A Methodology for Choosing Data Representations

Philip K. Robertson
CSIRO Division of Information Technology

This methodology helps you choose the best visualization of your data. With objective, directed display design methods, you can match representations to the intrinsic characteristics of data and goals for its interpretation.

Advanced data presentation and management tools are becoming more widely incorporated into information and decision support systems. We need a range of data representation options to cater for different data types, interpretation goals, and contextual requirements of such systems. Choosing the appropriate representation can provide the key to critical and comprehensive appreciation of the data, thus benefiting subsequent analysis, processing, or decision making.

To choose the best way to represent data, we need to match the type of information we are interested in with the ability of the different representation properties to convey that information. This article focuses on formalizing this matching process in a generic framework.

Visualization approaches
We can use many approaches to represent data, and different representations can reveal quite different characteristics. For example, Figure 1 shows...
then it might not be obvious to a data analyst how to choose or, once having chosen, how to modify or interact to best effect. But clear, effective interpretation of data depends critically on whether the visual representations are appropriate.

The task of choosing an appropriate representation becomes even more complex if we want to portray several data variables or compare them for correlations. Consider, for instance, the numerous ways of portraying two variables in Figures 3 through 8. Which, if any, of these approaches is best for a given purpose?

![Image of Figure 3. Two scalar variables in a gray-scale representation. I portray measurements of magnetic field intensity (MFI) and elevation (ELEV) over the same geographic 2D field.]

If we want to portray additional qualifying information, such as the binary variable indicating areas of low data confidence (or “null” regions shown in black in Figure 9) should we superimpose it on the data presentation or show it in an adjunct display (as in the central right relief-shaded view)? We might ask why the binary variable was not chosen for the surface height or the surface color, or why the tin (Sn) distribution was not shown in color and the bismuth (Bi) as surface height. Or we might ask what coloring schemes are more appropriate, and so on. The answers depend on the type of information the analyst was interested in and the capability of the different representation properties to convey that information.

### Need for a visualization methodology

More often than not, researchers base present methods of choosing representations on ad hoc approaches, relying on the experience and advice of a “visualization expert.” This is unsatisfactory when we consider the broad range of applications to which visualization techniques are applicable and the small number of visualization experts. Instead we need a methodology, based on some appropriate “information theory” of visualization, for choosing data representations to best achieve any specific visualization aims. Visualization software would then...
Aims of visualization

Underlying the concept of visualization is the idea that an observer can build a mental model, the visual attributes of which represent data attributes in a definable manner. This raises several questions:

1. What mental models most effectively carry various kinds of information?
2. Which definable and recognizable visual attributes of these models are most useful for conveying specific information either independently or in conjunction with other attributes?
3. How can we most effectively induce chosen mental models in the mind of an observer?
4. How can we provide guidance on choosing appropriate models and their attributes to a human or automated display designer?

**Natural scene paradigm**

Artists and scientists alike have long considered the first three questions, if not explicitly, certainly implicitly. One approach involves:

1. using clearly and easily understood models such as 3D structures or scenes,
2. representing data variables by the recognizable properties of the objects or scenes, and
3. inducing them in the observer's mind by using graphics scene simulation techniques.

This we can call the *natural scene paradigm*. It is based on our ability to glance at a scene and gain an immediate appreciation of its 3D surface structure, what the surface is covered with, and even the condition or state of that surface covering. This ability to shift attention, or reconfigure mentally, from appreciating one physical property of the scene to another is critical in allowing us to assess scene properties either independently or in relation to one another.

Several researchers have studied this paradigm, and their studies have resulted in an understanding of many of the perceptual processes involved. The various ways of representing...
be more soundly based (and more widely used), but at present, as Levine pointed out, there is a distinct lack of these tools for application and system builders.

Researchers such as Haber and Wilkinson have helped establish the aims of visualization, but the development of methodologies remains limited. Much work on designing data displays has consisted largely of sets of guiding rules. Although valuable, these do not constitute a methodology on which we can base automatic display generation methods. Robertson and O'Callaghan and other researchers have developed partial methodologies for the use of color in various fields. Hudson's logic-based approach constitutes one of few approaches to establishing a formal and systematic methodology to generate display representations for data. Mackinlay described a framework for developing presentation tools for 2D graphical displays, based on expressiveness and effectiveness criteria applied to precisely defined graphical languages. Senay's sets of presentation rules and a knowledge-based visualization tool also offer a broad framework for generating visualizations.

Developing a comprehensive and systematic treatment for choosing and generating display representations requires a complete formalism. This article describes a methodology to underpin visualization and illustrates its application. A natural scene paradigm, developed more fully in an earlier version of this article, helps illustrate the construction of the methodology and its extensibility to deal with progressively more sophisticated displays. Specific examples drawn from geoscientific data sets illustrate the development of the approach and methodology. The methodology itself is general; it is independent of the display paradigm, the data sets, or specific interpretation measures associated with them.
Figure 10. Variables of interest to an analyst in an ecological data set. Variables are represented in the form of a tree structure indicating relationships between them and the characteristics of each.

Table 1. Identification of dimensionality, parametric relationships, and data types of three variables in an ecological data set.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Variable 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall</td>
<td>2D-p1</td>
<td>Ordinal, continuous</td>
</tr>
<tr>
<td>Predominant fauna type</td>
<td>2D-p2((\text{sp}1))</td>
<td>Nominal, multivalued</td>
</tr>
<tr>
<td>Predominant fauna density</td>
<td>2D-p2((\text{sp}2))</td>
<td>Ordinal, continuous</td>
</tr>
</tbody>
</table>

provides a basis for the development, which we could then apply over a broader range of models or paradigms.

This approach, reflected in subsequent sections of this article, is based on

- establishing the nature of the data to be displayed and the interpretation aims of the analyst;
- establishing the capability of various visual representations, as components of natural scenes, to convey information about the attributes of data variables, and
- choosing an appropriate representation, or set of representations, for the data by matching the representation capability to the interpretation aims.

Data types, interpretation aims

Scientific and statistical data sets can involve a wide range of data with varying dimensionality. Display requirements can involve isolating data anomalies, portraying gradations in the data, or emphasizing abrupt variations. Often, depicting the spatial nature of correlations between these characteristics in two or more data variables can be important.

The examples here treat multiple scalar variables defined over a 2D field. The underlying methodology developed depends not on the dimensionality but rather on its objective determination. Multiple 2D scalar fields provide a demanding display problem widely applicable in geoscientific and other contexts.

Data variables and dimensionality

We can treat a data variable as lying within a multidimensional space, the dimensions of which can be spatial, spectral, temporal, or of another nature. Two variables that have identical distributions in \((n-1)\)-dimensional space, but different distributions in \(n\)-dimensional space, can be treated as two parameters of the \((n-1)\)-dimensional space. Practical examples include RGB components of a 2D image (which can be treated as a 2D three-parameter data set) or a bivariate choropleth map (which can be treated as a 2D two-parameter data set).
Table 2. Identification of attributes of interest, and correlations between attributes of interest, in the ecological data set variables. The attribute before the colon is row-indexed while the attribute after the colon is column-indexed; the colon itself means “in relation to.”

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Variable 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable 1</td>
<td>point</td>
<td>global:global</td>
</tr>
<tr>
<td>Variable 2</td>
<td>point:local</td>
<td>point:global</td>
</tr>
<tr>
<td>Variable 3</td>
<td>local</td>
<td>global</td>
</tr>
</tbody>
</table>

Table 3. Identification of attributes of interest, and correlations between attributes of interest, in the chemical data set variables.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Variable 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable 1</td>
<td>point:local</td>
<td>point:point</td>
</tr>
<tr>
<td>Variable 2</td>
<td>point:local</td>
<td></td>
</tr>
<tr>
<td>Variable 3</td>
<td>local</td>
<td></td>
</tr>
</tbody>
</table>

**Data types**

To construct the methodology, we distinguish between nominal (classification) and ordinal (intrinsic progressive relationship between successive values) variables. An ordinal variable can be continuous or discretely valued, while a nominal variable can be single valued or multivalued.

Figure 10 shows the structure of an ecological data set comprising three variables defined over a common 2D field, with two of the variables being parameters of a given delineation over that field. We term the rainfall and the predominant fauna as parameters (p1 and p2), and the type and density of the predominant fauna as substrata (s1 and s2) of the predominant fauna parameter. We can thus describe this data set more compactly as 2D – p1, p2(s1, s2).

The three variables from the ecological data set example appear in Table 1. The table also shows the dimensionality and parametric relationships of the variables, as well as their nature.

Another example of a data set might be two chemical samples taken at regular intervals and interpolated onto a regular grid, with a confidence level (samples taken or missing) as an additional data set. The chemical concentrations are ordinal continuous parameters of a 2D space spanned by time and depth, while the confidence level is a binary nominal parameter. (See later for a description of its visualization, shown in Figure 9.)

**Interpretation aims**

An analyst’s aims for interpretation might be defined clearly, but might also be broader or context specific. Nevertheless, even context-specific aims should be reducible to a set of more generic attributes. In fact, performing such an exercise can help us focus the interpretation aims.

This raises the question of what constitutes an appropriate set of general attributes. Here I use a particular set that provides a basis for the development of the methodology without necessarily making any claims for its uniqueness or optimality (although recognizing its prior use). We can thus distinguish between

- values at a point (point)
- local distribution of values, such as gradients and features (local)
- global distribution of values, such as trends and structure (global)

This type of distinction is not new in visual perception or indeed in broader human communication. The significance, and exploitation, of the difference between local and global interpretation (or concentration) is particularly clear, for example, in the graphic works of M.C. Escher, in the music of J.S. Bach, and in many other artistic fields.

The point, local, and global data attributes in fact correspond closely to the elementary, intermediate, and superior levels at which Bertin suggests that we can understand a graphic data variable. We can use different attributes for different tasks; the importance lies in isolating specific, sufficiently distinct attributes to avoid confusion between their scope for a given task.

We can also express correlations between two data variables in these terms. For example, the co-occurrence of specific values at a point in one variable with steep gradients or specific local distributions (such as circular features) in another variable might be significant.

We can represent in a matrix structure data attributes of interest to an analyst, both for interpretation of each variable independently and for interpretation of correlations between variables. Table 2 shows such a matrix for the ecological data set example.

Thus, in the on-diagonal entries the analyst indicates interest in the independent variables: amount of rainfall (variable 1) at a point, the species type (variable 2) at a point and its global distribution, and the local and global distribution of the density of the dominant species type (variable 3). In the off-diagonal entries the analyst indicates interest in the correlations between variables: the local rainfall distribution in relation to the species type at a point, global rainfall distribution in relation to global species density distribution, and the species type at a point in relation to its global density distribution.

Using the chemical data set example, Table 3 shows interest expressed in the correlation between (high) point values of one
Table 4. Identification of dimensionality, parametric relationships, and types of a set of natural scene properties.

| Property 1 | 2D-$p_1$ | Ordinal, continuous |
| Surface height |
| Property 2 | 2D-$p_2(sp_1)$ | Nominal, multi-valued |
| Surface cover/materia l type |
| Property 3 | 2D-$p_2(sp_2(ssp_1))$ | Ordinal, continuous |
| Surface cover/condition/density |
| Property 4 | 2D-$p_2(sp_2(ssp_2))$ | Ordinal, four-valued |
| Surface cover/condition/phase |

Using a natural paradigm

A simple real-world surface exhibits two (possibly independent) parameters over the same 2D spatial field: its height structure and its covering structure (such as undulating terrain covered by various forests). In practice, such parameters often relate to each other in some physical sense (for example, vegetation growth depends on surface slope and aspect), but they can also be independent. The surface covering structure can in turn exhibit independent sub-parameters—for example, its nature or type of material (such as conifer or deciduous forest) and the condition of the material (such as overall growth condition). Further, the condition might have several subparameters, such as its density (for example, mature trees per unit area) or its phase (for example, annual stage of plant growth), each indicated by aspects of color and texture.

Clearly, we can observe many properties, and subtle relationships between these properties, by rapidly, easily, and selectively focusing our attention on the physical components of a scene. These are exactly the types of relationship that an analyst often wishes to extract from related data variables. Providing such paradigms for model building—an essential role of visualization tools in an exploratory context—thus becomes a critical element in choosing appropriate representations.

We can also observe other, more complex, effects in natural scenes. Mixed coverings (more than one material, moisture, etc.), temporal variations in a property, shadow properties and regions, and spatial frequency (or spatial scale) of some scene properties can all allow appreciation of different processes or variables. The natural world thus offers the model of a very rich set of physical properties, many of them evident either independently or in co-occurrence with others.

To exploit this potential source of interrelated representations in a natural scene paradigm, we must look at each of these basic scene properties in greater detail. We must determine the type of information conveyed and any accompanying constraints on interpretation. The nominal/ordinal and discrete/continuous descriptors and dimension/parameter relationships can also describe the natural components, or physical properties, of a simple 3D scene. Table 4 demonstrates this for a set of basic scene properties. The tree structure corresponding to these properties appears in Figure 11.

We distinguish between the terms variable and property. Variable describes a data variable to be displayed; property describes a physical property of a 3D scene. We can thus use scene properties to represent data variables in a data display. Attribute describes the point, local, and global distribution information measures. The extent to which specific properties of a scene can also convey these attributes, and the analyst’s requirements for portraying specific attributes of a variable, determine the suitability of the representations chosen.

We can also construct a table showing the attributes conveyed by each scene property (see Table 5). Here, I have treated only the first three scene properties. We could include the fourth
Figure 11. A representation of a coastline profile showing the interaction of flexible materials with solid objects to depict material stiffness.

Figure 13. A representation of cutting profiles and paths, showing the intersection of two surfaces.

Implementation and display

The following steps form the basis for the guidance system:

1. Extract the structure and the nature of each variable from the data, including:
   - the dimensionality and, where known, the parameter/subparameter relationships; and
   - the type of data (ordinal or nominal, discrete or continuous).

2. Substantiate this information by asking the analyst for classification of:
   - the relationships between parameters and subparameters; and
   - the type of each variable.

3. Ask the analyst about the important attributes for interpretation, including:
   - individual data variables (point, local, global);
   - correlations between them (point, local, global relative to each other attribute for each variable);
   - their relative importances (priority ordering or weighing of these attributes); and
   - any display constraints (for example, analyst requirements forcing particular representations).

4. Match the representations to the interpretation aims.
• choose (and display) representations optimizing the match, based on strict analyst-specified priorities and constraints;
• choose (and display) alternative representations optimizing the match on a broader basis;
• indicate which interpretation aims are satisfied; and
• allow interactive change of priorities and display regeneration.

The structure and nature of data variables are generally available either explicitly or implicitly in the data structures or databases used to carry data sets. For example, multiple channels, spectral components, spatial or temporal references, and sampling intervals provide dimensionality and parametricity information; data formats and specified characteristics provide data-type information. This information is generally carried in header or ancillary data structures even if a database system is not used to store the data, and it is certainly maintained in any large-scale or production facility.

Practical application has demonstrated the substantial value of going through these steps when choosing appropriate representations. Encouraging analysts to think about the interpretation problem at a slightly more abstract level also helps them recognize the scope of techniques commonly applied to other data sets. And, as with many complex tasks (such as income tax determination), providing structured forms (such as tax returns) for human interaction provides enough of a cue to make the subsequent use of the display tools more straightforward.

A specified set of display demands might be impossible to meet with a given set of property representations in a single scene. Multiple and composite displays nevertheless offers substantial scope for resolving such demands, at least to a partial extent. Aligned adjacency with tiles constructs, such as traces or grids, can make evident specific attributes that would otherwise not be supported. For example, vertically stacked surfaces in perspective can show correlations in global distribution between individual surfaces that give point, local, and global attributes for each variable independently (see Figure 8, for example). Horizontally aligned surfaces in perspective can show correlations in point values between individual surfaces. Intersecting surfaces in perspective can show correlations in local distribution between the individual surfaces (as in Figures 12 and 13).

**Discussion**

The methodology presented here makes considerable progress over ad hoc display approaches. Using only two data types, three information measure attributes, and the natural scene paradigm does not limit the methodology. This approach accommodates other data types, attributes, and paradigms if we define the data types and attribute-of-interest matrix for a paradigm. In practice, these aspects have not been a constraint because the advantages gained outweigh any limitations imposed by this set of parameters. The scope for interaction constitutes the limiting factor, in terms of the view generation rate and manipulation of the attribute table. We use massively parallel architectures and algorithms to address the former. The latter requires a better understanding of the perceptual and mental expectations associated with use of the tools.

Interaction is clearly a key aspect of the interpretation process, just as in the real world, where an observer continually interacts through eye, head, and other movements. Interaction can take several forms: with the visualization itself, in terms of scene parameter adjustment (such as color scales, viewpoint, or height emphasis); or with metavisualizations, such as the tables that indicate attributes of interest or associated attribute priorities. In addition, providing information about the attributes supported as well as those requested in a given representation, either directly or in relation to the tables constructed, can benefit the opportunistic aspect of visualization.

It is important to recognize that using this methodology in no way reduces the opportunities for serendipitous interpretation. Rather, these opportunities are enhanced by allowing prediction (in the metavisualization of control objects) of benefits not necessarily obvious from the visualization itself. Researchers have not yet explored this enhanced role of metavisualizations systematically.
Several nontrivial perceptual aspects of the capability do require better understanding. For example, the relationship between attributes made evident when going from static to dynamic representations is complex and affected by many of the issues that have driven "graph-haptic" or "virtual reality" environments. We also need to understand the perceptual trade-off between spatial, spectral, and temporal resolution in different phases of exploring and interpreting different types of data under different representations. Within the methodology we can incorporate knowledge of the available computational power and other performance parameters. This would allow, for example, an improved interaction rate at the expense of spatial resolution if the attributes of greatest importance require dynamic representation.

Another complex question involves the distinction between the attributes evident through implicit or explicit representations. For example, we can detect the gradient of a variable in a surface representation using intuitive scene analysis—it is in implicit form, evident as local and global attributes but not as a point attribute. If we apply a gradient operator to the variable, we can represent the resulting new variable, which shows gradient as a point attribute. The type of information provided, and its relationship in each case to other data variables, either implicitly or explicitly displayed, differs markedly. Accounting for this type of possibility in the methodology could help reduce to objective levels the kind of display choices typically made on the basis of experience—but this work is in its early stages. Aiming for intuitive interpretation suggests using implicit representations where possible; deliberate extraction of specific information suggests explicit representations.

Part of the value of natural representation paradigms lies in our ability to base our analysis of attributes conveyed on a sound understanding of our perceptions of the natural world. Often a display represents one aspect of a variable explicitly (such as value by surface height and color) and another aspect implicitly (such as gradient). One of the aims of the methodology is to embed subtleties of interpretation within the attribute tables so that analysts need not explicitly understand them, meaning an analyst need not be a visualization expert.

Multiple or composite displays also pose the question of whether the attribute table can adequately cover the distinctions between directed and intuitive comparison under attention shifts. Indeed, some support exists for separating directed and intuitive comprehension, which arise from explicit and implicit representation mechanisms respectively and maintaining distinct attribute tables for each. This does not preclude overall tables in addition. Such a separation might relate to the focused versus serendipitous interpretation modes analysts often employ, frequently relying on some cues intrinsic or extrinsic to the display to perform a mental emphasis shift.

More generally, interpretation of any data display involves building a perceptual hypothesis. The building of such hypotheses is a process of following rules from base assumptions supporting high-level knowledge. This knowledge encompasses not only aspects of the data and its analysis, but also cues associated with visual representations within a given paradigm. It affects both the assumptions and the rules. Mechanisms to incorporate contextual knowledge lie beyond the scope of the article, as does a fuller treatment of the relationship between visual and mental models. But maintaining a known display paradigm—and limiting options within this paradigm to those that we can characterize objectively in the manner described in this methodology—is important and perhaps necessary for unambiguous data visualization.

**Summary and conclusions**

To develop a systematic approach to choosing data representations, we must establish objective measures of how a representation conveys information. I propose a practical set of informative attributes: the data values at any point; the local distribution of the data values around a point; and the global distribution of the data values. To choose compatible representations, we formulate our interpretation aims in terms of these attributes. The scene properties (and correlations between them) must match the characteristics of the data they represent.

I chose the interpretation measures and display paradigm used here because they are robust and well understood and have been used widely in less formalized contexts. With this type of methodology, you can also exploit the 3D scene paradigm, which covers most visualizations currently used.

This is a systematic approach to choosing data representations, based on direct matching of representation capability to interpretation aims. Considerations underlying the matching process include simultaneous depiction of both independent and correlated information about data variables, and the potential for exploiting the rich set of natural associations inherent in the interrelationships of physical properties of scenes.

Current research on this work includes incorporating better inferencing to help generate display representations based on analyst-specified priorities and scene-comprehension rules; developing optimal logic-based approaches to computational tree-structure matching and resolution; investigating the perceptual basis for implicit and explicit representations; and incorporating interaction into the methodology. Effective visualization depends on gaining a better understanding of information and developing methodologies that encapsulate it. This methodology is a step in that direction.
Acknowledgments

This work has benefited from discussions with research staff, students, and visitors at the Visualization Project at the CSIRO Division of Information Technology’s Centre for Spatial Information Systems, and from the referees’ suggestions and comments. I am particularly grateful for its detailed review by Lisa DeFerrari. Permission to use the data sets in the examples is acknowledged.

References

9. F. Senn, "Compositional Analysis and Synthesis of Scientific Data Visualizations," SIGGRAPH 91 Course Notes (State of the Art in Data Visualization) Section V.

Philip K. Robertson is a senior research scientist at the CSIRO Division of Information Technology’s Centre for Spatial Information Systems in Canberra, Australia. He leads the center’s Visualization Group and also lectures in computer graphics at Australian National University. His research interests cover conceptual and computational approaches to visualization, including visualization methodologies, perceptual color interfaces, parallel and multidimensional algorithms, and the design of kernel image processing software. Robertson holds a BEng and MS in electrical engineering and a PhD in computer science from the Australian National University. He is a member of IEEE and ACM, and a member of the editorial board of IEEE CG&A.

Readers can contact Robertson at CSIRO Div. of Information Technology, GPO Box 664, Canberra ACT 2601, Australia. His e-mail address is robertson@csis.dit.csiro.au.

IEEE COMPUTER SOCIETY NEW RELEASE

FORMAL VERIFICATION OF HARDWARE DESIGN

by Michael Yoeli

Formal Verification of Hardware Design examines the techniques for the determination of functional correctness of hardware prior to production. It concentrates on the means by which behavioral design errors can be detected and on the use of mathematically precise methods. Formal hardware verification methods are now being developed for use as practical tools in the detection of functional design errors that conventional simulation might not detect.

This tutorial introduces the most important and promising approaches in formal hardware verification, and focuses on a number of relevant topics, rather than completely surveying all aspects of the field. A large part of the book deals with hardware verification based on formal logic and discusses suitable, automated proof systems. Some other key issues include: Basics of Formal Verification, Abstraction Mechanisms, Formal Logic and Theorem Provers, Hybrid Approaches, Hardware Description Languages, Timing Verification.


SEND YOUR ORDERS TO: 1951-1991

IEEE COMPUTER SOCIETY PRESS
10662 LOS VAQUEROS CIRCLE
P.O. BOX 3014
LOS ALAMITOS, CA 90720-1264

* ADD $5.00 FOR HANDLING CHARGES

OR CALL TOLL-FREE
1800-CS-BOOKS
714-821-8380
OR IN CALIFORNIA
714-821-4010
OR FAX 714-821-4010