

Choosing the Right Graph

Abstract—When it comes to graphing data, most professionals show little method or creativity. They typically limit themselves to a small repertoire of graph types and select from it on the basis of habit, if not sheer ease of production. Similarly, the many books on graphing devote much attention to graphical integrity and readability, but little or none to graph selection. We developed a methodology to help engineers, scientists, and managers choose the “right graph” on the basis of three criteria: the structure of the data set in terms of number and type of variables, the intended use of the graph, and the research question or intended message. The first and third criteria allow one to construct an effective two-entry selection table.

—JEAN-LUC DOUMONT,
SENIOR MEMBER, IEEE,
AND PHILIPPE VANDENBROECK

Index Terms—Data sets, graphs, variables.

Manuscript received November 12, 2001;
revised November 28, 2001.
J.-L. Doumont is with JL Consulting,
B-1950 Kraainem,
Belgium
(email: JL@JLConsulting.be).
P. Vandenbroeck is with Epiphany,
B-3001 Leuven,
Belgium
(email: philippe@wholesys.org).
IEEE PII S 0361-1434(02)02338-X.

Written documents and oral presentations are essentially sequential. Even if they are constructed along a hierarchical (tree-like) structure, they have a beginning, a middle, and an end, either in space (documents) or in time (presentations). As a result, they lend themselves well to methodologies that specify “what goes where.” Introductions, for example, provide some context first, then establish the problem or need, state what was done to address the need, and finally announce what the document attempts or contains—all in a systematic sequence.

Graphical representations, by contrast, are in essence nonsequential. While the most quantitative of them usually embody a sequence of numbers, they do not suggest a viewing sequence, with a beginning and an end. On the contrary, they are meant to be perceived and interpreted globally, all at once. There lies their specific “competitive advantage,” in comparison to verbal communication (text). If there is a viewing sequence, it is in the level of granularity (first

the global trend, then the local variations) rather than in any spatial arrangement (for example, first the top, then the bottom).

Graphs, being nonsequential, seem to resist methodologies altogether. The well-known books of Edward R. Tufte, such as *The Visual Display of Quantitative Information* [1], offer authoritative guidelines on graphical integrity and readability, illustrated by very diverse examples, but no method to go from data to graph. More quantitative books, such as those of William S. Cleveland in the United States [2] or Jacques Bertin in France [3], attach similar importance to a statistically sound encoding of data and propose powerful graphical representations, but they still fail to help readers choose the right graph in a given situation. Clearly inspired by programming languages, Leland Wilkinson’s recent book, *The Grammar of Graphics* [4], takes an original object-oriented approach to (re)constructing graphs and, as such, reviews a repertoire of representational “objects,” but still lays no explicit link between graph type and, for example, intended message.

Our own training and consulting experience reveals a poor graph literacy on the part of engineers, scientists, and managers. These professionals and others typically use the same few graph types for all their data sets, regardless of the amount and nature of their data. When asked how else they could graph the same data, they usually do not have a clue. Yet when shown a different graphical representation (new to them or not) of the same data, they recognize it as insightful; they just “didn’t think of graphing it that way.”

As part of a training effort about the visual representation of data, then, we developed a methodology to help engineers, scientists, and managers go from data to graph. The training aimed at broadening their repertoire of graph types, but especially at enabling them to choose the “right graph” from that repertoire in any given situation. This selection method is the object of the present paper.

THREE CRITERIA FOR CHOOSING A GRAPH

We have found that the effectiveness of a visual representation can be gauged against the following three criteria that can usefully guide the choice of graph type:

- the structure of the data set, that is, the number and type of variables;
- the intended use of the graph, from analysis to communication; and
- the research question or, conversely, the intended message.

The choice of representation can also be influenced by the tools used to produce the data, such as the physical layout of the experiment, or used to produce the graph, such as hardware (printers) or software (graphing applications).

The Structure of the Data Set

In quantity (number of variables)

and in quality (type of variables), the structure of the data set is an obvious first criterion for choosing the optimal graph. Yet many professionals seem to lack the vocabulary to describe the structure of their data, let alone use this structure to elect a graph type. A useful reference on data structure is Pyle’s *Data Preparation for Data Mining* [5].

Variables can be either continuous or discrete. Continuous variables represent series of numbers that can assume (in theory) all possible values, such as measurements of the temperature. They run along either an interval scale, with an arbitrary zero, such as degree Celsius, or a ratio scale, with an absolute zero, such as Kelvin. Discrete (also called “grouping”) variables represent series of “labels,” dividing the data into groups. They are located along either a nominal scale, such as gender (the values male and female cannot be ordered), or an ordinal scale, such as dosage (the values control, low, medium, and high can meaningfully be ordered).

The structure of the data set, in terms of number of continuous and of discrete variables, is not a given. Like any structure, it is a view of the mind. For example, if measurements have been made at both 30°C and 80°C, the variable temperature can be considered either continuous (with, as it happens, two actual values only) or discrete (with “30°C” and “80°C” being then labels more than numerical values). Similarly, concentrations of substances *a*, *b*, and *c* in a given solvent could be considered either three continuous variables or a combination of one continuous variable (concentration) and one discrete one (substance, with labels *a*, *b*, and *c*). The other two criteria will dictate which of the possible structures is most effective in a given situation.

The Intended Use of the Graph

The intended use of the

graph, in particular its intended audience, is a second criterion for selecting a graph. Audiences are sometimes described as the three Ps—personal, peer, and publication—but we prefer to think of the corresponding use of the graph, ranging from analysis or answering questions for oneself (personal) to communication or conveying messages to others (publication), possibly with discussion (peer) somewhere between pure analysis and pure communication. In practice, graphs that allow a rich analysis may not excel at conveying a message effectively and vice versa.

The intended use of the graph also influences the level of care bestowed upon its final production. Graphs designed for communication usually require or deserve more care. Realistically, graphs should not be perfect: they should be **optimal** for their intended use.

The Research Question At the analysis end, the research question or, conversely, the intended message at the communication end is another obvious criterion. Professionally, graphs are not drawn to store (or, worse, decorate) data, but to answer questions, either for oneself or for an audience. Again, no graph type is absolute; each makes answering some questions easier and other questions harder.

Research questions are complex and multiple, yet they can be grouped in the following four generic categories:

- **comparison** among individual data,
- **distribution** of data along a scale,
- **correlation** between variables, and
- **evolution** over time of a variable.

One fifth category, almost at a metalevel compared to the above four, is the comparison among **groups** of data. As such,

it is a comparison of different comparisons, distributions, correlations, or evolutions, and usually involves more complex displays. It results, of course, from the presence, in the data set, of a discrete variable.

Graphical displays usually encode a discrete variable in the form of either subsets or categories. Graphs with subsets distinguish among the various groups of data on a single view, using a visual difference such as color, plotting symbol, line thickness, or dash pattern (Fig. 1(a)). Graphs with categories display the various groups of data in separate, juxtaposed views (Fig. 1(b)); these views then use the same scales to allow meaningful comparisons.

When the data set involves more than one discrete variable, the resulting graphs can use multiple subsets (for example, with both different colors and different plotting symbols), multiple categories (for example, with views juxtaposed horizontally and vertically), or both subsets and categories.

THE IMPACT OF THE DATA MODEL

In two-dimensional (2-D) space, such as a sheet of paper or a computer screen, the data set is typically rendered as a table of values, whether numbers or labels. The way this table is built always reflects an underlying data model,

which we may sometimes impose, but which is usually hard-wired in the software application we use. The data model has a major impact on the way we structure our data set and, therefore, on the graphs we will be able to construct or, conversely, the research questions these graphs will be able to answer. The three most common data models encountered in graphing applications focus on cells, columns, and variables, respectively.

The cell-oriented data model, the archetype of which is Microsoft Excel, considers all “cells” of a “spreadsheet” homogeneously: it allows users to write any information in any cell without specifying any a priori relationship between cells (Fig. 2(a)). With some work, it can produce compact and orderly tables, yet it does not lend itself easily to graphing several views of the same data set or to accommodating additional data. (Excel’s predecessor, *Microsoft Chart*, was clearly designed for graphing data sets limited to a single continuous variable [6].)

The column-oriented data model, as used for example by SPSS Science’s SigmaPlot, is a constrained spreadsheet. It organizes data in columns, each identified by a column head, but does not associate each column univocally with a variable (Fig. 2(b)). Like the cell-oriented model, it encodes discrete variables implicitly, by repeating

existing columns as many times as necessary. While it allows easier graphing of continuous variables than an unconstrained spreadsheet, it does not encourage users to think in terms of discrete variables and thus to compare groups of data.

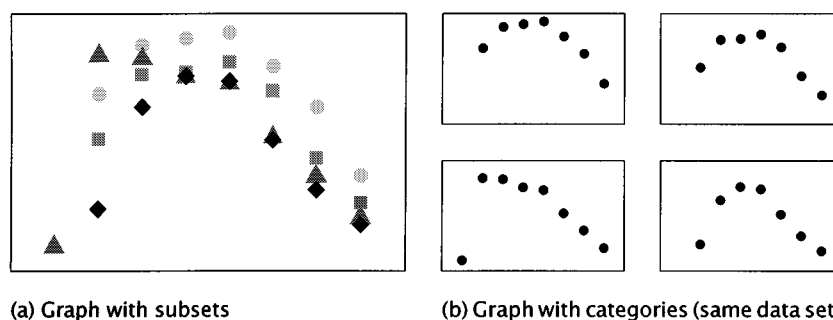
The variable-oriented data model, at the heart of statistical applications such as Insightful’s S-Plus, strictly associates one column of the table with one variable of the data set (Fig. 2(c)). While it yields longer, more redundant, and possibly less insightful tables than the other models, it encourages users to think in terms of data structure, not graph structure, and thus allows a more flexible analysis of the data set. Equally important, it accommodates additional continuous or discrete variables easily, by simply adding as many columns to the table.

Without surprise, graphing applications designed on cell- or column-oriented models have felt the limitations of their initial model as they attempted to add more capabilities. Recent versions of Microsoft Excel, for example, have thus incorporated more variable-oriented features, such as pivot tables.

A SIMPLE SELECTION TABLE

Using the first and third selection criteria above—structure of the data set and research question—as

Fig. 1. Graphical displays typically distinguish among the groups of data defined by a discrete variable with either subsets (a) or categories (b).



entries, one can turn a sequential repertoire of graph types into a much more useful selection table. A simple example of such a table is shown as Fig. 3, with graph types commented below. It could easily be extended to include more graph types or to show explicitly the corresponding graphs with subsets or with categories.

Bar charts and dot charts are maybe the two most basic representations of quantities along a numerical scale. Bar (or column) charts encode the data as lengths. To allow a meaningful comparison, they must be drawn along a linear (not logarithmic) ratio scale, starting from zero. Partial bars indeed mislead the viewer, even when accompanied by an explicit scale. Dot charts, by contrast, encode the data as positions along a scale, marked by dots. While somewhat less intuitive than (properly drawn) length representations, they can be used with any scale and can thus better resolve closely grouped data. They

also more easily accommodate additional information, such as subsets or whiskers (error bars). They have been much promoted by William S. Cleveland [2].

Histograms encode data as positions along a scale in the horizontal direction and corresponding frequencies as lengths in the vertical direction. While fairly intuitive to interpret, they are very sensitive to the origin and width of the intervals used to group data: a different choice of grouping intervals (wider, narrower, or simply shifted) may yield a very different picture of the distribution.

Box plots and related graphical representations provide a summary of the distribution of the data. Traditional boxes, with whiskers and a central point, are five-point summaries, corresponding for example (definitions indeed vary) to percentiles of 10%, 25%, 50% (median), 75%, and 90%, but

they can easily be extended to be nine-point summaries or to show individual extremes or outliers. Summaries are limiting, of course, especially for complex distributions, such as multimode ones. For small data sets, they are best replaced by individual data (sometimes called point plots). For large data sets, by contrast, they allow easy comparisons between groups of data, each summarized by one box.

Scatter plots, encoding the data as positions along two scales, reveal the shape and strength of the relationship between two continuous variables, as well as the presence of possible outliers. Three-dimensional (3-D) scatter plots, using three scales in a perspective view, are direct and sometimes useful generalizations, but are usually more difficult to visualize. A better alternative for three or more continuous variables may be the matrix plot, a juxtaposition of 2-D scatter plots—one for each pair of

Fig. 2. A data set structured along two continuous variables (XY) and one discrete variable (Z , with values a, b , and c) can be rendered in tabular form along three different data models. The cell- and column-oriented models encode discrete variables implicitly, as additional lines or columns. The variable-oriented model, while heavier to read as a table, encodes discrete variables explicitly; as a consequence, it is a more powerful and more flexible starting point for graphing the data set and, especially, for comparing groups of data.

(a) Cell-oriented data model <i>(cells are undifferentiated, a priori unrelated)</i>	(b) Column-oriented data model <i>(columns group data, do not match variables)</i>	(c) Variable-oriented data model <i>(each column matches a single variable)</i>																																																																																																																																																											
<table border="0" style="border-collapse: collapse;"> <tr> <td></td> <td colspan="3" style="text-align: center;">X</td> <td colspan="3" style="text-align: center;">Y</td> </tr> <tr> <td></td> <td style="text-align: center;">a</td> <td style="text-align: center;">b</td> <td style="text-align: center;">c</td> <td style="text-align: center;">a</td> <td style="text-align: center;">b</td> <td style="text-align: center;">c</td> </tr> <tr> <td style="text-align: center;">1</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">2</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">3</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">4</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">5</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> </table>		X			Y				a	b	c	a	b	c	1							2							3							4							5							<table border="0" style="border-collapse: collapse;"> <tr> <td style="text-align: center;">#</td> <td style="text-align: center;">X_a</td> <td style="text-align: center;">Y_a</td> <td style="text-align: center;">X_b</td> <td style="text-align: center;">Y_b</td> <td style="text-align: center;">X_c</td> <td style="text-align: center;">Y_c</td> </tr> <tr> <td style="text-align: center;">1</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">2</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">3</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">4</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> <tr> <td style="text-align: center;">5</td> <td></td><td></td><td></td><td></td><td></td><td></td> </tr> </table>	#	X_a	Y_a	X_b	Y_b	X_c	Y_c	1							2							3							4							5							<table border="0" style="border-collapse: collapse;"> <tr> <td style="text-align: center;">#</td> <td style="text-align: center;">X</td> <td style="text-align: center;">Y</td> <td style="text-align: center;">Z</td> </tr> <tr> <td style="text-align: center;">1</td> <td></td><td></td><td style="text-align: center;">a</td> </tr> <tr> <td style="text-align: center;">2</td> <td></td><td></td><td style="text-align: center;">a</td> </tr> <tr> <td style="text-align: center;">3</td> <td></td><td></td><td style="text-align: center;">a</td> </tr> <tr> <td style="text-align: center;">4</td> <td></td><td></td><td style="text-align: center;">a</td> </tr> <tr> <td style="text-align: center;">5</td> <td></td><td></td><td style="text-align: center;">a</td> </tr> <tr> <td style="text-align: center;">1</td> <td></td><td></td><td style="text-align: center;">b</td> </tr> <tr> <td style="text-align: center;">2</td> <td></td><td></td><td style="text-align: center;">b</td> </tr> <tr> <td style="text-align: center;">3</td> <td></td><td></td><td style="text-align: center;">b</td> </tr> <tr> <td style="text-align: center;">4</td> <td></td><td></td><td style="text-align: center;">b</td> </tr> <tr> <td style="text-align: center;">5</td> <td></td><td></td><td style="text-align: center;">b</td> </tr> <tr> <td style="text-align: center;">1</td> <td></td><td></td><td style="text-align: center;">c</td> </tr> <tr> <td style="text-align: center;">2</td> <td></td><td></td><td style="text-align: center;">c</td> </tr> <tr> <td style="text-align: center;">3</td> <td></td><td></td><td style="text-align: center;">c</td> </tr> <tr> <td style="text-align: center;">4</td> <td></td><td></td><td style="text-align: center;">c</td> </tr> <tr> <td style="text-align: center;">5</td> <td></td><td></td><td style="text-align: center;">c</td> </tr> </table>	#	X	Y	Z	1			a	2			a	3			a	4			a	5			a	1			b	2			b	3			b	4			b	5			b	1			c	2			c	3			c	4			c	5			c
	X			Y																																																																																																																																																									
	a	b	c	a	b	c																																																																																																																																																							
1																																																																																																																																																													
2																																																																																																																																																													
3																																																																																																																																																													
4																																																																																																																																																													
5																																																																																																																																																													
#	X_a	Y_a	X_b	Y_b	X_c	Y_c																																																																																																																																																							
1																																																																																																																																																													
2																																																																																																																																																													
3																																																																																																																																																													
4																																																																																																																																																													
5																																																																																																																																																													
#	X	Y	Z																																																																																																																																																										
1			a																																																																																																																																																										
2			a																																																																																																																																																										
3			a																																																																																																																																																										
4			a																																																																																																																																																										
5			a																																																																																																																																																										
1			b																																																																																																																																																										
2			b																																																																																																																																																										
3			b																																																																																																																																																										
4			b																																																																																																																																																										
5			b																																																																																																																																																										
1			c																																																																																																																																																										
2			c																																																																																																																																																										
3			c																																																																																																																																																										
4			c																																																																																																																																																										
5			c																																																																																																																																																										

variables—not unlike a chart of distances between cities.

Line plots, which are, in essence, sequenced scatter plots with connected dots, reveal the evolution of one variable versus another, typically time. Multiline plots compare the evolution of several variables expressed in the same units, so they can be graphed along the same scale, while multipanel plots relate the

evolution of several variables along different scales.

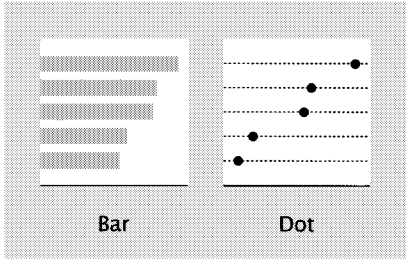
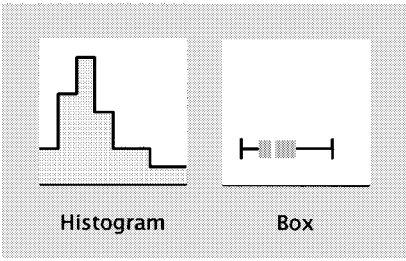
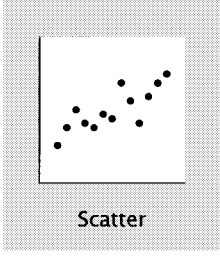
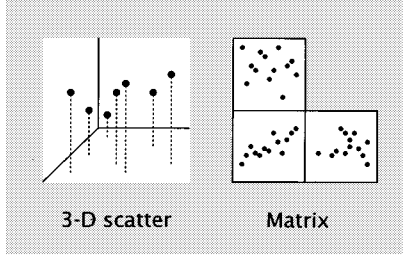
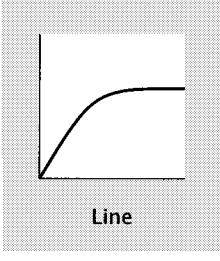
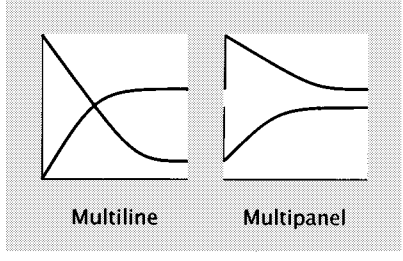
CONCLUSION

The methodology we developed to help engineers, scientists, and managers chose the “right graph” for their contents, audience, and purpose proved successful in the companies where we introduced it. Training participants, many of whom graph data daily for analysis and regularly for publication,

found it simple, innovative, and useful.

Still, the usefulness of the selection table depends largely on the relevance of the proposed graph types for the intended audience. We believe the success of our training programs comes partly from adapting each time the repertoire of graph types and the corresponding selection table to the specific graphing needs of the client company.

Fig. 3. A simple two-entry table to select candidate graph types on the basis of the structure of the data set (columns) and the research question (lines). The table can easily be extended to include more graph types.

	<i>One continuous variable</i>	<i>Two...</i>	<i>Three continuous variables</i>
Comparison	 <p>Bar Dot</p>		
Distribution	 <p>Histogram Box</p>		
Correlation		 <p>Scatter</p>	 <p>3-D scatter Matrix</p>
Evolution		 <p>Line</p>	 <p>Multiline Multipanel</p>

REFERENCES

- [1] E. R. Tufte, *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics, 1983.
- [2] W. S. Cleveland, *The Elements of Graphing Data*. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1985.
- [3] J. Bertin, *Sémiologie Graphique*. Paris, France: Mouton-Gauthiers-Villars, 1973.
- [4] L. Wilkinson, *The Grammar of Graphics*. New York: Springer-Verlag, 1999.
- [5] D. Pyle, *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann, 1999.
- [6] S. Lambert, "Presentation graphics primer," *MacWorld*, May/June 1984.

Jean-luc Doumont (S'90-M'93-SM'00) teaches and provides advice on professional speaking, writing, and graphing. He also trains trainers and facilitates any process that requires structuring and effective communication. For over 15 years, he has helped audiences of all ages, backgrounds, and nationalities structure their thoughts and construct their communication. He graduated as an engineer from the Université Catholique de Louvain, and obtained a Ph.D. in applied physics from Stanford University.

Philippe Vandebroeck consults widely in high-tech research organizations on data quality, information management, project management, and strategy. His interventions are designed to increase the client organization's problem-solving repertoire through the development of relevant and strong conceptual frameworks and their translation into focused action. He graduated as an agricultural engineer from the Katholieke Universiteit Leuven and obtained an M.A. in philosophy from the same university.