Chapter 3

Theoretical Foundations of Information Visualization

Helen C. Purchase[1], Natalia Andrienko[2], T.J. Jankun-Kelly[3], and Matthew Ward[4]

[1] Department of Computing Science, University of Glasgow,
17 Lilybank Gardens, G12 8QQ, UK,
hcp@dcs.gla.ac.uk

[2] Fraunhofer Institut Intelligent Analysis & Information Systems (FhG IAIS),
Schloss Birlinghoven, D-53754 Sankt-Augustin, Germany,
natalia.andrienko@iais.fraunhofer.de

[3] Department of Computer Science and Engineering,
Bagley College of Engineering, Mississippi State University,
Mississippi State, MS 39762, USA,
tjk@acm.org

[4] Computer Science Department, Worcester Polytechnic Institute,
100 Institute Road, Worcester, MA 01609-2280, USA,
matt@cs.wpi.edu

The field of Information Visualization, being related to many other diverse disciplines (for example, engineering, graphics, statistical modeling) suffers from not being based on a clear underlying theory. The absence of a framework for Information Visualization makes the significance of achievements in this area difficult to describe, validate and defend. Drawing on theories within associated disciplines, three different approaches to theoretical foundations of Information Visualization are presented here: data-centric predictive theory, information theory, and scientific modeling. Definitions from linguistic theory are used to provide an over-arching framework for these three approaches.

1 Introduction

Information Visualization suffers from not being based on a clearly defined underlying theory, making the tools we produce difficult to validate and defend, and meaning that the worth of a new visualization method cannot be predicted in advance of implementation. There is much unease in the community as to the lack of theoretical basis for the many impressive and useful tools that are designed, implemented and evaluated by Information Visualization researchers.

The purpose of a theory is to provide a framework within which to explain phenomena. This framework can then be used to both evaluate and predict events, in this case, users' insight or understanding of visualization, and their use of it. An Information Visualization theory would enable us to evaluate visualizations with

reference to an established and agreed framework, and to predict the effect of a novel visualization method.

This is not to say that a single theory would be able to encapsulate the whole of the Information Visualization field; it may be that multiple theories at different levels are needed. We already make use of many existing cognitive and perceptual theories, as well as established statistical methods. It might be that the complexity of Information Visualization as relating to engineering, cognition, design and science requires the use of several theories each taking a different perspective.

As a starting point, we liken the understanding of visualization to the understanding of ideas expressed in language. We draw on two perspectives in linguistic theory: language as representation and language as process.

In considering the representation of language, lexical tokens are syntactically ordered to produce a semantic concept which the reader understands with reference to a learned code. The concept is understood within a context and the reader responds pragmatically.

We can take this approach a step further and consider the semiotic theory of Saussure [1] wherein a sign is a relation between a perceptible token (signifier, referrer) and a concept (signified, referent) – giving us another useful term: the referent of the token, i.e. what it actually means in a given context. For example, a pair of numbers (the referrer) may mean a geographical location in one context (one possible referent), yet may mean a student's examination and coursework marks in another (a different referent).

Similarly, we can extend our consideration of pragmatics to include stylistics: the style in which language is written. The same pragmatic response may be stimulated by a different set of tokens (or the same set arranged with a different syntax) – this may produce a different emotional response.

This Saussurian view of language as a static representation of meaning can be contrasted with the view of Bhaktin [2] who considers language to be a dynamic process whereby a text is interacted with and manipulated, and its meaning constructed dynamically. This active and engaged understanding, Bhaktin says, creates new meanings: "… establishes a series of complex interrelationships, consonances and dissonances … [and] various different points of view, conceptual horizons … come to interact with one another" [2].

While Bhaktin's theory of the dynamic interpretation and negotiation of linguistic texts was primarily based around the social context of language interpretation and the construction of ideologies within cultures and institutions, it is a useful complement to the "language as representation" perspective presented above. We can use this alternate view of "language as a process" in our framework for Information Visualization: a model of data embodied in a visualization must be explored, manipulated and adapted within a investigatory process resulting in enhanced understanding of its meaning. When there are two processing agents in a human-computer interaction context (the human and the computer), either or both can perform this processing.

Thus, we can relate discussion of theoretical approaches to Information Visualization to the concepts of

- Interpretation of a visualization through its external physical form (referents, and lexical, syntactic, semantic, pragmatic and stylistic structures), an activity typically performed by a reader;
- Exploration and manipulation of the external representation by the reader so as to discover more about the underlying model, typically done through interaction facilities provided by a visualization tool; and
- Exploration and manipulation of the internal data model by the system in order to discover interrelationships, trends and patterns, so as to enable them to be represented appropriately.

The three sections that follow each take a different approach to suggesting a theory for Information Visualization. While they were not originally developed with the above linguistic model in mind, each can be related in some way to this framework.

Natalia Andrienko takes a data-centric view, focusing on the dataset itself, and the tokens that describe it. She considers how the characteristics of the dataset and the requirements of the visualization for a task may be matched to determine patterns, thus predicting the most appropriate visualization tool for the given task. Thus, this section describes the exploration of the data model so as to identify the best syntax to use for given tokens (taking into account their referents and the desired semantics). She highlights the usefulness of systems which can explore the data model, predict the patterns in datasets, and facilitate the perception of these patterns.

Matthew Ward's starting point is communication theory, and this section is clearly focused on information content – the meaning of the visualization and maintaining the flow of information through all stages of the visualization pipeline. He discusses how we may assess our progress in designing and enhancing visualizations through considering measurements of information transfer, content or loss, thus providing a useful theoretical means for validating visualizations. In this case, there is no internal exploration of the data, but it is the validity of the data after transfer from internal model to external representation that is considered important.

T.J. Jankun-Kelly introduces two useful models for a scientific approach to visualization, both of which are in their infancy. The visual exploration model describes and captures the dynamic process of user exploration and manipulation of visualization in order to affect its redesign, thus using the pragmatic response of the user to determine a new syntactical arrangement. The second model, visual transformation design, uses transformation functions applied to the data model to provide design guidance based on visualization parameters, thus performing an initial exploration of the data model to suggest syntax to enhance the pragmatic response of the user.

The chapter concludes with a summary, and suggestions for future research.

2 Predictive Data-centered Theory

Among other theories, Information Visualization requires a theory that could serve as a basis for instructing Information Visualization users how to select the right tools for their data and do data exploration and analysis with the use of these tools. The same

theory could also help tool designers in finding right solutions. The following argumentation is meant to clarify what kind of theory this could be.

Most Information Visualization researchers agree that the primary purpose for using Information Visualization tools is to explore data in order to gain understanding of the data and the phenomena behind. Gaining understanding may be thought of as constructing a concept, or mental model, of the data or phenomenon. A model, in turn, can be considered as a parsimonious representation capturing essential features of the data rather than listing all individual data items; this means that a model necessarily involves abstraction. For example, from observing morning temperatures over several days, a person may build a concept of the increase or decrease of the temperature.

Such an abstraction is based on a holistic grasp of characteristic features embracing multiple data items. We shall use the term "pattern" to refer to such features. Increase and decrease are examples of patterns. A model may be a synthesis of several patterns each representing some part or aspect of the data. Thus, when the observation of the morning temperatures is performed over a sufficiently long period, the model will probably incorporate the patterns of both increase and decrease of the temperature. Furthermore, patterns may also be composed of sub-patterns. For instance, the behavior of the temperature may be conceptualized as a repeated "wave" where increase is followed by decrease. Here, increase and decrease are basic, or atomic, patterns, the "wave" is a composite pattern including the increase and decrease patterns, and the repetition of the "wave" is a pattern of a yet higher level, which incorporates the "wave" pattern.

The main role of Information Visualization tools can be understood as helping the user to perceive patterns that could be used for building an appropriate model. This means, in particular, that a tool should facilitate the perception of (sub)sets of data items as units. For an appropriate support of the detection of patterns, a tool designer should know in advance what *types* of patterns need to be perceived (or otherwise detected) with the use of the tool. Then, after the tool is ready, it will be easy to explain to the users the purpose of the tool and instruct them how to detect the types of patterns the tool is oriented to.

The types of patterns that may be meaningful for the user depend on the structure and properties of the data under analysis. Thus, in the analysis of a temporal series of numeric measurements (such as temperatures) it makes sense to look for such basic patterns as increase, decrease, stability, fluctuation, peak, and low point. However, when numeric measurements refer to a discrete unordered set as, for example, melting temperatures of various substances, the possible types of patterns may be groupings of elements with close values of the measurements and frequency-related patterns: prevalence of certain values or value intervals, frequent values or exceptional values (outliers).

To support the designers and users of Information Visualization tools in the way described above, there is a need for a theory that could enable the possibility to predict, for a given dataset or a given class of datasets, what *types* of patterns may be found there. We specially emphasize the term *types* to exclude the possible impression of attempting to predict (and on this basis automatically detect) all specific patterns hidden in specific data. Thus, a prediction that a dataset may contain groups (clusters) of objects with similar characteristics does not define what specific clusters

are there. However, it orients tool designers, who will know that the tool must help the users to detect clusters, and users, who will know that they need a tool facilitating the detection of clusters. Then, if each Information Visualization tool and technique is supplied with an appropriate signature (i.e. what kind of data it is suitable for and what types of patterns it is oriented to), the user will be able to choose the right tool.

The theory we are advocating in this section can be called data-centered predictive theory. The theory needs to include

1. an appropriate generic framework for the characterization of various data types and structures;
2. a general typology of patterns;
3. a mechanism for deriving possible pattern types from data characterizations.

Here, we present some preliminary ideas concerning these components of the theory.

2.1 Data characterization framework

Data may be viewed abstractly as a set of records with a common structure, each record being a sequence of elements (such as numbers or strings) which either reflect the results of some observations or measurements or specify the context in which the observations or measurements were obtained. The context may include, for example, the place and the time of observation or measurement, and the object or group of objects observed. The elements that a data record consists of are called *values*.

All records of a dataset are assumed to have a common structure, with each position having its specific meaning, which is common to all values appearing in it. These positions may be named to distinguish between them. The positions are usually called *component*s of the data.

Definition: *Characteristic component*, or *attribute*, is a data component corresponding to a measured or observed property of the phenomenon reflected in the data. *Characteristic* is a value of a single attribute or a combination of values of several dataset attributes.

Definition: *Referential component*, or *referrer*, is a data component reflecting an aspect of the context in which the observations or measurements were made. *Reference* is the value of a single referrer or the combination of values of several referrers that fully specifies the context of some observation(s) or measurement(s).

Definition: *Reference set* of a dataset is the set of all references occurring in this dataset.

Definition: *Characteristic set* of a dataset is the set of all possible characteristics, (i.e. combinations of values of the dataset attributes).

Definition: *Multidimensional dataset* is a dataset having two or more referrers. Depending on the number of referrers, a dataset may be called one-dimensional, two-dimensional, three-dimensional, and so on.

For example, the geographical location and the time are referrers for measurements of properties of the climate such as air temperature or wind direction, which are attributes. Each combination of location and time is a reference, and the

corresponding combination of air temperature and wind direction is a characteristic. This is a two-dimensional dataset as it has two referrers; the attributes are not counted as dimensions. Referrers are *independent* components and attributes are *dependent* since the values of attributes depend on the context in which they are observed. In data analysis, it is possible to deal with selected attributes independently from the others; however, all referrers present in a dataset need to be handled simultaneously.

Data may be viewed formally as a function, in the mathematical sense, with the referrers being independent variables and the attributes being dependent variables. The function defines the correspondence between the references and the characteristics where for each combination of values of the referential components there is at most one combination of values of the attributes.

The structure of a dataset is characterized by specifying which components it includes, which of them are referrers, and which ones are attributes. Additionally to this, it is necessary to specify the properties of the components. The relevant properties are:

- whether distances exist between the elements. Any continuous set such as time, space, and values of temperature has distances, but there may be distances also in discrete sets such as a set of integer values denoting numbers of some items. The discrete set of substances has no distances.
- whether and how the elements are ordered. Thus, time moments are linearly ordered and may also be cyclically ordered, depending on the time span of observations.

It should be noted that a set consisting of combinations of values of several components does not inherit the properties of the individual components. Thus, a set of combinations of values of melting temperature and atomic weight is only partly ordered although the value sets of the original attributes are fully ordered. This data characterization framework is presented in more detail in [3].

2.2 Patterns

Definition: *Pattern* is an artifact that represents some arrangement of characteristics corresponding to a (sub)set of references in a holistic way, i.e. abstracted from the individual references and characteristics.

This is a more generic definition than is given in data mining: "a pattern is an expression E in some language L describing facts in a subset F_E of a set of facts F [i.e. a dataset, in our terms] so that E is simpler than the enumeration of all facts in F_E" [4]. In our definition, we mean any kind of representation, for example, graphical or mental.

We posit that all existing and imaginable patterns may be considered as instantiations of certain archetypes (or, simply, types). It is quite reasonable to assume that such archetypes may exist in the mind of a data analyst and drive the process of visual data analysis, which is commonly believed to be based on pattern recognition: the analyst looks for constructs that can be regarded as instances of the existing archetypes.

A pattern-instance may be characterized by referring to its type and specifying its individual properties, in particular, the reference (sub)set on which the pattern is based. Properties may be type-specific (for example, amount and rate of increase).

2.3 Pattern-by-data typology

The following table defines the basic types of patterns in relation to the characteristics of data for which such pattern types are relevant. We cannot guarantee at the present moment that this typology is complete; further work is obviously needed. Note that neither the columns nor the rows of the table are mutually exclusive. Thus, when the characteristic set is ordered and has distances, the pattern types from all columns are relevant. Similarly, when the reference set is linearly and cyclically ordered, the patterns from all rows are possible.

Characteristic set / Reference set	Any	Ordered	Has distances
Any	Even frequencies of the values, prevailing values, rare values	Tendency toward high, low or medium values	Groups (clusters) of references with close characteristics
Linearly ordered	Constancy, change, specific value order	Increase, decrease, peak, low point	Gradual change, sharp change
Cyclically ordered	Frequency of value appearing in certain positions of the cycle	Cyclical increase and decrease	Gradual or sharp changes within the cycle and between cycles
Has distances	Homogeneity or heterogeneity, large or small regions of congruency	Flatness, elevation, depression, peak, depth, plateau, valley	Smoothness (small differences between characteristics of neighboring references), abruptness (big differences)

These basic pattern types may be included in composite patterns. The types of composite patterns depend on the properties of the reference set:

1. For any kind of reference set: repeated pattern, frequent pattern, infrequent pattern, prevailing pattern;
2. For a linearly ordered reference set: specific sequence of patterns, alternation;
3. For a cyclically ordered reference set: cyclically repeated pattern;
4. For a reference set with distances: constant distance between repetitions of a pattern, patterns occurring close to each other.

Any composite pattern may, in turn, be included in a bigger composite pattern, for example, a frequently repeated pattern where increase is followed by decrease.

2.4 Directions of further work

What is presented in this section is only an initial sketch of the data-centered predictive theory. Further work is required to ensure the comprehensiveness of the pattern typology. Particular attention needs to be paid to multi-dimensional data. It is also necessary to define pattern types used to represent relationships between attributes or between phenomena (represented by several datasets differing in structure) such as correlation (co-occurrence) or influence.

Then, it is necessary to evaluate Information Visualization techniques according to the types of data they are suited for and the types of patterns they help to elicit. This can form an appropriate basis for instructive books and courses for users of Information Visualization tools.

3 Information Theory

3.1 Visual Communication

Information visualization can be viewed as a communication channel from a dataset to the cognitive processing center of the human observer. This suggests that it might be possible to employ concepts from the theories of data communication as a mechanism for evaluating and improving the effectiveness of information visualization techniques. While there are several early papers that tried to establish linkages between HCI in general to information theory [5], it might be time to revisit this concept in light of all the progress that has been made in information visualization in the past two decades.

We must start with defining information, as it is the core of information visualization.

Schneider defined information as “always a measure of the decrease of uncertainty at a receiver” [6] while Cherry stated “Information can only be received where there is doubt; and doubt implies the existence of alternatives where choice, selection, or discrimination is called for” [7]. Measuring information is a topic found in many fields, including science, engineering, mathematics, psychology, and linguistics. Information theory, which primarily evolved out of the study of hardware communication channels, defines entropy as the loss of information during transmission; it is also referred to as a measure of disorder or randomness. Another important term is bandwidth, which is a measure of the amount of information that can be delivered over a communication channel in a given period of time. We will attempt to analyze information visualization using this terminology.

Information can be categorized in a number of ways. MacKay [8] identifies three types of information content:

- Selective information-content: This is information that helps the receiver make a selection from a set of possibilities, or narrows the range of possibilities. An example might be a symptom that helps make a diagnosis.
- Descriptive information-content: This type of information helps the user build a mental model. Two types of descriptive information content have been identified:
  - Metrical: this type of observation increases the reliability of a pattern, e.g., a new member of an existing cluster (sometimes termed a metron).
  - Structural: this type of observation adds new features or patterns to a model/representation, e.g., a new cluster (termed a logon).
- Semantic information-content: This type of information is not covered in classical information theory. It lies between the physical signals of communications (called the syntactic) and the users and how they respond to the signals (called the pragmatics). The pragmatics are the domain of psychology, both perceptual and cognitive.

While the first two classes of information content lend themselves well to measurement, it is much harder to determine measures of semantic content, as this in general is very specific to individuals.

3.2 Measuring the Amount of Information

There have been many efforts to date to quantify the amount of information in a communication stream. If we think of plain text, there are numerous quantifiable features, including:

- The total number of words per minute
- The occurrence of specific words
- The frequency of occurrence for each word
- The occurrence of word pairs, triples, phrases, and sentences.

There are problems, however, with such simplistic, syntax-only measurement. Words can have variable significance; some are unnecessary or redundant. Many words can encode the same concept. In fact, reading text or hearing speech may have no affect on one's uncertainty regarding the subject of the text, e.g., you may already have known it, or you don't understand the meaning of the words or their implied concepts. This implies that the measurement of information content or volume can be specific to the individual receiver and, as we'll see later, the task that is being performed based on the communication.

Can we perform similar analysis on a dataset? Consider a table of numeric values. Features of potential interest in the dataset include:

- The count of number of entries or dimensions
- The values
- Clusters and their attributes (number, size, relations, …)
- Trends and their attributes (size, rate of change, …)
- Outliers and their attributes (number, degree of outlierness, relation to dense regions, …)

- Associations, correlations and any features between records, dimensions, or individual values.

In fact, we can observe that a featureless dataset is not differentiable from random noise: all values are equally likely. Features and relations can also vary in their magnitude, certainty, complexity, and importance. Clusters may differ in size; outliers may vary in their distance to the main body of data; features may be comprised of many sub-features; in many cases, a feature that is significant to one observer may be considered noise by another. Recently, researchers have proposed measuring and counting insights [9], which are new knowledge gained during visual analysis. These insights are generally specific to a particular task, some of which include [10]:

- Identify data characteristics
- Locate boundaries, critical points, other features
- Distinguish regions of different characteristics
- Categorize or classify
- Rank based on some order
- Compare to find similarities and differences
- Associate into relations
- Correlate by classifying relations.

For each of these tasks, we might have different accuracy requirements as well, which can influence the resolution at which feature extraction is accomplished during communication. Thus, for example, the tasks of detecting, classifying, and measuring a particular phenomenon each have their own accuracy demands. The tasks to be performed also have an implication on the types of information that the visualization must be able to convey; categorization and ranking imply that the visualization must have high selective information content, while identifying characteristics and boundaries are part of building a mental model and thus require good descriptive information content.

Returning to our dataset and the simplistic features and relations that are contained in it, we can try to quantify the volume of information and then measure how much of this volume a visualization technique is capable of effectively conveying. If we assume a table of scalar values (M records, N dimensions or variables), the number of individual values to be communicated is M*N, and the maximum resolution required is the number of significant digits. Often, however, the available visual resolution is far less than that of the data. We can then count all the pairwise relations between records, or dimensions, or even values. For records, this would be $M*(M-1)/2$, and similar for dimensions and values. Then there are relations that are 3-way, 4-way, or even among an arbitrary number of elements, e.g., in clustering tasks. Clearly, there are too many possibilities to consider them all, so perhaps we need a different tactic.

3.3 Measuring Information Loss

Perhaps it is easier to measure loss of information (entropy) during the visualization process than the total information content of a dataset. There are several techniques in common use for data transformation for visualization that provide an implicit measure of information loss. For example, multidimensional scaling, a process commonly used for dimensionality reduction, provides a measure of stress, which is the difference

between the distances between points in the original dimensioned space and the corresponding distances in the reduced dimension space. Similarly, when using principal component analysis for performing this reduction, the loss can be measured from the dropped components. Cui et al. [11] developed measures of representativeness when using processes such as sampling and clustering to reduce the number of data records in the visualization. These measures, based on nearest neighbor computations, histogram comparisons, and statistical properties, give the analysts control over what was termed abstraction quality, so they are aware of the trade-offs between speed of rendering, display clutter, and information loss. They, however, did not consider the perceptual issues, which are very dependent on the particular visual encoding used.

Distortion techniques such as lens effects and occlusion reduction also provide the analyst with trade-offs between accuracy and visual clarity. Each results in a transformation (typically of an object's position on the screen) that is meant to improve the local interpretability at the cost of accuracy of global relations. It would be interesting to see measures of these competing processes to gauge the overall implications.

Another transformation that can impact on the information being communicated in a visualization is the ordering of records and dimensions. Ordering can reveal trends, associations, and other types of relations, and is useful for many tasks. There are many possible orderings of a table of M records and N dimensions. The key is to determine which are the most useful. An ordering can convey many pairwise relations. If there are M records, an ordering can communicate M-1 of the $M*(M-1)/2$ possible pairwise relations. Many researchers have studied ways of selecting a useful ordering, including Bertin's reorderable matrix [12], Seo and Shneiderman's rank-by-feature techniques [13], and Peng et al.'s reordering for visual clutter reduction [14]. In all cases, a user should be able to prune orderings to emphasize those that show trends, groupings, or other discernable patterns. Thus far most research has focused on simple 1-D orderings, but higher level orderings and structures (e.g., hierarchies) have also been studied.

3.4 Hardware and Perceptual Limitations

This discussion of information content would not be complete without also considering the limitations imposed by the visual communication channel (i.e., the display) and receiver (the human visual perception system). Regarding the channel and its capacity, modern displays are limited to somewhere on the order of one to nine million pixels, although tiled displays can increase this substantially. The color palette generally has a size of 2^{24} possible values, although the limitations of human color perception take a big chunk out of this. Finally, the refresh rate of the system, typically between 20 and 30 frames per second, limits how fast the values on the screen change, though again the human limitations of change detection mean that much of this capability is moot.

Regarding these human limitations, from the study of human physiology we know that there are approximately 800k fibers in the optic nerve. We can perceive 8-9 levels of intensity graduation, and require a 0.1 second accumulation period to register a visual stimulus. In addition, we have a limited viewable area at any particular time,

and a variable density of receptors (much less dense in the peripheral vision). Studies have shown we have a limited ability to distinguish and measure size, position, and color, and the duration of exposure affects our capacity. Finally, it has been shown that our abilities are also related to the task at hand; we are much better at relative judgment tasks than absolute judgment ones.

3.5 Measuring Information Content on Visualizations

We now look at methods that have been used in the past for measuring the information content in a data or information visualization. For completeness sake, some of these are quite trivial. For example, simply counting the number of data values shown is a valid measure. The issue in this case would be how to deal with partial occlusion. In some cases this would be acceptable if sufficient information remains visible to make identification or recognition possible. Tufte [15] suggested the data-ink ratio as an indicator of information content, though tick marks, labels, and axes are often essential for appropriate identification. Many researchers have used counts of the number of features or patterns found in a particular amount of time. Ward and Theroux [16] counted the number of clusters and outliers found by users in different visualizations, while Suraiya et al. [9] counted insights discovered. In each case, a ground truth is needed to verify that what was found was really present. Similar experiments have been used to measure classification, measurement, and recall accuracy.

There are many other issues when attempting to measure information in a visualization. As mentioned earlier, distortion and other transformations can improve the readability of a visualization, but introduce errors in the data themselves. Data may have uncertainty attributes associated with them, which can interfere with the measurement. On the other hand, there are numerous examples of improving information content by using novel layout, shape, and color strategies or augmenting the visualization with links, boundaries, and even white space. The amount of information contained may also be enhanced using redundant mappings, which improves the chances of successful reception by the viewer. Finally, the use of animation to communicate information in an incremental fashion can be quite effective; it is lossy communication, as viewers quickly forget some of what they have seen, but the ability to replay the animation can replace some of this lost information.

3.6 Case Study: Parallel Coordinates

As an example of this information measurement activity, let us consider parallel coordinates, a popular multivariate visualization technique. The first question is how well does this technique present the values of a dataset? For individual values, this method has very high resolution, given most of the screen height can be used to convey the value. This implies the technique possesses high selective information content, at least on individual dimensions, as separation into sub-ranges is facilitated by the amount of screen space allocated to convey values. However, the loss due to occlusion can be high, especially for nominal variables. This loss is somewhat mitigated by the varying slopes of lines ending/starting at axes, which allows some

degree of differentiation. Relationships among data along an axis can be emphasized via embedded histograms, as found in some implementations of parallel coordinates. In terms of relationships between dimensions, this method is limited to showing pairwise relations, with N-1 out of the N*(N-1)/2 possible relations shown. Automated dimension ordering can help reveal interesting relationships, as seen in [14]. Relationships between records are problematic due to the ambiguous continuity of records that intersect on one or more axes. Coloring the lines based on a record ID can help with modest sized datasets. We can also use animation along an axis or based on an order to expose some inter-record relations. Intersections and near-parallel edges can reveal partial structures (between dimension pairs), and techniques such as Hierarchical Parallel Coordinates [17] can show grouping relations, though there is loss of individual data values. Each of these methods enhances the descriptive information content of the visualization, thus helping analysts form mental models of the data.

Through analyzing these augmentations of parallel coordinates we see that many recent innovations to parallel coordinates target different types of information loss resulting from this means of mapping data. For example, we see efforts to preserve and emphasize outliers in a paper by Novotny and Hauser [18]. However, many other issues still exist; there are still many data features that cannot be readily perceived and tasks that are difficult to perform using parallel coordinates.

3.7 Conclusions

To summarize, we feel there are many aspects of information visualization research that can find analogies in the concepts of information theory - it is all about communication. Perhaps finding such a formal structure on which to ground our efforts can potentially reduce the amount of ad hocness in the field. The key is to define measures of information transfer, content, or loss at all stages of the pipeline as a means of assessing our progress in the development of new visualization techniques and enhancement of existing ones.

4 Formal Models for a Science of Information Visualization

Information visualization utilizes computer graphics and interaction to assist humans in solving problems. As such, it incorporates elements of a constructive, formal science (the algorithmic and development aspects) with aspects of an empirical science (for measuring effectiveness and validity). Surrounding both are engineering efforts to improve the overall system. This section discusses the relationship between these three parts of the visualization discipline, and suggests that a deeper exploration of formal, scientific models is needed for a strengthening of the field.

4.1 The Need for Models

Traditionally, visualization has focused on the engineering aspects while importing “scientific” elements as needed. However, even this borrowing has not been

sufficiently utilized. To illustrate this, consider archetypical topics from an information visualization course syllabus:

- *Exploration:* The process of visual exploration in a larger context
- *Perception:* Fundamental mechanisms for human visual perception
- *Visual Cognition*: How perception translates to thought and action
- *Color:* Aspects of color for visualization
- *Techniques:* Specific visualization metaphors, including interaction
- *Evaluation:* Measuring the effectiveness of a visualization design

A survey of visualization education programs has found that most such programs focus on visualization techniques (the engineering core) in detriment to the foundational aspects (the scientific core) [19]. As a further example, consider that rainbow colormaps are still entrenched in visualization research [20] while ample scientific evidence demonstrates their muddling effect [20,21,22]. The ease of utility for providing rainbow colormaps does not outweigh the costs in terms of a user's time - the primary currency of users [23].

Examining the previous list, it is apparent that only the Techniques topic deals extensively with the engineering aspects of visualization design. While these efforts are vital to providing actual tools to users, the other elements are needed to provide a solid foundation to guide those efforts. For example, perceptual literature is grounded in empirical results with a strong scientific pedigree. A key aspect of these results is the *formal models* which are generated to explain the results. Such models are both *descriptive* - they encapsulate the factors of the empirical experiment and describe a mechanism for their operation - and *predictive* - they generalize the description to a larger context by predicting future behavior. The predictive nature of the models facilitates visualization design: it is the predictive nature of color perception models that explains the limitations of rainbow colormaps [22].

4.2 A Move to Visualization Formalisms: The Two Models

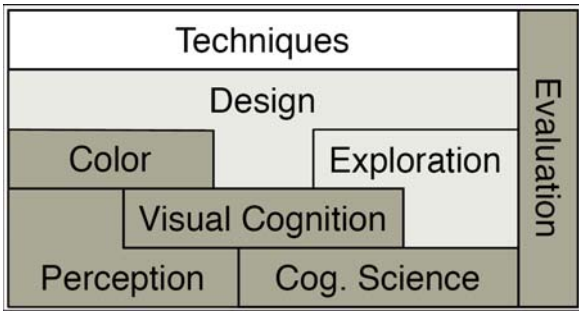


Figure 1. Topics in a Formalized Information Visualization Course. Dark grey topics are based upon formal foundations in other disciplines; light grey topics are yet-to-be-developed visualization-specific formal foundations.

There have been several recent calls for an establishment of a “theory” and “science” behind visualization [24,25]; this need can be partially addressed via formal scientific

models. If we accept that information visualization needs a formal foundation, the question remains whether the existing models from perceptual psychology and cognitive science are sufficient. The problem with these formalisms is they do not address the specific problems of visualization. While they provide general guidelines, models from non-visualization fields do not consider the context of the visualization environment - the user *and* the computer. What is needed is a set of formal foundations that bridges the gap between the general human experience and the visualization domain (Figure 1). We propose two models for this purpose: an *exploration* model that incorporates the user's interaction with the visualization and the dynamic aspects of their analysis, and a *transform design model* which encapsulates the depiction and constructive aspects of the visualization. These models would abstract fundamental principles of visualization science and design, and thus proscribe (via their predictive power) empirically driven practices.

4.3 Visualization Exploration Model

Visualization exploration is a goal-driven task incorporating visual search and information seeking. It is an iterative process - a user creates a visualization result, evaluates its worth, and then manipulates the visual parameters (e.g. color maps, selection regions) creating new results until satisfied. Thus, any formal model of computer-mediated visual exploration must capture the cognitive operations and how those realized actions manipulate the visualization. Cognitive operations are the domain of cognitive science, and several methods exist to model the human analysis process [26]. To bridge this work to visualization, two additional levels are needed: first, a description of the visual information search process and how it affects human cognition; second, a model of how the visualization session evolves due to human interaction. Visual sensemaking models such as the information foraging work of Pirolli and Card [27] begin to address the first need. Formalisms that capture the range of human-visualization interactions are targeted at the second [28,29,30].

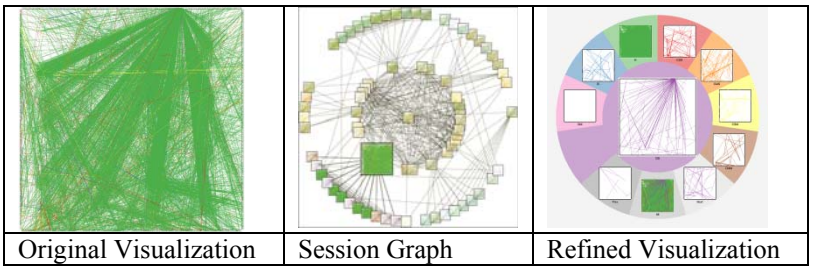
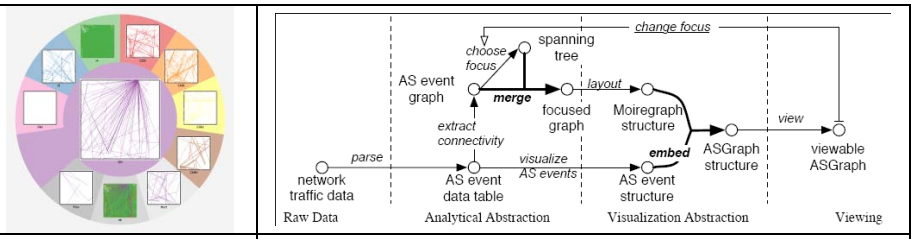


Figure 2. Analysis and evolution of a network traffic visualization. The original interface (a) uses colored lines connected to the edges of a square to depict changes. A formal model was used to capture interaction with the tool, and these sessions were analyzed to improve the interface (b). The redesigned interface (c) makes the exploration more efficient by displaying all event types individually and combined.

There are several benefits to a complete visualization exploration model. An understanding of how humans process visual cues in order to make exploration decisions can inform visualization design. For example, this "information scent" has been used to understand the cost-benefit trade-offs of different focus+context visualizations and to formally understand efficient web interfaces [31,32]. Similarly, a formalism for the visualization process lends itself to analysis to measure the user's efficiency of exploration [33,34]. Clusters of similar results (based upon metrics) during the session suggest redundant exploration; analysis of sessions based upon these metrics illuminate the path to more effective design [29,35] (Figure 2).

4.4 Visualization Transform Design Model

An exploration model describes and predicts a human's interaction with a visualization system based upon its design. This model neglects to describe the components that compose the design or provide initial design guidance. To provide this guidance, visualization transform design models are needed. A *visualization transform* is the function that computes the depicted result from visualization parameters – elements such as brushed graph nodes or opacity maps that dictate the rendered result. Significant work has expressed different mechanisms for constructing such transforms [36,37,38,39,40] (see Figure 3 for an extended example), but these categorizing efforts lack two things to provide a formal foundation. First, they do not utilize perceptual and cognitive literature to suggest and evaluate design decisions; second, most do not address the evolution of transforms and the utilization of extra-visualization tools important to the analysis (e.g., statistical packages). Efforts in providing design guidance for data display has been investigated [41,42,43] and formal algebras for transform modification have been recently presented [44,45]. However, these distinct contributions require effort to unify and to validate for a complete and cohesive scientific foundation.



(a) Routing Visualization

(b) Visualization Transform Components

Figure 3. A depiction of the revised network routing visualization transform. Nodes represent the state of the data (e.g., a table of events) while edges represent operators or interactions (e.g., parsing the data). In this example, the network visualization is combined with a graph visualization by embedding the results of the former within the latter. The graph itself is a composition, merging a spanning tree and the original graph to layout the selected sub-graph. Back-propagation of state due to interaction is included. The depiction is based upon an extended Data State transform model.

A complete, predictive transform design model will yield several benefits. Toolkits for visualization creation benefit by providing guidance on suitable, less suitable, and unsuitable component choices. Visualization pedagogy will improve due to a validated foundation for techniques. Further, formal models will lead to objective metrics for evaluating a transform's effectiveness. All of these enhancements feed back into a visualization system, improving its potential utilization.

4.5 The Value of Visualization Craft

A research program investigating scientific grounding for visualization is not meant to diminish the importance of the engineering component of visualization. Visualization is a tool for humans; engineering efforts form the basis of providing such tools. Formal exploration and design models can guide the creation of a visualization system; however, as is the case now, multiple comparable techniques will often solve the same problem. Thus, the craft of visualization - the confidence in design choices gained through experience - will still be needed to decide between the choices a formal model provides. Inspiration and creativity will not be eliminated by a more rigorous foundation; the foundation will serve as a springboard for such endeavours.

4.6 What is Left to be Done

Formal foundations for a science of information visualization are still in a nascent stage. Elements of complete exploration and transform models exist; however, they have neither been reconciled with each other nor validated for their correctness. A close collaboration between perceptual psychologists, cognitive scientists, visualization researchers, and practitioners is needed to drive research into foundations: perceptual and cognitive scientists provide the human-based foundations, practitioners provide the case studies for observation and validation of models, and visualization researchers will form the domain-specific bridge between them.

Humans and computers play an integrated role in the development and utilization of visualization. A formal foundation would measure the efficiency of the former, and guide the design of the latter in order to create a more effective whole.

5 Conclusion

The three discussions of Information Visualization presented here draw on existing theories of data-centric prediction, information communication and scientific modeling, and relate in different ways to the linguistic framework defined in the introduction. A single uniting theory of Information Visualization may be impossible due to its strong relationship to and use of several other diverse disciplines (e.g. psychology (perception, cognition and learning), graphic design and aesthetics).

Investigating theoretical approaches used in other disciplines, and their relation to Information Visualization, is an obvious way forward, and can provide a useful way for researchers in the area to present, discuss and validate their ideas; it is hoped that the over-arching linguistics-based framework of representation, user exploration and manipulation, and system exploration and manipulation will prove useful in linking

the constituent theories together. The more solid theoretical analyses that Information Visualization researchers or tool designers can call on in defending or validating their work, the more secure the discipline will be.

References

1. de Saussure, F.: Writings in General Linguistics. Bouquet, S., Engler, R., Sanders, C. and Pires, M. (eds) Oxford University Press (2006)
2. Bakhtin, M.: The Dialogic Imagination, University of Texas Press (1981) 282. Quoted in Ball, A.F and Freedman, S.W.: Bhaktinian Persepectives on Language, Literacy, and Learning, Cambridge University Press (2004) 174.
3. Andrienko, N. and Andrienko, G.: Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach, Springer (2006)
4. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine, 17 (1996) 37–54.
5. Card, S. K., Moran, T. P. and Newell, A.: The Psychology of Human-Computer Interaction. Erlbaum Associates, Hillsdale, New Jersey (1983)
6. Schneider, T.D.: Information Theory Primer. http://www.lecb.ncifcrf.gov /~toms/paper/primer (April 14, 2007)
7. Cherry, C.: On Human Communication (Second Edition). MIT Press, Cambridge, MA (1966)
8. MacKay, D.: Information, Mechanism and Meaning. MIT Press, Cambridge, MA (1969)
9. Saraiya, P., North, C., Duka, K.: An evaluation of microarray visualization tools for biological insight. Proc. IEEE Symposium on Information Visualization (2004) 1-8
10. Keller, P., Keller, M.: Visual cues: Practical Data Visualization. IEEE Computer Society Press, Los Alamitos, CA (1993)
11. Cui, Q., Ward, M., Rundensteiner, E., Yang, J.: Measuring data abstraction quality in multiresolution visualization. Proc. IEEE Symposium on Information Visualization (2006) 709-716
12. Bertin, J.: Matrix theory of graphics. Information Design , 10:1 (2001) 5-19
13. Seo, J., Shneiderman, B.: A rank-by-feature framework for interactive exploration of multidimensional data. Proc. IEEE Symposium on Information Visualization (2005) 96-113
14. Peng, W.,Ward, M., Rundensteiner, E.: Clutter reduction in multi-dimensional data visualization using dimension reordering. Proc. IEEE Symposium on Information Visualization (2004) 89-96
15. Tufte, E.: The Visual Display of Quantitative Information. Computer Graphics Press, Cheshire, CT (1983)
16. Ward, M., Theroux, K.: Perceptual benchmarking for multivariate data visualization. Proc. Dagstuhl Seminar on Scientific Visualization (1997) 314-328

17. Fua, Y.-H., Ward, M., Rundensteiner, E.: Hierarchical parallel coordinates for exploration of large datasets. Proc. IEEE Conference on Visualization (1999) 43-50
18. Novotny, M., Hauser, H.: Outlier-preserving focus+context visualization in parallel coordinates. IEEE Trans. Visualization and Computer Graphics 12 (2006) 893-900
19. Dommik, G.: Do We Need Formal Education in Visualization?, IEEE Computer Graphics and Applications, IEEE Computer Society Press, 20:4, (2000) 16-19.
20. Borland D. and Taylor R.M.: Rainbow Color Map (Still) Considered Harmful. IEEE Computer Graphics and Applications, 27:2 (2007) 14-17.
21. Brewer, C.A.: Color use guidelines for data representation. Proceedings of the Section on Statistical Graphics, American Statistical Association (1999) 50-60.
22. Ware C.: Information Visualization: Perception for Design. Morgan Kaufmann (2004)
23. van Wijk, J.J.: The Value of Visualization. Proceedings of IEEE Visualization, IEEE Computer Society (2005) 79-86
24. Johnson C., Moorehead R., Munzner T., Pfister H., Rheingans P. and Yoo T.S.: NIH-NSF Visualization Research Challenges Report (1st ed) , IEEE Press (2006)
25. Thomas J.J. and Cook K.A.: Illuminating the Path: The Research and Development Agenda for Visual Analytics, IEEE Computer Society Press (2005)
26. Anderson, J.R.: Cognitive Psychology and its Implications (6th ed.) Worth (2005)
27. Pirolli P. and Card S.K.: Information Foraging. Psychological Review (4) (1999) 643-674.
28. Brodlie K., Poon A., Wright, H. Brankin, L., Banecki G and Gray A.: Problem Solving Environment Integrating Computation And Visualization, Proceedings of the 4th IEEE Conference on Visualization, Nielson G.M. and Bergeron R.D. (eds) (1993) 102-109
29. Jankun-Kelly T.J., Ma K-L and Gertz, M.: A Model and Framework for Visualization Exploration. IEEE Transactions on Visualization and Computer Graphics, 13 (2007) 357-369
30. Lee J.P. and Grinstein G.G.: An Architecture For Retaining And Analyzing Visual Explorations Of Databases, Proceedings of the 6th IEEE Conference on Visualization. Nielson G.M. and Silver D (eds) (1995) 101-108
31. Pirolli P., Card S.K. and Van Der Wege, M.M.: Visual information foraging in a focus + context visualization. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press (2001) 506-513
32. Pirolli, P.: Rational Analyses of Information Foraging on the Web. Cognitive Science 29:3 (2005) 343-373.
33. Jankun-Kelly T.J., Ma K-L and Gertz M.:A Model for the Visualization Exploration Process. Proceedings of the the 13th IEEE Conference on Visualization (Vis '02). Moorhead R.J., Gross M. and Joy K.I. (eds) (2002) 323-330
34. Lee P.J.: A Systems and Process Model for Data Exploration. PhD thesis, U. of Massachuesetts Lowell (1998)
35. Teoh S.T., Jankun-Kelly T.J., Ma K-L. and Wu S.F.: Visual Data Analysis for Detecting Flaws and Intruders in Computer Network Systems, IEEE Computer Graphics and Applications, IEEE Computer Society Press (2004) 24

36. Abram G. and Treinish L.: An Extended Data-Flow Architecture For Data Analysis And Visualization, Proceedings of the IEEE Conference on Visualization 1995 (Vis '95), IEEE Computer Society Press, Nielson G.M. and Silver D. (eds) (1995) 263-270
37. Chi E.H. and Riedl, J.T.: An Operator Interaction Framework For Visualization Systems, Proceedings of the IEEE Symposium on Information Visualization, Dill J. and Wills G. (eds) (1998) 63-70
38. Haber R.B. and McNabb D.A.: Visualization Idioms: A Conceptual Model for Scientific Visualization Systems, Visualization in Scientific Computing, eds. Nielson G.M., Shriver B. and Rosenblum, L. IEEE Computer Society Press, (1990) 74-93
39. Hibbard W.L., Dyer C.R. and Paul B.E.: A Lattice Model for Data Display. Proceedings of the 5th IEEE Conference on Visualization (Vis '94), Bergeron R.D. and Kaufman A.E.(eds) (1994) 310-317
40. Schroeder W.J., Martin K.M. and Lorensen W.E.: The Design and Implementation of an Object-Oriented Toolkit for 3D Graphics and Visualization, Proceedings of the 7th IEEE Conference on Visualization. Yagel R and Nielson G.M. (eds) (1996) 93-100.
41. Casner, S.M.: Task-analytic approach to the automated design of graphic presentations, ACM Transactions on Graphics, 10:2 (1991) 111-151.
42. Mackinlay, M.: Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics, 5:2 (1986) 110-141
43. Roth S.F. and Mattis J.: Data Characterization for Intelligent Graphics Presentation, Proceedings on Human Factors in Computing Systems (CHI'90) (1990) 193—200.
44. Bavoli L., Callahan S.P., Crossno P.J., Freire J., Scheidegger C.E., Silva C.T. and Vo, H.T.: VisTrails: Enabling Interactive Multiple-View Visualizations. Proceedings of the 16th IEEE Conference on Visualization (2005)
45. Weaver C.: Building Highly-Coordinated Visualizations in Improvise. Proceedings 2004 IEEE Symposium on Information Visualization, IEEE Computer Society (2004) 159-166.