* **2017**, *36*, 73–85. [[CrossRef](#)]
36. Chen, M.; Gaither, K.; John, N.W.; McCann, B. cost–benefit analysis of visualization in virtual environments. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 32–42. [[CrossRef](#)]
37. Chen, M.; Ebert, D.S. An ontological framework for supporting the design and evaluation of visual analytics systems. *Comput. Graph. Forum* **2019**, *38*, 131–144. [[CrossRef](#)]
38. Streeb, D.; El-Assady, M.; Keim, D.; Chen, M. Why visualize? Untangling a large network of arguments. *IEEE Trans. Vis. Comput. Graph.* **2019**, *27*, 2220–2236. [[CrossRef](#)]
39. Viola, I.; Chen, M.; Isenberg, T. Visual Abstraction. In *Foundations of Data Visualization*; Springer: Berlin, Germany, 2020.
40. Tennekes, M.; Chen, M. Design Space of Origin-Destination Data Visualization. *Computer Graphics Forum* **2021**, *40*, 323–334. [[CrossRef](#)]
41. Chen, M.; Grinstein, G.; Johnson, C.R.; Kennedy, J.; Tory, M. Pathways for Theoretical Advances in Visualization. *IEEE Comput. Graph. Appl.* **2017**, *37*, 103–112. [[CrossRef](#)] [[PubMed](#)]
42. Chen, M. Cost–benefit Analysis of Data Intelligence—Its Broader Interpretations. In *Advances in Info-Metrics: Information and Information Processing across Disciplines*; Oxford University Press: Oxford, UK, 2020.
43. Chen, M. A Short Introduction to Information-Theoretic cost–benefit Analysis. *arXiv* **2021**, arXiv:2103.15113.
44. Chen, M.; Sbert, M. On the Upper Bound of the Kullback–Leibler Divergence and Cross Entropy. *arXiv* **2019**, arXiv:1911.08334.
45. Moser, S.M. *A Student's Guide to Coding and Information Theory*; Cambridge University Press: Cambridge, MA, USA, 2012.
46. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [[CrossRef](#)]
47. Österreicher, F.; Vajda, I. A new class of metric divergences on probability spaces and its statistical applications. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653. [[CrossRef](#)]
48. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [[CrossRef](#)]
49. Van Erven, T.; Harremoës, P. Rényi Divergence and Kullback–Leibler Divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [[CrossRef](#)]
50. Klein, H.A. *The Science of Measurement: A Historical Survey*; Dover Publications: Mineola, NY, USA, 2012.
51. Pedhazur, E.J.; Schmelkin, L.P. *Measurement, Design, and Analysis: An Integrated Approach*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1991.
52. Haseli, G.; Sheikh, R.; Sana, S.S. Base-criterion on multi-criteria decision-making method and its applications. *Int. J. Manag. Sci. Eng. Manag.* **2020**, *15*, 79–88. [[CrossRef](#)]
53. Chen, M.; Sbert, M. Is the Chen-Sbert Divergence a Metric? *arXiv* **2021**, arXiv:2101.06103.
54. Newton, I. Scala graduum caloris. *Philos. Trans.* **1701**, *22*, 824–829.
55. Golin, M.J. Lecture 17: Huffman Coding. Available online: <http://home.cse.ust.hk/faculty/golin/COMP271Sp03/Notes/MyL17.pdf> (accessed on 15 March 2020).