An Analysis of Information in Visualisation

MIN CHEN[1] AND LUCIANO FLORIDI[2,1]

[1]*University of Oxford* and [2]*University of Hertfordshire*

Address for correspondence: Professor Luciano Floridi, School of Humanities, University of Hertfordshire, de Havilland Campus, Hatfield, Hertfordshire AL10 9AB, UK; l.floridi@herts.ac.uk

ABSTRACT

Philosophers have relied on visual metaphors to analyse ideas and explain their theories at least since Plato. Descartes is famous for his system of axes, and Wittgenstein for his first design of truth table diagrams. Today, visualisation is a form of 'computer-aided seeing' information in data. Hence, information is the fundamental 'currency' exchanged through a visualisation pipeline. In this article, we examine the types of information that may occur at different stages of a general visualization pipeline. We do so from a quantitative and a qualitative perspective. The quantitative analysis is developed on the basis of Shannon's information theory. The qualitative analysis is developed on the basis of Floridi's analysis in the philosophy of information. We then discuss in detail how the condition of the 'data processing inequality' can be broken in a visualisation pipeline. This theoretic finding underlines the usefulness and importance of visualisation in dealing with the increasing problem of data deluge. We show that the subject of visualisation should be studied using both qualitative and quantitative approaches, preferably in an interdisciplinary synergy between information theory and the philosophy of information.

KEYWORDS

Information map, Information theory, Philosophy of information, Visualisation.

Introduction

Visualisation is a form of 'computer-aided seeing' information in data. As a technical term, 'visualising' refers to different aspects of a visualisation process, primarily in two semantic contexts. *Viewing* concerns the process of specifying significant or noteworthy information, creating appropriate visual representations, and conveying visual representations to viewers. In the literature on computer visualisation, this is explained intuitively in terms of *making visible to one's eyes*. *Seeing* concerns viewers' thought processes and cognitive experiences of interpreting received information and converting the information to mental representations of what the information intends to convey. In the aforementioned literature, this is explained intuitively in terms of *making visible to one's mind*.

The two contexts of viewing and seeing correspond to different parts of a visualisation pipeline, as shown in Figure 1.

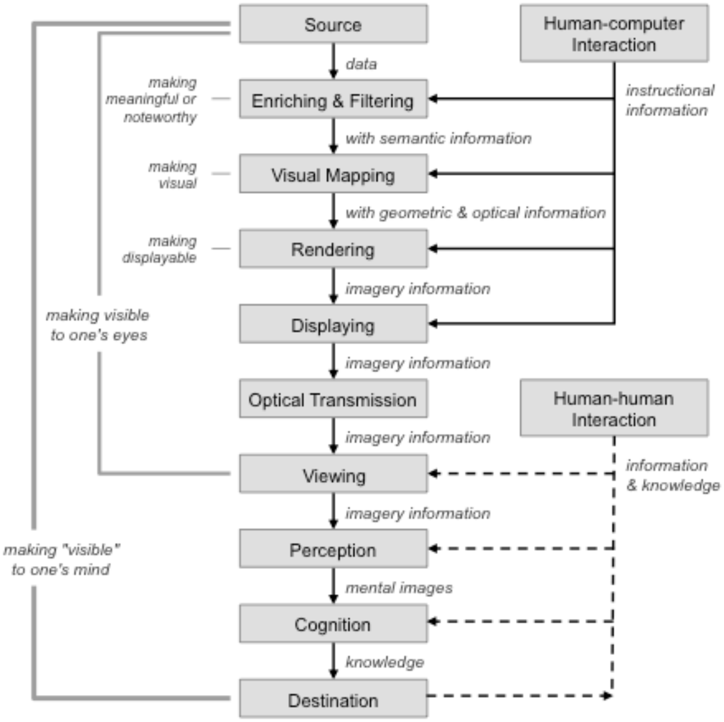

Figure 1 A typical visualisation pipeline.

In *viewing*, one focuses on the parts of a visualisation process that are mediated by some Information and Communication Technology (ICT) system, typically a computer. These include computational algorithms for filtering, visual mapping and rendering, as well as display systems and user interfaces. In *seeing*, one seeks to optimise the usefulness and effectiveness of a visualisation process. Issues addressed in this context typically include the creation of visual metaphors, design of visual representations, and evaluation of visualisation results and user experience.

Consider, for example, how visualization was created in a real-world application (Drocourt 2011). A team of glaciologists compiled a dataset that consisted of 10-year records of seasonal and inter-annual changes in frontal position (advance/retreat) of some 200 marine terminating glaciers in Greenland. A team found that conventional visual representations, such as time-series plots and topographic maps, could not provide an effective overview of the changes of all glaciers while maintaining both the spatial and temporal contexts. A few visualization scientists were thereby asked to help design a more effective visualization. They first *enriched* the data by connecting the names of glaciers with the actual geospatial locations in relation to the geography of Greenland. After observing the glaciologists for a period, they realised that these glaciologists knew the geography of Greenland extremely well. Viewing a Greenland map was mainly for providing a spatial context to the identities of the glaciers rather than geographical information about Greenland herself. The visualization scientists took advantages of this finding to reduce the dimension of the map by *filtering* out some spatial information. This was achieved by *mapping* the coastline of Greenland to a circle, and then *mapping* the spatial location of each glacier to a position on the circle. The two dimensional Cartesian coordinates of a glacier thus became a one-dimensional angular coordinate on the circle. This enabled the temporal dimension to be *mapped* to a spatial dimension represented by radial coordinates in the polar coordinate system. In addition, the visualization scientists and glaciologists worked together to choose a *visual mapping* in which status of advance and retreat of each glacier is *mapped* to two different colours and the levels of changes to the thickness of the circular rings corresponding to different years. When the visualization was first *displayed* to the glaciologists who were able to *view* the whole dataset in a single glance, the new *perceptual* experience stimulated some strong *cognitive* reactions, including new hypotheses about the correctness of some data records, the patterns of changes in different regions, and so forth.

Information is the fundamental 'currency' exchanged through a visualisation pipeline. In this paper, we consider two theoretic frameworks of information. The most well-known, formal theory of information is Shannon's *information theory* (Shannon 1948), which provides a framework for quantifying uninterpreted information, and optimising information coding and communication. Recently, Chen and Jänicke (2010) showed how information theory can explain many phenomena in visualisation processes, including overview-zoom-details interaction, logarithmic visual design, and the use of motion parallax in volume visualisation.

In philosophy, there have been some studies on the topics of information(Floridi, 2011), although the literature is still rather limited when compared to similar efforts about knowledge in epistemology. Floridi (2002) defined the *philosophy of information* as follows:

> DEFINITION Philosophy of information (PI) $=_{\text{def.}}$ the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilisation and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems.

These two theoretic frameworks—information theory and the philosophy of information—encompass our quantitative and qualitative understanding of information respectively. This paper focuses on the *taxonomies of information* in the context of visualisation. We examine how those technical categorisations of information in visualisation are related to the *information map* proposed byFloridi (2010). We then present a scheme that enables the application of the information map to visualisation in a qualitative manner, while accommodating information-theoretic measures quantitatively, through Shannon's theory.

Existing Taxonomic Maps for Visualization

There are many ways in which information in visualisation can be categorised. As shown in Figure 1, one may categorise the input data before it reaches the stage of *Enriching & Filtering*, the graphical models at the intermediate stage between *Visual Mapping* and *Rendering*, or the output imagery information appearing on a display. In addition, instead of categorising information directly, one may consider the tasks and operations for visually processing information, or the interactions allowed in visualisation. It is also common to provide a hybrid scheme, where different categorisations are organised into a hierarchical classification tree, hence a taxonomy.

Many taxonomies proposed for visualisation include categorisation of input data featuring data types, data attributes and application contexts (Wehrend and Lewis 1990; Shneiderman 1996). Tory and Möller (2004) divide input data broadly into two classes: (a) *spatial data* and (b) *non-spatial data*. The former have an inherent spatial component, such as a computed tomography dataset, or a collection of geographic information. The latter typically are not associated with a precise geometric or geographic specification, and require a visual mapping process before the data can be rendered. For example, given a family tree as the input data to the pipeline in Figure 1, the *Visual Mapping* stage has to assign a pair of 2D coordinates to every node in the tree.

For *spatial data*, one tends to consider the dimensions of a spatial domain (e.g., 1D, 2D, 3D, etc.), the presence of a temporal dimension, the measured or computed quantities associated with each spatio-temporal location (e.g., scalars, vectors, tensors, etc.), the underlying data model (e.g., continuous or discrete), and the ways in which data quantities are organized (regular grids, meshes, scattered points, etc.).

For *non-spatial data*, one may categorise data based on the primitive data types (e.g., nominal, ordinal, interval, ratio, etc.), the composite types (enumerated sets, strings, objects, documents, web contents, pictures, voice and sound, videos, etc.), the organization and connectivity of data (e.g., sequences, tabular data, trees, networks, etc.), and the cardinality of attribute space.

A number of taxonomies are based on the tasks of information processing in visualisation (Buja et al. 1996; Zhou and Feiner 1998; Chi 2000;Pfitzner et al. 2003). In visualisation, user operations and tasks can be grouped broadly into three main categories: *information retrieval*, *information analysis*, and *information dissemination*. The category of information retrieval encompasses operations for exploring the data space through overview, browsing, navigation, zooming, observing derived quantities such as data ranges, distributions, errors, certainty and sensitivity, inspecting extracted features such as iso-contours and iso-surfaces, performing deformation on object space, and viewing animated sequences representing spatial navigation or temporal progression. The category of information analysis serves perhaps the most important goal of visualisation for gaining insight from the data. It includes a wide range of analytical tasks, such as finding extrema, anomalies and clusters, sorting, filtering, combining and partitioning data, making comparisons and identifying correlation, and evaluating hypotheses. The category of information dissemination includes operations for presenting information, hence helping others to comprehend the data, such as summarisation, annotation, illustration and animation. One operation common to all three categories is memory externalisation, providing users with efficient means to support future cognitive operations and tasks in information processing with visual representations closer (than the data itself) to users' mental models of the data.

Interestingly, categorisation based on output visual information has not been as common as that for input data and visualisation tasks. Keim and Kriegel (1996) proposed a categorisation based on some common visual representations, including geometric projection, pixel-based, icon-based, and tree and graph. This categorisation has not been widely adopted, partly because the category of geometric projection encompasses a very large collection of visual representations. A meaningful way to divide this category into sub-classes is yet to be found. One alternative approach is to characterise a visualisation output by the visual channels that are used meaningfully in the visualisation. These visual channels include:

- □ Geometric Channels:
  - ○ size / length / width / depth
  - ○ orientation
  - ○ shape
  - ○ curvature
  - ○ smoothness
- □ Optical Channels:
  - ○ intensity / brightness
  - ○ colour / hue / saturation
  - ○ opacity / transparency
  - ○ texture (partly geometric)
  - ○ line styles (partly geometric)
  - ○ shape / blur
  - ○ shading and lighting effects
  - ○ shadow
  - ○ depth (implicit / explicit cues)
  - ○ implicit motion / motion blur
  - ○ explicit motion / animation / flicker
- □ Topological and Relational Channels:
  - ○ connection
  - ○ node / internal node / terminator
  - ○ intersection / overlap
  - ○ depth ordering / partial occlusion
  - ○ closure
  - ○ distance / density
- □ Semantic Channels:
  - ○ number
  - ○ text
  - ○ symbol / ideogram
  - ○ sign / icon / logo / glyph / pictogram
  - ○ isotype

Each visual representation usually makes use of several visual channels. It is also common to use a combination of visual channels to encode concepts and metaphors (e.g., pie and division, stream and flow, safe and dangerous, maps, and so on).

An Information Map for Visualisation

There has been no general agreement on a unified definition of *information*. Shannon ‘philosophically’ commented on the lack of an agreement (Shannon 1993, 180) without much hope:

> The word “information” has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible application of this general field.

Floridi (2005) studied a large collection of definitions of *information*. A popular definition may be paraphrased thus: *information* is *data* + *meaning* (Davis and Olson 1985, 200; Checkland and Scholes, 1990, 303). As we shall see presently, this corresponds in the philosophy of information to the following *weak* definition of information (Floridi, 2011):

DEFINITION information $=_{\text{def.}}$ well-formed and meaningful data.

A stronger definition includes the further condition of truthfulness. In the rest of this article, we shall use information in the previous weak sense, unless specified otherwise.

Although the various taxonomic maps described above have many practical uses in visualisation, it would be contentious to refer to any of them as an *information map*. The categorisation based on input data types captures very little about the *meaning* of the information contained in the data. While it semantically distinguishes one type of data from another, it does not semantically separate one data set from another. Likewise, the categorisation based on operations and tasks encodes the actions on information, but it is totally insensitive to its semantics. Finally, the categorisation based on output visual information is concerned primarily with the forms of visualisation or the mechanisms for delivering information. It also appears to be insensitive to the meaning of the data and hence the information being displayed.

Floridi (2010) proposed an information map by categorising information into several types, as shown in Figure 2. In the rest of this article, we shall adopt it as a taxonomy based on meaning in order to develop a new categorisation of visualisations.

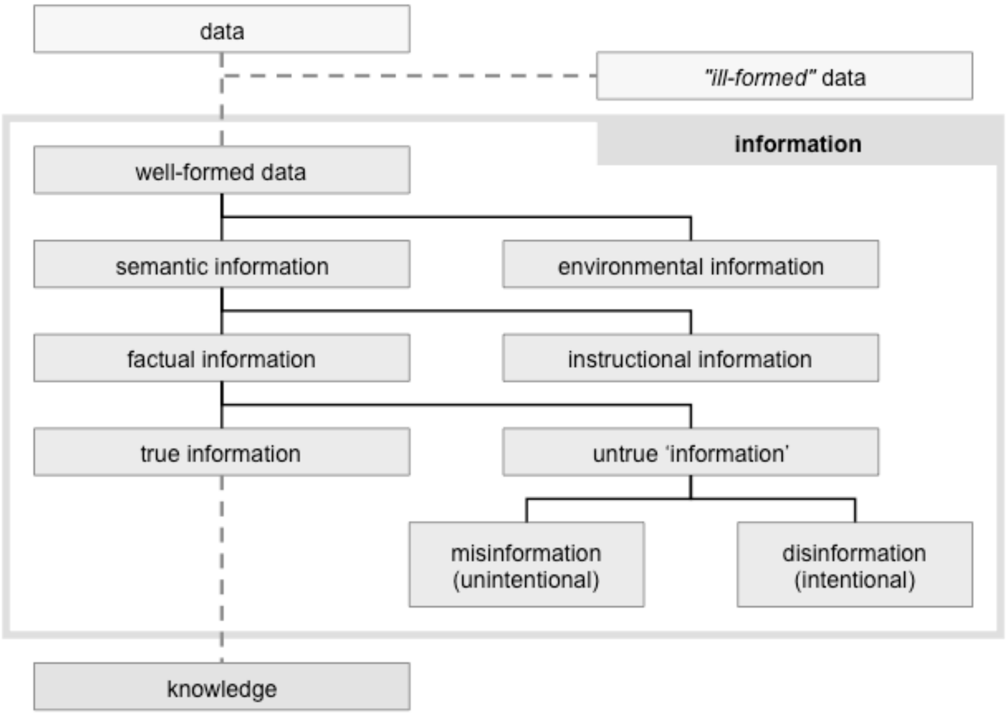

Figure 2 Floridi's original information map, redrawn based on (Floridi, 2011).

Note that the map also indicates how *information* relates *a parte ante* to *data* and *a parte post* to *knowledge* in a hierarchical manner. Strictly speaking, in order to become information, data need to be *well-formed*, *meaningful*, and *truthful*. The first requirement implies that a collection of data has been put together correctly in one or more data sets according to the rules (*syntax*) of the chosen code (usually a combination of natural and formal languages). The second requirement implies that the data must also comply with the meanings (*semantics*) of the chosen code. The third requirement allows one to distinguish, in a strict sense, between information and mis- or disinformation (untruthful data).New knowledge can then be built upon available information and existing knowledge through various cognitive processes, such as learning, association, and reasoning.

As defined in(Floridi, 2010), the different categories or sub-categories of information are:

□ *Environmental* (also known as *natural*) *information*—this is well-formed data (patterns) *as* something, e.g., the series of concentric rings visible in the wood of a cut tree trunk correlated to its age;

□ *Semantic information*—this is well-formed and meaningful data, that can be analysed as

- *Instructional information*—this is semantic information *for* something, e.g., 'open the door!';or
- *Factual information*—this is semantic information *about* something, e.g., 'the door is open'; this in turn can be
    - *True information*—this is semantic content (well-formed and meaningful data), which is also truthful; also known as semantic information, or simply information. The lack of precision may generate confusion, but contexts often resolve the ambiguity. As indicated above, in this paper 'information' is used both in its weak and in its strong sense, with further specifications whenever the distinction is unclear;
    - *Untrue 'information'*—this is pseudo information (cf. false friend, who is not a friend at all), equivalent to semantic content (well-structured and meaningful data) which is not truthful; it is further analysable as
        - *Misinformation*—pseudo information accidentally or unintentionally untruthful, e.g., a mistake; and
        - *Disinformation*—pseudo information purposefully or intentionally untruthful, e.g., a lie.

An interesting question is how the previous categorisation by meaning is related to various categorisations by input, output and process. Comparing Figures 1 and 2, we can make the followings observations.

Once the data have entered into the visualisation pipeline in Figure 1, we can assume that the data have been parsed correctly. In other words, the data is well-formed, since it would have otherwise been thrown out by the syntactic parser. We can also assume that the pipeline does not generate syntactic errors within the system, or has a mechanism to detect and correct such syntactic errors. Hence, with conditions (1) and (2), all data in the pipeline may be assumed to be well-formed. After the first stage of processing, meaningless data will either be filtered out or enriched with additional semantic tagging. All data at the end of this stage are both well-formed and meaningful. Hence, they constitute information according to the definition by Davis and Olson (1985, 200), or semantic content (information in the weak sense), according to Floridi (2011). Arguably, none of the processing stages afterwards will deliberately remove semantic associations. Even when some processing stages do remove some semantic associations by mistake, we assume that such associations can be recalled from the stage of *Enriching & Filtering*. If one can assume that the data are truthful, then this ensures that all data in the pipeline are information (in the strong sense) from that stage onwards. However, we do include the possibility that initial errors, or computational errors, or sampling noise may be introduced at each processing stage. Because human-computer interaction is allowed in the pipeline, and the software involved might be faulty and unreliable, cases of *misinformation* and *disinformation* may occur and information (in the strong sense) may become corrupted. The interested reader may wish to consult Tufte (2001) for a collection of interesting examples.

It is possible, although neither intuitive nor useful in this context, to consider the digital information in the pipeline as *environmental information*, that is, as mere patterns to be interpreted. Rather, it is preferable to consider all information after the stage of *Enriching & Filtering* as *semantic information*. We call this *the principle of presumed informativeness*: data are considered well-formed, meaningful, and truthful until proven otherwise. It is also possible to classify some information in the pipeline as *instructional information*, since one of the goals of a class of visualisation techniques, namely illustrative visualisation, is usually instructional (think, for example, of the visual instructions usually accompanying ready-to-assemble furniture). Finally, in visualisation, it is common to introduce various forms of abstraction for more effective perception and cognition, which usually involve omission of some information in the resultant visual representations. In many applications, the size of the data set concerned is too large for a visualisation to depict all the information contained. A decision will have to be made, either by the system or by the users, to leave out some information of the resultant visual representations.

Therefore, it is useful to underline a specific category of information that gets lost during the process.

Based on the above observations, we can conclude that, in general, Floridi's categorisation by meaning is applicable to the information in a visualisation pipeline. Some may suggest that it might be helpful to introduce new sub-categories of information, such as geometric and optical information, into the information map. We consider this unnecessary, mainly because such information is in a transitional status before a visual representation is produced by the *Rendering* stage. The visualisation being viewed by the users, that is, the imagery information as labelled in Figure 1, is in the most important as well as stationary state of the pipeline. So we can, and should, focus on the information in this particular state. Figure 3 shows an information map, which has been slightly modified based on Figure 2, in order to illustrate its relationship with the pipeline as well as to highlight the category of lost information.

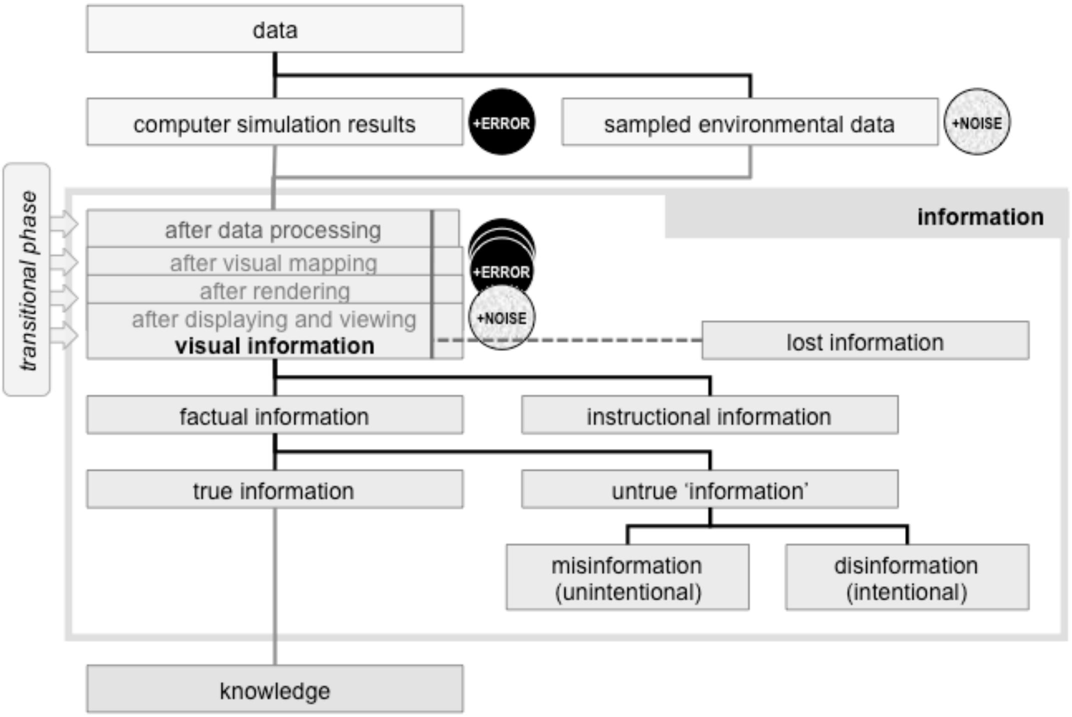

Figure 3 An information map for visualisation.

An Information-theoretic Framework for Visualisation

Information theory is a branch of probability theory. It was first developed by Claude E. Shannon (1948) in the context of communication systems, focusing on data compression, error detection and correction. Since then, this theoretic framework has been further developed and has found applications in many disciplines, including image processing and computer vision. Recently, Chen and Jänicke (2010) showed how information theory can be fruitfully applied to many aspects of visualisation, and they made a case for information theory to be an underpinning theoretic framework of visualisation. However, they also pointed out areas of visualisation where information theory cannot be naïvely applied without adaptation and extension, due to the semantic nature of information. *Visualisation* concerns visually coding and communicating *meaningful* data. When we consider only the central path of the visualisation pipeline in Figure 1, that is, without human-computer and human-human interaction, we can observe that it is very similar to a communication system or a data processing pipeline. Although human-computer and human-human interaction is an ordinary phenomenon and is highly valuable in visualisation, it is not absolutely compulsory. In general, our affordability for human-computer and human-human interaction will always be limited, whereas we will continue to increase our access to more computational power in the central path, up to the *Displaying* stage. Therefore, it is not an over-simplification to take a first look at the central path of the visualisation pipeline in Figure 1.

Let us first examine the similarities and differences between a communication system and the central path of a visualisation pipeline without human-computer and human-human interaction. The two are similar in the following ways.

System Structure. A communication system is typically composed of a series of sub-systems. Some sub-systems may modify the messages for various reasons, such as conversion between different standards, strengthening the security, and so on. Others may perform a simple function of relaying messages, which are referred to as 'store and forward'. A visualisation pipeline as shown in Figure 1 can also be seen as a series of sub-systems.

System Abstraction. Shannon (1948) defines a basic communication system as a pipeline connecting *source*, *encoder*, *channel*, *decoder* and *destination*. Chen and Jänicke (2010) show that a visualisation pipeline, without interaction, can also be abstracted into five basic components as a basic communication system. These two abstract models are illustrated in Figure 4.

Objective. The primary objective of a communication system is to transfer messages from a source to a destination as accurately and quickly as possible. Visualisation has a comparable objective. By transforming information contained in the original data into an appropriate visual representation, the goal is to enable the viewers to gain an insight about the data quickly and accurately.

Information Loss. Although many forms of communication are lossless (e.g., emails and file transfer), some are lossy (e.g., voice over internet protocol, and video conferences). For small data sets, it is possible to preserve all the information from the source in the resultant visualisation. For large data sets, visualisation is usually a lossy process.

Errors and Noise. Both communication systems and visualisation pipelines are subject to errors and noise.

Probabilistic Nature. Many aspects of a visualisation pipeline feature events and phenomena with probabilistic certainty or uncertainty, bearing a striking resemblance with a modern communication pipeline. For example, messages in a communication system are not guaranteed to reach their destination, while information in a visual representation is not guaranteed to be received by a viewer. The quality of a communication system is typically measured by sampled probabilistic quantities, while the quality of visualisation is often measured by probabilistic experiments.

Semantic Awareness. There has often been a misconception that a communication system does not require any understanding of the information being transmitted, whereas the visualisation pipeline does so. First, both involve some responsiveness to the semantic content passing through the system. A modern communication system usually applies different compression algorithms to different types of messages (e.g., text, voice and video). In some cases, some basic forms of meaning are detected. For instance, a piece of text may be compressed using a dictionary-based method, or a piece of phone conversation may be compressed using salience detection and removal. Similarly, a visualisation pipeline usually expresses a good knowledge of input data types (e.g., volume data, network data). Sometimes, data may be further classified by using a feature classifier or a transfer function. In fact, the goal of the *Visual Mapping* stage is to encode the semantics made available using geometric and optical information. Second, given the functional nature of a communication system and a visualisation pipeline, it is not appropriate for either to have to focus too much on the semantic content passing through the system. For a communication system, handling too much semantic content will undermine the privacy and security requirements for such a system while seriously affecting its performance. As an enabling technology, the goal of visualisation is to help viewers to interpret data, especially in situations where analytical tools are not ‘smart’ enough to draw useful and reliable conclusions from the data. For instance, given a 3D computed tomography, it may be acceptable for a medical visualisation system to highlight the regions of interest. However, at the moment it is simply not feasible for a system to detect a tumour automatically and then show it to a doctor instructively.

The communication systems and visualisation pipelines are also different in several ways:

Type Compatibility. A communication system normally ensures that the messages received from the source will reach the final destination more or less in the same data type. As shown in Figure 1, a visualisation pipeline almost always transforms the information from the source to imagery information, if we consider the viewers as the destination. In most situations, the information in the source data and the generated visual information will not be of the same form. There are of course some exceptions, such as tag cloud visualisation. In fact, the type of the information at any stage of the pipeline after *Visual Mapping* is expected to differ significantly from the original type.

Compactness. The encoding scheme in communication system places a huge emphasis on data compression. In particular, source coding for noiseless channels focus almost solely on the compactness of the messages to be transmitted. In a visualisation pipeline, the visual encoding stages (i.e., *Enriching & Filtering*, *Visual Mapping*, and *Rendering*) often result in output that requires more space than the input, especially for a small data set from the source. For example, when a pie chart shows a dozen of percentage values, the space requirement for the display is much more than that for the numerical representation of the percentage values.

Human Involvement. Though we focus here on the central path of Figure 1, it is important to point out that the most significant difference between typical visualisation pipelines and communication systems is the involvement of humans. Shannon’s model of communication assumes that humans do not participate in any operations of the three main components, *encoder*, *channel* and *decoder*. However, the decoder stages of a visualisation pipeline (i.e., *Perception* and *Cognition*) are essentially human-centred components. In addition, human-computer and human-human interaction brings about further human involvement in the visualisation.

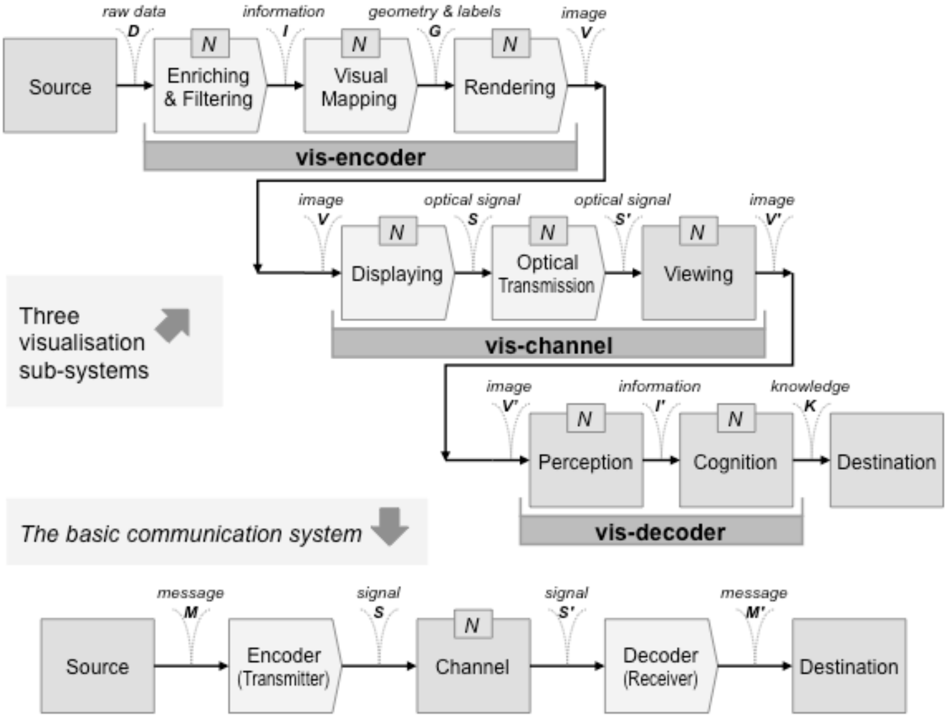

Figure 4 Abstract models of visualisation and communication.

To summarise, on the one hand, the *similarity* between communication systems and visualisation pipelines is significant enough to warrant the consideration of information theory as a theoretic framework to underpin visualisation. The role of such a theoretic framework is to house a collection of fact-based theorems consistently for explaining observed phenomena or events in visualisation, to provide quantitative means for measuring the properties and attributes of observed phenomena or events in visualisation, to enable the discovery of new theorems inferentially, and to test and falsify conjectures proposed for visualisation. In fact, up to the time of writing this paper, there has not been any serious proposal for other alternative theoretic framework in the field of visualisation, though attempts have been made to draw inspiration from other theoretic or conceptual frameworks in computer science (including logic, AI, and software engineering), psychology, and linguistics. An exception is Chen and Jänicke (2010), whose work gave several examples where theorems, rules and measures in information theory can be used to explain phenomena and events in visualisation.

These examples include logarithmic plots, overview and details-on demand, redundancy, and motion parallax in visualisation. They also outlined a collection of possible correspondence between information theory and visualisation, from which one may find further examples of using information theory in visualisation. To avoid repetition, readers are encouraged to consult the original paper (Chen and Jänicke, 2010) for details.

On the other hand, the *difference* between communication systems and visualisation pipelines is also significant enough to suggest that one needs to be cautious when applying information theory to visualisation. We should not indiscriminately apply information measures to quantifying properties and attributes in visualisation without considering the underlying probabilistic space, which often encodes some human factors within the pipeline or external factors that enter the pipeline with data. Because of the involvement of humans, some of such factors are intrinsically semantic and very difficult to capture and measure quantitatively or syntactically. Hence, the probabilistic space underlying probability mass functions becomes undefined. At the same time, we should not be deterred by any situation where information theory cannot currently offer a satisfactory explanation or measurement. We know that much of information theory assumes that an information source or a channel is memoryless. Such an assumption is critical to the derivation of many theorems in information theory. Nevertheless, this does not imply that all components in a communication system are memoryless. It merely suggests that an application of information theory to a system is valid when the memory factor is negligible in the system. By carefully defining the underlying probabilistic space, one can also mitigate the effect of memory. However, it is highly desirable to broaden information theory by combining it with a philosophy of information that can provide better analytical methods for memoryful systems and fully semantic features in the visualisation system. Information theory is a scientific subject that is continuously being developed. Theoretic research in the context of visualisation will no doubt contribute to its expansion.

One of the well-known concepts in information theory is the *data processing inequality* (Cover and Thomas, 2006). This is a widely accepted principle among scientists and researchers in data processing, and many other areas of computer science. Without being drawn into the mathematical details of the data processing inequality, we can explain it with the aid of Figure 5.

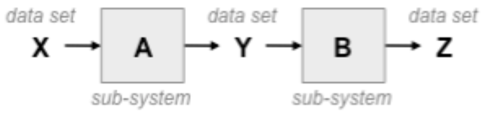

Figure 5 A typical pipeline that may meet the Markov chain condition

Consider two data processing sub-systems, A and B. Sub-system A takes an input data set X, processes it according to a pre-defined algorithm, and generates an output data set Y. The next sub-system B in the pipeline takes data set Y as the input, processes it according to another pre-defined algorithm and generates data set Z as an output. In information theory, *mutual information* is a quantity that measures how much information one data set holds about another data set. Mathematically, it is trivial to prove that the measurement is symmetric, that is, the two data sets contain the same amount of information about each other. This is why it is called ‘mutual’ information. The data processing inequality theorem states that the information contained in Z about X cannot be more than the information contained in Y about X. In other words, the information about X can only decrease or be kept the same after it has been processed. This is an intuitive and sensible conclusion that is consistent with common sense. Mathematically, the proof of the data processing inequality theorem is based on the assumption that the pipeline, X to A to Y to B to Z in Figure 5 is a *Markov chain*. This implies that X and Z are conditionally independent, given Y. In other words, sub-system B does not have any direct information about X except for the information included in data set Y. Alternatively, even if B has direct access to such information, for some reason, it is assumed that it cannot, or will not, make use of such information in producing output Z.

In visualisation and, in fact, in many data processing systems in which semantic information has a role to play, the Markov chain condition is usually broken. When it is, the data processing inequality may not hold. When it does not, we are then able to increase the information in Z about input data set X.

There are three different ways in which one may break the Markov chain condition:

(a) One may allow users to interact with sub-system B. If the users have some knowledge about X, which is not encoded in the intermediate data set Y, such knowledge can be used to influence the production of Z, and hence improve the inferences that can be made from data set Y.

For example, let X be an itemised list of five products sold by a store in December. Sub-system A computes a statistical summary based on X and outputs data set Y containing [Jackets:15, Trousers:18, Shirts:28, Ties:5, Shoes: 34]. Sub-system B offers several visual representations (e.g., bar chart, bubble chart, pie chart) to display Y. The user interactively selects the pie chart to display the summary data, and types in a caption 'proportion of sales in £ in December'. In this case, the interaction has brought back some information lost in data set Y. The pie chart is a well-understood metaphor for proportional partition, from which one can infer that the 5 values are percentage values. The caption adds further information that the partitioning is in terms of sales values rather than numbers, and is about December rather than other months.

(b) One may encode some knowledge about X in sub-system B.

For example, data set X may contain a 400x400 grid of sampled temperature values in a range between −40°C and 40°C. Data set Y is a set of contour lines computed from Y, representing temperatures {−40°C, −20°C, 0°C, +20°C, +40°C} respectively. Sub-system B assigns purple, blue, white, yellow and red colours to the contour lines at the five different temperatures, and then computes colours for pixels between each pair of neighbouring contour lines by smoothly transforming one contour colour to another. Hence the resultant visualisation Z includes information between contour lines that was removed by sub-system A. As long as the colour interpolation algorithm in sub-system B matches reasonably well with the transitional patterns of temperatures in X, Z has more information about X than Y.

(c) One may allow sub-system B to seek extra information about X, based on the information contained in Y, from an external sub-system.

For example, X is a simulation model with a control parameter $0 \leq t \leq 1$. Sub-system A runs the model with $t = 1$ and produces a data set Y. Sub-system B is for generating a visualisation from Y. B is aware that the same model has been run previously with $t = 0, 0.1, 0.2, ..., 0.9$, and retrieves the visualisations generated previously for other t values from a database. B then compares the current visualisation with the previous ones, and illustratively highlights those parts of visualisation most different from the previous ones. This is resulting in a final visualisation Z with sensitivity annotation. Z thereby contains information about X that is not in Y.

One may object that, even though we are able to increase the information to the same level as it is contained in data set X, we cannot gain more information than that which enters into a visualisation pipeline. The original input data set X therefore should set the upper limit. However, let us consider the sub-system B as a visualisation system, and the data set Y as the input data set to the visualisation system. Sub-system A merely represents a process that obtains a data set to be visualised from a data source. For example, the process can be a computational simulation of a dynamic model, which generates a set of values representing the changes of some attributes at some discrete time intervals. In this case, the dynamic model is X, and the set of values is Y. Can the visualisation system B generate more information than that contained in the input data set Y? Of course, the answer is yes if we can break the Markov chain condition. Note that this reasoning can be extended to more complex situations by adding more sub-systems at the beginning of the pipeline. For instance, X may be a special case of a more general model W, which may be a model approximating a natural phenomenon V. Can the visualisation system B potentially help the users to gain an insight about the natural phenomenon V? Again, the obvious answer is yes. This is because the users who interact with sub-system B have the knowledge about the general model W and the phenomenon V. Hence the output Z does not solely dependent on Y, but also on X, W and V.

A difficult point to address concerns how the increase in information, gained through visualization by breaking (the conditions of) a Markov Chain, may be measured. The new information gained largely depends on (a) the human users' interaction with the visualization system, (b) the information that is hardcoded in the system, or (c) the information in a knowledgebase that can be assessed dynamically by the system. Now, (a) is potentially measurable by considering the parameter space of a system as the information space. There have been theoretically attempts to measure (b)[1] but, so far, there is no practical solution to the problem of measuring information quantity in a program. Finally, (c) features a combination of (a) and (b), and thus inherently is as difficult as (b).

The usefulness of human-computer interaction in a visualisation pipeline has been widely appreciated in the field of visualisation, but it has never been explained in terms of information theory until the work by Chen and Jänicke (2010). Despite the fact that breaking the essential condition of data processing inequality is an everyday phenomenon in visualisation,[2] we only recently realised that breaking such a condition in significant and substantive ways may hold the key to address the increasing problem of data deluge. Since human-computer interaction requires time and human resources, the amount of interaction that we can afford will always be limited. This indicates that the aforementioned approach (a) does not scale well in the long run. Meanwhile, hard-coding too much application-specific or data-specific information in a visualisation system will be

[1] On measuring information in a program, see works on algorithmic information theory such as Chaitin (1975) and Claude (1996).

[2] Floridi (forthcoming) argues that breaking the condition of data processing inequality is essential in order to explain non-natural (i.e., conventional, artificial, synthetic) meanings, thus complementing the naturalist tradition, which seeks to account for non-natural meanings by reducing them entirely to natural ones through signalling or information theory.

costly, in terms of software engineering, and restrictive, in terms of software deployment. This places an engineering constraint on approach (b). In comparison with (a) and (b), the adoption of approach (c) is clearly to be preferred. Chen *et al.* (2009) introduced the notion of *knowledge-assisted visualisation* to highlight the potential merits of capturing and reusing knowledge in a visualisation pipeline. Their results are consistent with approach (c) for breaking the condition of the data processing inequality and with the analysis of semantic information proposed by Floridi (2011).

This brings us back to the information map illustrated in Figure 2. While it is useful to break the condition of data processing inequality, it is also necessary to realise that the semantic information added into the visualization pipeline through the abovementioned ways can be true as well as untrue. Furthermore, there is limited control over the perception and cognition stages of the pipeline, and hence over the insight gained by individual viewers of a visualization. This poses a fundamental as well as a practical question about the *quality of visualization*, which is a part of a more general philosophical question about the *quality of information*. On the one hand, visualization has a crucial role in dealing with data deluge. On the other hand, like almost all mechanisms for information processing and communication, there will be opportunity for mis- and disinformation. These are important but still open questions, which we are investigating in our current research.[3]

Conclusion

The information map presented in Figure 3 introduces a qualitative outlook of the visualisation pipeline. The first half of the central path (up to *Optical Transmission*) is traditionally studied in a quantitative manner using various measures, ranging from information-theoretic measures (e.g.,

[3] See the AHRC-funded project "Understanding Information Quality Standards and their Challenges (2011-2013)", directed by Luciano Floridi as PI: http://www.philosophyofinformation.net/IQ/AHRC_Information_Quality_Project/Home.html.

entropy, mutual information) to algorithmic measures (e.g., complexity, speed, space usage). More and more emphasis has been placed on quantitative analysis of the second half of the central path, for instance through user studies. Nevertheless, what the users are really interested in is the semantic content of the information, which is traditionally studied in a qualitative manner. This presents us with two options, which are not mutually exclusive. We could conduct more qualitative research on the subject of visualisation. While there have already many case studies on specific applications of visualisation, there is no reason to suggest that we should not study human subjects in depth to gain a better understanding of how a form of visualisation is learned and used in a common or specialised contextual setting. For instance, many of us have been enlightened by Oliver Sack's books (Sacks, 1986; 2010). We would equally be informed by case studies such as 'the man who mistook a treemap for ...' It is also necessary to state that qualitative research does involve data collection, data analysis, and validation. At the same time, we could introduce more quantitative measures to describe meanings and related concepts. Many abstract concepts already have parallel concepts that are quantifiable, for example, accuracy, Kullback-Leibler information, and so on. Many abstract concepts that were not quantifiable one or two millennia ago (e.g., force, heat, etc.) are quantifiable today. It would certainly be useful if scientists could devise an information map consisting of quantifiable concepts corresponding to those in Figure 3. *We should pursue both options*.

Acknowledgements

We would like to acknowledge the useful comments by the two anonymous reviewers. Penny Driscoll skilfully copyedited the final version.

Bibliography

Buja, A., D. Cook, D. F. Swayne. 1996. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5:78–99.

Claude, C.S. (1996) Algorithmic information theory: Open problems, Journal of Universal Computer Science, 2: 439–441.

Chaitin, G. J. (1975) A Theory of Program Size Formally Identical to Information Theory, Journal of the Association for Computing Machinery, 22(3): 329–340.

Checkland, P. B., J. Scholes. 1990. *Soft Systems Methodology in Action*. John Wiley & Sons, New York.

Chen, M., D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann and D. Silver. 2009. Data, Information and Knowledge in Visualization, *IEEE Computer Graphics and Applications*, 29(1):12-19.

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

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

Cover, T. M., J. A. Thomas. 2006.*Elements of Information Theory*. John Wiley & Sons, 2006.

Davis, G. B., M. H. Olson. 1985. *Management Information Systems: Conceptual Foundations, Structure, and Development*, 2nd ed. McGraw-Hill, New York.

Drocourt, Y., R. Borgo, K. Scharrer, T. Murray, S. I. Bevan and M. Chen. 2011. Temporal visualization of boundary-based geo-information using radial projection, *Computer Graphics Forum*, 30(3):981-990.

Floridi, L. 2002. What is the philosophy of information? *Metaphilosophy*, 33(12): 123-145.

Floridi, L. 2010. *Information: A Very Short Introduction*. Oxford University Press.

Floridi, L. 2011. *The Philosophy of Information*. Oxford University Press.

Floridi, L. Forthcoming. Perception and Testimony as Data Providers. *Logique et Analyse*.

Keim, D. A., H.-P. Kriegel. 1996. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(16):923–936.

Sacks, O. 1986. *The Man Who Mistook His Wife for a Hat.* Pan Macmillan.

Sacks, O. 2010. *The Mind's Eye*. Pan Macmillan.

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

Shannon, C. E. 1993. *The Lattice Theory of Information*.

Shneiderman, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symposium on Visual Languages*, 336–343.

Tory, M., T. Möller. 2004. Rethinking visualization: A high-level taxonomy. In *Proc. IEEE Information Visualization*, 151–158.

Tufte, E. R. *The Visual Display of Quantitative Information*. 2nd Ed. Graphics Press USA, 2001.

Wehrend, S., C. Lewis. 1990. A problem-oriented classification of visualization techniques. In *Proc. IEEE Visualization*, 139–143.

Zhou, M. X., S. K. Feiner.1998.Visual task characterization for automated visual discourse synthesis. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 392–399.