Article

A Bounded Measure for Estimating the Benefit of Visualization (Part II): Case Studies and Empirical Evaluation

Min Chen [1,]*, Alfie Abdul-Rahman [2], Deborah Silver [3] and Mateu Sbert [4]

[1] Department of Engineering Science, University of Oxford, Oxford OX1 3QG, UK

[2] Department of Informatics, King's College London, London WC2R 2LS, UK; alfie.abdulrahman@kcl.ac.uk

[3] Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ 08901, USA; silver@jove.rutgers.edu

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

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

Abstract: Many visual representations, such as volume-rendered images and metro maps, feature a noticeable amount of information loss due to a variety of many-to-one mappings. At a glance, there seem to be numerous opportunities for viewers to misinterpret the data being visualized, hence, undermining the benefits of these visual representations. In practice, there is little doubt that these visual representations are useful. The recently-proposed information-theoretic measure for analyzing the cost–benefit ratio of visualization processes can explain such usefulness experienced in practice and postulate that the viewers' knowledge can reduce the potential distortion (e.g., misinterpretation) due to information loss. This suggests that viewers' knowledge can be estimated by comparing the potential distortion without any knowledge and the actual distortion with some knowledge. However, the existing cost–benefit measure consists of an unbounded divergence term, making the numerical measurements difficult to interpret. This is the second part of a two-part paper, which aims to improve the existing cost–benefit measure. Part I of the paper provided a theoretical discourse about the problem of unboundedness, reported a conceptual analysis of nine candidate divergence measures for resolving the problem, and eliminated three from further consideration. In this Part II, we describe two groups of case studies for evaluating the remaining six candidate measures empirically. In particular, we obtained instance data for (i) supporting the evaluation of the remaining candidate measures and (ii) demonstrating their applicability in practical scenarios for estimating the cost–benefit of visualization processes as well as the impact of human knowledge in the processes. The real world data about visualization provides practical evidence for evaluating the usability and intuitiveness of the candidate measures. The combination of the conceptual analysis in Part I and the empirical evaluation in this part allows us to select the most appropriate bounded divergence measure for improving the existing cost–benefit measure.

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

1. Introduction

This two-part paper is concerned with the measurement of the benefit of visualization and viewers' knowledge used in visualization. The history of measurement science shows that the development of measurements in different fields has not only stimulated scientific and technological advancements but also encountered some serious contentions due to instrumental, operational, and social conventions [1]. While the development of measurement systems, methods, and standards for visualization may take decades of research, one can easily imagine their impact on visualization as a scientific and technological subject.

"Measurement ... is defined as the assignment of numerals to objects or events according to rules" [2].

Rules may be defined based on physical laws (e.g., absolute zero temperature), observational instances (e.g., the freezing and boiling points of water), or social traditions (e.g., seven days per week). Without exception, measurement development in visualization aims to discover and define rules that will enable us to use mathematics in describing, differentiating, and explaining phenomena in visualization, as well as in predicting the impact of a design decision, diagnosing shortcomings in visual analytics workflows, and formulating solutions for improvement.

In 2016, Chen and Golan proposed an information-theoretic measure for quantifying the cost–benefit of visualization [3]. However, this measure consists of an unbounded divergence term, making the numerical measurements difficult to interpret. In the first part of this paper [4], Chen and Sbert:

- Reviewed the related work that prepared for this cost–benefit measure, provided the measure with empirical evidence, and featured the application of the measure.
- Identified a shortcoming of using the Kullback–Leibler divergence (KL-divergence) in the cost–benefit measure and demonstrated the shortcoming using practical examples.
- Presented a theoretical discourse to justify the use of a bounded measure for finite alphabets.
- Proposed a new bounded divergence measure, while studying existing bounded divergence measures.
- Analyzed nine candidate measures using seven criteria reflecting desirable conceptual or mathematical properties, and narrowed the nine candidate measures to six measures.

In this second part of the paper, we focus on the remaining six candidate measures and evaluate them based on empirical evidence. In particular:

- We report several case studies for collecting practical instances to evaluate the remaining candidate measures.
- We demonstrate 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.
- We report the discovery of 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.
- Finally, we bring the multi-criteria decision analysis (MCDA) in Parts I and II together and offer a recommendation to revise the information-theoretic measures proposed by Chen and Golan [3].

In addition, we use the data collected in two visualization case studies to explore the relationship between the benefit of visualization and the viewers' knowledge used in visualization. As shown in Figure 1, in one case study, we asked participants to perform tasks for estimating the walking time (in minutes) between two underground stations indicated by a pair of red or blue arrows. Although the deformed London underground map was not designed to perform visualization tasks, many participants performed rather well, including those who had very limited experience of using the London underground. This suggests that with the presence of knowledge, a seemingly-tiny amount of visual information can be very useful.

We proposed two different ways of estimating viewers' knowledge that has been used in the visualization process to alleviate the potential distortion. When we use different candidate measures to estimate viewers' knowledge, we evaluate these candidate measures using the collected practical instances, while demonstrating that we are getting closer to be able to estimate the "benefit" of and "knowledge" used in practical visualization processes.

Readers are encouraged to consult the related reports on the cost–benefit analysis [5] and the Part I of this paper [4]. Nevertheless, this part of the paper is written in a self-contained manner.
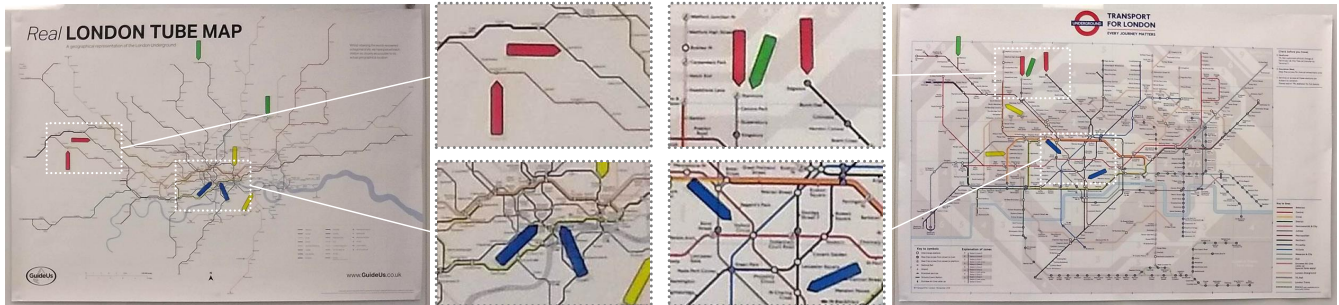
Figure 1. The London underground map (**right**) is a deformed map. In comparison with a relatively more faithful map (**left**), there is a significant amount of information loss due to many-to-one mappings in the deformed map, which omits some detailed variations among different connection routes between pairs of stations (e.g., distance and geometry). One common rationale is that the deformed map was designed for certain visualization tasks, which likely excluded the task for estimating the walking time between a pair of stations indicated by a pair of red or blue arrows. In one of our experiments, when asked to perform such tasks using the deformed map, some participants with little knowledge about London or London Underground performed these tasks well. Can information theory explain this phenomenon? Can we quantitatively measure relevant factors in this visualization process?

2. Related Work

This two-part paper is concerned with information-theoretic measures for quantifying aspects of visualization, such as benefit, knowledge, and potential misinterpretation. The first part [4] focuses its review on previous information-theoretic work in visualization. In this section, we focus our review on previous measurement work in visualization.

2.1. Measurement Science

There is currently no standard measurement scale for measuring the benefit of visualization, levels of visual abstraction, the human knowledge used in visualization, or the potential to misinterpret visual abstraction. While these are considered to be complex undertakings, many scientists in the history of measurement science would have encountered similar challenges [1].

In their book [6], Boslaugh and Watters described measurement as "the process of systematically assigning numbers to objects and their properties, to facilitate the use of mathematics in studying and describing objects and their relationships." They emphasized that measurement is not limited to physical qualities (e.g., height and weight) but also includes abstract properties (e.g., intelligence and aptitude). Pedhazur and Schmelkin [7] asserted the necessity of an integrated approach for measurement development, involving data collection, mathematical reasoning, technology innovation, and device engineering. Tal [8] pointed out that measurement is often not totally "real", involves the representation of ideal systems and reflects conceptual, metaphysical, semantic, and epistemological understandings. Schlaudt [9] went one step further, referring measurement as a cultural technique.

This work is particularly inspired by the historical development of temperature scales and seismic magnitude. The former attracted the attention of many well-known scientists, benefited from both experimental observations (e.g., by Newton, Fahrenheit, Delisle, Celsius, etc.) and theoretical discoveries (e.g., by Boltzmann, Thomson (Kelvin), etc.). The latter started not long ago as the Richter scale was outlined in 1935. Since then, there have been many schemes proposed relating different physical properties. Many scales in both applications are related to logarithmic transformations.

Figure 2 depicted a number of instances that are quantified in different temperature scales. Isaac Newton proposed one of the first temperature scales based on his observation of over 20 instances [10]. Nine of them are shown in Figure 2, where the corresponding data in other scales were obtained based on Grigull's study of the Netwon scale [11]. Although the Newton scale has not been adopted, his approach to mark and observe different data points at his proposed scale has been considered as "the first attempt to introduce an objective way of measuring ... temperature" [12].

From that first step, it took more than 40 years and many other proposals to developed the Celsius scale (with two of Newton's data points as the reference points), which is most commonly used today. It took another century to develop the Kelvin scale with absolute zero as a new reference point. History motivates us to collected practical instances and conduct data-driven evaluation of candidate measures for estimating the benefit of visualization.
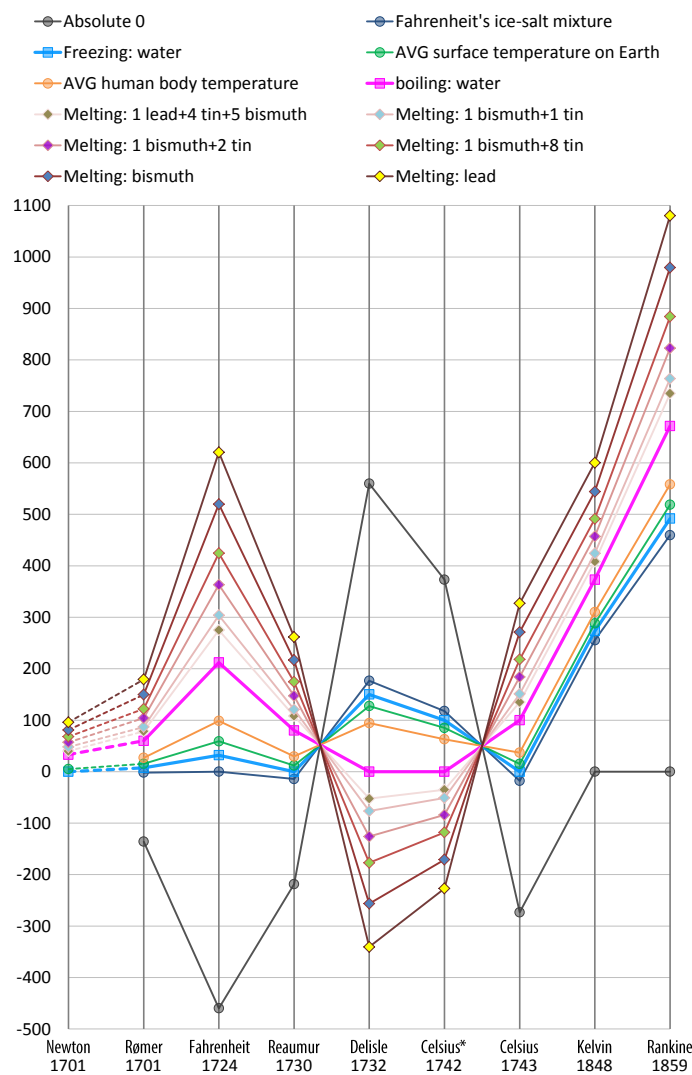
Figure 2. Major temperate scales proposed in history. Different lines show instances used as observation points, some of which became major reference points. Note: "Celsius* 1742" indicates the original scale proposed by Anders Celsius, while "Celsius 1743" indicates the revised Celsius scale used today that was proposed by Jean-Pierre Christin. The Newton scale is not linearly related to the others (shown as dash lines).

2.2. Metrics Development in Visualization

Behrisch et al. [13] presented a survey of quality metrics for information visualization. Bertini et al. [14] described a systemic approach of using quality metrics for evaluating

high-dimensional data visualization focusing on scatter plots and parallel coordinates plots. A variety of quality metrics have been proposed to measure many different attributes, such as abstraction quality [15–17], quality of scatter plots [18–24], quality of parallel coordinates plots [25], cluttering [26–28], aesthetics [29], visual saliency [30], and color mapping [31–34].

In particular, Jänicke et al. [30] first considered a metric for estimating the amount of original data that is depicted by visualization and may be reconstructed by viewers. Chen and Golan [3] used the abstract form of this idea in defining their cost–benefit ratio. While the work by Jänicke et al. [30] relied on computer vision techniques for reconstruction, this work focused on collecting and analyzing empirical data because human knowledge has a major role to play in information reconstruction.

2.3. Measurement in Empirical Experiments

Almost all controlled empirical studies in visualization involve measuring the participants' performance in visualization processes, typically in terms of accuracy and response time (e.g., [35]). Many uncontrolled empirical studies also collect participants' experience and opinions qualitatively. Such collected data allow us to assess the benefit of visualization or potential misinterpretation. The empirical studies particularly relevant to this work are those on the topics of visual abstraction and human knowledge in visualization.

Isenberg [36] presented a survey of evaluation techniques on non-photorealistic and illustrative rendering. Isenberg et al. [37] reported an observational study comparing hand-drawn and computer-generated non-photorealistic rendering. Cole et al. [38] performed a study evaluating the effectiveness of line drawing in representing shape. Mandryk et al. [39] evaluated the emotional responses to non-photorealistic generated images. Liu and Li [40] presented an eye-tracking study examining the effectiveness and efficiency of schematic designs for depicting 30° and 60° directions in underground maps. Hong et al. [41] evaluated the usefulness of distance cartograms map "in the wild". These studies confirmed that visualization users can deal with significant information loss due to visual abstraction in many situations.

Tam et al. [42] reported an observational study comparing automated and semi-automated machine learning (ML) workflows. Their information-theoretical analysis showed that ML developers entered a huge amount of knowledge (measured in bits) into a visualization-assisted ML workflow. Kijmongkolchai et al. [43] reported a study designed for detecting and measuring human knowledge used in visualization, and translated the traditional accuracy values to information-theoretic measures. They encountered an undesirable property of the Kullback–Leibler divergence in their calculations. In this work, we collect empirical data to evaluate the mathematical solutions proposed to address the issue encountered in [43].

If we can address this mathematical issue successfully, we will be able to complement qualitative methods for assessing the value of visualization (e.g., by Wall et al. [44]) with quantitative measurement; we will be able to carry out many experiments (e.g., those by Cleveland and McGill [35] and Saket et al. [45]) to examine the trade-off between alphabet compression and potential distortion [3]; we will be able to estimate the knowledge used (or gained) by the users in (or from) visualization as discussed by Sacha et al. [46]; and we will be able to transform the current qualitative methods for optimizing visual analytics workflow (e.g., [47]) to quantitative methods.

3. Overview, Notations, and Problem Statement

3.1. Brief Overview

Whilst hardly anyone in the visualization community would support any practice intended to deceive viewers, there have been many visualization techniques that inherently cause distortion to the original data. The deformed London underground map in Figure 1 shows such an example. The distortion in this example is largely caused by many-to-one

mappings. A group of lines that would be shown in different lengths in a faithful map are now shown with the same length.

Another group of lines that would be shown with different geometric shapes are now shown as the same straight line. In terms of information theory, when the faithful map is transformed to the deformed map, a portion of information has been lost due to the many-to-one mappings. In this work, we follow the Shannon's definition of information. Many-to-one mappings result in the reduction of Shannon entropy [48].

The common phrase that "the appropriateness of information loss depends on tasks" is not an invalid explanation. Partly by a similar conundrum in economics "what is the most appropriate resolution of time series for an economist", Chen and Golan proposed an information-theoretic cost–benefit ratio for measuring various factors involved in visualization processes [3]. Its qualitative version is:

$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\text{Alphabet Compression} - \text{Potential Distortion}}{\text{Cost}} \tag{1}$$

This cost–benefit ratio was described and discussed in the first part of the paper [4]. Appendix A provides a more detailed explanation of this measure in the context of visualization, while Appendix B explains in detail how tasks and users are considered by this measure in the abstract. A more comprehensive introduction can be found in an arXiv report [5].

3.2. Mathematical Notations

Consider a simple metro map consisting of only two stations in Figure 3. We consider three different grid resolutions, with 1×1 cell, 2×2 cells, and 4×4 cells, respectively. The following set of rules determine whether a potential path is allowed or not:

- The positions of the two stations are fixed on each grid and there is only one path between the red station and the blue station.
- As shown on the top-right of Figure 3, only horizontal, and diagonal path-lines are allowed.
- When one path-line joins another, it can rotate by up to $\pm 45°$.
- All joints of path-lines can only be placed on grid points.

For the first grid with the 1×1 cell, there is only one possible path. We define an alphabet $\mathbb{A}$ to contain this option as its only letter a_1, i.e., $\mathbb{A} = \{a_1\}$. For the second grid with 2×2 cells, we have an alphabet $\mathbb{B} = \{b_1, b_2, b_3\}$, consisting of three optional paths. For the third grid with 4×4 cells, there are 15 optional paths, which are letters of alphabet $\mathbb{C} = \{c_1, c_2, \ldots, c_{15}\}$. When the resolution of the grid increases, the alphabet of options becomes bigger quickly. We can imagine it gradually allows the designer to create a more faithful map.

To a designer of the underground map, at the 1×1 resolution, there is only one choice regardless of how much the designer would like to draw the path to reflect the actual geographical path of the metro line between these two stations. At the 2×2 and 4×4 resolutions, the designer has 3 and 15 options, respectively. Increasing the number of options is one factor that causes the increasing uncertainty about the selection of a specific option. The other factor is the provability of each option being selected. This *uncertainty* can be measured by Shannon entropy, which is defined as:

$$\mathcal{H}(\mathbb{Z}) = - \sum_{i=1}^{n} p_i \log_2 p_i \quad \text{where } p_i \in [0,1], \sum_{i=1}^{n} p_i = 1$$

where $\mathbb{Z}$ is an alphabet, and can be replaced with $\mathbb{A}$, $\mathbb{B}$, or $\mathbb{C}$. To calculate Shannon entropy, the alphabet $\mathbb{Z}$ needs to be accompanied by a *probability mass function* (PMF), which is written as $P(\mathbb{Z})$. Each letter $z_i \in \mathbb{Z}$ is thus associated with a probability value $p_i \in P$.

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., $\mathcal{H}(P(\mathbb{Z})) =$

$\mathcal{H}(P) = \mathcal{H}(\mathbb{Z})$. Readers can find more mathematical background about information theory in [49] in general, and [5] in relation to this paper.
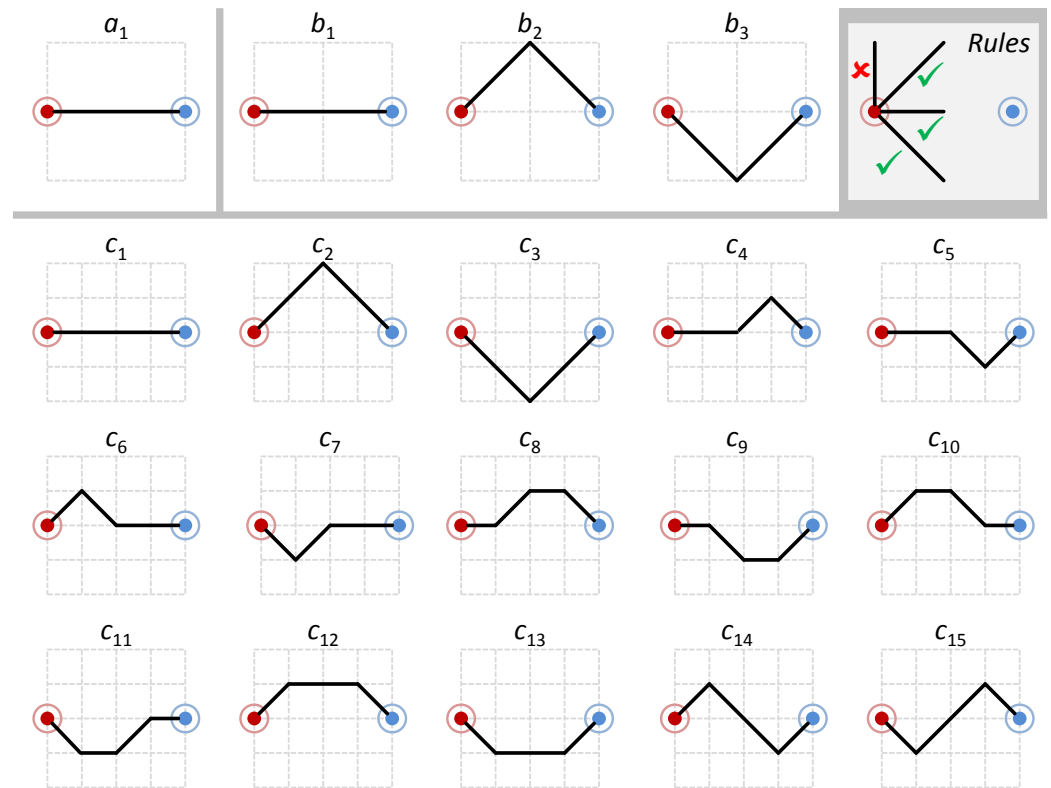
Figure 3. Three alphabets illustrate possible metro maps (letters) in different grid resolutions. Increasing the resolution enables the depiction of more reality, while reducing the resolution compels more abstraction.

To ensure the calculation is easy to follow, we consider only the first two grids below. Let us first consider the single-letter alphabet $\mathbb{A}$ and its PMF Q. As $n = 1$ and $q_1 = 1$, we have $\mathcal{H}(\mathbb{A}) = 0$ bits. $\mathbb{A}$ is 100% certain, reflecting the fact that the designer has no choice.

The alphabet $\mathbb{B}$ has three design options b_1, b_2, and b_3. If they have an equal chance to be selected by the designer, we have a PMF Q_u with $q_1 = q_2 = q_3 = 1/3$, and thus $\mathcal{H}(Q_u(\mathbb{B})) \approx 1.585$ bits. When we examine the three options in Figure 3, it is not unreasonable to consider a second scenario that the choice may be in favor of the straight line option b_1 in designing a metro map according to the real geographical data. If a different PMF Q_v is given as $q_1 = 0.9, q_2 = q_3 = 0.05$, we have $\mathcal{H}(Q_v(\mathbb{B})) \approx 0.569$ bits. The second scenario features less entropy and is thus of more certainty.

Consider that the designer is given a metro map designed using alphabet $\mathbb{B}$, and is asked to produce a more abstract map using alphabet $\mathbb{A}$. To the designer, it is a straightforward task, since there is only one option in $\mathbb{A}$. When a group of viewers is visualizing the final design a_1, we could give these viewers a task to guess what may be the original map designed with $\mathbb{B}$. If most viewers have no knowledge about the possible options b_2 and b_3, and almost all choose b_1 as the original design, we can describe their decisions using a PMF P such that $p_1 = 0.998, p_2 = p_3 = 0.001$. Since P is not the same as either Q_u or Q_v, the viewers' decisions diverge from the actual PMF associated with $\mathbb{B}$. This divergence can be measured using the Kullback–Leibler divergence (KL-divergence):

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

Using $\mathcal{D}_{\mathrm{KL}}$, we can calculate (i) if the original design alphabet $\mathbb{B}$ has the PMF Q_u, we have $\mathcal{D}_{\mathrm{KL}}(P||Q_u) \approx 1.562$ bits; and (ii) if the original design alphabet $\mathbb{B}$ has the PMF Q_v, we have $\mathcal{D}_{\mathrm{KL}}(P||Q_v) \approx 0.138$ bits. There is more divergence in case (i) than case (ii). Intuitively, we can guess this as P appears to be similar to Q_v.

Recall the qualitative formula in Equation (1). In the original mathematical defini-tion [3], the benefit of a visual analytics process is defined as:

$$\text{Benefit} = \text{AC} - \text{PD} = \mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1}) - \mathcal{D}_{\mathrm{KL}}(\mathbb{Z}_i'||\mathbb{Z}_i) \qquad (2)$$

where $\mathbb{Z}_i$ is the input alphabet to the process and $\mathbb{Z}_{i+1}$ is the output alphabet. $\mathbb{Z}_i'$ is an alphabet reconstructed based on $\mathbb{Z}_{i+1}$. $\mathbb{Z}_i'$ has the same set of letters as $\mathbb{Z}_i$ but likely a different PMF. In Equation (2), the first two terms, $\mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1})$, directly measure the amount of information loss in terms of Shannon entropy, while the third term, $\mathcal{D}_{\mathrm{KL}}(\mathbb{Z}_i'||\mathbb{Z}_i)$, measures the consequence of the information loss.

In terms of Equation (2), we have $\mathbb{Z}_i = \mathbb{B}$ with PMF Q_u or Q_v, $\mathbb{Z}_{i+1} = \mathbb{A}$ with PMF Q, and $\mathbb{Z}_i' = \mathbb{B}'$ with PMF P. We can thus calculate the benefit in the two cases as:

$$\begin{aligned}
\text{Benefit of case (i)} &= \mathcal{H}(\mathbb{B}) - \mathcal{H}(\mathbb{A}) - \mathcal{D}_{\mathrm{KL}}(\mathbb{B}'||\mathbb{B}) \\
&= \mathcal{H}(Q_u) - \mathcal{H}(Q) - \mathcal{D}_{\mathrm{KL}}(P||Q_u) \\
&\approx 1.585 - 0 - 1.562 = 0.023 \text{bits}
\end{aligned}$$

$$\begin{aligned}
\text{Benefit of case (ii)} &= \mathcal{H}(Q_v) - \mathcal{H}(Q) - \mathcal{D}_{\mathrm{KL}}(P||Q_v) \\
&\approx 0.569 - 0 - 0.138 = 0.431 \text{bits}
\end{aligned}$$

In case (ii), because the viewers' expectation is closer to the original PMF Q_v, there is more benefit in the visualization process than case (i) though case (ii) has less AC than case (i).

However, $\mathcal{D}_{\mathrm{KL}}$ has an undesirable mathematical property. If we consider a third case, (iii), where the original PMF Q_w is strongly in favor of b_2, such as $q_1 = \epsilon, q_2 = 1 - 2\epsilon, q_3 = \epsilon$, where $0 < \epsilon < 1$ is a small positive value. If $\epsilon = 0.001$, $\mathcal{D}_{\mathrm{KL}}(P||Q_w) = 9.933$ bits. If $\epsilon \to 0$, $\mathcal{D}_{\mathrm{KL}}(P||Q_w) \to \infty$. Since the maximum entropy (uncertainty) for $\mathbb{B}$ is only about 1.585 bits, it is difficult to interpret that viewers' divergence can be more than that maximum, not to mention the infinity.

3.3. Problem Statement

When using $\mathcal{D}_{\mathrm{KL}}$ in Equation (1) in a relative or qualitative context (e.g., [47,50]), the unboundedness of the KL-divergence does not pose an issue. However, this does be-come an issue when $\mathcal{D}_{\mathrm{KL}}$ is used to measure the PD in an absolute and quantitative context.

In the first part of this paper [4], Chen and Sbert showed that, conceptually, it is the unboundedness that is not consistent with a conceptual interpretation of KL-divergence for measuring the inefficiency of a code (alphabet) that has a finite number of codewords (letters). They proposed to find a suitable bounded divergence measure to replace the $\mathcal{D}_{\mathrm{KL}}$ term in Equation (2). They examined nine candidate measures, analyzed their mathematical properties with the aid of visualization, and narrowed these down to six measures using multi-criteria decision analysis (MCDA) [51].

In this work, we continue their MCDA process by introducing criteria based on the analysis of instances obtained when using the remaining six candidate measures in different case studies, which correspond to criteria $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{R}_1$, and $\mathbf{R}_2$ in Table 1 that is presented in Section 7.

For self-containment, we give the mathematical definition of the six candidate mea-sures below. In this second part of the paper, we treat them as black-box functions, since they have already undergone the conceptual evaluation in the first part of this paper. For more detailed conceptual and mathematical discourse on these six candidate measures, please consult that part [4].

Table 1. A summary of the multi-criteria decision analysis (MCDA). Each measure is scored against a criterion using an integer in [0, 5] with 5 being the best. Scores are calculated as: starting with a full score of 5. For each "good" deduct 1, each "adequate" deduct 2, and each "inadequate" deduct 3. The top table summarize the empirical scores obtained from the two synthetic case studies (S_1 and S_2) in Section 5 and two experimental case studies (R_1 and R_2) in Section 6. The bottom table presents the final results of MCDA by combining the subtotals of the seven conceptual criteria in the first part of the paper, the subtotals of the empirical criteria in this second part of the paper, and the scores of the extra conceptual criterion discussed in Section 5.3.

A Summary of the Empirical Scores Obtained of the Four Case Studies						
Criteria	$\mathcal{D}_{JS}$	$\sqrt{\mathcal{D}_{JS}}$	$\mathcal{D}_{new}^{k=1}$	$\mathcal{D}_{new}^{k=2}$	$\mathcal{D}_{ncm}^{k=1}$	$\mathcal{D}_{ncm}^{k=2}$
S_1 : *order*	adequate	adequate	excellent	excellent	good	good
benefit	adequate	good	excellent	good	excellent	good
knowledge	excellent	excellent	excellent	excellent	excellent	excellent
score	1	2	5	4	4	3
S_2 : *order*	excellent	excellent	good	excellent	adequate	good
benefit	adequate	excellent	good	adequate	good	adequate
knowledge	excellent	excellent	excellent	excellent	excellent	excellent
score	3	5	3	3	2	2
R_1 : *order*	excellent	excellent	good	excellent	inadequate	good
benefit	excellent	excellent	excellent	excellent	excellent	excellent
knowledge	excellent	excellent	good	good	adequate	good
score	5	5	3	4	0	3
R_2 : *order*	excellent	excellent	excellent	excellent	adequate	excellent
benefit	adequate	adequate	adequate	excellent	adequate	adequate
knowledge	excellent	excellent	excellent	excellent	excellent	excellent
score	3	3	3	5	1	3
Empirical Subtotal:	12	15	14	16	7	11
Combining All Scores Obtained from the Conceptual and Empirical Evaluation						
Criteria	$\mathcal{D}_{JS}$	$\sqrt{\mathcal{D}_{JS}}$	$\mathcal{D}_{new}^{k=1}$	$\mathcal{D}_{new}^{k=2}$	$\mathcal{D}_{ncm}^{k=1}$	$\mathcal{D}_{ncm}^{k=2}$
Conceptual Subtotal [4]:	30	30	28	30	26	29
Empirical Subtotal:	12	15	14	16	7	11
Componentization (extra criterion):	5	1	5	5	5	5
Total without the extra criterion:	**42**	**45**	**42**	**46**	**33**	**40**
Total with the extra criterion:	**47**	**46**	**47**	**51**	**38**	**45**

The first candidate measure is Jensen–Shannon divergence [52], which is defined as:

$$\mathcal{D}_{JS}(P||Q) = \frac{1}{2}\big(\mathcal{D}_{KL}(P||M) + \mathcal{D}_{KL}(Q||M)\big) = \mathcal{D}_{JS}(Q||P)$$
$$= \frac{1}{2}\sum_{i=1}^{n}\left(p_i \log_2 \frac{2p_i}{p_i + q_i} + q_i \log_2 \frac{2q_i}{p_i + q_i}\right) \tag{3}$$

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 (3), $\mathcal{D}_{JS}(P||Q)$ is bounded by 0 and 1.

The second candidate measure is the square root of $\mathcal{D}_{JS}$. The conceptual evaluation gave both $\mathcal{D}_{JS}$ and $\sqrt{\mathcal{D}_{JS}}$ the same promising score 30 as shown in Table 1. The third and fourth candidate measures are two instances of a new measure $\mathcal{D}_{new}^{k}$ proposed by Chen

and Sbert [4]. The two instances are $\mathcal{D}_{\text{new}}^k(k=1)$ and $\mathcal{D}_{\text{new}}^k(k=2)$. They received scores of 28 and 30, respectively, in the conceptual evaluation. $\mathcal{D}_{\text{new}}^k$ is defined as follows:

$$\mathcal{D}_{\text{new}}^k(P||Q) = \frac{1}{2}\sum_{i=1}^{n}(p_i + q_i)\log_2(|p_i - q_i|^k + 1) \tag{4}$$

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

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

Since $\log_2 1 = 0$, $\log_2 2 = 1$, $\sum p_i = 1$, $\sum q_i = 1$, $\mathcal{D}_{\text{new}}^k(P||Q)$ is thus bounded by 0 and 1.

The fifth and sixth candidate measures are two instances of a non-commutative version of $\mathcal{D}_{\text{new}}^k$. It is denoted as $\mathcal{D}_{\text{ncm}}^k$, and the two instances are $\mathcal{D}_{\text{ncm}}^k(k=1)$ and $\mathcal{D}_{\text{ncm}}^k(k=2)$, which also received scores of 26 and 29, respectively, in the conceptual evaluation. $\mathcal{D}_{\text{ncm}}^k$ is defined as follows:

$$\mathcal{D}_{\text{ncm}}^k(P||Q) = \sum_{i=1}^{n}p_i\log_2(|p_i - q_i|^k + 1), \tag{5}$$

which captures the non-commutative property of $\mathcal{D}_{\text{KL}}$.

As $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{D}_{\text{new}}^k$, and $\mathcal{D}_{\text{ncm}}^k$ are bounded by [0, 1], if any of them is selected to replace $\mathcal{D}_{\text{KL}}$, Equation (2) can be rewritten as

$$\text{Benefit} = \mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1}) - \mathcal{H}_{\max}(\mathbb{Z}_i)\mathcal{D}(\mathbb{Z}_i'||\mathbb{Z}_i) \tag{6}$$

where $\mathcal{H}_{\max}$ denotes maximum entropy, while $\mathcal{D}$ is a placeholder for $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{D}_{\text{new}}^k$, 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 $\mathbb{Z}_i$.

4. Evaluation Methodology and Criteria

Historically, developing different temperature scales is motivated by the need for defining and quantifying the divergence between any pair of values representing two instances of different temperatures. Isaac Newton approached this problem by collecting over 20 instances, nine of which are shown in Figure 2.

Given two PMFs P and Q associated with an alphabet $\mathbb{Z}$ with n letters, measuring the divergence between P and Q involves the definition and quantification of the interaction between n pairs of probability values. Hence, the measuring function is n-dimensional and is likely more difficult to define. Nevertheless, we can adapt Isaac Newton's approach of using data points with practical meanings. Unlike the Newton scale, we do not need these data points to specify a scale, but only to evaluate candidate measures. Analogously, this is similar to use Newton's data points to evaluate other temperature scales in Figure 2.

Consider a real world phenomenon being visualized by a user or a group of users for a specific task. Let alphabet $\mathbb{Z}_w$, $\mathbb{Z}_v$, and $\mathbb{Z}_t$ be the information spaces of the phenomenon, the visualization, and the task concerned, respectively. There are two major transformations, one from $\mathbb{Z}_w$ to $\mathbb{Z}_v$ and another from $\mathbb{Z}_v$ to $\mathbb{Z}_t$. The first major transformation may contain processes for data capture, data processing, and data visualization, while the second major transformation may contain all cognitive processes from viewing to task performance. Both transformations may feature alphabet compression and potential distortion.

Most visualization tasks (including confirmation, categorization, recognition, search, estimation, etc.) can be abstracted as a decision to select from two or more options. A decision alphabet $\mathbb{Z}_t$ essentially contains all valid options with a PMF. In some cases, there can be numerous options (e.g., counting). There is a ground truth PMF Q that reflects the ideal task performance when users have full access to perfect data sampled in the information

space of the phenomenon $\mathbb{Z}_w$, have an infinite amount of time to view the data with or without visualization, and do not have any cognitive bias in selecting the correct option. Although an accurate Q may be difficult to obtain, one can estimate it in a synthetic or real-world case study, which will be demonstrated in the next two sections.

As the transformation from $\mathbb{Z}_w$ to $\mathbb{Z}_v$ will lose a fair amount of information, users, who have a different amount of knowledge about the phenomenon, the visual representation, and the task concerned, will perform differently. Such difference will be captured in the PMF, P, compiled according to the actual task performance. For example, consider three typecasting cases:

(a)   *P is close to a uniform PMF P_{uniform}, while the ground truth Q is dissimilar to a uniform PMF*—This suggests that the users may not have adequate knowledge and may have been making random guesses. In such a case, their task performance would lead to a PMF similar to P_{uniform}.

(b)   *P is close to a PMF P_{visinfo} that characterizes the available visual information while the ground truth Q differs from P_{visinfo} noticeably*—This suggests that the users may not have adequate knowledge and may have been reasoning about the options in $\mathbb{Z}_t$ entirely based on what is depicted visually. In such a case, their performance would result in a PMF similar to P_{visinfo}.

(c)   *P is close to the ground truth Q, while Q differs from P_{uniform} and P_{visinfo} noticeably*—This suggests that the users may have been able to make the perfect combination of the available visual information and their knowledge. In such a case, their task performance could lead to a PMF similar to the ideal PMF Q.

One obvious method to determine whether a visual design is suitable for a group of users is to ask these users to perform some tasks. For a particular task, the users' task performance can be sampled and approximated using a PMF P_{sampled}. All the candidate measures can quantify the divergence between P_{sampled} and the ground truth PMF Q. We can also use such a candidate measure $\mathcal{D}_*$ to quantify the benefit of visualization as:

$$\text{Benefit} = \mathcal{H}(Q) - \mathcal{H}(P_{\text{visinfo}}) - \mathcal{H}_{\max}(Q)\mathcal{D}_*(P_{\text{sampled}}\|Q)$$

Before we are able to reach the final conclusion, we consider that $\mathcal{D}_*$ may be any one of the six candidate measures given in Section 3.3, i.e., $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{D}_{\text{new}}^{k=1}$, $\mathcal{D}_{\text{new}}^{k=2}$, $\mathcal{D}_{\text{ncm}}^{k=1}$, and $\mathcal{D}_{\text{ncm}}^{k=2}$. In addition, we can also estimate the impact of the human knowledge used in performing a visualization task as:

$$K_v = \mathcal{H}_{\max}(Q)\big(\mathcal{D}_*(P_{\text{visinfo}}\|Q) - \mathcal{D}_*(P_{\text{sampled}}\|Q)\big)$$
$$K_\psi = \mathcal{H}_{\max}(Q)\big(\mathcal{D}_*(P_{\text{uniform}}\|Q) - \mathcal{D}_*(P_{\text{sampled}}\|Q)\big)$$

where K_v is an estimation against the scenario where users rely only on visual information without using any knowledge, and K_ψ is that against the scenario of random guesses. If $K_v > 0$ and $K_\psi > 0$, they suggest a positive impact of human knowledge. If $K_v < 0$ or $K_\psi < 0$, they suggest some biases.

Given some instance data in the form of PMFs P_{sampled}, we would like to observe how different candidate measures would (i) order these instances in terms of their divergence against an estimated ground truth PMF Q, (ii) quantify the benefit at a scale meaningful to visualization scientists, and (iii) assign the sensible signs to K_v and K_ψ.

5. Synthetic Case Studies

We first consider two synthetic case studies, $\mathbf{S}_1$ and $\mathbf{S}_2$, which allow us to define idealized situations, from which collected data do not contain any noise. In many ways, this is similar to testing a piece of software using pre-defined test cases. Nevertheless, these test cases feature more complex alphabets than those considered by the conceptual evaluation presented in the first part of this paper [4].

5.1. Synthetic Case S_1

Let $\mathbb{Z}_w$ be a phenomenon alphabet with two letters, *good* and *bad*, for describing a scenario (e.g., an object or an event), where the ground truth probability of *good* is $q_1 = 0.8$, and that of *bad* is $q_2 = 0.2$. In other words, $Q = \{0.8, 0.2\}$. Imagine that a biased process (e.g., a distorted visualization, faulty data collection, an incorrect algorithm, or a misleading communication) conveys the information about the scenario always *bad*, i.e., a visualization alphabet $\mathbb{Z}_v$ with a PMF $P_{\text{visinfo}} = R_{\text{biased}} = \{0, 1\}$. Users at the receiving end of the process may have different knowledge about the actual scenario, and they will make a decision, $\mathbb{Z}_t$, after receiving the output of the process. For example, there are five users, and we obtained the probability of their decisions (with different P_{sampled}) as follows:

- LD—The user has a little doubt about the output of the process, and decides the letter of *bad* 90% of the time, and the letter of *good* 10% of the time, i.e., with PMF $P_{\text{LD}} = \{0.1, 0.9\}$.
- FD—The user has a fair amount of doubt, with $P_{\text{FD}} = \{0.3, 0.7\}$.
- RG—The user makes a random guess, with $P_{\text{RG}} = \{0.5, 0.5\}$.
- UC—The user has adequate knowledge about $\mathbb{Z}_w$ but under-compensates it slightly, with $P_{\text{UC}} = \{0.7, 0.3\}$.
- OC—The user has adequate knowledge about $\mathbb{Z}_w$ but over-compensates it slightly, with $P_{\text{OC}} = \{0.9, 0.1\}$.

We can use different candidate measures to compute the divergence between Q and each P_{sampled}. The bar chart in Figure 4 shows different divergence values returned by these measures, while the transformations from Q to R_{biased} and then to different P_{sampled} are illustrated on the right margin of the figure. Each value is decomposed into two parts, one for *good* and one for *bad*, except that the candidate measure $\sqrt{\mathcal{D}_{\text{JS}}}$ cannot distinguish the component measures for individual letters since it is a global transformation after $\mathcal{D}_{\text{JS}}$ is calculated. This shortcoming of $\sqrt{\mathcal{D}_{\text{JS}}}$ was not noticed in the conceptual analysis in the Part I of this paper [4].
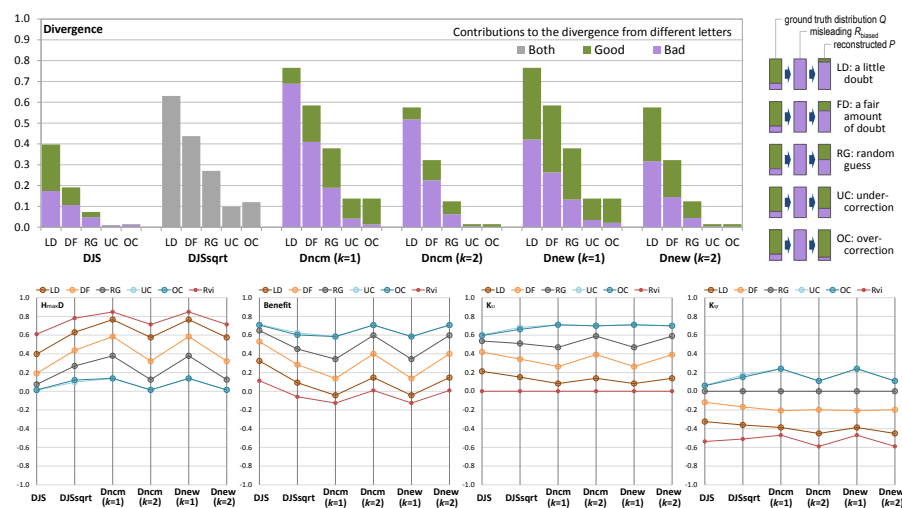
Figure 4. An example scenario with two states *good* and *bad* has a ground truth PMF $Q = \{0.8, 0.2\}$. From the output of a biased process that always informs users that the situation is *bad*. Five users, LD, DF, RG, UC, and OC, have different knowledge and thus different divergence. The five candidate measures return different values of divergence. We would like to see which sets of values are more intuitive. The illustration on the top-right shows two transformations of the alphabets and their PMFs, one by the misleading communication and the other by the reconstruction. The bar chart shows the divergence values calculated by each candidate measure, while the four parallel coordinate plots (PCPs) show the values of $\mathcal{H}_{\text{max}}\mathcal{D}$ (divergence scaled by SE_{max}), benefit, K_v (impact of knowledge against relying solely on visual information), and K_ψ (against random guess).

All these measures can order these five users reasonably well. The users UC (under-compensate) and OC (over-compensate) have the same values with $\mathcal{D}_{\text{new}}^{k}$ and $\mathcal{D}_{\text{ncm}}^{k}$, while $\mathcal{D}_{\text{JS}}$ and $\sqrt{\mathcal{D}_{\text{JS}}}$ consider OC has slightly more divergence than UC. For $\mathcal{D}_{\text{JS}}$, UC:OC = 0.010:0.014 and for $\sqrt{\mathcal{D}_{\text{JS}}}$, UC:OC = 0.098:0.120.

$\mathcal{D}_{\text{ncm}}^{k=1}$ and $\mathcal{D}_{\text{ncm}}^{k=2}$ show strong asymmetric patterns between *good* and *bad*, reflecting the probability values in P_{sampled}. In other words, the more decisions on *good*, the more *good*-related divergence. This asymmetric pattern is not in any way incorrect, as the KL-divergence is also non-commutative and would also produce much stronger asymmetric patterns. An argument for supporting commutative measures would point out that the higher probability of *good* in Q should also influence the balance between the *good*-related divergence. We are slightly in favor of commutativity as it is easier to interpret. In terms of **ordering**, we consider $\mathcal{D}_{\text{new}}^{k}$ "excellent", $\mathcal{D}_{\text{ncm}}^{k}$ "good" due to asymmetry, and $\mathcal{D}_{\text{JS}}$ and $\sqrt{\mathcal{D}_{\text{JS}}}$ "adequate" as the non-equal UC and OU measures are not so intuitive.

As $\mathcal{H}(Q) = 0.722$ and $\mathcal{H}(R_{\text{biased}}) = 0$, the amount of alphabet compression (AD) is 0.722 bits. $\mathcal{H}_{\text{max}}(Q) = 1$ bit. We can compute the benefits of the visualization to the six users, which are shown in the left parallel coordinate plot (PCP) in Figure 4. From these PCPs, we notice that $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{D}_{\text{ncm}}^{k=2}$, and $\mathcal{D}_{\text{new}}^{k=2}$ give positive benefits to all five users, with $\mathcal{D}_{\text{JS}}$ returning the highest values. $\mathcal{D}_{\text{ncm}}^{k=1}$, and $\mathcal{D}_{\text{new}}^{k=1}$ yield negative benefit values for user LD, which is consistent with our expectation. In terms of **benefit quantification**, we consider $\mathcal{D}_{\text{ncm}}^{k=1}$ and $\mathcal{D}_{\text{new}}^{k=1}$ are "excellent", $\sqrt{\mathcal{D}_{\text{JS}}}$, $\mathcal{D}_{\text{ncm}}^{k=2}$, and $\mathcal{D}_{\text{new}}^{k=2}$ "good"; and $\mathcal{D}_{\text{JS}}$ "adequate".

With $P_{\text{visinfo}} = \{0, 1\}$ and $P_{\text{uniform}} = \{0.5, 0.5\}$, we can calculate K_v and K_ψ, which are shown in the middle and right PCPs in Figure 4, respectively. From the two PCPs, we cannot observe any major issue in categorizing positive and negative impact by any candidate measure. Hence, for **knowledge impact**, we consider all "excellent".

To continue the multi-criteria decision analysis (MCDA) [51] in the first part of the paper [4], we rate the six candidate measures using the same scoring system, i.e., using ordinal values between 0 and 5 (0 unacceptable, 1 fall-short, 2 inadequate, 3 mediocre, 4 good, and 5 best). By combining our evaluation of the ordering of divergence, benefit quantification, and the sign of knowledge impact, we give a 5 score to $\mathcal{D}_{\text{new}}^{k=1}$, a 4 to $\mathcal{D}_{\text{new}}^{k=2}$ and $\mathcal{D}_{\text{ncm}}^{k=1}$, a 3 to $\mathcal{D}_{\text{ncm}}^{k=2}$, a 2 to $\sqrt{\mathcal{D}_{\text{JS}}}$ and a 1 to $\mathcal{D}_{\text{JS}}$. The qualitative rating and numerical scores are given in Table 1.

5.2. Synthetic Case S_2

We now consider a slightly more complicated scenario with four pieces of data, A, B, C, and D, which can be defined as an alphabet $\mathbb{Z}_w$ with four letters. The ground truth PMF is $Q = \{0.1, 0.4, 0.2, 0.3\}$. Consider two processes that combine these into two classes AB and CD, each resulting in a two-letter alphabet $\mathbb{Z}_v$. These typify clustering algorithms, downsampling processes, discretization in visual mapping, and so on. One process is considered to be *correct*, which has a PMF for AB and CD as $R_{\text{correct}} = \{0.5, 0.5\}$, and another *biased* process with $R_{\text{biased}} = \{0, 1\}$.

Let CG, CU, and CB be three users at the receiving end of the *correct* process, and BG, BS, and BM be three other users at the receiving end of the *biased* process. The users with different knowledge exhibit different abilities to reconstruct the original $\mathbb{Z}_w$ featuring A, B, C, and D from aggregated information about AB and CD in $\mathbb{Z}_v$. Similar to the *good-bad* scenario, such abilities can be captured by a PMF P_{sampled}. For example, we have:

- CG makes random guess, $P_{\text{CG}} = \{0.25, 0.25, 0.25, 0.25\}$.
- CU has useful knowledge, $P_{\text{CU}} = \{0.1, 0.4, 0.1, 0.4\}$.
- CB is highly biased, $P_{\text{CB}} = \{0.4, 0.1, 0.4, 0.1\}$.
- BG makes guess based on R_{biased}, $P_{\text{BG}} = \{0.0, 0.0, 0.5, 0.5\}$.
- BS makes a small adjustment, $P_{\text{BS}} = \{0.1, 0.1, 0.4, 0.4\}$.
- BM makes a major adjustment, $P_{\text{BM}} = \{0.2, 0.2, 0.3, 0.3\}$.

Figure 5 compares the divergence values returned by the candidate measures for these six users, while the transformations from Q to R_{correct} or R_{biased}, and then to P_{sampled} are

illustrated on the right. Different measures provided slightly different ordering of the six users as:

$$\mathcal{D}_{\text{JS}}, \sqrt{\mathcal{D}_{\text{JS}}}, \mathcal{D}_{\text{new}}^{k=2} : \text{CU} < \text{CG} < \text{BM} < \text{BS} < \text{CB} < \text{BG}$$

$$\mathcal{D}_{\text{ncm}}^{k=2} : \text{CU} < \text{CG} < \text{BM} < \text{BS} < [\text{CB}, \text{BG}]$$

$$\mathcal{D}_{\text{ncm}}^{k=1} : \text{CU} < \text{BM} < \text{CG} < \text{BS} < [\text{CB}, \text{BG}]$$

$$\mathcal{D}_{\text{new}}^{k=1} : \text{CU} < \text{BM} < \text{CG} < \text{BS} < \text{CB} < \text{BG}$$

$$\text{collective votes} : \text{CU} < \text{CG} < \text{BM} < \text{BS} < \text{CB} < \text{BG}$$

The **order of divergence** can be observed in the bar chart as the first PCP where the divergence values are scaled by $\mathcal{H}_{\text{max}} = 2$ bits. Using the collective votes as the benchmark, we consider $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, and $\mathcal{D}_{\text{new}}^{k=2}$ "excellent", $\mathcal{D}_{\text{ncm}}^{k=2}$ and $\mathcal{D}_{\text{new}}^{k=1}$ "good", and $\mathcal{D}_{\text{ncm}}^{k=1}$ "adequate".

The PCPs in Figure 5 also depict two additional sets of values for P_{visinfo} when a user relies solely on visual information. For CG, CU, and CB, the benchmark is Cvi that corresponds to $R_{\text{corrected}}$. For BG, BS, and BM, the benchmark is Bvi that corresponds to R_{biased}. From the first PCP, we can observe that Bvi causes more distortion than Cvi.
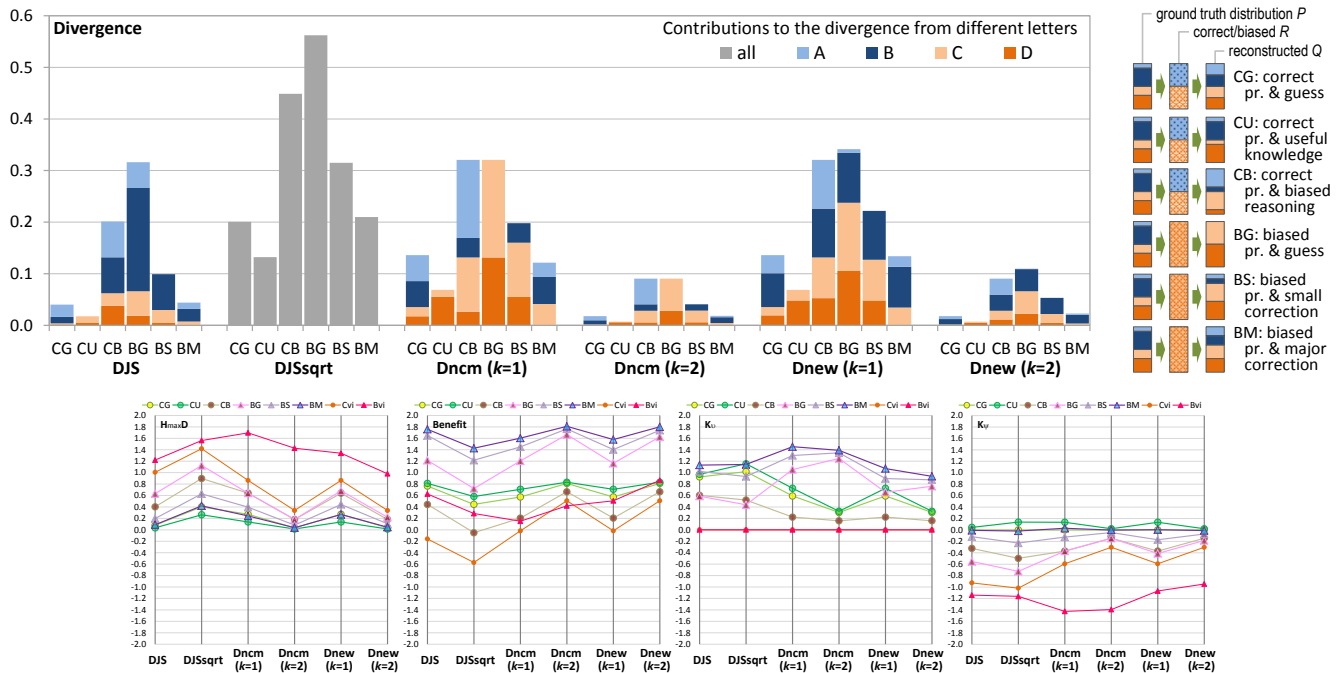
Figure 5. An example scenario with four data values: A, B, C, and D. Two processes (one correct and one biased) aggregated them to two values AB and CD. Users CG, CU, CB attempt to reconstruct [A, B, C, D] from the output [AB, CD] of the correct process, while BG, BS, and BM attempt to do so with the output from the biased processes. The bar chart shows the divergence values of the six users computed using the five candidate measures. The illustration on the right shows two transformations of the alphabets and their PMFs, one by the correct or biased process (pr.) and the other by the reconstruction. The bar chart shows the divergence values calculated by each candidate measure, while the four PCPs show the values of $\mathcal{H}_{\text{max}}\mathcal{D}$ (i.e., divergence scaled by SE_{max}), benefit, K_v and K_ψ. The values for Cvi and Bvi correspond to R_{correct} and R_{biased}, respectively.

However, because the entropy of the ground truth alphabet $\mathcal{H}(Q) = 1.84$, and the entropy values of $R_{\text{corrected}}$ and R_{biased} are 1 and 0 bits, R_{biased} results in more alphabet compression. The second PCP shows that if a user relies solely on visual information, R_{biased} leads to more benefit, except that $\mathcal{D}_{\text{ncm}}^{k=2}$ thinks otherwise. We cannot find major issues with other benefit values in the second PCP, though we consider that the negative

values produced by $\sqrt{\mathcal{D}_{JS}}$, $\mathcal{D}_{ncm}^{k=1}$, and $\mathcal{D}_{new}^{k=1}$ are intuitive. In terms of **benefit quantification**, we consider $\sqrt{\mathcal{D}_{JS}}$ to be "excellent", $\mathcal{D}_{ncm}^{k=1}$ and $\mathcal{D}_{new}^{k=1}$ "good", and the others "adequate".

Observing the third and fourth PCPs is interesting. The clustering algorithm changes $Q = \{0.1, 0.4, 0.2, 0.3\}$ to $R_{correct} = \{0.5, 0.5\}$ for users CG, CU, and CB and $R_{biased} = \{0, 1\}$ for users BG, BS, and BM. The random guess of Q with a uniform distribution is not that bad. Only CU's knowledge has a positive impact against random guess as shown in the last PCP.

Against the less-ideal visual information characterized by $R_{correct}$ and R_{biased}, the knowledge of all six users has a positive impact. It is important to state here the **knowledge** can be gained from other visualization. For example, we can postulate that the reason CU, BS, and BM can make adjustments against what the clustering algorithm says is because they have seen some visualizations of the raw data without clustering at an early stage of a workflow. In general, we cannot find any major issues with the PCPs for K_v and K_ψ. We thus rate all candidate measures as "excellent".

By combining our evaluation of the ordering of divergence, benefit quantification, and the sign of knowledge impact, we give a 5 score to $\sqrt{\mathcal{D}_{JS}}$, a 3 to $\mathcal{D}_{JS}$, $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$, and a 2 to $\mathcal{D}_{ncm}^{k=1}$ and $\mathcal{D}_{ncm}^{k=2}$.

5.3. An Extra Conceptual Criterion

The square root of JS-divergence, i.e., $\sqrt{\mathcal{D}_{JS}}$, is the only candidate measure that is not the probabilistic mean of its component measures, which correspond to the letters of the alphabet concerned. From the perspective of visualization, it cannot be depicted in the same way as the other five measures. As demonstrated in Figures 4 and 5, $\sqrt{\mathcal{D}_{JS}}$ values are depicted in grey bars, and one cannot view the individual contributions of its components to the overall divergence quantity as intuitively as others. Although this shortcoming of $\sqrt{\mathcal{D}_{JS}}$ may not affect the deployment of $\sqrt{\mathcal{D}_{JS}}$ in numerical applications, it will hinder its deployment in applications of visual analytics, making it difficult to observe, analyze, and explain the relationships between a divergence value and its component measures and the contributions of different component measures.

We encountered this issue after we considered the synthetic cases in this section. In order to avoid the complication of introducing any synthetic case in the first part of the paper [4], we report this issue as an extra conceptual criterion in this second part of the paper. For this extra conceptual criterion, we give a 1 score to $\sqrt{\mathcal{D}_{JS}}$, and a 5 score to each of other five candidate measures.

6. Experimental Case Studies

To complement the synthetic case studies in Section 5, we conducted two surveys to collect some realistic examples that feature the use of knowledge in visualization. In addition to providing instances of criteria $\mathbf{R}_1$ and $\mathbf{R}_2$ for selecting a bounded measure, the surveys were also designed to demonstrate that one could use a few simple questions to estimate the cost–benefit of visualization in relation to individual users.

It is necessary to note that these surveys are not intended for evaluating any hypothesis related to the application concerned. They are designed to collect data that may be similar to the results of a controlled, semi-controlled, or uncontrolled empirical study, or to the estimation by a visual designer after an interview with potential users.

6.1. Volume Visualization (Criterion R_1)

This survey, which involved ten surveyees, was designed to collect some real-world examples that reflect the use of knowledge in viewing volume visualization images. We invited surveyees with different levels of knowledge about volume visualization and medical imagining. They all volunteered their time as technical advisers without any financial reward. The full set of questions was presented to surveyees in the form of slides, which are included in the Supplementary Materials.

The full set of survey results is given in Appendix C. The featured volume datasets were from "The Volume Library" [53], and visualization images were either rendered by the authors or from one of the four publications [54–57]. The transformation from a volumetric dataset to a volume-rendered image typically features a noticeable amount of alphabet compression.

Some major algorithmic functions in volume visualization, e.g., iso-surfacing, transfer function, and rendering integral, all facilitate alphabet compression, hence information loss. As a rendering integral, maximum intensity projection (MIP) incurs a huge amount of information loss in comparison with the commonly-used emission-and-absorption integral [58]. As shown in Figure 6, the surface of arteries are depicted more or less in the same color.

The accompanying question intends to tease out two pieces of knowledge, "curved surface" and "with wrinkles and bumps". Among the ten surveyees, one selected the correct answer B, eight selected the relatively plausible answer A, and one selected the doubtful answer D. Among the participants, four rated their knowledge of medical imaging and volume visualization at 4 or 5 (out of 5). We consider them as an expert group. For this particular question (Figure 6), three selected answer A and one selected B.

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 6. 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. This example was first described in Part I of this two-part paper [4] for demonstrating the role of human knowledge in dealing with information loss due to many-to-one mappings in such a visualization image. Similar to Figure 3 (Section 3) in this part, the example was used in Part I to illustrate the difficulty to interpret the unboundedness of the KL-divergence when considering a binary alphabet $\mathbb{A} = \{curved, flat\}$ with maximum entropy of 1 bit.

Let alphabet $\mathbb{Z} = \{A, B, C, D\}$ contain the four optional answers. Based on our observation of photographs online and consultation with medical doctors, we first assume a ground truth PMF $Q_1 = \{0.1, 0.878, 0.002, 0.02\}$ since there might still be a small probability for a section of an artery to be flat or smooth. The rendered image depicts a misleading impression, implying that answer C is correct or a misleading PMF $R_C = \{0, 0, 1, 0\}$. The amount of alphabet compression is thus $\mathcal{H}(Q_1) - \mathcal{H}(R_C) = 0.628$ bits.

The top four PCPs in Figure 7 show the measurements returned by the six candidate measures in a way similar to the PCPs in Figures 4 and 5. In terms of **divergence ordering**, we notice a major anomaly that $\mathcal{D}_{\text{ncm}}^{k=1}$ returns divergence values indicating the "experts" group has the most divergence, followed by "all" and then "rest". Looking at some marginal difference in detail, $\mathcal{D}_{\text{ncm}}^{k=2}$ indicates that "all" has the highest divergence, followed by "rest" and then "experts". $\mathcal{D}_{\text{new}}^{k=1}$ indicates the group giving answer D has marginally more divergence than that giving answer C. These ordering conclusions are not intuitive.

$\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, and $\mathcal{D}_{new}^{k=2}$ returned the expected ordering, i.e., "rest" > "all" > "experts", and C > D > A > B.
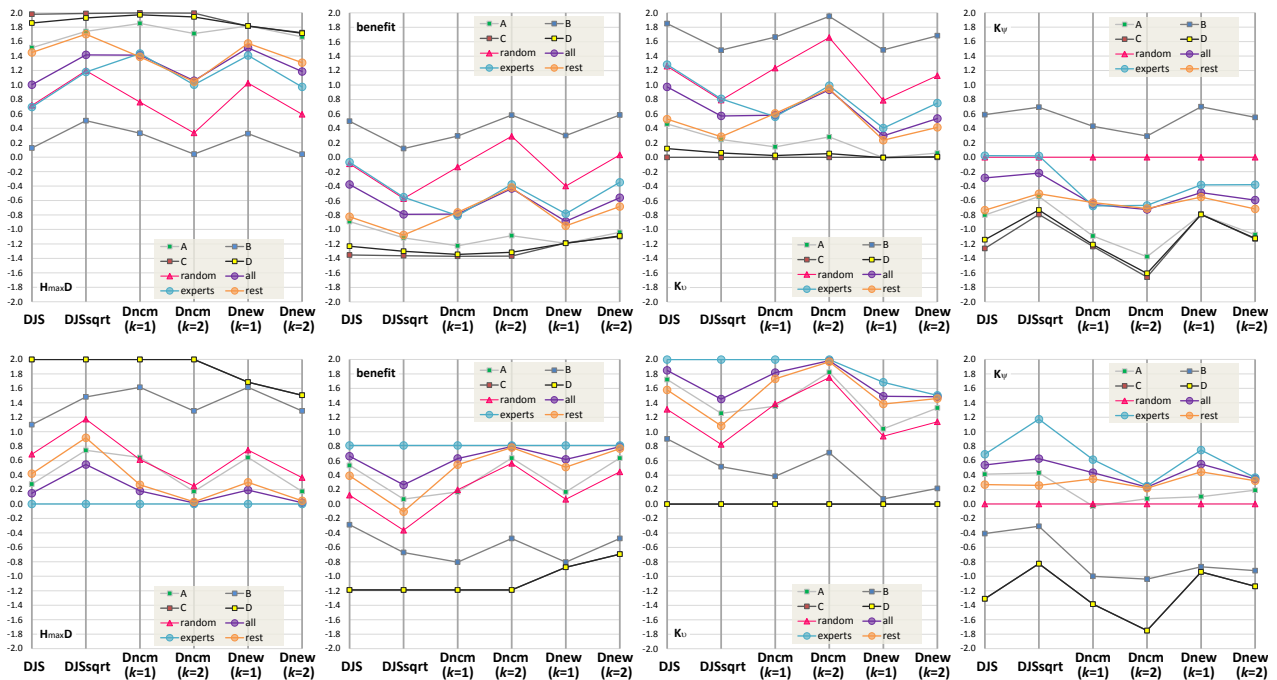
Figure 7. For the survey question shown in Figure 6, our survey of 10 participants returned 8 answers for A, 1 for B, 0 for C, and 1 for D. Among them, more knowledgeable participants (referred to as experts) returned 3 answers for A and 1 for B, and none for C or D. We consider two possible ground truth PMFs. $Q_1 = \{0.1, 0.878, 0.002, 0.02\}$ is based on our observations of photographs of arteries, and $Q_2 = \{0.75, 0.25, 0.0, 0.0\}$ is based on the experts' survey results. The top four PCPs show the values of $\mathcal{H}_{max}\mathcal{D}$, benefit, K_v, and K_ψ calculated based on Q_1, while the bottom four PCPs are measured based on Q_2. In addition, we also consider five other groups that make a random guess or always answer A, B, C, or D.

In terms of **benefit quantification**, $\mathcal{D}_{JS}$ and $\sqrt{\mathcal{D}_{JS}}$ suggest that "expert" is similar to making random guesses and "rest" is similar to the A group. $\mathcal{D}_{ncm}^k$ and $\mathcal{D}_{new}^k$ all consider that making random guesses is more beneficial than "expert". This becomes a question about how to interpret the difference between Q_1 and $\{0.25, 0.25, 0.25, 0.25\}$, and that between Q_1 and $\{0.75, 0.25, 0, 0\}$, i.e., which is the more meaningful difference?

We thus introduce a second possible ground truth PMF based on the answers of "experts", i.e., $Q_2 = \{0.75, 0.25, 0, 0\}$. The calculation results are depicted in the bottom four PCPs in Figure 7. In terms of **divergence order**, $\mathcal{D}_{ncm}^{k=1}$ shows an outlier, indicating the A group has more divergence than random guesses. With the observation of two PCPs in the first column of Figure 7, we consider $\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, and $\mathcal{D}_{new}^{k=2}$ "excellent", $\mathcal{D}_{ncm}^{k=2}$ and $\mathcal{D}_{new}^{k=1}$ "good", and $\mathcal{D}_{ncm}^{k=1}$ "inadequate".

$\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, $\mathcal{D}_{ncm}^{k=1}$, and $\mathcal{D}_{ncm}^{k=2}$ all rate the C and D groups with the maximum divergence, while $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$ do not. Following a careful reading of the intermediate calculation results, we notice that $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$ would rate the divergence between $\{1, 0, 0, 0\}$ and $\{0, 0, 1, 0\}$ as the maximum divergence, but not for the divergence between $\{0.75, 0.25, 0, 0\}$ and $\{0, 0, 1, 0\}$. This is an interesting feature of $\mathcal{D}_{new}^{k=1}$ and $\mathcal{D}_{new}^{k=2}$. We cannot decide whether to reward or penalize this feature. We hope to conduct future studies to examine the relative merits of this feature in detail.

In terms of **benefit quantification**, we cannot observe any major issues in the second column of Figure 7. We thus rate all candidate measures "excellent".

From the PCPs in the third and fourth columns, we notice that with Q_2, more groups show a positive **impact of knowledge**. This is understandable, as Q_1 deviates more from

the survey results. If we assume Q_1 is correct, then participants clearly do not have the necessary knowledge to answer the question in Figure 6 with the misleading MIP visualization. If Q_2 is correct, not only do the "experts" have the knowledge, but the "rest" group also seems to have some useful knowledge. In the Q_1-K_ψ PCP, only $\mathcal{D}_{\text{JS}}$ and $\sqrt{\mathcal{D}_{\text{JS}}}$ indicate a positive knowledge impact for the "experts". This is intuitive. In the Q_2-K_ψ PCP, only $\mathcal{D}_{\text{new}}^{k=1}$ indicates a negative knowledge impact for the A group.

This is not intuitive. We thus consider $\mathcal{D}_{\text{JS}}$ and $\sqrt{\mathcal{D}_{\text{JS}}}$ "excellent", $\mathcal{D}_{\text{ncm}}^{k=2}$, $\mathcal{D}_{\text{new}}^{k=1}$, and $\mathcal{D}_{\text{new}}^{k=2}$ "good", and $\mathcal{D}_{\text{ncm}}^{k=1}$ "adequate". By combining our evaluation of the ordering of divergence, benefit quantification, and the sign of knowledge impact, we give a 5 score to $\mathcal{D}_{\text{JS}}$, $\sqrt{\mathcal{D}_{\text{JS}}}$, a 4 to $\mathcal{D}_{\text{new}}^{k=2}$, a 3 to $\mathcal{D}_{\text{ncm}}^{k=2}$, and a 0 to $\mathcal{D}_{\text{ncm}}^{k=1}$.

6.2. London Underground Map (Criterion R_2)

This survey was designed to collect some real-world data that reflects the use of knowledge in viewing different London underground maps. It involved sixteen surveyees, twelve at King's College London (KCL) and four at University of Oxford. Surveyees were interviewed individually with the stimuli shown in Figure 1. Each surveyee was asked to answer 12 questions using either a geographically-faithful map or a deformed map, followed by two further questions about their familiarity of a metro system and London. A £5 Amazon voucher was offered to each surveyee as an appreciation of their effort and time. The survey sheets and the full set of survey results is given in Appendix D.

Harry Beck first introduced a geographically-deformed design of the London underground maps in 1931. Today, almost all metro maps around the world adopt this design concept. Information-theoretically, the transformation of a geographically-faithful map to such a geographically-deformed map causes a significant loss of information. Naturally, this affects some tasks more than others.

For example, the distances between stations on a deformed map are not as useful as in a faithful map. The first four questions in the survey asked surveyees to estimate how long it would take to walk (i) from *Charing Cross* to *Oxford Circus*, (ii) from *Temple* and *Leicester Square*, (iii) from *Stanmore* to *Edgware*, and (iv) from *South Rulslip* to *South Harrow*. On the deformed map, the distances between the four pairs of the stations are all about 50 mm. On the faithful map, the distances are (i) 21 mm, (ii) 14 mm, (iii) 31 mm, and (iv) 53 mm, respectively. According to the Google map, the estimated walk distances and times are (i) 0.9 miles, 20 min; (ii) 0.8 miles, 17 min; (iii) 1.6 miles, 32 min; and (iv) 2.2 miles, 45 min, respectively.

The average range of the estimations about the walk time by the 12 surveyees at KCL are: (i) 19.25 [8, 30], (ii) 19.67 [5, 30], (iii) 46.25 [10, 240], and (iv) 59.17 [20, 120] minutes. The estimations by the four surveyees at Oxford are: (i) 16.25 [15, 20], (ii) 10 [5, 15], (iii) 37.25 [25, 60], and (iv) 33.75 [20, 60] minutes. The values correlate better to the Google estimations than what would be implied by the similar distances on the deformed map. Clearly some surveyees were using some knowledge to make better inference.

Let $\mathbb{Z}$ be an alphabet of integers between 1 and 256. The range is chosen partly to cover the range of the answers in the survey, and partly to round up the maximum entropy $\mathbb{Z}$ to 8 bits. For each pair of stations, we can define a PMF using a skew normal distribution peaked at the Google estimation ξ. As an illustration, we coarsely approximate the PMF as $Q = \{q_i \mid 1 \le i \le 256\}$, where

$$q_i = \begin{cases} 0.01/236 & \text{if } 1 \le i \le \xi - 8 & (\textit{wild guess}) \\ 0.026 & \text{if } \xi - 7 \le i \le \xi - 3 & (\textit{close}) \\ 0.12 & \text{if } \xi - 2 \le i \le \xi + 2 & (\textit{spot on}) \\ 0.026 & \text{if } \xi + 3 \le i \le \xi + 12 & (\textit{close}) \\ 0.01/236 & \text{if } \xi + 13 \le i \le 256 & (\textit{wild guess}) \end{cases}$$

Using the same way in the previous case study, we can estimate the divergence and the benefit of visualization for an answer in each range. Recall our observation of the phenomenon in Section 6.1 that the measurements by $\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, $\mathcal{D}_{new}^{k=1}$, $\mathcal{D}_{new}^{k=2}$, $\mathcal{D}_{ncm}^{k=1}$ and $\mathcal{D}_{ncm}^{k=2}$ occupy different ranges of values, with $\mathcal{D}_{new}^{k=2}$ be the most generous in measuring the benefit of visualization. With the entropy of the alphabet as $\mathcal{H}(Q) \approx 3.6$ bits and the maximum entropy being 8 bits, the benefit values obtained for this example exhibit a compelling pattern:

Benefit for:	$\mathcal{D}_{JS}$	$\sqrt{\mathcal{D}_{JS}}$	$\mathcal{D}_{new}^{k=1}$	$\mathcal{D}_{new}^{k=2}$	$\mathcal{D}_{ncm}^{k=1}$	$\mathcal{D}_{ncm}^{k=2}$
spot on	-1.765	-2.777	-0.418	**0.287**	-3.252	-2.585
close	-3.266	-3.608	-0.439	**0.033**	-3.815	-3.666
wild guess	-3.963	-3.965	-0.416	-0.017	-3.966	-3.965

Only $\mathcal{D}_{new}^{k=2}$ has returned positive benefit values for *spot on* and *close* answers. Since it is not intuitive to say that those surveyees who gave good answers benefited from visualization negatively, clearly only the measurements returned by $\mathcal{D}_{new}^{k=2}$ are intuitive. In terms of **benefit quantification**, we consider thus $\mathcal{D}_{new}^{k=2}$ "excellent" and the other five measures "adequate".

In addition, the ordering resulting from $\mathcal{D}_{new}^{k=1}$ is inconsistent with others. For **divergence order**, we consider $\mathcal{D}_{new}^{k=1}$ "adequate" and the other five measures "excellent". We have not detected any major issues with the values for K_v and K_ψ. For the **impact of knowledge**, we thus rate all candidate measures "excellent".

More detailed discussions with further computational results and PCPs can be found in Appendix D. By combining all these observational ratings, we give a 5 score to $\mathcal{D}_{new}^{k=2}$, a 3 score to $\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, $\mathcal{D}_{ncm}^{k=1}$, and $\mathcal{D}_{ncm}^{k=2}$, and a 1 score to $\mathcal{D}_{new}^{k=1}$.

7. Conclusions

This two-part paper aims 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 [3]. The concern about the original measure is its unbounded term based on the KL-divergence. The conceptual analysis in the first part of the paper [4] examined nine candidate measures and narrowed the options down to six, providing important evidence to the multi-criteria decision analysis (MCDA) of these candidate measures.

In the second part of the paper, we used two synthetic and two experimental case studies to obtain some data, which allowed us to observe the behaviors of the remaining candidate measures. Building on the MCDA in the first part, the case studies provided two additional aspects of the MCDA with important evidence.

From the top table in Table 1, we can observe that the empirical data helps identify the strengths and weaknesses of each candidate measures considered in this paper. The empirical data suggests that $\mathcal{D}_{new}^{k}(k = 2)$ is slightly ahead of $\sqrt{\mathcal{D}_{JS}}$ (i.e., 16 vs. 15). Since the conceptual analysis in the first part of this paper [4] gives a subtotal of 30 to $\mathcal{D}_{JS}$, $\sqrt{\mathcal{D}_{JS}}$, and $\mathcal{D}_{new}^{k}(k = 2)$. We cannot separate $\mathcal{D}_{new}^{k}(k = 2)$ and $\sqrt{\mathcal{D}_{JS}}$ conclusively.

However, it is necessary to consider the extra conceptual criteria discussed in Section 5.3. From a visualization perspective, we cannot ignore the shortcoming of $\sqrt{\mathcal{D}_{JS}}$ discovered during the analysis of empirical data (i.e., its value is not the probabilistic mean of the entropic measures of its components). This places $\mathcal{D}_{new}^{k}(k = 2)$ in a favorable position. We therefore propose to revise the original cost–benefit ratio in [3] to the following:

$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\text{Alphabet Compression} - \text{Potential Distortion}}{\text{Cost}}$$
$$= \frac{\mathcal{H}(\mathbb{Z}_i) - \mathcal{H}(\mathbb{Z}_{i+1}) - \mathcal{H}_{max}(\mathbb{Z}_i)\mathcal{D}_{new}^2(\mathbb{Z}_i'||\mathbb{Z}_i)}{\text{Cost}} \tag{7}$$

This cost–benefit measure was developed in the field of visualization for optimizing visualization processes and visual analytics workflows. Its broad interpretation may include data intelligence workflows in other contexts [59]. The measure has now been improved by using visual analysis and with the empirical data collected in the context of visualization applications.

The history of measurement science [1] informs us that proposals for metrics, measures, and scales will continue to emerge in visualization, typically following the arrival of new theoretical understanding, new observational data, new measurement technology, and so on. As measurement is one of the driving forces in science and technology, we shall welcome such new measurement development in visualization.

The work presented in the first part of this paper [4] and this second part does not indicate a closed chapter but an early effort to be improved frequently in the future. For example, future work may discover measures that have better mathematical properties than $\mathcal{D}_{new}^{k=2}$, $\sqrt{\mathcal{D}_{JS}}$, and $\mathcal{D}_{JS}$, or future experimental observations may provide evidence that $\mathcal{D}_{JS}$ or $\sqrt{\mathcal{D}_{JS}}$ offer more intuitive explanations than $\mathcal{D}_{new}^{k=2}$ in other case studies. In particular, we would like to continue our theoretical investigation into the mathematical properties of $\mathcal{D}_{new}^{k}$.

"Measurement is not an end but a means in the process of description, differentiation, explanation, prediction, diagnosis, decision making, and the like" [7]. Having a bounded cost–benefit measure offers many new opportunities of developing tools for aiding the measurement and optimization of data intelligence workflows and for using such tools in practical applications, especially in visualization and visual analytics.

Supplementary Materials: The following are available at https://www.mdpi.com/article/10.3390/e24020282/s1, File S1: Questionnaire.

Institutional Review Board Statement: The surveys were conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of King's College London (Reference MRA-19/20-14614, 31 January) and the Ethics Committee of University of Oxford (Reference CS_C1A_19_048, 26 November 2019).

Informed Consent Statement: Informed consent was obtained from all participants involved in the surveys.

Data Availability Statement: The survey results are reported in Appendices C and D.

Abbreviations

The following abbreviations are used in this manuscript:

BG	Biased process in conjunction with faithful Guess
BM	Biased process in conjunction with a Major adjustment in decision
BS	Biased process in conjunction with a Small adjustment in decision
CB	Correct process in conjunction with Biased reasoning
CG	Correct process in conjunction with Random guess
CU	Correct process in conjunction with Useful knowledge
FD	a Fair amount of Doubt
KL	Kullback–Leibler
LD	Little Doubt

MCDA	Multi-Criteria Decision Analysis
ML	Machine Learning
PCP	Parallel Coordinates Plot
PMF	Probability Mass Function
OC	Over-Compensate
RG	Random Guess
UC	Under-Compensate

Appendix A. Explanation of the Original Cost-Benefit Measure

This appendix is not an independent paper but supports Part II of this two-part paper by providing some background information about the cost–benefit ratio proposed in [3] and its role in visualization. Mathematical knowledge of information theory is not a prerequisite for reading this appendix.

The first section of this appendix contains an extraction from a previous publication [47], offering a relatively concise but informative description of the cost–benefit ratio. The inclusion of this is to minimize the readers' effort to locate such an explanation. The extraction has been slightly modified. In the second section of this appendix, we provide a relatively informal and somehow conversational discussion about using this measure to explain why visualization is useful.

Appendix A.1. An Information-Theoretic Measure for Cost-Benefit Analysis

Chen and Golan introduced an information-theoretic formula for measuring the cost–benefit ratio of a visual analytics (VA) workflow or any of its component processes [3]. The formula consists of three fundamental measures that are abstract representations of a variety of qualitative and quantitative criteria used in practice, including operational requirements (e.g., accuracy, speed, errors, uncertainty, provenance, and automation), analytical capability (e.g., filtering, clustering, classification, and summarization), cognitive capabilities (e.g., memorization, learning, context-awareness, and confidence), and so on. The abstraction results in a formula with the desirable mathematical simplicity [3]. The qualitative form of the formula is as follows:

$$\frac{Benefit}{Cost} = \frac{Alphabet\ Compression - Potential\ Distortion}{Cost} \tag{A1}$$

The formula describes the trade-off among the three fundamental measures:

- *Alphabet Compression* (AC) measures the amount of entropy reduction (or information loss) achieved by a process. As it was noticed in [3], most visual analytics processes (e.g., statistical aggregation, sorting, clustering, visual mapping, and interaction) feature many-to-one mappings from input to output, hence losing information. Although information loss is commonly regarded as harmful, it cannot be all bad if it is a general trend of VA workflows. Thus, the cost–benefit ratio makes AC a positive component.
- *Potential Distortion* (PD) balances the positive nature of AC by measuring the errors typically due to information loss. Instead of measuring mapping errors using some third party metrics, PD measures the potential distortion when one reconstructs inputs from outputs. The measurement takes into account humans' knowledge that can be used to improve the reconstruction processes. For example, given an average mark of 62%, the teacher who taught the class can normally guess the distribution of the marks among the students better than an arbitrary person.
- *Cost* (Ct) of the forward transformation from input to output and the inverse transformation of reconstruction provides a further balancing factor in the cost–benefit ratio in addition to the trade-off between AC and PD. In practice, one may measure the cost using *time* or a monetary measurement.

Appendix A.2. An Information-Theoretic Reasoning about Why Visualization Is Useful

There have been many arguments about why visualization is useful. Streeb et al. collected a large number of arguments and found that many arguments were in conflict with each other [60]. Chen and Edwards presented an overview of schools of thought in the field of visualization, and showed that the "why" question was a bone of major contention [61].

The most common argument about the "why" question is because visualization offers insight or helps humans to gain insight. When this argument is used outside the visualization community, there are often counter-arguments that statistics and algorithms can offer insight automatically and often with better accuracy and efficiency. There are also concerns that visualization may mislead viewers, which casts further doubts about the usefulness of visualization, while leading to a related argument that "visualization must be accurate" in order for it to be useful.

The accuracy argument itself is not bullet-proof since there are many types of uncertainty in a visualization process, from uncertainty in data, to that caused by visual mapping, and to that during perception and cognition [62]. Nevertheless, it is easier to postulate that visualization must be accurate, as it seems to be counter-intuitive to condone the idea that "visualization can be inaccurate," not mentioning the idea that "visualization is normally inaccurate," or "visualization should be inaccurate."

The word "inaccurate" is itself an abstraction of many different types of inaccuracy. Misrepresentation truth is a type of inaccuracy. Such acts are mostly wrong, but some (such as wordplay and sarcasm) may cause less harm. Converting a student's mark in the range of [0, 100] to the range of [A, B, C, D, E, F] is another type of inaccuracy. This is a common practice and must be useful. From an information-theoretic perspective, these two types of inaccuracy are information loss.

In their paper [3], Chen and Golan observed that statistics and algorithms usually lose more information than visualization. Hence, this provides the first hint about the usefulness of visualization. They also noticed that like wordplay and sarcasm, the harm of information loss can be alleviated by knowledge. For someone who can understand a workplay (e.g., a pun) or can sense a sarcastic comment, the misrepresentation can be corrected by that person at the receiving end. This provides the second hint about the usefulness of visualization because any "misrepresentation" in visualization may be corrected by a viewer with appropriate knowledge.

On the other hand, statistics and algorithms are also useful, and sometimes more useful than visualization. As statistics and algorithms usually cause more information loss, some aspects of information loss must be useful. One important merit of losing information in one process is that the succeeding process has less information to handle and thus incurs less cost. This is why Chen and Golan divided information loss into two components, a positive component called *alphabet compression* and a negative component called *potential distortion* [3].

The positive component explains why statistics, algorithms, visualization, and interaction are useful because they all lose information. The negative component explains why they are sometimes less useful because information loss may cause distortion during information reconstruction. Both components are moderated by the cost of a process (i.e., statistics, algorithms, visualization, or interaction) in losing information and reconstructing the original information. Hence, given a dataset, the best visualization is the one that loses the most information while causing the least distortion. This also explains why visual abstraction is effective when the viewers have adequate knowledge to reconstruct the lost information and may not be effective otherwise [63].

The central thesis by Chen and Golan [3] may appear to be counter-intuitive to many as it suggests "inaccuracy is a good thing", partly because the word "inaccuracy" is an abstraction of many meanings and itself features information loss. Perhaps the reason for the conventional wisdom is that it is relatively easy to think that "visualization must be accurate". To a very small extent, this is a bit like the easiness to think that "the

earth is flat" a few centuries ago, because the evidence for supporting that wisdom was available everywhere, right in front of everyone at that time. Once we step outside the field of visualization, we can see the phenomena of inaccuracy everywhere, in statistics and algorithms as well as in visualization and interaction. All these suggest that "the earth may not be flat," or "inaccuracy can be a good thing".

In summary, the cost–benefit ratio proposed by Chen and Golan [3] explains that when visualization is useful, it is because visualization has a better trade-off than simply reading the data, simply using statistics alone, or simply relying on algorithms alone.

The ways to achieve a better trade-off include: (i) visualization may lose some information to reduce the human cost in observing and analyzing the data, (ii) it may lose some information since the viewers have adequate knowledge to recover such information or can acquire such knowledge at a lower cost, (iii) it may preserve some information because it reduces the reconstruction distortion in the current and/or succeeding processes, and (iv) it may preserve some information because the viewers do not have adequate knowledge to reconstruct such information or it would cost too much to acquire such knowledge.

Appendix B. How Tasks and Users Are Featured in the Cost-Benefit Ratio?

This appendix is not an independent paper but supports Part II of this two-part paper by explaining how information theory can explain mathematically represent the common wisdom in visualization that *visualization is user- and task-dependent*. Mathematical knowledge of information theory is not a prerequisite for reading this appendix.

Whilst hardly anyone in the visualization community would support any practice intended to deceive viewers, there have been many visualization techniques that inherently cause distortion to the original data. The deformed London underground map in Figure 1 shows such an example. The distortion in this example is largely caused by many-to-one mappings. A group of lines that would be shown in different lengths in a faithful map is now shown with the same length. Another group of lines that would be shown with different geometric shapes is now shown as the same straight line. In terms of information theory, when the faithful map is transformed to the deformed map, a good portion of information has been lost because of these many-to-one mappings.

In fact, there are many other forms of information loss. For example, when a high-resolution data variable (e.g., an integer in the range [0, 10,000]) is visually encoded as a bar in a bar chart that is restricted to a height of 1000 pixels, about every 10 values are mapped onto the same height in terms of pixels. Furthermore, it is unlikely that humans can precisely identify the height of each bar at the pixel resolution. Likely a viewer may perceive a height of 833 pixels to be the same as one with 832 pixels or 834 pixels, which is also a many-to-one mapping. When multivariate data records are encoded as glyphs, there is usually a significant amount of information loss. As we have discussed in the first part of this paper [4], in volume visualization, when a sequence of n voxel values are transformed to a single pixel value, as long as n is a reasonably large value, a huge amount of information loss is almost guaranteed to happen.

Despite the ubiquitous phenomenon of information loss in visualization, it has been difficult for many of us to contemplate the idea that information loss may be a good thing. In particular, one theory based on an algebraic framework defines three principles that formalize the notion of graphical integrity to prevent such information loss [64]. When one comes across an effective visualization but featuring noticeable information loss, the typical answer is that it is task-dependent, and the lost information is not useful to the task concerned. When a visualization is evaluated, common critiques are about information loss, such as inadequate resolution, view obstruction, distorted representation, which are also characteristics of the aforementioned glyphs, volume rendering, and deformed metro map.

The common phrase that "the appropriateness of information loss depends on tasks" is not an invalid explanation. On its own, this explanation is not adequate, because:

- The appropriateness depends on many attributes of a task, such as the selection of variables in the data and their encoded visual resolution required to complete a task satisfactorily, and the time allowed to complete a task.
- The appropriateness depends also on other factors in a visualization process, such as the original data resolution, the viewer's familiarity of the data, the extra information that is not in the data but the viewer knows, and the available visualization resources.
- The phrase creates a gray area as to whether information loss is allowed or not, and when or where one could violate some principles such as those principles in [64].

Partly inspired by the above puzzling dilemma in visualization and partly by a similar conundrum in economics "what is the most appropriate resolution of time series for an economist", Chen and Golan proposed an information-theoretic cost–benefit ratio for measuring various factors involved in visualization processes [3]. As this cost–benefit ratio can measure some abstract characteristics of "data", "visualization", "information loss", "knowledge", and "task" using the most fundamental information-theoretic unit *bit*, it provides a means to define their relationship coherently. In this appendix, we continue to use the qualitative version of this cost–benefit ratio as given in Equation (A1) in Appendix A, making it more accessible to readers who are not familiar with information theory.

Chen and Golan noticed that not only do visualization processes lose information but also other data intelligence processes also lose information. For example, when statistics is used to down-sample a time series, or to compute its statistical properties, there is a substantial amount of information loss; when an algorithm groups data points into clusters or sort them according to a key variable, there is information loss; and when a computer system asks a user to confirm an action, there is information loss in the computational processes [65].

They also noticed that almost all decision tasks, the number of decision options is usually rather small. In terms of information theoretic quantities, the amount of information (i.e., in terms of Shannon entropy) associated with a decision task is usually much lower than the amount of information associated with the data entering a data intelligence workflow. They concluded that this general trend of information reduction must be a positive thing for any data intelligence workflows. They referred to the amount of information reduction as *Alphabet Compression* (AC) and made it a positive contribution to the *benefit* term in Equation (A1).

Figure A1 shows an example of a simple visual analytics workflow, where at the moment, the visual analytics process is simply a visualization process, (a_1), for viewing a deformed London underground map. There can be many possible visualization tasks, such as counting the number of stops between two stations, searching for a suitable interchange station, and so on. From the workflow in Figure A1, one can easily observe that the amount of information contained in the world around the entire London underground system must be much more than the information contained in the digital data describing the system.

The latter is much more than the information depicted in the deformed map. By the time when the workflow reaches a task, the number of decision options is usually limited. For example, counting the number stops may have optional values between 0 and 50. The amount of information contained in the counting result is much smaller than that in the deformed map. This evidences the general trend observed in [3].
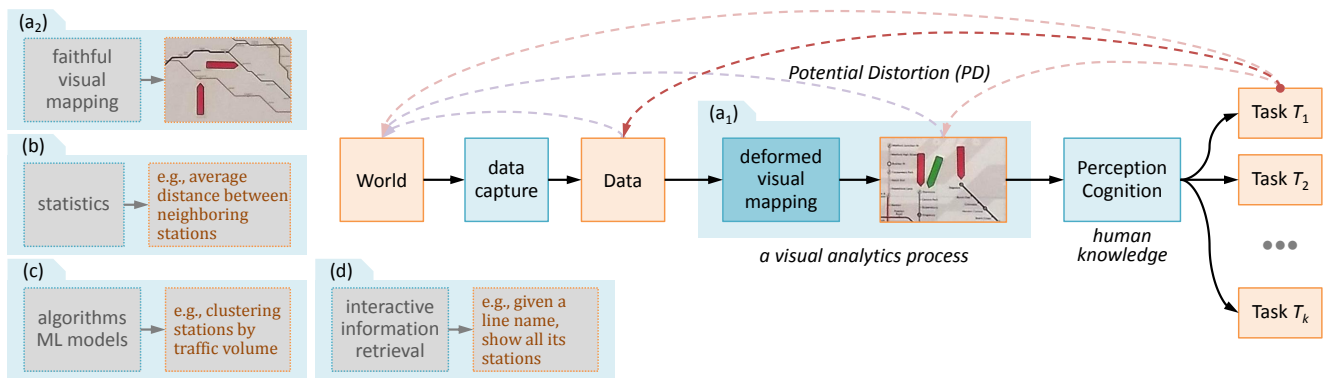
Figure A1. A visual analytics workflow features a general trend of alphabet compression from left (World) to right (Tasks). The potential distortion compares at an information space reconstructed based on the output with the original input information space. When we place different processes (i.e., (**a₁,a₂,b–d**)), in the workflow, we can appreciate that statistics, algorithms, visualization, and interaction have different levels of alphabet compression, potential distortion, and cost.

After considering the positive contribution of information, we must counterbalance AC by the the term *Potential Distortion* (PD), which describes, in abstract, the negative consequences that may be caused by information loss. In the past, one typically uses a third-party metric to determine whether a chosen decision option is good or not. This introduces a dilemma that one needs a fourth-party metric to determine if the third-party metric is good or not, and this can go on forever.

At least, mathematically, this unbounded reasoning paradigm is undesirable. This third-party metric was avoided in Equation (A1) by imagining if a viewer would have to reconstruct the original data that is visualized, how much the reconstructed data would diverge from the original data. In [3], this divergence is measured using the well-known Kullback–Leibler divergence (KL-divergence) [66]. As this divergence measure is unbounded, Chen and Sbert proposed to replace it with a bounded measure in the first part of thus paper [4], where they have detailed the concerns about the unboundedness.

As shown in Equation (A1), the AC term makes a positive contribution, the PD term makes a negative contribution, reflecting the two sides of the same coin of information loss. Both terms have the same unit *bit* and are moderated by the term *Cost*. The term AC characterizes many useful approaches in visualization and visual analytics, such as data filtering and visual abstraction, while the term PD characterizes many undesirable shortcomings, such as rendering errors and perceptual errors.

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). The term is defined as an energy measure, but can be approximated using time, monetary, and other appropriate measures.

The cost–benefit ratio in Equation (A1) can also be used to measure other processes in a visual analytics workflow. One can simply imagine replacing the block (a₁) in Figure A1 with one of the other four blocks on the left, (a₂) for faithful visual mapping, (b) for statistics, (c) for algorithms, and (d) for interactive information retrieval. This exercise allows us to compare the relative merits among the four major components of visual analytics, i.e., statistics, algorithms, visualization, and interaction [67].

For example, statistics may be able to deliver a set of indicators about the London underground map to a user. In comparison with the deformed map, these statistical indicators contain much less information than the map, offering more AC contribution. If a user is asked to imagine what the London underground system looks like, having these statistical indicators will not be very helpful. Hence, statistics may cause more PD.

Of course, whether to use statistics or visualization may be task-dependent. Mathematically, this is largely determined by both the PD and *Cost* associated with the perception and cognition process in Figure A1. If a user tries to answer a statistical question using

the visualization, it is likely to cost more than using statistics, provided that the statistical answer has already been computed or statistical calculation can be performed easily and quickly.

Whether to use statistics or visualization may also be user-dependent. Consider a user **A** that has a fair amount of prior knowledge about the London underground system and another user **B** that has little. If both are shown some statistics about the system (e.g., the total number of stations of each line), **A** can redraw the deformed map more accurately than **B** and more accurately than without the statistics, even though the statistical information is not meant to support the users' this task. Hence, to **A**, having a deformed map to help appreciate the statistics may not be necessary, while to **B**, viewing both statistics and the deformed map may help reduced the PD but may also incur more cost in terms of effort. Hence, visualization is more useful to **B**.

This example echos the scenario presented in Figure 1, where we asked two questions: Can information theory explain this phenomenon? Can we quantitatively measure some factors in this visualization process? If prior knowledge can explain the trade-off among AC, PD, and *Cost* in comparing statistics and deformed map. We can also extrapolate this reasoning to analyze the trade-off in comparing viewing the deformed map (more AC) and viewing the faithful map (less AC) as in Figure 1. Perhaps we can now be more confident to say that information theory can explain such a phenomenon. In the main body of this second part of the paper, we demonstrate the potential answer to the second question, i.e., we can quantitatively measure some relevant factors in such a visualization process.

To some readers, it may still be counter-intuitive to consider that information loss has a positive side. It is essential for asserting why visualization is useful as well as asserting the usefulness of statistics, algorithms, and interaction since they all usually cause information loss [47]. Further discourse on this topic can be found in Appendix A.

Table A1. The answers by ten surveyees to the questions in the volume visualization survey. The surveyees are ordered from left to right according to their self-ranking about the knowledge of volume visualization. In rows 1–8, the dataset used in each question is indicated in square brackets. Correct answers are indicated by letters in round brackets. The upper case letters are the most appropriate answers, while the lower case letters with brackets are acceptable answers as they are correct in some circumstances. The lower case letters without brackets are incorrect answers. In rows 9 and 10, the self-assessment scores are in the range of [1 lowest, 5 highest].

Questions with (Correct Answers) and [Database] in Brackets	Surveyee's ID									
	S1	S2	S3	S4	S5	S6	S7	S8	P9	P10
1. Use of different transfer functions (D), [Carp]	(D)	(D)	(D)	(D)	(D)	c	b	(D)	a	c
2. Use of translucency in volume rendering (C), [Engine Block]	(C)	(C)	(C)	(C)	(C)	(C)	(C)	(C)	d	(C)
3. Omission of voxels of soft tissue and muscle (D), [CT head]	(D)	(D)	(D)	(D)	b	b	a	(D)	a	(D)
4. sharp objects in volume-rendered CT data (C), [CT head]	(C)	(C)	a	(C)	a	b	d	b	b	b
5. Loss of 3D information with MIP (B, a), [Aneurysm]	(a)	(B)	(a)	(a)	(a)	(a)	D	(a)	(a)	(a)
6. Use of volume deformation (A), [CT head]	(A)	(A)	b	(A)	(A)	b	b	(A)	b	b
7. Toenails in non-photo-realistic volume rendering (B, c), [Foot]	(c)	(c)	(c)	(B)	(c)	(B)	(B)	(B)	(B)	(c)
8. Noise in non-photo-realistic volume rendering (B), [Foot]	(B)	(B)	(B)	(B)	(B)	(B)	a	(B)	c	(B)
9. Knowledge about 3D medical imaging technology	4	3	4	5	3	3	3	3	2	1
10. Knowledge about volume rendering techniques	5	5	4–5	4	4	3	3	3	2	1

Table A2. Summary statistics of the survey results in Table A1, where we classified experts simply based on their self-assessment with an average rate ($\geq$4) in answering Q9 and Q10. They are S1, S2, S3, and S4.

Question		All Participants				Experts				The Rest			
		A	B	C	D	A	B	C	D	A	B	C	D
1. (Carp)	numbers:	1	1	2	6	0	0	0	4	1	1	2	2
	probability:	0.10	0.10	0.20	0.60	0.00	0.00	0.00	1.00	0.17	0.17	0.33	0.33
2. (Engine Block):	numbers:	0	0	1	9	0	0	0	4	0	0	1	5
	probability:	0.00	0.00	0.10	0.90	0.00	0.00	0.00	1.00	0.00	0.00	0.17	0.83
3. (CT head)	numbers:	2	2	0	6	0	0	0	4	2	2	0	2
	probability:	0.20	0.20	0.00	0.60	0.0	0.0	0.0	1.00	0.33	0.33	0.00	0.33
4. (CT head)	numbers:	2	4	3	1	1	0	0 3	0	1	4	0	1
	probability:	0.20	0.40	0.30	0.10	0.25	0.00	0.75	0.00	0.17	0.67	0.00	0.17
5. (Aneurism)	numbers:	8	1	0	1	3	1	0	0	5	0	0	1
	probability:	0.80	0.10	0.00	0.10	0.75	0.25	0.00	0.00	0.83	0.00	0.00	0.17
6. (CT head)	numbers:	5	5	0	0	3	1	0	0	2	4	0	0
	probability:	0.50	0.50	0.00	0.00	0.75	0.25	0.00	0.00	0.33	0.67	0.00	0.00
7. (Foot)	numbers:	0	5	5	0	0	1	3	0	0	4	2	0
	probability:	0.00	0.50	0.50	0.00	0.00	0.25	0.75	0.00	0.00	0.67	0.33	0.00
8. (Foot)	numbers:	1	8	1	0	0	4	0	0	1	4	1	0
	probability:	0.10	0.80	0.10	0.00	0.00	1.00	0.00	0.00	0.17	0.67	0.17	0.00

Appendix C. Survey Results of Useful Knowledge in Volume Visualization

This appendix is not an independent paper but supports Part II of this two-part paper by providing additional details about the survey on volume visualization described in Section 6.1 in this part of the paper.

This survey consists of eight questions presented as slides. The questionnaire is given in the Supplementary Materials. The ten surveyees were primarily colleagues from the UK, Spain, and the USA. They include doctors and experts in medical imaging and visualization, as well as several persons who are not familiar with the technologies of medical imaging and data visualization. Table A1 summarizes the answers from these ten surveyees.

There is also a late-returned survey form that was not included in the analysis. As a record, the answers in this extra survey form are: 1: c, 2: d, 3: (D), 4: a, 5: (a), 6: (A), 7: (c), 8: (B), 9: 5, 10: 4. The upper case letters (always in brackets) are the most appropriate answers, while the lower case letters with brackets are acceptable answers as they are correct in some circumstances. The lower case letters without brackets are incorrect answers.

The following example illustrates how to estimate the **benefit of visualization** and **knowledge impact** based on the survey result of Question 5. We first define the following:

- Ground truth PMF $Q = \{0.1, 0.878, 0.002, 0.02\}$.
- If one always answers A: $P_a = \{1, 0, 0, 0\}$.
- If one always answers B: $P_b = \{0, 1, 0, 0\}$.
- If one always answers C: $P_c = \{0, 0, 1, 0\}$.
- If one always answers D: $P_d = \{0, 0, 0, 1\}$.
- Survey results (all): $P_{\text{all}} = \{0.8, 0.1, 0, 0.1\}$.
- Survey results (expert): $P_{\text{expert}} = \{0.75, 0.25, 0, 0\}$.
- Survey results (rest): $P_{\text{rest}} = \{0.83, 0, 0, 0.17\}$.

We can roughly translate the survey results to the following PMFs:

- Survey results (all): $P_{\text{all}} = \{0.8, 0.1, 0, 0.1\}$.
- Survey results (expert): $P_{\text{expert}} = \{0.75, 0.25, 0, 0\}$.
- Survey results (rest): $P_{\text{rest}} = \{0.83, 0, 0, 0.17\}$.

The four sets of measured values returned by different candidate measures are shown in Figure A2, i.e., (a) divergence values, (b) benefit of visualization, (c) impact of knowledge K_v (against the scenario of relying on visual information only), and (d) impact of knowledge

K_ψ (against the scenario of random guess). In each table, the columns labeled with 1–6 are six candidate bounded divergence measures used for estimating the values concerned in each table (i.e., divergence, benefit, K_υ, or K_ψ).

Row "random" estimates the values for a viewer who makes random guesses. Rows "A"–"D" estimate the values for viewers for answering "A"–"D", respectively. Row "all" estimates the values associated with all viewers as a group. Row "experts" estimates the values associated with those experts as a group, while row "rest" estimates the values associated with the non-expert group.

Divergenc	1	2	3	4	5	6
	DJS	DJSsqrt	Dncm1	Dncm2	Dnew1	Dnew2
random	0.359	0.599	0.3808	0.168	0.5124	0.297
A	0.758	0.871	0.9260	0.856	0.9087	0.833
B	0.064	0.253	0.1661	0.021	0.1631	0.021
C	0.990	0.995	0.9986	0.997	0.9066	0.862
D	0.929	0.964	0.9855	0.971	0.9086	0.858
all	0.502	0.708	0.7066	0.529	0.7572	0.593
experts	0.348	0.589	0.7176	0.501	0.7039	0.487
rest	0.725	0.851	0.6942	0.522	0.7879	0.655

(**a**) divergence values

Benefit	1	2	3	4	5	6
	DJS	DJSsqrt	Dncm1	Dncm2	Dnew1	Dnew2
random	-0.090	-0.570	-0.134	0.292	-0.397	0.034
A	-0.889	-1.114	-1.224	-1.084	-1.190	-1.038
B	0.500	0.122	0.296	0.585	0.302	0.586
C	-1.351	-1.362	-1.369	-1.366	-1.185	-1.097
D	-1.230	-1.300	-1.343	-1.314	-1.189	-1.088
all	-0.376	-0.789	-0.785	-0.431	-0.886	-0.559
experts	-0.067	-0.551	-0.807	-0.375	-0.780	-0.345
rest	-0.822	-1.075	-0.761	-0.417	-0.948	-0.681

(**b**) benefit of visualization

K_upsilon	1	2	3	4	5	6
	DJS	DJSsqrt	Dncm1	Dncm2	Dnew1	Dnew2
random	1.262	0.792	1.236	1.658	0.788	1.130
A	0.463	0.248	0.145	0.282	-0.004	0.059
B	1.851	1.484	1.665	1.952	1.487	1.683
C	0.000	0.000	0.000	0.000	0.000	0.000
D	0.121	0.062	0.026	0.052	-0.004	0.009
all	0.976	0.573	0.584	0.935	0.299	0.538
experts	1.284	0.811	0.562	0.992	0.405	0.751
rest	0.529	0.287	0.609	0.949	0.237	0.415

(**c**) knowledge impact K_υ

K_psi	1	2	3	4	5	6
	DJS	DJSsqrt	Dncm1	Dncm2	Dnew1	Dnew2
random	0.000	0.000	0.000	0.000	0.000	0.000
A	-0.799	-0.544	-1.090	-1.376	-0.793	-1.071
B	0.590	0.692	0.429	0.293	0.699	0.553
C	-1.262	-0.792	-1.236	-1.658	-0.788	-1.130
D	-1.141	-0.730	-1.209	-1.606	-0.792	-1.122
all	-0.286	-0.219	-0.652	-0.723	-0.489	-0.593
experts	0.022	0.019	-0.674	-0.666	-0.383	-0.379
rest	-0.732	-0.505	-0.627	-0.709	-0.551	-0.715

(**d**) knowledge impact K_ψ

Figure A2. Estimating the benefit of visualization and knowledge impact in relation to the survey result of Question 5 (Figure 6).

Appendix D. Survey Results of Useful Knowledge in Viewing London Underground Maps

This appendix is not an independent paper but supports Part II of this two-part paper by providing additional details about the empirical study on viewing London Underground maps described in Section 6.2 in this part of the paper.

It is necessary to note that this empirical study is not a hypothesis-based study. We can easily anticipate that some participants can use a very small or illusive amount of information shown on a map to answer questions that seem to require some information that is not on the map. However, this study is not designed to draw a conclusion about this phenomenon, but to collect some data about the phenomenon. As such a phenomenon suggests that there is a knowledge input to the visualization process, we would like to use the collected data to evaluate a few information-theoretic measures that have been proposed for quantifying the impact of such knowledge. Figure A3 shows the set up for this empirical study.

Figure A3. A survey for collecting data that reflects the use of some knowledge in viewing two types of London underground maps.

Figures A4–A6 show the questionnaire used in the survey about two types of London Underground maps. Table A3 summarizes the data from the answers by the 12 surveyees at King's College London, while Table A4 summarizes the data from the answers by the four surveyees at the University Oxford.

In Section 6.2, we discussed Questions 1–4 in some detail. In the survey, Questions 5–8 constitute the second set. Each question asks surveyees to first identify two stations along a given underground line, and then determine how many stops between the two stations. All surveyees identified the stations correctly for all four questions, and most also counted the stops correctly. In general, for each of these cases, one can establish an alphabet of all possible answers in a way similar to the example of walking distances. However, we did not observe any interesting correlation between the correctness and the surveyees' knowledge about metro systems or London.

With the third set of four questions, each question asks surveyees to identify the closest station for changing between two given stations on different lines. All surveyees identified the changing stations correctly for all questions.

The design of Questions 5–12 was also intended to collect data that might differentiate the deformed map from the faithful map in terms of the time required for answering questions. As shown in Figure A7, the questions were paired, such that the two questions feature the same level of difficulties.

Although the comparison seems to suggest that the faithful map might have some advantage in the setting of this survey, we cannot be certain about this observation as the sample size is not large enough. In general, we cannot draw any meaningful conclusion about the cost in terms of time. We hope to collect more real world data about the timing cost of visualization processes for making further advances in applying information theory to visualization.

The space cost is a valid consideration. While both maps have a similar size (i.e., deformed map: 850 mm × 580 mm, faithful map: 840 mm × 595 mm), their font sizes for station labels are very different. For long station names, "High Street Kensington" and "Totteridge & Whetstone", the labels on the deformed map are of 35 mm and 37 mm in length, while those on the faithful map are of 17 mm and 18 mm long. Taking the height into account, the space used for station labels in the deformed map is about four times of that in the faithful map. In other words, if the faithful map were to display its labels with the same font size, the cost of the space would be four times of that of the deformed map.

Table A3. The answers by twelve surveyees at King's College London to the questions in the London underground survey.

Questions		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	Mean
								Surveyee's ID						
Q1:	answer (min.)	8	30	12	16	20	15	10	30	20	20	20	30	19.25
	time (sec.)	06.22	07.66	09.78	11.66	03.72	04.85	08.85	21.12	12.72	11.22	03.38	10.06	09.27
Q2:	answer (min.)	15	30	5	22	15	14	20	20	25	25	25	20	19.67
	time (sec.)	10.25	09.78	06.44	09.29	12.12	06.09	17.28	06.75	12.31	06.85	06.03	10.56	09.48
Q3:	answer (min.)	20	45	10	70	20	20	20	35	25	30	20	240	46.25
	time (sec.)	19.43	13.37	10.06	09.25	14.06	10.84	12.46	19.03	11.50	16.09	11.28	28.41	14.65
Q4:	answer (min.)	60	60	35	100	30	20	45	35	45	120	40	120	59.17
	time (sec.)	11.31	10.62	10.56	12.47	08.21	07.15	18.72	08.91	08.06	12.62	03.88	24.19	11.39
Q5:	time 1 (sec.)	22.15	01.75	07.25	03.78	14.25	37.68	06.63	13.75	19.41	06.47	03.41	34.97	14.29
	time 2 (sec.)	24.22	08.28	17.94	05.60	17.94	57.99	21.76	20.50	27.16	13.24	22.66	40.88	23.18
	answer (10)	10	10	10	9	10	10	10	10	9	10	10	10	
	time (sec.)	06.13	28.81	08.35	06.22	09.06	06.35	09.93	12.69	10.47	05.54	08.66	27.75	11.66
Q6:	time 1 (sec.)	02.43	08.28	01.97	08.87	05.06	02.84	06.97	10.15	18.10	21.53	03.00	07.40	08.05
	time 2 (sec.)	12.99	27.69	04.81	10.31	15.97	04.65	17.56	16.31	20.25	24.69	15.34	20.68	15.94
	answer (9)	9	10	9	9	4	9	9	9	8	9	9	9	
	time (sec.)	07.50	06.53	04.44	16.53	19.41	05.06	13.47	07.03	12.44	04.78	07.91	16.34	10.12
Q7:	time 1 (sec.)	17.37	08.56	01.34	03.16	08.12	01.25	21.75	15.56	02.81	07.84	02.22	46.72	11.39
	time 2 (sec.)	17.38	13.15	02.34	03.70	08.81	02.25	22.75	26.00	17.97	10.37	03.18	47.75	14.64
	answer (7)	7	7	7	7	6	7	7	7	6	7	7	7	
	time (sec.)	07.53	06.34	03.47	03.87	02.75	04.09	02.16	04.94	26.88	05.31	06.63	12.84	07.23
Q8:	time 1 (sec.)	12.00	08.50	06.09	02.88	08.62	14.78	19.12	08.53	12.50	10.22	12.50	20.00	11.31
	time 2 (sec.)	13.44	10.78	23.37	09.29	13.03	36.34	23.55	09.50	13.53	10.23	32.44	22.60	18.18
	answer (6)	6	6	6	6	6	6	6	6	6	6	6	6	
	time (sec.)	02.62	05.94	02.15	04.09	04.94	07.06	07.50	04.90	04.37	04.53	05.47	09.43	05.25
Q9:	answer (P)	P	P	P	P	P	P	P	P	P	P	P	P	
	time (sec.)	35.78	02.87	07.40	13.03	06.97	52.15	13.56	02.16	08.13	09.06	01.93	08.44	13.46
Q10:	answer (LB)	LB	LB	LB	LB	LB	LB	LB	LB	LB	LB	LB	LB	
	time (sec.)	05.50	03.13	12.04	14.97	07.00	26.38	11.31	03.38	06.75	07.47	06.50	09.82	09.52
Q11:	answer (WP)	WP	WP	WP	WP	WP	WP	WP	WP	WP	WP	WP	WP	
	time (sec.)	06.07	05.35	07.72	05.00	04.32	23.72	05.25	03.07	10.66	05.37	02.94	17.37	08.07
Q12:	answer (FP)	FP	FP	FP	FP	FP	FP	FP	FP	FP	FP	FP	FP	
	time (sec.)	05.16	02.56	11.78	08.62	03.60	19.72	11.28	03.94	20.72	01.56	02.50	06.84	08.19
live in metro city		>5 yr	>5 yr	mths	1–5 yr	>5 yr	1–5 yr	weeks	>5 yr	1–5 yr	>5 yr	mths	mths	
live in London		>5 yr	>5 yr	mths	1–5 yr	1–5 yr	mths	mths	mths	mths	mths	mths	mths	

Survey Questions for the London Underground Study

Participant's Anonymised ID: _____   Time and Date: _____

Survey Coordinator's ID: _____   Survey Location: _____

Q1. Please use the **Conventional London Underground Map** to answer this question as accurately as possible.
Consider these two stations, **Charing Cross** and **Oxford Circus** (as indicated by blue arrows on the map). Estimate how long it would take (in minutes) for an ordinary healthy adult to walk from **Charing Cross** to **Oxford Circus**.

    How Long (mins)? _____   Response Time (mins & secs): _____

Q2. Please use the **Other London Underground Map** to answer this question as accurately as possible.
Consider these two stations, **Temple** and **Leicester Square** (as indicated by blue arrows on the map). Estimate how long it would take for an ordinary healthy adult to walk from **Temple** to **Leicester Square**.

    How Long (mins)? _____   Response Time (mins & secs): _____

Q3. Please use the **Conventional London Underground Map** to answer this question as accurately as possible.
Consider these two stations, **Stanmore** and **Edgware** (as indicated by red arrows on the map). Estimate how long it would take (in minutes) for an ordinary healthy adult to walk from **Stanmore** to **Edgware**.

    How Long (mins)? _____   Response Time (mins & secs): _____

Q4. Please use the **Other London Underground Map** to answer this question as accurately as possible.
Consider these two stations, **South Rulslip** and **South Harrow** (as indicated by red arrows on the map). Estimate how long it would take for an ordinary healthy adult to walk from **South Rulslip** to **South Harrow**.

    How Long (mins)? _____   Response Time (mins & secs): _____

Figure A4. London underground survey: question sheet 1 (out of 3).

Q5. Please use the **Conventional London Underground Map** to answer this question as quickly as possible.
(a) Where are station **Russell Square** and station **Barons Court** on the *Piccadilly* line (navy colour or dark blue)?
(b) How many stops between **Russell Square** and **Barons Court** (excluding the source and destination, i.e., **Russell Square** and **Barons Court**)?

(a) Response Time (mins & secs), first station: _____ total: _____

(b) How many stops? _____ Response Time (mins & secs): _____

Q6. Please use the **Other London Underground Map** to answer this question as quickly as possible.
(a) Where are station **Piccadilly Circus** and station **Queen's Park** on the *Bakerloo* line (brown colour)?
(b) How many stops between **Piccadilly Circus** and **Queen's Park** (excluding the source and destination, i.e., **Piccadilly Circus** and **Queen's Park**)?

(a) Response Time (mins & secs), first station: _____ total: _____

(b) How many stops? _____ Response Time (mins & secs): _____

Q7. Please use the **Conventional London Underground Map** to answer this question as quickly as possible.
(a) Where are station **Richmond** and station **West Kensington** on the *District* line (green colour)?
(b) How many stops between **Richmond** and **West Kensington** (excluding **Richmond** and **West Kensington**)?

(a) Response Time (mins & secs), first station: _____ total: _____

(b) How many stops? _____ Response Time (mins & secs): _____

Q8. Please use the **Other London Underground Map** to answer this question as quickly as possible.
(a) Where are station **Epping** and station **Snaresbrook** on the *Central* line (red colour)?
(b) How many stops between **Epping** and **Snaresbrook** (excluding **Epping** and **Snaresbrook**)?

(a) Response Time (mins & secs), first station: _____ total: _____

(b) How many stops? _____ Response Time (mins & secs): _____

Figure A5. London underground survey: question sheet 2 (out of 3).

Q9. Please use the **Conventional London Underground Map** to answer this question as quickly as possible.
Consider these two stations, <u>**Ladbroke Grove**</u> on the *Circle* line and <u>**North Wembley**</u> on the *Bakerloo* line (as indicated by yellow arrows on the map). Find the closest station for changing between <u>**Ladbroke Grove**</u> and <u>**North Wembley**</u>.

    Where to change? _____   Response Time (mins & secs): _____

Q10. Please use the **Other London Underground Map** to answer this question as quickly as possible.
Consider these two stations, <u>**Old Street**</u> on the *Northern* line and <u>**Canada Water**</u> on the *Jubilee* line (as indicated by yellow arrows on the map). Find the closest station for changing between <u>**Old Street**</u> and <u>**Canada Water**</u>.

    Where to change? _____   Response Time (mins & secs): _____

Q11. Please use the **Conventional London Underground Map** to answer this question as quickly as possible.
Consider these two stations, <u>**Watford**</u> on the *Metropolitan* line and <u>**Stanmore**</u> on the *Jubilee* line (as indicated by green arrows on the map). Find the closest station for changing between <u>**Watford**</u> and <u>**Stanmore**</u>.

    Where to change? _____   Response Time (mins & secs): _____

Q12. Please use the **Other London Underground Map** to answer this question as quickly as possible.
Consider these two stations, <u>**Cockfosters**</u> on the *Piccadilly* line and <u>**Walthamstow Central**</u> on the *Victoria* line (as indicated by green arrows on the map). Find the closest station for changing between <u>**Cockfosters**</u> and <u>**Walthamstow Central**</u> .

    Where to change? _____   Response Time (mins & secs): _____

How long have you lived or stayed in a city with a metro/underground system?

    ○ a. Never   ○ b. a few days   ○ c. a few weeks
    ○ d. a few months   ○ e. between 1 and 5 years   ○ f. more than 5 years

How long have you lived or stayed in London?

    ○ a. Never   ○ b. a few days   ○ c. a few weeks
    ○ d. a few months   ○ e. between 1 and 5 years   ○ f. more than 5 years

Figure A6. London underground survey: question sheet 3 (out of 3).

Table A4. The answers by four surveyees at the University of Oxford to the questions in the London underground survey.

Questions		P13	P14	P15	P16	Mean
				Surveyee's ID		
Q1:	answer (min.)	15	20	15	15	16.25
	time (sec.)	11.81	18.52	08.18	07.63	11.52
Q2:	answer (min.)	5	5	15	15	10.00
	time (sec.)	11.10	02.46	13.77	10.94	09.57
Q3:	answer (min.)	35	60	30	25	37.50
	time (sec.)	21.91	16.11	10.08	22.53	17.66
Q4:	answer (min.)	20	30	60	25	33.75
	time (sec.)	13.28	16.21	08.71	18.87	14.27
Q5:	time 1 (sec.)	17.72	07.35	17.22	09.25	12.89
	time 2 (sec.)	21.06	17.00	19.04	12.37	17.37
	answer (10)	10	8	10	10	
	time (sec.)	04.82	02.45	02.96	15.57	06.45
Q6:	time 1 (sec.)	35.04	38.12	11.29	07.55	23.00
	time 2 (sec.)	45.60	41.32	20.23	40.12	36.82
	answer (9)	9	10	9	8	
	time (sec.)	03.82	13.57	08.15	34.32	14.97
Q7:	time 1 (sec.)	01.05	02.39	09.55	11.19	06.05
	time 2 (sec.)	02.15	05.45	09.58	13.47	07.66
	answer (7)	10	6	7	7	
	time (sec.)	01.06	01.60	02.51	14.06	04.81
Q8:	time 1 (sec.)	08.74	26.14	20.37	15.01	17.57
	time 2 (sec.)	16.50	30.55	27.01	17.91	22.99
	answer (6)	6	6	6	6	
	time (sec.)	09.30	03.00	02.11	04.94	04.48
Q9:	answer (P)	P	P	P	P	
	time (sec.)	05.96	09.38	04.56	05.16	06.27
Q10:	answer (LB)	LB	LB	LB	LB	
	time (sec.)	12.74	07.77	01.30	09.94	07.94
Q11:	answer (WP)	WP	WP	WP	WP	
	time (sec.)	09.84	04.43	03.39	07.18	06.21
Q12:	answer (FP)	FP	FP	FP	FP	
	time (sec.)	06.22	10.46	06.78	05.10	07.14
live in metro city		never	days	days	days	
live in London		never	days	days	days	
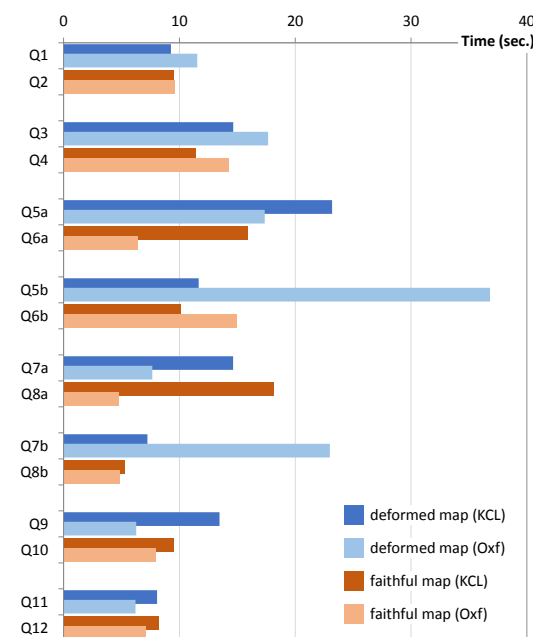
Figure A7. The average time used by surveyees for answering each of the 12 questions. The data does not indicate any significant advantage of using the geographically-deformed map.

Divergenc	1	2	3	4	5	6
	DJS	DJSsqr	Dncm1	Dncm2	Dnew1	Dnew2
spot on	0.7249	0.8514	0.9107	0.8273	0.5565	0.4684
close	0.9126	0.9553	0.9811	0.9625	0.5591	0.5001
wild gues	0.9997	0.9998	1.0000	0.9999	0.5563	0.5064
Benefit	1	2	3	4	5	6
	DJS	DJSsqr	Dncm1	Dncm2	Dnew1	Dnew2
spot on	-1.765	-2.777	-3.252	-2.585	-0.418	0.287
close	-3.266	-3.608	-3.815	-3.666	-0.439	0.033
wild gues	-3.963	-3.965	-3.966	-3.965	-0.416	-0.017

Figure A8. The original table of numerical values for the text in the main paper.
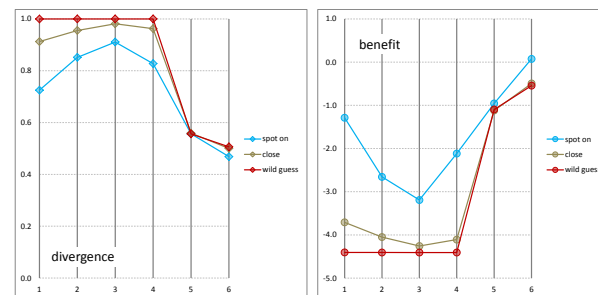

Figure A9. The PCPs of the data in Figure A8.

References

1. Klein, H.A. *The Science of Measurement: A Historical Survey*; Dover Publications: Mineola, NY, USA, 2012.
2. Stevens, S.S. On the Theory of Scales of Measurement. *Science* **1946**, *103*, 677–680. [CrossRef] [PubMed]
3. Chen, M.; Golan, A. What May Visualization Processes Optimize? *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2619–2632. [CrossRef] [PubMed]
4. Chen, M.; Sbert, M. A Bounded Measure for Estimating the Benefit of Visualization (Part I): Theoretical Discourse and Conceptual Evaluation. *Entropy* **2022**, *24*, 228. [CrossRef]
5. Chen, M. A Short Introduction to Information-Theoretic Cost-Benefit Analysis. *arXiv* **2021**, arXiv:2103.15113.
6. Boslaugh, S.; Watters, P.A. *Statistics in a Nutshell: A Desktop Quick Reference Paperback*; O'Reilly: Akron, OH, USA, 2008.
7. Pedhazur, E.J.; Schmelkin, L.P. *Measurement, Design, and Analysis: An Integrated Approach*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1991.
8. Tal, E. Measurement in Science. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Palo Alto, CA, USA, 2020; ISSN 1095-5054.
9. Schlaudt, O. Measurement. *Online Encycl. Philos. Nat.* **2020**, *5*, 11. [CrossRef]
10. Newton, I. . Scala graduum caloris. *Philos. Trans.* **1701**, *22*, 824–829.
11. Grigull, U. Newton's temperature scale and the law of cooling. *Wärme Stoffübertragung* **1984**, *18*, 195–199. [CrossRef]
12. Wikipedia. Newton Scale. Available online: https://en.wikipedia.org/wiki/Newton_scale (accessed on 20 November 2021).
13. Behrisch, M.; Blumenschein, M.; Kim, N.W.; Shao, L.; El-Assady, M.; Fuchs, J.; Seebacher, D.; Diehl, A.; Brandes, U.; Pfister, H.; et al. Quality Metrics for Information Visualization. *Comput. Graph. Forum* **2018**, *37*, 625–662. [CrossRef]
14. Bertini, E.; Tatu, A.; Keim, D. Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2203–2212. [CrossRef]
15. Tufte, E.R. *The Visual Display of Quantitative Information*; Graphics Press: Cheshire, CT, USA, 1986.
16. Cui, Q.; Yang, J.; Ward, M.; Rundensteiner, E. Measuring Data Abstraction Quality in Multiresolution Visualizations. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 709–716. [CrossRef]
17. Johansson, J.; Cooper, M. A Screen Space Quality Method for Data Abstraction. *Comput. Graph. Forum* **2008**, *27*, 1039–1046. [CrossRef]
18. Friedman, J.H.; Tukey, J.W. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. Comput.* **1974**, *23*, 881–890. [CrossRef]
19. Tukey, J.W.; Tukey, P. Computer Graphics and Exploratory Data Analysis: An Introduction. In Proceedings of the 6th Annual Conference and Exposition: Computer Graphics, Dallas, TX, USA, 14–18 April 1985; pp. 773–785.
20. Bertini, E.; Santucci, G. Quality Metrics for 2D Scatterplot Graphics: Automatically Reducing Visual Clutter. In *International Symposium on Smart Graphics*; Butz, A., Krüger, A., Olivier, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3031, pp. 77–89.

21. Wilkinson, L.; Anand, A.; Grossman, R. Graph-theoretic Scagnostics. In *Information Visualization, IEEE Symposium*; IEEE Computer Society: Washington, DC, USA, 2005; pp. 157–164.

22. Sips, M.; Neubert, B.; Lewis, J.P.; Hanrahan, P. Selecting Good Views of High-dimensional Data using Class Consistency. *Comput. Graph. Forum* **2009**, *28*, 831–838. [CrossRef]

23. Tatu, A.; Albuquerque, G.; Eisemann, M.; Schneidewind, J.; Theisel, H.; Magnor, M.; Keim, D. Combining Automated Analysis and Visualization Techniques for Effective Exploration of High-dimensional Data. In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, Atlantic City, NJ, USA, 2–13 October 2009; pp. 59–66.

24. Tatu, A.; Bak, P.; Bertini, E.; Keim, D.; Schneidewind, J. Visual Quality Metrics and Human Perception: An Initial Study on 2D Projections of Large Multidimensional Data. In Proceedings of the International Conference on Advanced Visual Interfaces, Bari, Italy, 7–10 June 2010; pp. 49–56.

25. Dasgupta, A.; Kosara, R. Pargnostics: Screen-Space Metrics for Parallel Coordinates. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1017–1026. [CrossRef]

26. Peng, W.; Ward, M.O.; Rundensteiner, E. Clutter Reduction in Multi-dimensional Data Visualization using Dimension Reordering. In Proceedings of the IEEE Symposium on Information Visualization, Austin, TX, USA, 10 October 2004; pp. 89–96.

27. Rosenholtz, R.; Li, Y.; Nakano, L. Measuring Visual Clutter. *J. Vis.* **2007**, *7*, 1–22. [CrossRef] [PubMed]

28. Yu, C.P.; Samaras, D.; Zelinsky, G.J. Modelling visual clutter perception using proto-object segmentation. *J. Vis.* **2014**, *14*, 4. [CrossRef]

29. Filonik, D.; Baur, D. Measuring Aesthetics for Information Visualization. In Proceedings of the IEEE Information Visualization, Atlantic City, NJ, USA, 11–16 October 2009; pp. 579–584.

30. Jänicke, H.; Chen, M. A Salience-based Quality Metric for Visualization. *Comput. Graph. Forum* **2010**, *29*, 1183–1192. [CrossRef]

31. Bernard, J.; Steiger, M.; Mittelstädt, S.; Thum, S.; Keim, D.A.; Kohlhammer, J. A survey and task-based quality assessment of static 2D color maps. In *Visualization and Data Analysis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2015.

32. Mittelstädt, S.; Jäckle, D.; Stoffel, F.; Keim, D.A. ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks. In Proceedings of the Conference on Visualization (EuroVis), Sardinia, Italy, 25–29 May 2015; pp. 115–119.

33. Mittelstädt, S.; Keim, D.A. Efficient Contrast Effect Compensation with Personalized Perception Models. *Comput. Graph. Forum* **2015**, *34*, 211–220. [CrossRef]

34. Gramazio, C.C.; Laidlaw, D.H.; Schloss, K.B. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 521–530. [CrossRef]

35. Cleveland, W.S.; McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc.* **1984**, *79*, 531–554. [CrossRef]

36. Isenberg, T. Evaluating and Validating Non-photorealistic and Illustrative Rendering. In *Image and Video-Based Artistic Stylisation*; Rosin, P., Collomosse, J., Eds.; Springer: London, UK, 2013; pp. 311–331.

37. Isenberg, T.; Neumann, P.; Carpendale, S.; Sousa, M.C.; Jorge, J.A. Non-Photorealistic Rendering in Context: An Observational Study. In Proceedings of the 4th International Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 5–7 June 2006; pp. 115–126.

38. Cole, F.; Sanik, K.; DeCarlo, D.; Finkelstein, A.; Funkhouser, T.; Rusinkiewicz, S.; Singh, M. How Well Do Line Drawings Depict Shape? In Proceedings of the ACM SIGGRAPH Conference, New Orleans, LA, USA, 3–7 August 2009; p. 28.

39. Mandryk, R.L.; Mould, D.; Li, H. Evaluation of Emotional Response to Non-Photorealistic Images. In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering, Vancouver, BC, Canada, 5–7 August 2011; pp. 7–16.

40. Liu, Z.; Li, Z. Impact of schematic designs on the cognition of underground tube maps. *ISPRS—Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *49*, 421–423. [CrossRef]

41. Hong, S.; Yoo, M.J.; Chinh, B.; Han, A.; Battersby, S.; Kim, J. To Distort or Not to Distort: Distance Cartograms in the Wild. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–12.

42. 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] [PubMed]

43. Kijmongkolchai, N.; Abdul-Rahman, A.; Chen, M. Empirically measuring soft knowledge in visualization. *Comput. Graph. Forum* **2017**, *36*, 73–85. [CrossRef]

44. Wall, E.; Agnihotri, M.; Matzen, L.; Divis, K.; Haass, M.; Endert, A.; Stasko, J. A heuristic approach to value-driven evaluation of visualizations. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 491–500. [CrossRef] [PubMed]

45. Saket, B.; Srinivasan, A.; Ragan, E.D.; Endert, A. Evaluating interactive graphical encodings for data visualization. *IEEE Trans. Vis. Comput. Graph.* **2017**, *24*, 1316–1330. [CrossRef]

46. Sacha, D.; Stoffel, A.; Stoffel, F.; Kwon, B.C.; Ellis, G.; Keim, D.A. Knowledge generation model for visual analytics. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1604–1613. [CrossRef]

47. 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]

48. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

49. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.

50. 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]

51. Ishizaka, A.; Nemery, P. *Multi-Criteria Decision Analysis: Methods and Software*; John Wiley & Sons: Hoboken, NJ, USA, 2013.

52. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]

53. Roettger, S. The Volume Library. Available online: http://schorsch.efi.fh-nuernberg.de/data/volume/ (accessed on 20 November 2021).

54. Nagy, Z.; Schneide, J.; Westerman, R. Interactive Volume Illustration. In Proceedings of the Vision, Modeling and Visualization, Erlangen, Germany, 20–22 November 2002.

55. Correa, C.; Silver, D.; Chen, M. Feature aligned volume manipulation for illustration and visualization. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 1069–1076. [CrossRef]

56. Wu, Y.; Qu, H. Interactive Transfer Function Design Based on Editing Direct Volume Rendered Images. *IEEE Trans. Vis. Comput. Graph.* **2007**, *13*, 1027–1040. [CrossRef] [PubMed]

57. Jung, Y. Instantreality 1.0. Available online: https://doc.instantreality.org/tutorial/volume-rendering/ (accessed on 20 November 2021).

58. Max, N.; Chen, M. Local and global illumination in the volume rendering integral. In *Scientific Visualization: Advanced Concepts*; Hagen, H., Ed.; Schloss Dagstuhl: Wadern, Germany, 2010.

59. Chen, M. Cost-Benefit Analysis of Data Intelligence—Its Broader Interpretations. In *Advances in Info-Metrics: Information and Information Processing across Disciplines*; Chen, M., Dunn, J.M., Golan, A., Ullah, A., Eds.; Oxford University Press: Oxford, UK, 2020.

60. 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] [PubMed]

61. Chen, M.; Edwards, D.J. 'Isms' in Visualization. In *Foundations of Data Visualization*; Chen, M., Hauser, H., Rheingans, P., Scheuermann, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2020.

62. Dasgupta, A.; Chen, M.; Kosara, R. Conceptualizing visual uncertainty in parallel coordinates. *Comput. Graph. Forum* **2012**, *31*, 1015–1024. [CrossRef]

63. Viola, I.; Chen, M.; Isenberg, T. Visual Abstraction. In *Foundations of Data Visualization*; Chen, M., Hauser, H., Rheingans, P., Scheuermann, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2020.

64. Kindlmann, G.; Scheidegger, C. An Algebraic Process for Visualization Design. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2181–2190. [CrossRef] [PubMed]

65. Chen, M. The value of interaction in data intelligence. *arXiv* **2018**, arXiv:1812.06051.

66. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]

67. Chen, M.; Trefethen, A.; Banares-Alcantara, R.; Jirotka, M.; Coecke, B.; Ertl, T.; Schmidt, A. From data analysis and visualization to causality discovery. *IEEE Comput.* **2011**, *44*, 84–87. [CrossRef]