What May Visualization Processes Optimize?

Min Chen, *Member, IEEE* and Amos Golan

Abstract—In this paper, we present an abstract model of visualization and inference processes, and describe an information-theoretic measure for optimizing such processes. In order to obtain such an abstraction, we first examined six classes of workflows in data analysis and visualization, and identified four levels of typical visualization components, namely disseminative, observational, analytical and model-developmental visualization. We noticed a common phenomenon at different levels of visualization, that is, the transformation of data spaces (referred to as alphabets) usually corresponds to the reduction of maximal entropy along a workflow. Based on this observation, we establish an information-theoretic measure of cost-benefit ratio that may be used as a cost function for optimizing a data visualization process. To demonstrate the validity of this measure, we examined a number of successful visualization processes in the literature, and showed that the information-theoretic measure can mathematically explain the advantages of such processes over possible alternatives.

Index Terms—Visualization, visual analytics, theory of visualization, model of visualization, information theory, entropy, cost-benefit ratio, workflow, pipeline, process optimization.

1 Introduction

Over the past 25 years, the field of *visualization* has developed to encompass three major subfields, namely *scientific visualization*, *information visualization* and *visual analytics* as well as many domain-specific areas, such as geo-information visualization, biological data visualization, software visualization, and others. A number of pipelines have been proposed for visualization in general (e.g., [43], [58], [59]) and for visual analytics in particular [28], [39]. In practice, a visualization workflow normally includes machine-centric components (e.g., statistical analysis, rule-based or policy-based models, and supervised or unsupervised models) as well as human-centric components (e.g., visualization, human-computer interaction, and human-human communication). The integration of these two types of components become more and more common since *visual analytics* [56], [70] has become a de facto standard approach for handling large volumes of complex data.

Given a visualization workflow in a specific context, it is inevitable that one would like to improve its cost-benefit ratio, from time to time, in relation to many factors such as accuracy, speed, computational and human resources, credibility, logistics, changes in the environment, data or tasks concerned, and so forth. Such improvement can typically be made through introducing new technologies, restructuring the existing workflow, or re-balancing the tradeoff between different factors. While it is absolutely essential to optimize each visualization workflow in a heuristic and case-by-case manner [44], it is also desirable to study the process optimization theoretically and mathematically through abstract reasoning. In many ways, this is similar to the process optimization in tele- and data communication, where each subsystem is optimized through careful design and customization but the gain in cost-benefit is mostly underpinned by information theory [18], [51]. In this paper, we study, in abstraction, the process optimization in visualization from an information-theoretic perspective.

Visualization is a form of information processing. Like other forms of information processing (e.g., statistical inferences), visualization enables transformation of information from one representation to another. The objective of such a transformation is typically to infer a finding, judgment or decision from the observed *data*, which may be incomplete and noisy. The input to the transformation may also include "soft" *information* and *knowledge*, such as known theories, intuition, belief, value judgment, and so on. Another form of input, which is often referred to as *priors*, may come from knowledge about the system where the data are captured, facts about the system or related systems, previous observations, experimentations, analytical conclusions, etc. Here we use the terms *data*, *information* and *knowledge* according to the commonly-used definitions in computational spaces [13].

All inferential processes are designed for processing a finite amount of information. In practice, they all encounter some difficulties, such as the lack of adequate technique for extracting meaningful information from a vast amount of data; incomplete, incorrect or noisy data; underdetermined, unreliable, or outdated computational models; biases encoded in computer algorithms or biases of human analysts; lack of computational resources or human resources; urgency in making a decision; and so on. All inferential problems are inherently under-determined problems [24], [25].

The traditional machine-centric solutions to the inferential problem address these difficulties by imposing certain assumptions and structures on the model of the system, reflecting the availability of data. If these assumptions were correctly specified and these structures were perfectly

- *M. Chen is with University of Oxford. Email: min.chen@oerc.ox.ac.uk.*
- *A. Golan is with American University and the Santa Fe Institute. Email: agolan@american.edu.*

observed, computed inference based on certain statistics (e.g., moments) would provide us with perfect answers. In practice, it is seldom possible to transform our theory, axioms, intuition and other soft information into such statistics. Hence optimization of a visualization process is not just about the best statistical method, the best analytical algorithm, or the best machine learning technique. It is also about the best human-centric mechanisms for enabling uses of "soft" information and knowledge.

In this paper, we propose to measure the cost-benefit of a visualization-assisted inference process within an information-theoretic framework. The work is built on a wealth of literature on visualization and visualization pipelines (e.g., [28], [39], [43], [58], [59]) and that on information theoretic measures and inference in statistics and econometrics [26], [27], [33]. It is a major extension of the information-theoretic framework for visualization proposed by Chen and Jänicke [15], and a major extension of statistical inference and information processing in general (e.g., [24]). Our contributions are:

- We propose a new categorization of visualization workflows and identify four levels of visualization commonly featured in different data analysis and visualization processes (Section 3).
- We present an information-theoretic abstraction of visualization processes as transformation of alphabets along a workflow for data analysis and visualization, and identify a common trend of reduction of Shannon entropy (i.e., uncertainty) in such workflows (Section 4).
- We propose an information-theoretic measure of cost-benefit, which can be applied to the whole workflow as well as individual processing steps (Section 4).
- We demonstrate that this cost-benefit measure can explain the information-theoretic advantages of successful visualization workflows in the literature, suggesting that it can be used for optimizing a visualization-assisted inference process through a combination of quantitative and qualitative analysis (Section 5).

2 Related Work

In 2003, Grinstein *et al.* [30] posed an intriguing question about usability vs. utility when they considered visualization as an interface technology that draws from both machine- and human-centric capabilities. This is a question about optimization.

Pipelines and Workflows. In the field of visualization, many have considered pipelines or workflows that feature components such as analysis, visualization and interaction. Upson *et al.* provided one of the earliest abstraction of a pipeline with four main components, data source, filtering and mapping, rendering and output [58]. Wood et al. proposed an extension for collaborative visualization in the form of parallel pipelines [71]. van Wijk outlined a two-loop pipeline, bringing interaction and cognition into a visualization process [59]. Green *et al.* proposed a revision of this pipeline [28]. Keim *et al.* proposed a pipeline featuring two interacting parallel components for data mining models and visual data exploration respectively [39]. Jänicke *et al.* examined several pipelines for comparative visualization, and discussed quality metrics for evaluating reconstructibility of visualization [35]. Bertini *et al.* proposed an automated visualization pipeline driven by quality metrics [4]. Recently Moreland surveyed visualization pipelines mainly in the context of scientific visualization [43]. There are many other variations of visualization pipelines in the literature, such as [12], [14], [17], [31], [37]. All these discussions on visualization pipelines pointed out one common fact, i.e., visualization processes can be broken down to steps, which may be referred to as transformations or mappings. This work considers this ubiquitous feature of visualization in abstraction.

Design Methods and Processes. Abram and Treinish proposed to implement visualization processes on data-flow architectures [1]. Chi described visualization processes using a state reference model, involving data, visualization, and visual mapping transformation [17]. Jansen and Dragicevic proposed an interaction model in the context of visualization pipelines [38]. Munzner proposed a nested model for designing and developing visualization pipelines [44]. Wang *et al.* proposed a two-stage framework for designing visual analytics systems [67]. Ahmed *et al.* proposed to use purpose-driven games for evaluating visualization systems [2]. Scholtz outlined a set of guidelines for assessing visual analytics environments [49], and Scholtz *et al.* further developed them into an evaluation methodology [50]. The theoretic abstraction presented in this paper is built on these works, and complements them by offering a mathematical rationalization for good practices in designing and assessing visualization systems.

Theories of Visualization and their Applications. In developing theories of visualization, much effort has been made in formulating categorizations and taxonomies (e.g., [3], [57], [68]). Some 25 different proposals are listed in [15], [16]. In addition, a number of conceptual models have been proposed, including object-oriented model by Silver [52], feature extraction and representation by van Walsum *et al.* [63], visualization exploration by Jankun-Kelly *et al.* [37], distributed cognition model by Liu *et al.* [42], predictive data-centered theory by Purchase *et al.* [46], Visualization Transform Design Model by Purchase *et al.* [46], cognition model for visual analytics by Green *et al.* [29], sensemaking and model steering by Endert *et al.* [21], modelling visualization using semiotics and category theory by Vickers *et al.* [61], composition of visualization tasks by Brehmer and Munzner [10], and visual embedding by Demiralp *et al.* [20]. Recently, Sacha *et al.* proposed a knowledge generation model [48], introducing a visual analytics model with exploration and verification loops. The deliberations in these works represent qualitative abstraction of visualization processes.

Meanwhile, the development of mathematical frameworks is gathering its pace in recent years. One of these is the information theoretic framework, which was initially

suggested by Ward [46], then generalized and detailed by Chen and Jänicke [15], and further enriched by Xu *et al.* [72] and Wang and Shen [65] in the context of scientific visualization. Another is the algebraic framework proposed by Kindlmann and Scheidegger [40], who justifiably placed their focus on visual mappings, which are inherently the most important transformations from a visualization perspective. While an algebraic formulation typically describes mappings between set members (e.g., from a pair of datasets to a pair of visual representations in [40]), an information-theoretic formulation describes mappings between sets together with the probabilistic distributions of their members.

This holistic nature of information-theoretic reasoning has enabled many applications in visualization, including light source placement by Gumhold [32], view selection in mesh rendering by Vázquez *et al.* [60] and Feixas *et al.* [22], view selection in volume rendering by Bordoloi and Shen [6], and Takahashi and Takeshima [53], focus of attention in volume rendering by Viola *et al.* [62], multiresolution volume visualization by Wang and Shen [64], feature highlighting in unsteady multi-field visualization by Jänicke and Scheuermann [34], [36], feature highlighting in time-varying volume visualization by Wang *et al.* [66], transfer function design by Bruckner and Möller [11], and by Ruiz *et al.* [9], [47], multimodal data fusion by Bramon *et al.* [7], evaluating isosurfaces [69], measuring of observation capacity [8], measuring information content in multivariate data [5], and confirming the mathematical feasibility of visual multiplexing [16].

3 Workflows in Visualization

Consider a broad range of workflows in visualization, including those historically referred to as analysis, inference, simulation or visual analytics as well as those emerged recently, as long as they feature a *component of visualization*, i.e., mapping some data to alternative visual representations. There are some elementary workflows, each of which features a primary visualization task.

Complex workflows can be decomposed into these elementary workflows. In this section, we first provide a categorization of visualization tasks based on task complexity. We then examine six elementary workflows, and an example of a complex workflow. The discussions on workflows motivate further abstraction in Section 4, while the task categorization helps select case studies in Section 5 that substantiate the theoretical abstraction in Section 4.

3.1 Four Levels of Visualization Tasks

Visualization tasks can be categorized into the following four levels, reflecting the complexity of search space from the perspective of analysts.

- *Level 1: Disseminative Visualization* ($\mathbf{V_D}$) — Visualization is a presentational aid for disseminating information or insight to others. The analyst who created the visualization does not have a question about the data, except for informing others: "This is *A*!" Here *A* may be a fact, a piece of information, an understanding, etc. At this level, the complexity for the analyst to obtain an answer about the data is O(1). Here we make use the big O notation in algorithm and complexity analysis. However, instead of measuring the complexity of computation or storage costs, we focus on the search space for answers in performing a visualization task.
- *Level 2: Observational Visualization* ($\mathbf{V_O}$) — Visualization is an operational aid that enables intuitive and/or speedy observation of captured data. It is often a part of routine operations of an analyst, and the questions to be answered may typically be in the forms of "What has happened?" "When and where *A*, *B*, *C*, etc., happened?' At this level, the observation is usually sequential, and thus the complexity is generally O(n), where n is the number of data objects. Broadly speaking, a *data object* is a data record. We will give a more precise definition of it in Section 4.
- *Level 3: Analytical Visualization* ($\mathbf{V_A}$) — Visualization is an investigative aid for examining and understanding complex relationships (e.g., correlation, association, causality, contradiction). The questions to be answered are typically in the forms of "What does *A* relate to?" and "Why?" Given n data objects, the number of possible k-relationships among these data objects is at the level of O(n^k) ($k \geq 2$). For a small n, it may be feasible to examine all k-relationships using observational visualization. When n increases, it becomes necessary to use analytical models to prioritize the analyst's investigative effort. Most visual analytics processes reported in the recent literature operate at this level.
- *Level 4: Model-developmental Visualization* ($\mathbf{V_M}$) — Visualization is a developmental aid for improving existing models, methods, algorithms and systems, as well as for creating new ones. The questions to be answered are typically in the forms of "How does *A* lead to *B*?" and "What are the exact steps from *A* to *B*?" If a model has n parameters and each parameter may take k values, there are a total of k^n combinations. In terms of complexity, this is O(k^n). If a model has n distinct algorithmic steps, the complexity of their ordering is O($n!$). Model-developmental visualization is a great challenge in the field of visualization. Note that we have avoided the phrase "modelling visualization" here as it could be misread as an action "to model visualization". One day, there might be a new adjective, e.g., in the form of *modelative* or *modelary*.

Hence the levels correspond to the questions to be asked and the complexity of the space of optional answers. For example, given a financial prediction model, if an analyst uses visualization to demonstrate its effectiveness to an audience, this is level 1 visualization, as the analyst knows or assumes the model to be correct.

If the analyst sequentially observes a financial data stream and some basic statistics about the data in order

to capture some events, it is level 2 visualization.

If the analyst applies a prediction model to the input data streams, and then uses visualization to observe the input data and its basic statistics, to receive the predictions and recommendations computed by a machine process, and to reason about potential errors, it is more complex than observing events in a data stream sequentially. This is because the analysis of errors and noise typically involves examination of the relationships among different events in the input data streams, statistical indicators, computed trends and recommendations. The search space includes all possible relationships. This is level 3 visualization.

If the analyst identifies that a prediction model does not perform satisfactorily, the analyst may attempt to improve it by, for example, experimenting with various parameters in the model, or may wish to create a new prediction model based on a different economic theory. When visualization is used to assist the analyst in exploring the parameter space or the model space, this is level 4 visualization.

3.2 Six Classes of Workflows

As a process of abstraction, we consider six elementary classes of workflows as illustrated in Fig. 1. They are intended to contrast machine-centric and human-centric processes, as well as integrated visual analytics processes, while illustrating the roles of the four levels of visualization tasks. These workflows feature the following types of components:

- *Machine Processing* (M) — These are computational processes executed by computers including, for instance, computation of statistical indicators (e.g., mean, correlation index, etc.), data analysis (e.g., classification, anomaly detection, association analysis, etc.), simulation, prediction, recommendation and so on. Each computational process is defined by a program that may encode a theoretic or heuristic model, which we refer to generally as a *Model*.
- *Human Processing* (H) — These are human cognitive processes and related activities including, for instance, viewing, reasoning, memorizing, discussing, decision making and so on.
- *Visual Mapping* (V) — These are processes where data are transformed to alternative visual representations to be viewed by humans. We purposely treat these processes separately from M and H, and assume that visual representations can be generated by many means from hand-drawn plots and illustrations to automated generation of visualization.
- *Interaction* (I) — These are actions taken by humans to influence an M or V process. They include typical interactions in visualization [73], such as parameter adjustment, and model creation and refinement. In Fig. 1, they are not explicitly shown as a processing block, as the main cognitive processing for interaction is assumed to take place in H. Instead, they are indicated by a solid bar on a connection.

Workflow class W_1 encompasses perhaps some of the most common process in data analysis and visualization. In

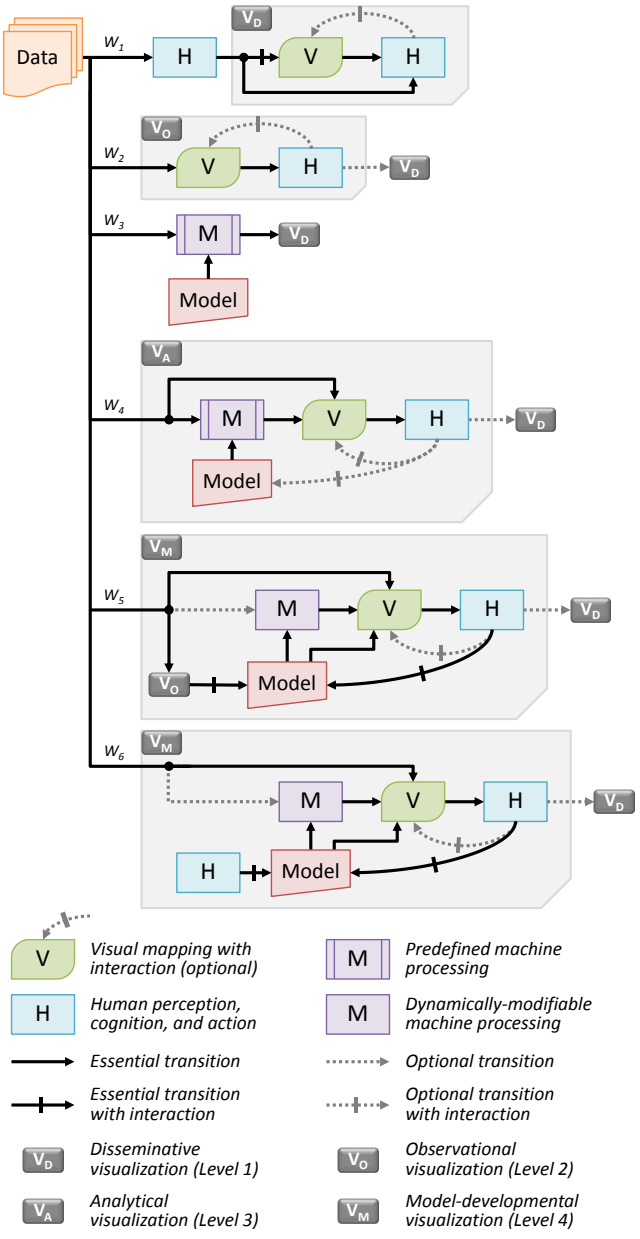


Fig. 1. Six typical workflows in data analysis and visualization. The subgraphs, $\mathbf{V_D}$, $\mathbf{V_O}$, $\mathbf{V_A}$, and $\mathbf{V_M}$ represent four levels of visualization.

this process, one or more human analysts (H) process the input data with or without the aid of computation, gain some understanding, create some visualization (V) and convey the understanding to others (H). Many visualization images in the media and visual illustration in user manuals fall into this class. The goal of visualization is to pass on known information and knowledge to others, and the dissemination process is almost always accompanied by written or verbal commentaries describing the understanding and/or opinions of analysts. The central task of such a workflow is *Disseminative Visualization*, and it is represented as a macro block $\mathbf{V_D}$ in Fig. 1.

The second class, W_2, encompasses many operational processes, where human analysts need to use visualization to observe data routinely. For examples, stock brokers

frequently glance at various time series plots, drivers glance at their GPS-navigation devices regularly, neurologists examine visual representations of various scans (e.g., electroencephalography, computed tomography, diffusion tensor imaging, etc.) of their patients, and computational scientists visualize simulation results after each run. The goal of visualization is to enable intuitive and speedy observation of features, phenomena and events captured in the data, and to provide external memorization of what have been observed. The central task of such a workflow is *Observational Visualization*, which is represented as a macro block $\mathbf{V_O}$. Although the two macro blocks $\mathbf{V_O}$ and $\mathbf{V_D}$ appear to be similar except an extra forward transition in $\mathbf{V_D}$, their fundamental difference is that in $\mathbf{V_D}$ analysts have already gained the understanding to be conveyed before the visualization is generated, while in $\mathbf{V_O}$ visualization is generated (or dynamically updated) in order for analysts to gain a new understanding. Of course, $\mathbf{V_O}$ can be followed by $\mathbf{V_D}$ to disseminate such a new understanding, though the inclusion of $\mathbf{V_D}$ is optional.

Workflow W_3 depicts a class of processes where automated data analysis plays a dominant role, and humans are only the destination of dissemination. In many ways, W_3 is almost identical to W_1, except that in W_3 the understanding and/or opinions conveyed to humans through $\mathbf{V_D}$ are from machine processing. Such a workflow has its place in data analysis and visualization, when the machine is always or almost always correct about what is being conveyed. When such a high level of correctness is not assured, it is necessary to increase humans' involvement.

This leads to workflow class W_4, where human analysts are able to observe input data in conjunction with the machine's "understanding". In many ways, this workflow is similar to the parallel pipeline proposed by Keim *et al.* [39]. It allows analysts to receive computational results from machine processing, while evaluating the correctness of the results and identify possible false positives and negatives. For example, in much investigative analysis for examining and understanding complex relationships among data objects, the amount of input data often makes direct observation time-consuming. The machine-processing hence enables the analysts to prioritize their effort and structure their reasoning and decision-making process. At the same time, analysts are able to explore the data and adjust the model depending on the analysts' judgment about the quality of the computed results. The central task of such a workflow is thus *Analytical Visualization*, which is represented as a macro block $\mathbf{V_A}$.

When the correctness or accuracy of a model is the main concern, the focus of visualization is shifted to assisting analysts in improving an existing model or creating a new model. Both workflow classes W_5 and W_6 represent such a focus. In W_5, analysts first observe some input data and identify an existing model or formulate a new one for processing the data. Tasks for such processing may include, but not limited to, computing statistical indicators; detecting features, objects, and events; identifying patterns, associations, and rules; and making predictions and recommendations. In many cases, W_5 may represent a long-term process for developing a theory and its applications, such as physical laws and their applications in computer simulation.

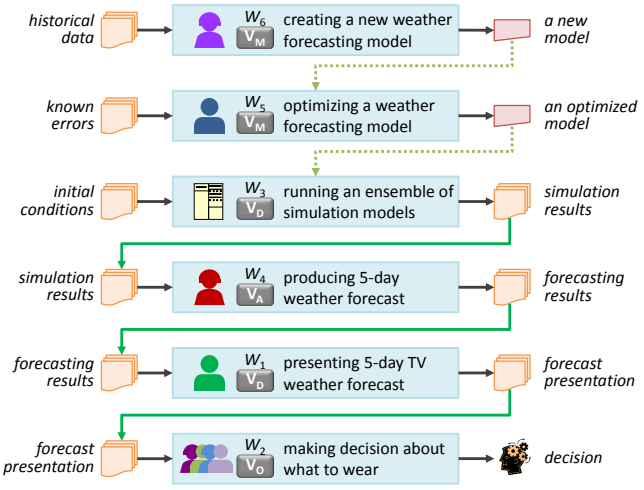


Fig. 2. A complex work flow can be decomposed into different visualization tasks, each of which may make use a workflow in one of the six classes.

W_6 represents a class of commonly-occurred workflows where analysts deploy known theories to specify a model without the initial observational visualization for establishing these theories. In practice, to create, test and optimize a model, analysts often make use of W_5 and W_6 for different parts of a model. For example, in developing a simulation model, major computational steps are defined according to known quantitative laws, while initial and boundary conditions are defined based on observations and experiments. Although W_5 and W_6 feature two structurally-different workflows, we represent the visualization tasks using the same macro block $\mathbf{V_M}$ as a combined representation for *Model-developmental Visualization*.

The above six classes of work flows are abstractions based on a single visualization task usually performed by one data analyst. They are not intended as a taxonomy of all workflows in practical applications, but for providing necessary abstraction for us to observe the characteristics of four levels of visualization as discussed in Section 3.1. Many data analysis and visualization processes feature different tasks and different types of analysts. Nevertheless, a complex workflow can be decomposed into a set of elementary workflows, each falling into one of the six classes. Fig. 2 shows an example of such a complex workflow.

4 Information-Theoretic Abstraction

In this section, we first provide an information-theoretic abstraction of data spaces as *alphabets*, and an abstraction of data processing as *alphabet transformation*. Building on the abstraction, we propose an information-theoretic metric for evaluating the cost-benefit of processes in a data analysis and visualization workflow.

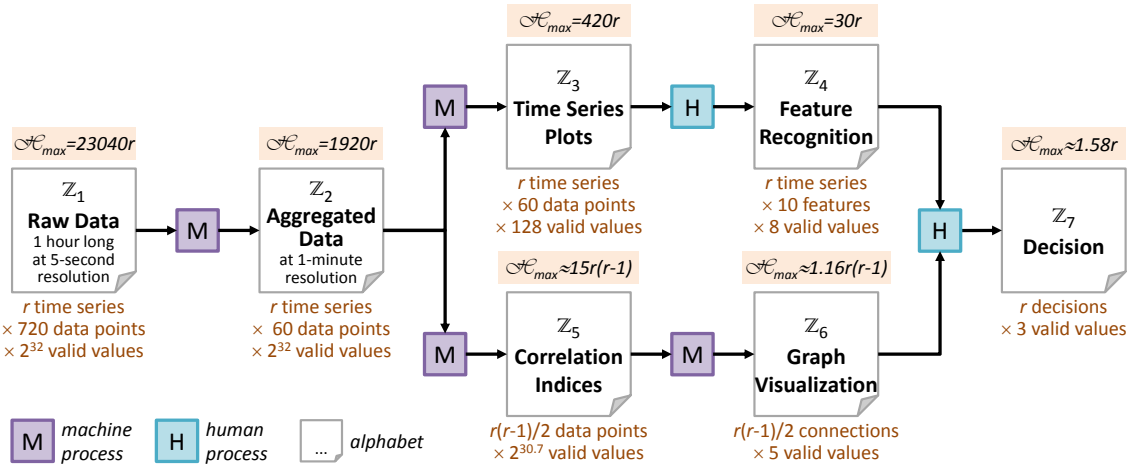


Fig. 3. An example transformation of alphabets during a data analysis and visualization process. From left to right, the initial alphabet corresponds to r time series each capturing a share price at 5 second interval within an hour. For each time series, the 12 data points in every minute are then aggregated into a mean value. The r time series is then visualized as line plots. The analyst identifies various features during the visualization, such as different levels of rise or fall, different intensity, etc. Meanwhile, the analyst computes the correlation indices between each pair of time series and visualize these using, for instance, a circular graph plot, where correlation indices are mapped to five different colors. The analyst finally makes a decision for each of the r shares as to buy, sell or hold. The maximal entropy $\mathscr{H}_{MAX}$ shows a decreasing trend from left to right.

4.1 Alphabets and Letters

The term *data object* is an encompassing generalization of datum, data point, data sample, data record and dataset. It contains a finite collection of quantitative and/or qualitative measures that are values of a finite set of variables. For example, consider a univariate variable X for recording the population of a country. A value representing the UK population in 2010 is a datum, and thus a data object. A collection of the individual population figures of N countries in 2010 is also a data object, where the N values may be considered as a sample of data points of X, or separate records of N variables X_i $(i = 1, 2, \ldots, \mathrm{N})$. Similarly, a time series recoding the UK annual population between 1900 and 2010 is a data object. The 111 values in the time series may be considered as data points of the same univariate variable X, or a multivariate record for time-specific variables X_t $(t = 1900, 1901, \ldots, 2010)$. Of course, the term data object can also refer to a multivariate data point that consists of values of conceptually-different variables, e.g., area, population and GDP of a country.

The generalization also encompasses datasets that are often regarded as "unstructured". For example, a piece of text may be treated as a multivariate record of M characters, each of which is a value of a variable C_j for encoding a letter, digit or punctuation mark at a specific position j $(j = 1, 2, \ldots, \mathrm{M})$ within the text. Hence, the multivariate record is a data object. Alternatively, we can consider a composite variable, Y, which encodes all possible variations of texts with M or fewer characters. A specific text with $1 \leq k \leq \mathrm{M}$ characters is thus a value of Y. This example also illustrates the equivalence between encoding a data object as a multivariate data record or encoding it as an instance of single composite variable.

In this generalized context, let Z be a variable, and $\mathbb{Z} = \{z_1, z_2, \ldots, z_{\mathrm{M}}\}$ be the set of all its valid values. Z may be a univariate, multivariate, or composite variable. When Z is a multivariate variable, each of its valid value, z_i, is a valid combination of valid values of individual univariate variables. When Z is a composite variable, we can flatten its hierarchy by encoding the hierarchical relationships explicitly using additional variables. The flattened representation thus represents a multivariate variable. Hereby z_i is a valid combination of valid values of individual variables including the additional ones. In information theory, such a set $\mathbb{Z}$ is referred to as an *alphabet*, and each of its member z_i as a *letter*.

When the probability of every letter, $\mathbf{p}(z_i)$, is known or can be estimated, $\mathbf{p}$ is the *probability mass function* for the set $\mathbb{Z}$. Shannon introduced the measure of *entropy*:

$$\mathscr{H}(Z) = -\sum_{1}^{\mathrm{M}} \mathbf{p}(z_i) \log_2 \mathbf{p}(z_i)$$

for describing the *level of uncertainty* of an alphabet. With the above $\log_2$-based formula, the unit of $\mathscr{H}(Z)$ is *bit*.

4.2 Transformation of Alphabets

In many data-intensive environments, the alphabet of raw input data may contain numerous letters. For example, consider all valid time series of share prices within one hour period. Assuming that the share price is updated every 5 seconds, there are 720 data points per time series within an hour. Assuming that we represent share price at USD \$0.01 resolution using 32-bit unsigned integers[1], the

1. By the end of 2014, the highest share price in the US is probably that of Berkshire Hathaway Inc. (BRK-A) at 22,937,400 cents, i.e., $2^{24} < 22,937,400 < 2^{25}$). Source: Yahoo Finance.

minimum and maximum values are thus 0 and $2^{32}-1$ cents respectively. If the probability of different time series were uniformly distributed, the entropy of this alphabet would be $23040 = 720 \times \log_2(2^{32})$ bits. This is the *maximal entropy* of this alphabet. In practice, as many high values in the range $[0, 2^{32}-1]$ are very unlikely, and sudden changes between a very low value and a very high value (or vice versa) during a short period are also rare, the actual entropy is lower than 23040 bits.

On the other hand, if we need to consider r of such time series in order to make a decision, the size of the new alphabet will increase significantly. Although some combinations of r time series may be highly improbable, they may still be valid letters. Hence the maximal entropy of this new alphabet is $23040r$ bits. Let us consider such r time series as the initial raw data for a data analysis and visualization process as illustrated in Fig. 3.

One may find that the resolution of 1 data point per 5 seconds is not necessary, and choose to reduce it to 1 data point every minute by computing the average of 12 data points in each minute. The average values may also be stored using 32-bit unsigned integers. This aggregation results in a new alphabet, whose maximal entropy of $1920r = r \times 60 \times \log_2(2^{32})$ bits. As indicated in Fig. 3, one may use line plots with 128 distinguishable pixels along the y-axis. When we use these plots to visualize these r time series, we may only be able to differentiate up to 128 data values per data point. In this case, the maximal entropy is reduced to $r \times 60 \times \log_2(128) = 420r$ bits.

When one observes these r time series, one may identify some specific features, such as [rise, fall, or flat], [slow, medium, or fast], [stable, uneven, or volatile] and so on. These features become a new set of variables defined at the level of an hour-long time series. If we construct a new alphabet based on these feature variables, its entropy will be much less than $23040r$ bits. For example, if there are 10 feature variables and each with 8 valid values, the maximal entropy of this "observational" alphabet is $30r$ bits.

When one analyzes the relations among these r time series, one may, for instance, compute the correlation indices between every pair of time series. This yields $r(r-1)/2$ numbers. Assuming that these are represented using 32-bit floating-point numbers, the maximal entropy of this "analytical" alphabet is around $15r(r-1)$ bits as the single precision floating-point format supports some $2^{30.7}$ values in $[-1,1]$. When we visualize these correlation indices by mapping them to, for instance, five colors representing $[-1,-0.5,0,0.5,1]$, the entropy is reduced to $\log_2(5)r(r-1)/2 \approx 1.16r(r-1)$ bits.

One may wish to make a decision with three options, [buy, sell, or hold]. In this case, this "decisional" alphabet for each time series has only three letters. The maximal entropy of this alphabet is less than 2 bits. If a decision has to be made for all r time series, we have less than $2r$ bits. Fig. 3 illustrates the abovementioned changes of alphabets with different maximal entropy values. The final alphabet ultimately defines the visualization task, while some intermediate alphabets may also capture subtasks in a data analysis and visualization process.

4.3 Measuring Cost-Benefit Ratio

From Fig. 3, one observation that we can make is that there is almost always a reduction of maximal entropy from the original data alphabet to the decisional alphabet. This relates to one of the basic objectives in statistical inference, i.e., to optimize the process between the initial alphabet and the final alphabet with minimal loss of information that is "important" to the decision based on the final alphabet. However, as visualization processes involve both machine-centric and human-centric mappings, it is necessary (i) to optimize both types of mapping in an integrated manner, (ii) to take into account "soft" information that can be introduced by human analysts during the process, (iii) to consider information loss as part of a cost-benefit analysis.

Let us consider a sequential workflow with L processing steps. There are $L+1$ alphabets along the workflow, Let $\mathbb{Z}_s$ and $\mathbb{Z}_{s+1}$ be two consecutive alphabets such that:

$$F_s : \mathbb{Z}_s \longrightarrow \mathbb{Z}_{s+1}$$

where F_s is a mapping function, which can be an analytical algorithm that extracts features from data, a visual mapping that transforms data to a visual representation, or a human decision process that selects an outcome from a set of options.

The cost of executing F_s as part of a visualization process can be measured in many ways. Perhaps the most generic cost measure is *energy* since energy would be consumed by a computer to run an algorithm or to create a visualization, as well as by a human analyst to read data, view visualization, reason about a possible relationship, or make a decision. We denote this generic measurement as a function $\mathscr{C}(F_s)$. While measuring energy usage by computers is becoming more practical [55], measuring that of human activities, especially cognitive activities may not be feasible in most situations. A more convenient measurement is *time*, $\mathscr{C}_{time}(F_s)$, which can be considered as an approximation of $\mathscr{C}(F_s)$. Another is a monetary measurement of computational costs or employment costs, which represent a subjective approximation from a business perspective. Without loss of generality, we will use $\mathscr{C}(F_s)$ as our cost function in this section.

DEFINITION 1 (Alphabet Compression Ratio). As shown in Fig. 3, a mapping function (i.e., a machine or human process) usually facilitates the reduction of data space at each stage of data processing though the reduction is not guaranteed. We can measure the level of reduction as the *alphabet compression ratio* (ACR) of a mapping F_s:

$$\Psi_{ACR}(F_s) = \frac{\mathscr{H}(\mathrm{Z}_{s+1})}{\mathscr{H}(\mathrm{Z}_s)} \tag{1}$$

where $\mathscr{H}$ is the Shannon entropy measure. In a closed machine-centric processing system that meets the condition of a Markov chain, we have $\mathscr{H}(\mathrm{Z}_s) \geq \mathscr{H}(\mathrm{Z}_{s+1})$. This is the *data processing inequality* [18]. In such a system, Ψ_{ACR} is a normalized and unitless entropy measure in $[0,1]$ as first

proposed by Golan in [23] (see also [26]). However, Chen and Jänicke pointed out that the Markov chain condition is broken in most visualization processes [15], and further examples were given in [14]. Hence, we do not assume that $\mathscr{H}(\mathbb{Z}_s) \geq \mathscr{H}(\mathbb{Z}_{s+1})$ here since F_s can be a human-centric transformation, unless one encodes all possible variants of "soft" information and knowledge in the initial data alphabet.

Meanwhile, given an output of an analytical process, F_s, an analyst will gain an impression about the input. Considering the time series transformation in Fig. 2, for example, learning the mean price value for each minute, an analyst may have a conjecture about the 12 original data values. Viewing a visualization of each time series plot in a resolution of 128 possible values per data point, an analyst may infer, estimate or guess the time series in its original resolution of 2^{32} possible values per data point. Let us denote an impression about Z_s as a variable Z'_s, which is a result of a mapping G_s such that:

$$G_s : \mathbb{Z}_{s+1} \longrightarrow \mathbb{Z}'_s$$

where $\mathbb{Z}'_s$ is the alphabet of this impression with a probability mass function representing the inferred or guessed probability of each letter in $\mathbb{Z}'_s$. Note that G_s is a reconstruction function, similar to what was discussed in [35]. In most cases, G_s is only a rough approximation of the true inverse function F^{-1}. The informational difference between such an impression about $\mathbb{Z}'_s$ obtained from observing letters in $\mathbb{Z}_{s+1}$ and the actual $\mathbb{Z}_s$ is defined by Kullback-Leibler divergence (or relative entropy) [18]:

$$\mathscr{D}_{KL}(Z'_s||Z_s) = \mathscr{D}_{KL}(G_s(Z_{s+1})||Z_s) = \sum_j \mathbf{p}(z'_{s,j}) \log_2 \frac{\mathbf{p}(z'_{s,j})}{\mathbf{q}(z_{s,j})}$$

where $z'_{s,j} \in \mathbb{Z}'_s$, and $z_{s,j} \in \mathbb{Z}_s$, and $\mathbf{p}$ and $\mathbf{q}$ are two probability mass functions associated with $\mathbb{Z}'_s$ and $\mathbb{Z}_s$ respectively. $\mathscr{D}_{KL} = 0$ if and only if $\mathbf{p} = \mathbf{q}$, and $\mathscr{D}_{KL} > 0$ otherwise. Note that $\mathscr{D}_{KL}$ is not a metric as it is not symmetric. The definition of $\mathscr{D}_{KL}$ is accompanied by a precondition that $\mathbf{q} = 0$ implies $\mathbf{p} = 0$.

DEFINITION 2 (Potential Distortion Ratio). With the $\log_2$ formula, $\mathscr{D}_{KL}$ is also measured in bits. The higher the number of bits is, the further is the deviation of the impression $\mathbb{Z}'_s$ from $\mathbb{Z}_s$. The *potential distortion ratio* (PDR) of a mapping F_s is thus:

$$\Psi_{PDR}(F_s) = \frac{\mathscr{D}_{KL}(Z'_s||Z_s)}{\mathscr{H}(\mathbb{Z}_s)} \tag{2}$$

Both $\Psi_{ACR}(F_s)$ and $\Psi_{PDR}(F_s)$ are unitless. They can be used to moderate the cost of executing F_s, i.e., $\mathscr{C}(F_s)$. Since $\mathscr{H}(\mathbb{Z}_{s+1})$ indicates the intrinsic uncertainty of the output alphabet and $\mathscr{D}_{KL}(Z'_s||Z_s)$ indicates the uncertainty caused by F_s, the sum of $\Psi_{ACR}(F_s)$ and $\Psi_{PDR}(F_s)$ indicates the level of combined uncertainty in relation to the original uncertainty associated with $\mathbb{Z}_s$.

DEFINITION 3 (Effectual Compression Ratio). The *effectual compression ratio* (ECR) of a mapping F_s from $\mathbb{Z}_s$ to $\mathbb{Z}_{s+1}$ is a measure of the ratio between the uncertainty before a transformation F_s and that after:

$$\Psi_{ECR}(F_s) = \frac{\mathscr{H}(\mathbb{Z}_{s+1}) + \mathscr{D}_{KL}(Z'_s||Z_s)}{\mathscr{H}(\mathbb{Z}_s)} \quad \text{for } \mathscr{H}(\mathbb{Z}_s) > 0 \tag{3}$$

When $\mathscr{H}(Z_s) = 0$, it means that variable Z_s has only one probable value, and it is absolutely certain. Hence, the transformation of F_s is unnecessary in the first place. The measure of ECR encapsulates the tradeoff between ACR and PDR, since deceasing ACR (i.e., more compressed) often leads to an increase of PDR (i.e., harder to infer $\mathbb{Z}_s$), and vice versa. However, this tradeoff is rarely a linear (negative) correlation. Finding the most appropriate tradeoff is thus an optimization problem, which is to be further enriched when we incorporate below the cost $\mathscr{C}(F_s)$ as another balancing factor.

DEFINITION 4 (Benefit). We can now define the *benefit* of a mapping F_s from $\mathbb{Z}_s$ to $\mathbb{Z}_{s+1}$ as:

$$\mathscr{B}(F_s) = \mathscr{H}(\mathbb{Z}_s) - \mathscr{H}(\mathbb{Z}_{s+1}) - \mathscr{D}_{KL}(Z'_s||Z_s) \tag{4}$$

The unit of this measure is *bit*. When $\mathscr{B}(F_s) = 0$, the transformation does not create any change in the informational structure captured by the entropy. In other words, there is no informational difference between observing variable Z_s and observing Z_{s+1}. When $\mathscr{B}(F_s) < 0$, the transformation has introduced more uncertainty, which is undesirable. When $\mathscr{B}(F_s) > 0$, the transformation has introduced positive benefit by reducing the uncertainty. This definition can be related to Shannon's grouping property [18].

THEOREM (Generalized Grouping Property). Let X be a variable that is associated with an N-letter alphabet $\mathbb{X}$ and a normalized N-dimensional discrete distribution $\mathbf{p}(x), x \in \mathbb{X}$. When we group letters in $\mathbb{X}$ to M subsets, we derive a new variable Y with an M-letter alphabet $\mathbb{Y}$ and a normalized M-dimensional discrete distribution $\mathbf{q}(y), y \in \mathbb{Y}$.

$$\mathscr{H}(X) = \mathscr{H}(Y) + \sum_{k=1}^{\mathrm{M}} \mathbf{q}(y_k)\mathscr{H}_k \tag{5}$$

where $\mathscr{H}_k$ is the entropy of the local distribution of the original letters within the k^{th} subset of $\mathbb{X}$. Comparing Eq. (4) and Eq. (5), we can see that the last term on the right in Eq. (5) is replaced with the Kullback-Leibler divergence term in Eq. (4). The equality in Eq. (5) is replaced with a measure of difference in Eq. (4). This is because of the nature of data analysis and visualization. After each transformation F_s, the analyst is likely to infer, estimate or guess the local distribution within each subset, when necessary, from the observation of X in the context of Eq. (5) or Z_{s+1} in the context of Eq. (4) in conjunction with some "soft" information and knowledge, as mentioned in Section 1.

DEFINITION 5 (Incremental Cost-Benefit Ratio). The *incremental cost-benefit ratio* (Incremental CBR) of a mapping F_s from $\mathbb{Z}_s$ to $\mathbb{Z}_{s+1}$ is thus defined as the ratio between benefit $\mathscr{B}(F_s)$ and cost $\mathscr{C}(F_s)$.

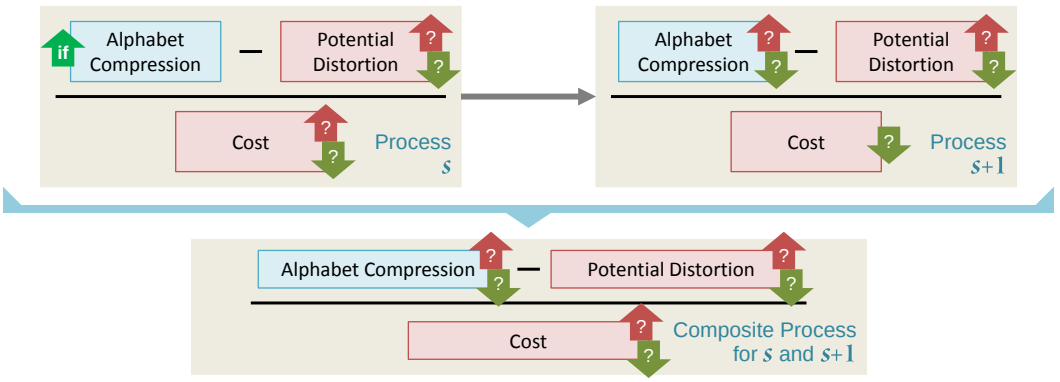


Fig. 4. Making changes to a transformation may result in changes the amount of alphabet compression, potential distortion and cost within the transformation (top-left). It may also have cascading effects on the succeeding transformation (top-right), and combined effects on the composite transformation (below). .

$$\Upsilon(F_s) = \frac{\mathscr{B}(F_s)}{\mathscr{C}(F_s)} = \frac{\mathscr{H}(\mathrm{Z}_s) - \mathscr{H}(\mathrm{Z}_{s+1}) - \mathscr{D}_{KL}(Z'_s||Z_s)}{\mathscr{C}(F_s)} \quad (6)$$

Note that we used cost as the denominator because (i) the benefit can be zero, while the cost of transformation cannot be zero as long as there is an action of transformation; (ii) it is better to associate a larger value to the meaning of more cost-beneficial.

At each transformation s, if one changes the method (e.g., a visual design or an analytical algorithm), the change will likely affect the amount of alphabet compression, $\mathscr{H}(\mathrm{Z}_s) - \mathscr{H}(\mathrm{Z}_{s+1})$, potential distortion $\mathscr{D}_{KL}(Z'_s||Z_s)$, and cost $\mathscr{C}(F_s)$. For example, as illustrated in the top-left block in Fig. 4, one may wish to increase alphabet compression by (a) using a more abstract visual representation or (b) rendering a visualization at a lower-resolution. Approach (a) may increase the cost while reducing (or increasing) the potential distortion. Approach (b) may reduce cost while increasing the potential distortion. Hence discovering the best method is an optimization process.

Furthermore, the change of a method at one transformation may likely trigger subsequent changes in the succeeding transformation. This cascading effect is illustrated in Fig. 4. For example, if one uses a visualization with a high alphabet compression ratio at transformation F_s, a human observer may be able to observe a different set of features, resulting in a change of the feature alphabet and thus the alphabet compression ratio in the following transformation F_{s+1}. Even when the feature alphabet remains the same, the human observer may (or may not) recognize various multivariate features more accurately (i.e., less potential distortion) or speedily (i.e., less cost). Hence it is necessary to optimize the combined cost-benefit ratio of a composite transformation.

Given a set of cascading mapping functions, $F_1, F_2, \ldots, F_\mathrm{L}$, which transform alphabets from $\mathbb{Z}_1$ to $\mathbb{Z}_{\mathrm{L}+1}$, we can simply add up their costs and benefits as:

$$\mathscr{C}_{total} = \sum_{s=1}^{\mathrm{L}} \mathscr{C}(F_s)$$

$$\mathscr{B}_{total} = \sum_{s=1}^{\mathrm{L}} \mathscr{B}(F_s) = \mathscr{H}(\mathrm{Z}_1) - \mathscr{H}(\mathrm{Z}_{\mathrm{L}+1}) - \sum_{s=1}^{\mathrm{L}} \mathscr{D}_{KL}(Z'_s||Z_s)$$

The *overall cost-benefit ratio* (Overall CBR) is thus $\mathscr{B}_{total}/\mathscr{C}_{total}$.

For workflows containing parallel data processing branches, the merge of CBR at a joint of two or more branches partly depends on the semantics of the cost and benefit measures. If we are concerned about the energy, or monetary cost, the simple summation of cost measures of the branches arrived at a joint makes sense. If we are concerned about the time taken, we may compute the maximum cost of all branches arriving at a joint. If all parallel branches arriving at a joint contain only machine-centric processes, the benefit is capped by the entropy at the beginning of the branching-out. The combined benefit can be estimated by taking into account the mutual information between the arriving alphabets. When these parallel branches involve human-centric processing, "soft" information will be added into the process. The combined benefit can be estimated in the range between the maximum and the summation of the arriving benefit measures.

In this paper, we largely focus on the workflows for conducting data analysis and visualization. Our formulation of cost-benefit analysis can be extended to include the cost of development and maintenance. It is more appropriate to address such an extension in future work.

5 Examples of Workflow Analysis

In this section, we consider several successful visualization processes in the literature. We analyze their cost-benefit ratios in comparison with possible alternative processes. The comparison serves as initial validation of the information-theoretic measures proposed in the previous section. Like most theoretic development, the practical validation of the proposed information-theoretic measures, e.g., Eq. (6),

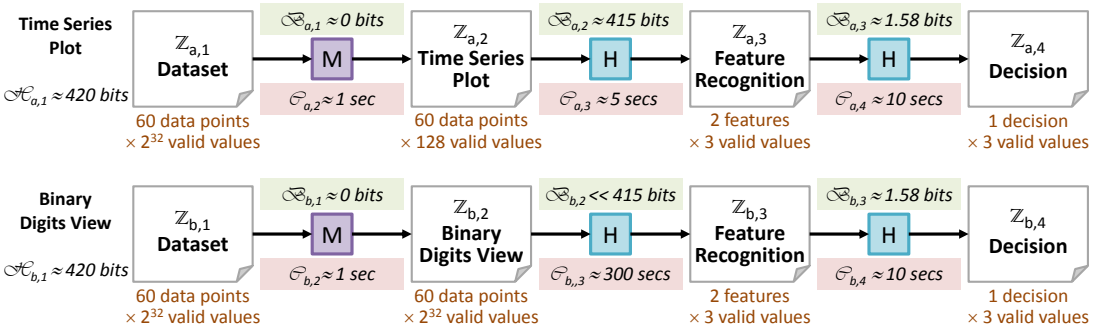


Fig. 5. Comparison between time series plot and binary digitals view for disseminative visualization. The same legend in Fig. 3 applies. The estimated benefit and cost values here are based on heuristic reasoning and for illustration only. For example, for $\mathscr{B}_{a,2}$, we consider two feature variables [stable, uneven, volatile] and [rise, fall, flat]. Hence the maximal entropy of $\mathbb{Z}_{a,3}$ is about 3.17 bits. As the $\mathscr{D}_{KL}$ term for $\mathscr{B}_{a,2}$ indicates some uncertainty, the estimated benefit is $420 - 3.17 - \mathscr{D}_{KL}(\mathbb{Z}'_{a,2}||\mathbb{Z}_{a,2}) \approx 415$ bits. Meanwhile, $\mathscr{D}_{KL}$ for $\mathscr{B}_{b,2}$ is much higher.

should be, and is expected to be, a long-term undertaking, along with the advancement of techniques and the increasing effort for collecting performance data about various human-centric processes, e.g., through empirical studies.

5.1 Disseminative Visualization

The history of time series plot can be traced back more than a millennium ago. If success is measured by usage, it is undoubtedly one of the most successful visual representations. However, its display space utilization is rather poor in comparison with a binary digits view [15]. Fig. 5 shows two such representations that are used as disseminative visualization for a scenario in Fig. 3. The dataset being displayed is a time series with 60 data points, i.e., an instance of $\mathbb{Z}_2$ in Fig. 3. Assume that the value of this particular share has been largely moving between 100 and 200 cents. Hence the entropy of $\mathbb{Z}_{a,1} = \mathbb{Z}_{b,1}$ is estimated to be about 420 bits, significantly below the maximal entropy of the data representation.

The binary digits view uses a 2×2 pixel-block per digit, and requires 32×60 blocks (7,680 pixels) for the plotting canvas. Using the same number of pixels, 128×60, the time series plot is an instance of $\mathbb{Z}_{a,2}$. During dissemination, the presenter (or analyst) points out "stable" and "rise" features to a viewer (or client), suggesting a decision "to hold". The overall CBRs for the two pipelines in Fig. 5 are:

$$\Upsilon_{plot} = \sum_{j=1}^{3} \frac{\mathscr{H}(Z_{a,j+1}) + \mathscr{D}_{KL}(Z'_{a,j}||Z_{a,j})}{\mathscr{C}(F_{a,j})} \tag{7}$$

$$\Upsilon_{binary} = \sum_{j=1}^{3} \frac{\mathscr{H}(Z_{b,j+1}) + \mathscr{D}_{KL}(Z'_{b,j}||Z_{b,j})}{\mathscr{C}(F_{b,j})} \tag{8}$$

To the presenter, the decision "to hold" has already been made, and the total CBR would be zero for either workflow. For a viewer unfamiliar with binary representations, the binary digits view is almost undecipherable. For a pair of untrained eyes, recognizing features such as "stable" and "rise" would take a while. The inverse mapping from the features pointed out by the presenter is also rather uncertain, hence a high value for the $\mathscr{D}_{KL}$ term in $\mathscr{B}_{b,2}$. The binary digits view thereby incurs a huge cost at the feature recognition step, while bringing lower benefit. This mathematically explains the merits of time series plot over a spatially-compact binary digits view.

5.2 Observational Visualization

The example in Section 5.1 can also be considered in the context of observational visualization, where an analyst creates visualization for him/herself. Similar abstract reasoning and step-by-step inference can be carried out, just as in the previous example, likely for a much larger input data alphabet (e.g., with r time series and t hours).

Let us consider a different example of observational visualization. Legg *et al.* reported an application of visualization in sports [41]. The Welsh Rugby Union required a visualization system for in-match and post-match analysis. One of the visualization tasks was to summarize events in a match, facilitating external memorization. The input datasets are typically in the form of videos including data streams during a match, and can be generalized to include direct viewing of a match in real-time. The alphabet is thus huge. The objective for supporting external memorization is to avoid watching the same videos repeatedly. Especially during a half-time interval, coaches and players cannot afford much time to watch videos.

The workflow can be coarsely divided into three major transformations, namely F_a: transforming real-world visual data to events data, F_b: transforming events data to visualization, and F_c: transforming observations to judgments and decisions. Clearly, transformation F_c should be performed by coaches and other experts.

As illustrated in Fig. 6, for transformation F_a, two options were considered: $F_{a,1}$ for computers to detect events, and $F_{a,2}$ for humans to detect events. For transformation F_b, two options were considered: $F_{b,1}$ statistical graphics, and $F_{b,2}$ glyph-based event visualization. Because $F_{b,1}$ and $F_{b,2}$ generate different output alphabets, the succeeding observation processes are thus different.

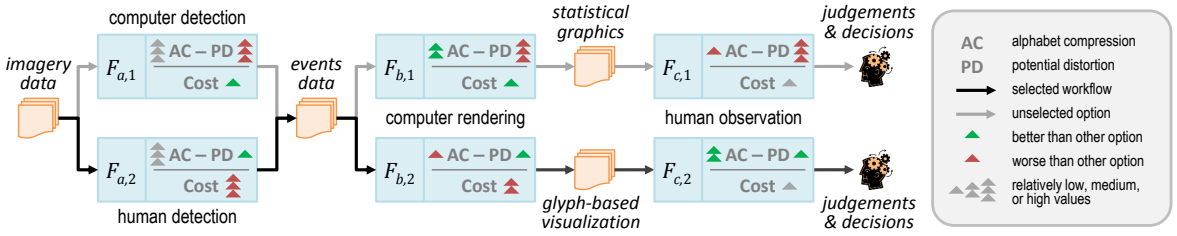


Fig. 6. Comparison between possible optional processes in a workflow for analysing and visualizing sports events. Although the levels of increments of alphabet compression, potential distortion and costs are roughly estimated, they offer qualitative indication of the tradeoffs between different options.

For $F_{a,1}$ and $F_{a,2}$, the letters of the output alphabet are multivariate data objects describing what type of event, when and where it happens, and who are involved. This alphabet is much smaller than the input alphabet for real-world visual data. The team did not find any suitable computer vision techniques that could be used to detect events and generate the corresponding data objects in this application. The accuracy of available techniques were too low, hence the $\mathscr{D}_{KL}$ term for $F_{a,1}$ will yield a high-level of uncertainty. Using a video annotation system, an experienced sports analyst can generate more accurate event data during or after a match.

For an 80 minute Rugby match, the number of data objects generated is usually in hundreds or thousands. Statistics can be obtained, and visualized using statistical graphics. However, it is difficult to connect statistics with episodic memory about events in $F_{c,1}$, and thereby difficult for coaches to make decisions based on statistical graphics. Such a difficulty corresponds to a high-level of uncertainty resulting from the $\mathscr{D}_{KL}$ term for $F_{b,1}$. On the other hand, the direct depiction of events using glyphs can stimulate episodic memory much better, yielding a much lower-level uncertainty in the $\mathscr{D}_{KL}$ term for $F_{b,2}$ and $F_{c,2}$. The team implemented $F_{a,2}$ and $F_{b,2}$ transformations as reported in [41], while $F_{b,1}$ was also available for other tasks.

5.3 Analytical Visualization

Oelke *et al.* studied a text analysis problem using visual analytics [45]. They considered a range of machine-centric and human-centric transformations in evaluating document readability. For example, the former includes 141 text feature variables, and their combinations. The latter includes four representations at three different levels of details. Since different combinations of machine-centric and human-centric transformations correspond to different visual analytics pipelines, their work can be seen as an optimization effort.

Consider feature extraction collectively as a transformation F_a, and the generation of all combinations of features and levels of representations as a transformation F_b. Since F_a and F_b can be automated, one would naturally consider using a ranking algorithm to determine which combination is the optimal. We denote this approach as a transformation $F_{c,1}$. A ranking algorithm would have to assess the usefulness of each type of features by examining hundreds of these features in all documents. The aggregation of such assessments would deliver very aggressive alphabet compression. However it would be almost impossible to infer how features affect readability in each document from a ranking score resulting from $F_{c,1}$. Hence, the uncertainty of the reserve mapping (i.e., the $\mathscr{D}_{KL}$ term) would be huge.

In [45], Oelke *et al.* reported an approach using pixel-based visualization to depict the results of F_b. We denote the rendering and viewing of the pixel-based visualization as a combined transformation $F_{c,2}$. In this transformation, analysts examined hundreds of features in each document. This enabled visual analysis of the relationships among local features and qualitative judgements of the effectiveness of different combinations. Through experimentation and analysis, they confirmed the need for enabling analysts to observe details at the sentence or block levels. Over-aggregation (e.g., assigning an average readability score to each document) is not cost beneficial, as the tradeoff between the alphabet compression ratio (ACR) and the potential distortion ratio (PDR) is in favor of PDR.

5.4 Model-developmental Visualization

In [54], Tam *et al.* compared a visualization technique and a machine learning technique in generating a decision tree as a model for expression classification. The input to this model development exercise is a set of annotated videos, each of which records one of four expressions [anger, surprise, sadness, smile]. The output is a decision tree that is to be used to classify videos automatically with reasonable accuracy. It is thus a *data analysis and visualization process* for creating a *data analysis model*. Although this sounds like a conundrum, it fits well within the scope of visualization. Tam *et al.* approached this problem through a series of transformations. The first transformation F_a identifies 14 different facial features in each video, and records it temporal changes using a geometric or texture measurement. This results in 14 different alphabets of time series. The second transformation F_b characterizes each time series using 23 different parameters. This results in a total of $322 = 14 \times 23$ variables. At the end of F_b, each video becomes a 322-variate data object.

For the visualization-based pipeline, the third transformation $F_{c,1}$ generates a parallel coordinate plot with 322 axes. This is followed by the fourth transformation $F_{d,1}$, where two researchers laid the big plot on the floor and

spent a few hours to select the appropriate variables for constructing a decision tree. For the machine-learning based pipeline, the team used a public-domain tool, C4.5, as the third transformation $F_{c,2}$, which generates a decision tree from a multivariate dataset automatically.

In terms of time cost, transformation $F_{c,2}$ took much less time than transformations $F_{c,1}$ and $F_{d,1}$ together. In terms of performance, the decision tree created by $F_{c,1}$ and $F_{d,1}$ was found slightly more accurate than that resulting from $F_{c,2}$. From further analysis, they learned that (i) handling real values has been a challenge in automatic generation of decision trees; (ii) the two researchers did not rely solely on the parallel coordinates plot to choose variables, their "soft" knowledge about the underlying techniques used in transformations F_a and F_b also contributed to the selection. Such "soft" knowledge reduces the uncertainty expressed by the $\mathscr{D}_{KL}$ term in Eq. 4. This example demonstrates the important role of visualization in model development.

6 Falsifiability

The quantitative metric given in Eq. (6) is a fitness metric for optimizing a data processing and visualization workflow. For any optimization problem, the "baseline" process is stochastic sampling of the design space of such a workflow. A design methodology better than trials-and-errors facilitates a more efficient and effective optimization process. Although developing new design methodologies is beyond the scope of this work, having an appropriate fitness metric represents an important step towards such development.

The three components in Eq. (6) – *alphabet compression*, *potential distortion*, and *cost* – are all quantifiable in principle, describing the quantities that may be optimized. Typically, one may optimize one component while constraining the other two, e.g., minimizing potential distortion while setting a maximum for the cost, and a minimum for alphabet compression. Because *alphabet compression* and *potential distortion* share a common unit (bits), it is feasible to examine the trade-off between the two quantitatively. Hence falsification of this metric would involve the discovery of counter-examples. For example, consider any two competing processes A and B in data analysis and visualization workflow, where A and B have the same input and output alphabets. If $\Upsilon(A) > \Upsilon(B)$, A should be better than B. If one could show that A were definitely worse than B, one would have a counter-example.

If there were a diverse range of counter-examples, it would indicate that the proposed metric is wrong. If the metric were to fail in some conditions but correct in others, the application of the metric may be restricted to certain conditions. Following the discovery of counter-examples, one might also discover a better metric than Eq. 6. The falsification process continues with the new metric.

The challenge in both application and falsification of this metric is the measurement of benefit $\mathscr{B}(F_s)$ in terms of the information-theoretic quantities. For each process F_s, this requires the knowledge of three probability distribution functions for alphabets $\mathbb{Z}_s$, $\mathbb{Z}_{s+1}$, and $\mathbb{Z}'_s$. For a decision alphabet with a small number of letters (e.g., [buy, sell, hold]), it is not difficult to obtain a probability distribution. However for a simple visualization alphabet, the number of letters increases rapidly. For example, a bar chart with three bars, each with 10 valid values, corresponds to an alphabet with 1,000 letters. Obtaining a probability distribution becomes a non-trivial undertaking, especially when the alphabet concerned results from a human-centric process.

However, such difficulties do not imply that information-theoretic quantities cannot be measured. Firstly, some contemporary methods for empirical studies can reach out to a huge number of participants, and collect a huge number of responses to stimuli. They may potentially enable an efficient and effective means for capturing probability distributions in human-centric processes. Secondly, information-theoretic quantities in thermodynamics are usually estimated from related qualities such as pressure, volume and energy. Hence the future advancement of data sciences may provide us with quantitative laws to facilitate indirect measurement of information-theoretic quantities in data analysis and visualization workflows.

In the short term, one can carry out qualitative assessment of the three components as illustrated in Fig. 4 and exemplified by case studies in Section 5. Given a machine-centric or human-centric process, one can use the notion of *alphabet compression* to assess collectively levels of abstraction, filtering, aggregation, summarization, etc.; use the notion of *potential distortion* to assess collectively levels of inference errors, context-awareness, internal and external memorization, provenance, reconstructability, etc.; and use the notion of *cost* to assess collectively levels of human resources, skill requirements, facility requirement, processing speed, impact of errors, etc.

Although such qualitative assessments cannot offer a definite means of falsification of the proposed metric, they will offer useful insight to further theoretic advancement in visualization. Perhaps more importantly, if the proposed metric is theoretically plausible, it means that we work towards the establishment of a common measurement unit for benefit attributes ranging from abstraction to context-awareness, and similarly a common measurement unit for cost attributes. With these two common units, we can develop new methodologies for systematic optimization of data analysis and visualization workflows.

7 Conclusions

In this paper, we have proposed an information-theoretic measure for offering a mathematical explanation as to what may have been optimized in successful visualization processes. We have used several examples from the literature to demonstrate its explanatory capability for both machine-centric and human-centric transformations in data analysis and visualization. One question that naturally occurs is how one may use such a theoretical measure in a practical environment. We consider this question in three stages.

(i) At present, it is important for us to recognize that the overall objective of data analysis and visualization

corresponds to the reduction of Shannon entropy from the original data alphabet to the decisional alphabet. There is a cost associated with this reduction process. It is also necessary to recognize that the benefit of such reduction at each incremental step is likely to be weakened by the uncertainty of an approximated inverse mapping, i.e., the $\mathscr{D}_{KL}$ term in Eq. 4. This uncertainty can be caused by inaccuracy or aggressive aggregation of a machine-centric transformation, as well as by human factors such as visual uncertainty [19] and lack of understanding and experience.

(ii) Next, we can learn from cost-benefit analysis in social sciences, where quantitative and qualitative methods are integrated together to optimize various business and governmental processes in a systematized manner. Once a visualization process is defined as a transformation-based pipeline, we can estimate the cost for each transformation. We should start to define alphabets and estimate the uncertainty measures associated with them.

(iii) Historically, theoretical advancements were often part of long-term co-evolution with techniques and processes for measurements. This suggests that in the future we will be able to optimize visualization processes in a more quantitative manner. It also suggests that in visualization, empirical studies are not only for evaluating hypotheses but also for collecting measurements that can potentially be used in process optimization.

ACKNOWLEDGMENTS

Both authors are members of *Info-Metrics Institute* and would like to thank the Institute for providing a stimulating environment and financial support for this interdisciplinary research. Golan thanks the *Santa Fe Institute* where parts of this work were discussed in numerous seminars.

REFERENCES

[1] G. Abram and L. Treinish. An extended data-flow architecture for data analysis and visualization. In *Proc. IEEE Visualization*, pages 263–270, 1995.

[2] N. Ahmed, Z. Zhang, and K. Mueller. Human computation in visualization: using purpose-driven games for robust evaluation of visualization algorithms. *IEEE Transactions on Visualization and Computer Graphics*, 18(2):2104–2113, 2012.

[3] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[4] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[5] A. Biswas, S. Dutta, H.-W. Shen, and J. Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2683–2692, 2013.

[6] U. Bordoloi and H.-W. Shen. View selection for volume rendering. In *Proc. IEEE Visualization*, pages 487–494, 2005.

[7] R. Bramon, I. Boada, A. Bardera, Q. Rodrguez, M. Feixas, J. Puig, and M. Sbert. Multimodal data fusion based on mutual information. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1574–1587, 2012.

[8] R. Bramon, M. Ruiz, A. Bardera, I. Boada, M. Feixas, and M. Sbert. An information-theoretic observation channel for volume visualization. *Computer Graphics Forum*, 32(3pt4):411–420, 2013.

[9] R. Bramon, M. Ruiz, A. Bardera, I. Boada, M. Feixas, and M. Sbert. Information theory-based automatic multimodal transfer function design. *IEEE Journal of Biomedical and Health Informatics*, 17(4):870–880, 2013.

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

[11] S. Bruckner and T. Möller. Isosurface similarity maps. *Computer Graphics Forum*, 29(3):773–782, 2010.

[12] S. Card and J. Mackinlay. The structure of the information visualization design space. In *Proc. IEEE Information Visualization*, pages 92–99, 1997.

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

[14] M. Chen and L. Floridi. An analysis of information in visualisation. *Synthese*, 190(16):3421–3438, 2013.

[15] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010.

[16] M. Chen, S. Walton, K. Berger, J. Thiyagalingam, B. Duffy, H. Fang, C. Holloway, and A. E. Trefethen. Visual multiplexing. *Computer Graphics Forum*, 33(3):241–250, 2014.

[17] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proc. IEEE Information Visualization*, pages 69–75, 2000.

[18] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

[19] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum*, 31(3):1015–1024, 2012.

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

[21] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012.

[22] M. Feixas, M. Sbert, and F. González. A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Transactions on Applied Perception*, 6(1):1–23, 2009.

[23] A. Golan. *A Discrete Stochastic Model of Economic Production and a Model of Fluctuations in Production Theory and Empirical Evidence*. PhD thesis, University of California, Berkeley, 1988.

[24] A. Golan. *Information and Entropy Econometrics – A Review and Synthesis*, volume 2 of *Foundations and Trends in Econometrics*. 2008.

[25] A. Golan. On the foundations and philosophy of info-metrics. In *How the World Computes: Turing Centenary Conference and 8th Conference on Computability in Europe*, Springer LNCS 7318, pages 273–244. 2012.

[26] A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley & Sons, 1996.

[27] A. Golan, L. Karp, and J. Perloff. Estimating firm's mixed price and advertising strategies: Coke and pepsi. *Journal of Business and Economic Statistics*, 18(4):398–409, 2000.

[28] T. M. Green, W. Ribarsky, and B. Fisher. Visual analytics for complex concepts using a human cognition model. In *Proc. IEEE VAST*, pages 91–98, 2008.

[29] T. M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *Information Visualization*, 8(1):113, 2009.

[30] G. Grinstein, A. Kobsa, C. Plaisant, and J. T. Stasko. Which comes first, usability or utility? In *Proc. IEEE Visualization*, pages 605–606, 2003.

[31] D. P. Groth and K. Streefkerk. Provenance and annotation for visual exploration systems. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1500–1510, 2006.

[32] S. Gumhold. Maximum entropy light source placement. In *Proc. IEEE Visualization*, pages 275–282, 2002.

[33] L. P. Hansen. Large sample properties of the generalized method of moments estimator. *Econometrica*, 50(4):1029–1054, 1982.

[34] H. Jänicke and G. Scheuermann. Visual analysis of flow features using information theory. *IEEE Computer Graphics and Applications*, 30(1):40–49, 2010.

[35] H. Jänicke, T. Weidner, D. Chung, R. S. Laramee, P. Townsend, and M. Chen. Visual reconstructiability as a quality metric for flow visualization. *Computer Graphics Forum*, 30(3):781–790, 2011.

[36] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1384–1391, 2007.

[37] T. Jankun-Kelly, K.-L. Ma, and M. Gertz. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):357–369, 2007.

[38] Y. Jansen and P. Dragicevic. An interaction model for visualizations beyond the desktop. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2396–2405, 2013.

[39] D. Keim, G. Andrienko, J. D. Fekete, C. Görg, J. Kohlhammer, and G. Melanon. Visual analytics: Definition, process, and challenges. In *Information Visualization: Human-Centered Issues and Perspectives*, Springer LNCS 4950, pages 154–175, 2008.

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

[41] P. Legg, D. Chung, M. Parry, M. Jones, R. Long, I. Griffiths, and M. Chen. MatchPad: interactive glyph-based visualization for real-time sports performance analysis. *Computer Graphics Forum*, 31(3):1255–1264, 2012.

[42] Z. Liu, N. Nersessian, and J. Stasko. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1173–1180, 2008.

[43] K. Moreland. A survey of visualization pipeline. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):367–378, 2013.

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

[45] D. Oelke, D. Spretke, A. Stoffel, and D. A. Keim. Visual readability analysis: how to make your writings easier to read. In *Proc. IEEE VAST*, pages 123–130, 2010.

[46] H. C. Purchase, N. Andrienko, T. Jankun-Kelly, and M. Ward. Theoretical foundations of information visualization. In *Information Visualization: Human-Centered Issues and Perspectives*, Springer LNCS 4950, pages 46–64, 2008.

[47] M. Ruiz, A. Bardera, I. Boada, I. Viola, M. Feixas, and M. Sbert. Automatic transfer functions based on informational divergence. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1932–1941, 2011.

[48] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014.

[49] J. Scholtz. Developing guidelines for assessing visual analytics environments. *Information Visualization*, 10(3):212–231, 2011.

[50] J. Scholtz, C. Plaisant, M. Whiting, and G. Grinstein. Evaluation of visual analytics environments: The road to the visual analytics science and technology challenge evaluation methodology. *Information Visualization*, 13(4):326–335, 2013.

[51] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[52] D. Silver. Object-oriented visualization. *IEEE Computer Graphics and Applications*, 15(3):54–62, 1995.

[53] S. Takahashi and Y. Takeshima. A feature-driven approach to locating optimal viewpoints for volume visualization. In *Proc. IEEE Visualization*, pages 495–502, 2005.

[54] G. K. L. Tam, H. Fang, A. J. Aubrey, P. W. Grant, P. L. Rosin, D. Marshall, and M. Chen. Visualization of time-series data in parameter space for understanding facial dynamics. *Computer Graphics Forum*, 30(3):901–910, 2011.

[55] J. Thiyagalingam, S. Walton, B. Duffy, A. Trefethen, and M. Chen. Complexity plots. *Computer Graphics Forum*, 32(3pt1):111–120, 2013.

[56] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[57] M. Tory and T. Moller. Rethinking visualization: A high-level taxonomy. In *Proc. IEEE Information Visualization*, pages 151–158, 2004.

[58] C. Upson, T. Faulhaber, Jr., D. Kamins, D. H. Laidlaw, D. Schlegel, J. Vroom, R. Gurwitz, and A. van Dam. The application visualization system: A computational environment for scientific visualization. *IEEE Computer Graphics and Applications*, 9(4):30–42, 1989.

[59] J. J. van Wijk. The value of visualization. In *Proc. IEEE Visualization*, pages 79–86, 2005.

[60] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modelling. *Computer Graphics Forum*, 22(4):689–700, 2004.

[61] P. Vickers, J. Faith, and N. Rossiter. Understanding visualization: A formal approach using category theory and semiotics. *IEEE Transactions on Visualization and Computer Graphics*, 19(6):1048–1061, 2013.

[62] I. Viola, M. Feixas, M. Sbert, and M. E. Gröller. Importance-driven focus of attention. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):933–940, 2006.

[63] T. V. Walsum, F. H. Post, D. Silver, and F. J. Post. Feature extraction and iconic visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2(2):111–119, 1996.

[64] C. Wang and H.-W. Shen. LOD Map - a visual interface for navigating multiresolution volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1029–1036, 2005.

[65] C. Wang and H.-W. Shen. Information theory in scientific visualization. *Entropy*, 13:254–273, 2011.

[66] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547–1554, 2008.

[67] X. Wang, W. Dou, T. Butkiewicz, E. A. Bier, and W. Ribarsky. A two-stage framework for designing visual analytics system in organizational environment. In *Proc. IEEE VAST*, pages 251–260, 2011.

[68] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2013.

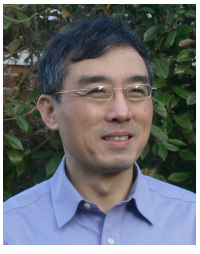
[69] T.-H. Wei, T.-Y. Lee, and H.-W. Shen. Evaluating isosurfaces with level-set-based information maps. *Computer Graphics Forum*, 32(3):1–10, 2013.

[70] P. C. Wong and J. Thomas. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[71] J. Wood, H. Wright, and K. Brodlie. Collaborative visualization. In *Proc. IEEE Visualization*, pages 253–259, 1997.

[72] L. Xu, T. Y. Lee, and H. W. Shen. An information-theoretic framework for flow visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1216–1224, 2010.

[73] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.



Min Chen received the PhD degree from University of Wales in 1991. He is currently a professor of scientific visualization at Oxford University and a fellow of Pembroke College. Before joining Oxford, he held research and faculty positions at Swansea University. His research interests include visualization, computer graphics and human-computer interaction. His services to the research community include papers co-chair of IEEE Visualization 2007 and 2008, IEEE VAST 2014 and 2015, and Eurographics 2011; co-chair of Volume Graphics 1999 and 2006, and EuroVis 2014; associate editor-in-chief of IEEE TVCG; and co-director of Wales Research Institute of Visual Computing. He is a fellow of BCS, EG and LSW.

Amos Golan received his PhD degree from UC Berkeley. He is currently a professor at American University, Washington DC, an external professor at the Santa Fe Institute and the Director of the Info-Metrics Institute (Washington, DC). Before joining the department of economics at American University he held research and faculty positions at the University of Haifa, Israel and UC Berkeley. His main research interests includes an interdisciplinary study of info-metrics (the science and practice of information processing) as well as information, information processing and optimal decision rules based on efficient use of information. He authored and coauthored three books on these topics, edited a number of special issues on information theory and inference in econometric journals, and Chaired (or Co-chaired) many conferences on info-metrics and entropic inference. He is an elected member of the International Statistical Institute (ISI).