DATA VISUALIZATION PAST, PRESENT, AND FUTURE

STEPHEN FEW

PERCEPTUAL EDGE

Wednesday, January 10, 2007

COGNOS | INNOVATION CENTER
for Performance Management™

INTRODUCTION

Data visualization, the use of images to represent information, is only now becoming properly appreciated for the benefits it can bring to business. It provides a powerful means both to make sense of data and to then communicate what we've discovered to others. Despite their potential, the benefits of data visualization are undermined today by a general lack of understanding. Many of the current trends in data visualization are actually producing the opposite of the intended effect, confusion rather than understanding. Nothing going on in the field of business intelligence today can bring us closer to fulfilling its promise of intelligence in the workplace than data visualization. But this will happen only if we understand it and use it properly. We must embrace what really works and jettison the silly stuff that undermines data visualization today.

HISTORY OF DATA VISUALIZATION

To understand current and future trends in the field of data visualization, it helps to begin with some historical context. Despite the fact that predecessors to data visualization date back to the 2nd century AD, most developments have occurred in the last two and a half centuries, predominantly during the last 30 years.

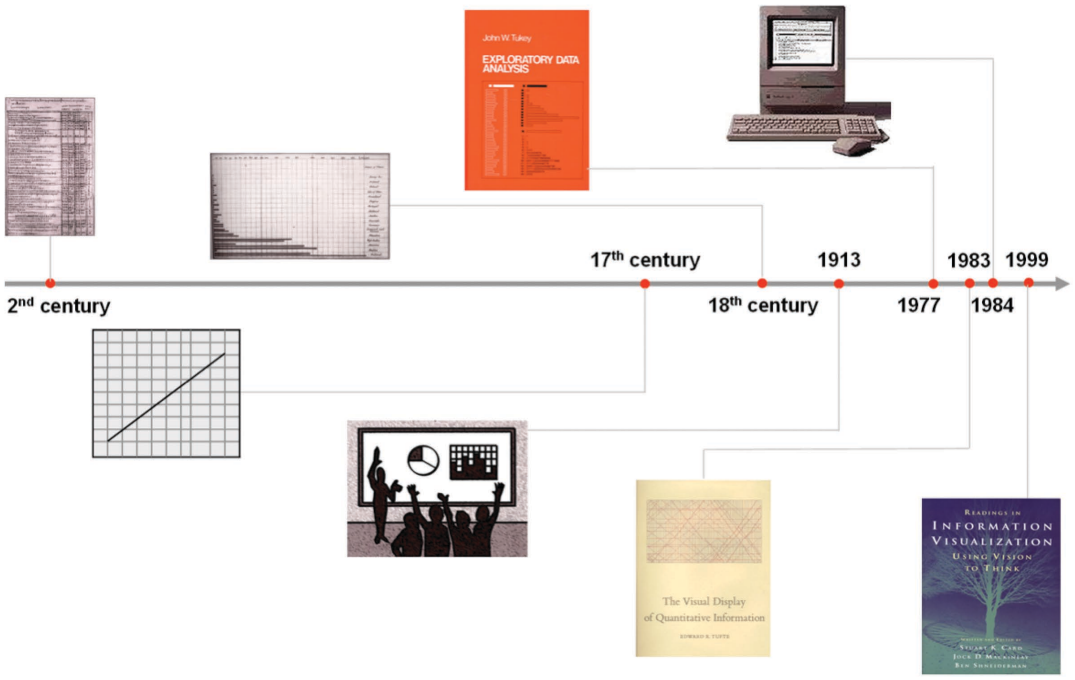


Figure 1: *History of data visualization timeline*

The earliest table that has been preserved was created in the 2nd century in Egypt to organize astronomical information as a tool for navigation. A table is primarily a textual representation of data, but it uses the visual attributes of alignment, white space, and at times rules (vertical or horizontal lines) to arrange data into columns and rows. Tables, along with graphs and diagrams, all fall into the class of data representations called charts. Although tables are predominantly

textual, their visual arrangement of data into columns and rows was a powerful first step toward later developments, which shifted the balance from textual and visual representations of data.

The visual representation of quantitative data in relation to two-dimensional coordinate scales, the most common form of what we call graphs, didn't arise until much later, in the 17th century. Rene Descartes, the French philosopher and mathematician probably best known for the words "Cogito ergo sum" ("I think therefore I am"), invented this method of representing quantitative data originally, not for presenting data, but for performing a type of mathematics based on a system of coordinates. Later, however, this representation was recognized as an effective means to present information to others as well.

Following Descartes' innovation, it wasn't until the late 18th and early 19th centuries that many of the graphs that we use today, including bar charts and pie charts, were invented or dramatically improved by a Scottish social scientist named William Playfair.

Over a century passed, however, before the value of these techniques became recognized to the point that academic courses in graphing data were finally introduced, originally at Iowa State University in 1913.

The person who introduced us to the power of data visualization as a means of exploring and making sense of data was the statistics professor John Tukey of Princeton, who in 1977 developed a predominantly visual approach to exploring and analyzing data called *exploratory data analysis*.

In 1983 data visualization aficionado Edward Tufte published his groundbreaking book *The Visual Display of Quantitative Information*, which showed us that there were effective ways of displaying data visually and then there were the ways that most of us were doing it, which were sadly lacking in effectiveness. One year later, in 1984, while we were watching the Super Bowl, Apple Computer introduced the first popular and affordable computer that focused on graphics as a mode of interaction and display. This paved the way for the use of data visualizations that we could view and interact with using a computer.

Given the availability of affordable computers with powerful graphics, a new research specialty emerged in the academic world, which was given the name "information visualization." In 1999 the book *Readings in Information Visualization: Using Vision to Think* collected this work into a single volume and made it accessible beyond the walls of academia.

In addition to these milestones in the development of data visualization, another event in the second half of the 20th century greatly influenced the quality of data visualization, but in the wrong direction: the proliferation of the IBM PC. Before the personal computer became commonplace in the workplace, if you needed to present data graphically, you were faced with a labor-intensive process involving the use of a T-square, draftsmen's triangles, and a collection of special pencils and pens. It sometimes took hours to produce a graph that could be displayed in a meeting or attached to a printed report. When the process took this much time and effort, people responsible for this work usually took time to develop graphical communication skills. But with the advent of the PC and the proliferation of business software such as the electronic spreadsheet, this changed. With the PC, the click of a mouse could transform a host of numbers into a graph,

and people who knew nothing about graph design suddenly became Rembrandts of graphical communication—or so they imagined. Despite Edward Tufte's efforts beginning in the 1980s, the quality of data visualization went largely ignored, especially in form of business graphs, despite their exponential growth.

Now that the stage has been set with the backdrop of history, let's take a look at what's happening today.

CURRENT TRENDS IN DATA VISUALIZATION

Today, data visualization is increasingly taking its rightful place as an important part of business intelligence. It is being talked about, investigated, requested by people who work with data, purchased by people who hold the purse strings, and used by a growing percentage of people in the workforce, especially analysts. That's the good news. The bad news is that, in the world of business, data visualization is still mostly ignored, largely misunderstood, used ineffectively, and too often undermined by the very vendors that produce and sell visualization software. The fact that you're reading this indicates that you want to learn about it and take full advantage of what it offers, so let's start with the good news and save the bad news as a warning about what to avoid for last.

Good Trends

Data visualization has in recent years become an established area of study in academia. Many universities now have faculty members who focus on visualization and a few have excellent programs that serve the needs of many graduate students who produce worthwhile research studies and prototype applications. This research community consists of people who are not just from computer science, but from many other disciplines as well, such as psychology and even business, which provides the context for a great deal of innovation while drawing on the robust practices of more mature disciplines.

We're beginning to see some data visualization products that actually work well. It still represents the minority, but a growing minority. Most of the best commercial visualization software has directly emerged from work that began as academic research. Efforts are currently under way, including my own, to bridge the gap between academic researchers with great ideas and business intelligence vendors who know how to build and sell commercially viable software products.

One of the encouraging new trends in business intelligence is the growing recognition that the greatest benefits of data visualization will come in the form of analytics. Visual analysis software allows us to not only represent data graphically, but to also interact with those visual representations to change the nature of the display, filter out what's not relevant, drill into lower levels of detail, and highlight subsets of data across multiple graphs simultaneously. This makes good use of our eyes and assists our brains, resulting in insights that cannot be matched by traditional approaches. Static graphs delivered on paper or electronically on a computer screen help us communicate information in a clear and enlightening way, which is a benefit that should not be undervalued, but it is from visual analytics that businesses will derive the greatest benefits.

One of the most powerful techniques of visual analysis involves the simultaneous display of multiple graphs, which feature either different subsets of data taken from a larger data set, or different views of a shared data set. Edward Tufte popularized a form of display that he calls *small multiples*, which uses a series of small graphs arranged together within eye span so they can be compared. Each graph represents a different subset of data belonging to a full data set, such as a series of line graphs that displays a company's expenses through time, with a separate graph per department. Small multiples greatly expand the number of variables (dimensions) that can be viewed together and compared. A different approach to the simultaneous display of multiple graphs uses each to examine a different aspect of a common data set. For instance, several graphs, perhaps of different types (bar graphs, line graphs, scatterplots, etc.), could be displayed together to simultaneously examine several aspects of a data set, allowing us to discover connections in the data that might not ever surface if the graphs were viewed separately. Visual analysis products that support displays such as these are rapidly becoming recognized for the rich analytical insights they make available to our eyes.

Despite my enthusiasm for the growing popularity of visual analytics, it is important to mention that something significant is also happening regarding the use of plain old graphs to communicate information. When you have something to say to others about data that you've examined, visual representations such as graphs or diagrams are often the best medium, but only if you know the language. Visual communication involves semantics and syntax, much like verbal language. You must know the rules to communicate effectively with graphs. Today, due in part to the pioneering work of Edward Tufte and William Cleveland beginning in the 1980s, and more recently to the efforts of Gene Zelazny, Naomi Robbins, and myself, the message is getting out that graphical communication requires fundamental skills that must be learned. I believe that these skills are quite easy to learn, but they aren't necessarily intuitive; it requires effort and the right resources.

No example of data visualization occupies a more prominent place in the consciousness of business people today than the dashboard. These displays, which combine the information that's needed to rapidly monitor an aspect of the business on a single screen, are powerful additions to the business intelligence arsenal. When properly designed for effective visual communication, dashboards support a level of awareness—a picture of what's going on—that could never be stitched together from traditional reports. Unfortunately, most dashboard products and most of the vendors that develop and sell them, fail to take full advantage of data visualization's power. Instead, these dashboards tend to look and function more like video games than serious information displays. In fact, many dashboards and dashboard products, while raising the visibility of data visualization, have only managed to give it a bad name due to poor design.

Another expression of data visualization that has captured the imagination of many in the business world in recent years is geo-spatial visualization. The popularity of Google Earth and other similar Web services have contributed a great deal to this interest. Much of the information that businesses must monitor and understand is tied to geographical locations. For instance, sometimes sales information can only be understood if you can see where those sales are occurring. In such cases, the ability to see measures such as sales revenues on a map adds a dimension of understanding that is critical. The ability to take advantage of location information that already resides in your systems, such as customers' zip codes, to display related information such as sales on a map is becoming increasingly available in business intelligence software and better integrated into the overall reporting and analysis experience every day.

Another trend that is only now beginning to find its way into business intelligence applications involves the use of visual animation (the movement of objects in charts) to show change through time. We have used line graphs for ages to effectively represent change through time. This works great when you are focusing directly on time-based information, for instance, measures of Web traffic taken at equal intervals of time, such as daily for the last month. But, what if you want to examine a different relationship between values, but also look at how it varies through time?

Consider the correlation between marketing expenses and resulting sales. The best way to examine this correlation at a particular point in time is by using a scatterplot, with marketing expenses measured along the X-axis (the horizontal axis), sales revenues along the Y-axis (the vertical axis), and a separate data point for individual items, such as one for each state, totaling fifty data points in all. If we want to see if the nature of this correlation has changed over the course of time, however, this correlation cannot be represented as effectively using a line graph, so what can we do to display it? The answer is that we can animate the scatterplot, allowing the data points representing marketing expenses and sales revenues for each state to move inside the scatterplot to show how these values have changed through time. Some of the best examples of using graph animations for this purpose have been developed by the folks at www.GapMinder.org, who use this technique to show important world data, such as the relationship between the income of countries and infant mortality, and how the world has changed in this respect over the last 30 years.

Another trend that has made the journey in recent years from the academic research community to commercial software tackles the problem of displaying large sets of quantitative data in the limited space of a screen. The most popular example of this is the treemap, which was initially created by Ben Shneiderman of the University of Maryland. Treemaps are designed to display up to two different quantitative variables at different levels of a hierarchy. For instance, you might be interested in examining all of the stocks that are traded on the New York Stock Exchange, arranged by industry, in a way that allows you to compare their prices and the amount of change in their prices since yesterday. That's a lot of data. You could try to use two horizontal bar graphs arranged side by side, one for stock prices and one for the change in prices, but you would quickly run into the limit to the number of bars that can be displayed in a bar graph. Treemaps provide a means of maximizing the amount of information that can be displayed on a screen, completely filling the available space with information. In Figure 2, you see a rectangle for each stock, arranged inside larger rectangles that group them into industries. The size (2-D area) of each rectangle represents its price and the change in price is represented by color, with negative values ranging from light to dark red, positive values ranging from light to dark green, and black for values that didn't change. Treemaps are not meant for precise comparisons between values, which is impossible when using 2-D area and color to encode them, but rather for quickly scanning a great deal of information to spot extremes (really large or small values) and predominant trends (for example, the fact that the greatest positive change occurred in the technology sector). When used for this purpose, the treemap is a newcomer to commercial data visualization software that has made quite a splash.

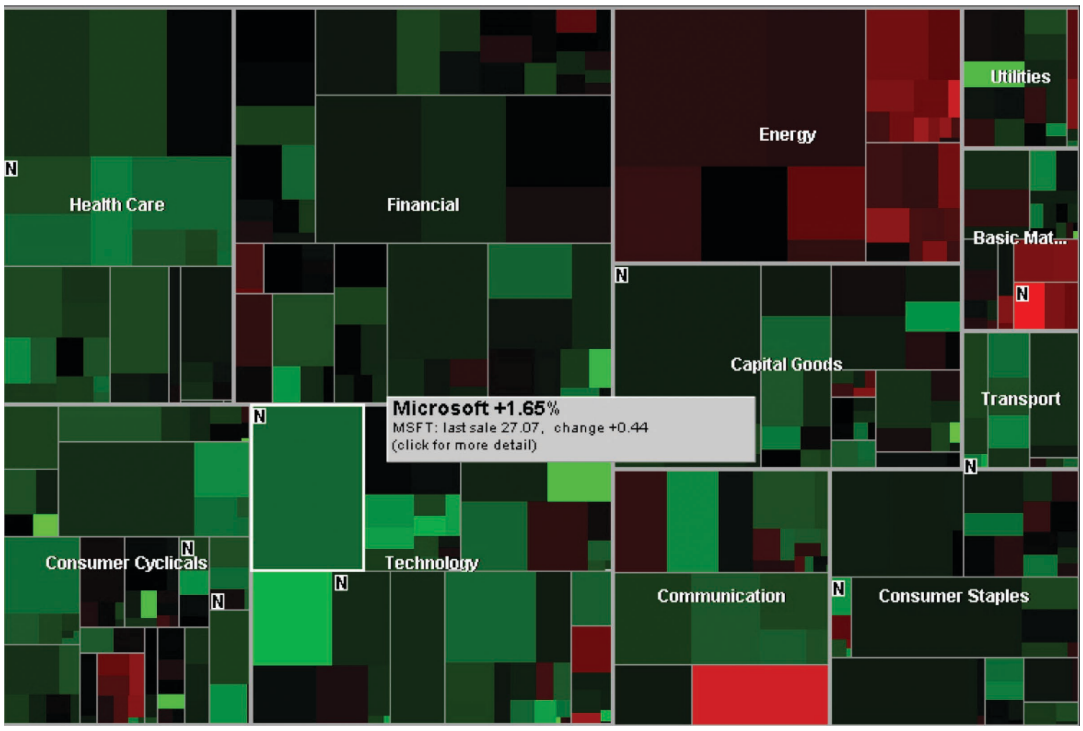


Figure 2: *This treemap displays information about the stock market (Source: www.SmartMoney.com). (Note: It would work even better if either the color green or red were replaced with another, because 10% of males and 1% of females cannot discriminate red and green, due to color blindness.)*

One final trend that has been making its way recently from the academic research lab to commercial software is a bit different from the other visualizations that I've mentioned because it focuses more on displaying relationships between entities (for example, companies or Web sites) than on quantitative values. Relationships between many things that interest us in business can be described as networks, with links connecting entities in a complex arrangement. For example, the Internet consists of Web sites that form a rich and complicated network of connections. Figure 3 displays the Internet in the form of a node and link visualization. Each node is a Web site and the lines that connect them represent hyperlinks between them. The thicker the line, the greater the number of links. Because this visualization is viewed on a computer screen and is designed to be interactive, you can easily focus on particular parts of the whole to examine them in greater detail, causing other entities to fade into the background. You could also filter out unwanted Web sites based on attributes such as the country it is located in or the type of site it is.

People in a company are also linked together in a complex network of connections, which could be examined based on emails between them. Besides network visualizations of the Internet, the other network visualization that has created quite a stir displays relationships between people, which are used to study social networks. This has been of particular interest to social scientists, but applications of interest to business are also beginning to emerge. Figure 3 shows a network of friends. Particular individuals have been highlighted, because two people were chosen and a feature in the application was invoked to highlight all the friends they have in common.

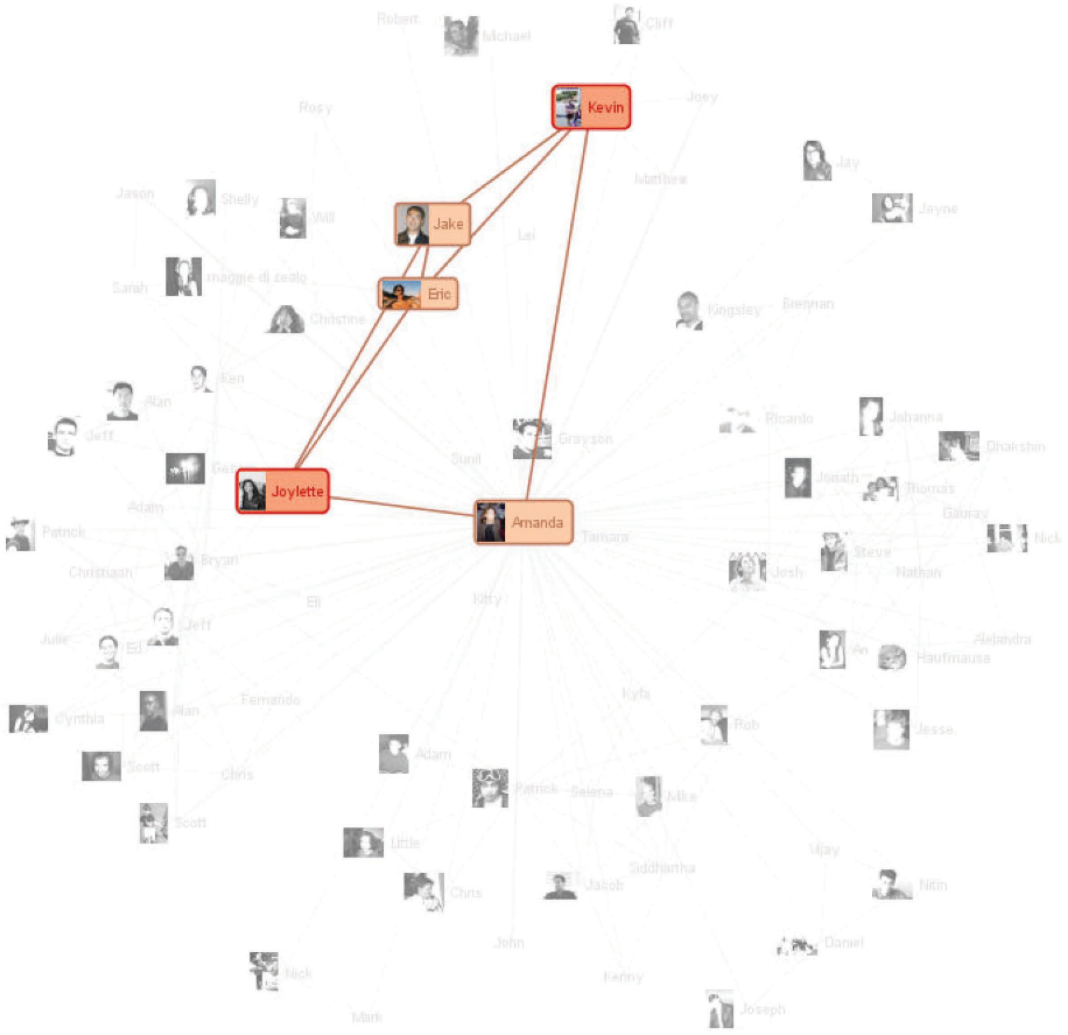


Figure 3: A node-link visualization of a social network by produced using Vizster

We've examined some of the recent positive trends in data visualization, but aspects of what's been going on recently are not all rosy. Let's note the bad trends as well to learn what to avoid.

Bad Trends

Whenever a new trend in information technology captures the interest of enough people to become popular, a great deal of confusion is created as everyone rushes to embrace it with little understanding of what it is and how it works. This has certainly been true of data visualization. Many vendors have rushed visualization products to market or rushed to add visualization functionality to existing products without taking the time to do it right. Marketing campaigns sometimes promote the worst in data visualization, substituting flash and dazzle for useful and effective functionality, which ends up

giving data visualization a bad name when people try to use it and fail. Dashboards are a good example. Most products look snazzy but work poorly. And because they are popular, every vendor calls whatever they sell that combines more than a single chart on a screen a dashboard. This misappropriation of the term robs dashboards of the clear definition that is required to say anything useful about them, such as how they ought to be designed to work effectively. Proper design for a true dashboard, which is used to monitor what's going on, is quite different from the design of an ad hoc collection of charts that you might view and interact with simultaneously on your computer screen to analyze data.

Data visualization is not about making things look cute or pretty. It is not about dressing up your presentations to dazzle your audience. We have a name for this kind of display: it is called "marketing." Data visualization has little to do with art, though aesthetics are involved in displaying data effectively. Rather, it mostly involves science, a set of rules based on what we know about visual perception and cognition, which we can follow to display information effectively. When these rules are understood and software is designed to support them, then, and only then, great things can happen.

FUTURE DIRECTIONS IN DATA VISUALIZATION

I'll begin this section with an expression of hope for the future, rather than a sure prediction of what will happen. Modern advances in data visualization have emerged from scientific research, rooted primarily in studies of visual perception and human cognition. These studies have explored the capacities and limitations of both to produce data visualization methods and applications that take advantage of our most powerful abilities and work around many of the limitations that hinder us. As such, data visualization is well equipped to assume a central role in business intelligence, for it is intelligence that it is tailored to foster.

This fact is obvious to anyone who has taken the time to understand the real needs and capacities of the business people who work with information to achieve the goal of business intelligence. For the promise of business intelligence to be fulfilled, we must shift our primary focus from technologies to human beings, from the machines that assist us to the human beings who alone possess the intelligence that businesses must rely on to succeed. With this as our primary focus, we will begin to see business intelligence software that is stripped of superfluous functionality and is designed to work so well that it becomes invisible, fading into the background of consciousness as our minds become immersed in the exploration of data. This is the future for which I hope and for which I struggle to usher in.

The data visualization research community and commercial software vendors have been involved in an interesting dynamic. Largely ignored by most vendors, researchers too often spend their time in work that ignores the real needs of business people, happily pursuing projects that have little or no real application. Without being faced with the demands of real people who need their help, why shouldn't researchers pursue whatever interests them, without any goal but exploration? A Catch-22 has been set in motion, because this state of affairs in turn reinforces the software vendors in their disinterest. It is up to us who actually work with information to understand it and to communicate our findings to others to break through this dynamic by making our demands heard. To the vendors we must make it clear that we won't settle for software that works poorly, focuses on what matters little, and is just plain hard to

use. We must stop being impressed with superficial features that add no real value to our work. When we do this, the vendors will then turn their attention to the research community for inspiration, and their focus will shift to work that addresses real needs.

In many matters involving technology, government tends to lag behind industry, but in this matter government is leading the way. Beginning in 2005, a United States government program involving National Visual Analytics Centers (NVACs) has established centers in several universities to explore and develop ways to apply data visualization to the analysis needs of various government agencies, especially intelligence agencies. These centers are forging partnerships with commercial software vendors to help bring these ideas to fruition. It is time for the business world to recognize its tardiness and add its weight and voice to the effort.

I expect that data visualization will continue for the next few years to pursue and mature those trends that have already begun. Dashboards, visual analytics, and even simple graphs will continue to develop and conform to best practices. I also have seen evidence that newer efforts are emerging that will soon develop into full-blown trends.

One of these new efforts involves the delivery of dashboards on smaller screens. As a tool for monitoring what's going on, dashboards often need to be available when you are not sitting in front of your computer, which means that they must be accessible through mobile devices such as cell phones and PDAs. The greater space limitations of these screens place greater demands on the design of dashboards, which will force even greater research into effective ways to display maximum content in a small space. Research groups are already working on this.

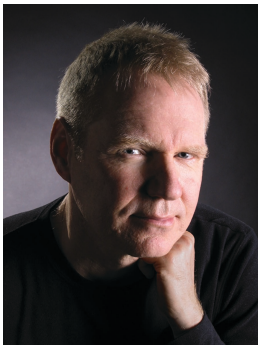
On the opposite end of the reduced form factor is the enlarged form factor: entire environments, such as specially equipped video rooms, that expand the amount of information that can be accessed for monitoring or analyzing information. If you saw the science fiction film Minority Report, you were probably fascinated when Tom Cruise donned a virtual reality headset and then faced a large display, which he was able to manipulate at the speed of thought to search through, examine, and make sense of information. Work to produce more immersive information environments of various types is taking place, some probably not too far from the vision shown in this movie. The objective is to make it easier to access more information and to assimilate and interact with it faster to gain understanding and then make decisions.

Several months ago, I watched a government-commissioned film, which presented a vision of how we might collaborate to analyze information in the future. The film was based on input from several information experts, including experts in data visualization. It reminded me a little of *Minority Report*. Everyone wore sensors that monitored several aspects of their physiology to detect conditions such as stress, which the system used to automatically tailor how it presented data on their displays to better match what they could handle. A team of analysts sat at a round table facing personal screens that were transparent from the rear so that the others could see through them without obstruction to maintain eye contact and fully view one another's expressions. Besides this particular project, a growing body of research is being done to find better ways for people to collaborate around the task of exploring and making sense of information. For instance, an entire team of data visualization experts at IBM Research are currently focused on this task. Collaboration, of course, doesn't require that people are in the same room with one another. Technology can provide ways to bring us together to share

information, discuss it, and make decisions even when we are in different parts of the world. I fully expect these efforts to pay off in better communication through the use of shared data visualizations.

One of the problems that plague data exploration and analysis involves large amounts of data that can be difficult to visualize in ways that don't overwhelm the viewer or hide what's important behind a wall of clutter. This is a problem that has been receiving a great deal of recent attention by the research community. Methods are being explored and sophisticated algorithms are being developed to tame the quantity of data either by reducing the amount in ways that avoid loss of meaning, or by reducing visual clutter in the visualization itself through novel approaches to the positioning of data objects, better uses of color, or other visual attributes such as transparency. This work is ideal for being included in commercial visualization software that is otherwise already effective.

Some of the greatest data visualization innovations of the next 10 years are without a doubt missing from this list, partly because they will grow from seeds that have not yet been planted. I am looking forward to being surprised and hope that you will join me in this exciting journey.



Stephen Few has worked for over 20 years as an IT innovator, consultant, and educator. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides consulting and training services, writes the monthly Visual Business Intelligence Newsletter, speaks frequently at conferences, and teaches in the MBA program at the University of California at Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and a new book entitled *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work at www.perceptualedge.com.

COGNOS®

THE NEXT LEVEL OF PERFORMANCE™