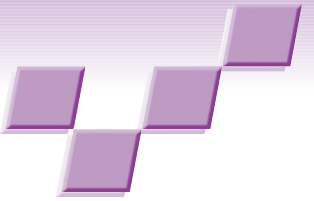

Discovering the Unexpected

Kris Cook
Pacific Northwest National Laboratory

Rae Earnshaw
University of Bradford, UK

John Stasko
Georgia Institute of Technology

> Visualization has been the cornerstone of scientific progress throughout history. Much of modern physics is a result of the superior abstract visualization abilities of a few brilliant men. Newton visualized the effect of gravitational force fields in three dimensional space acting on the center of mass. And Einstein visualized the geometric effects of objects in relative and uniform accelerated motion, with the speed of light a constant, time part of space, and acceleration indistinguishable from gravity. Virtually all comprehension in science, technology, and even art calls on our ability to visualize. In fact, the ability to visualize is almost synonymous with understanding. We have all used the expression "I see" to mean "I understand."[1]

The need to make sense of complex, conflicting, and dynamic information has provided the impetus for new tools and technologies that combine the strengths of visualization with powerful underlying algorithms and innovative interaction techniques; tools that make up the emerging field of visual analytics.[2] Visual analytics is the formation of abstract visual metaphors in combination with a human information discourse (usually some form of interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces. It is an outgrowth of the fields of scientific and information visualization but includes technologies from many other fields, including knowledge management, statistical analysis, cognitive science, decision science, and others.

This marriage of computation, visual representation, and interactive thinking supports intensive analysis. The goal is not only to permit users to detect expected events, such as might be predicted by models, but also to help users discover the unexpected—the surprising anomalies, changes, patterns, and relationships that are then examined and assessed to develop new insight.

The "Visualization Time Line" sidebar gives a brief summary of some of the key developments associated with visualization that have led to the current situation. In addition to introducing the articles in this special issue, this column sets out some of the key issues and challenges associated with discovering the unexpected.

Interfaces and interaction

In visual analytics, the key purpose of visualizations and interaction techniques is to help the user gain insight into complex data and situations where models alone are insufficient and human analytic skills must be employed. Visualizations must not only support the representation of critical data features but also provide sufficient contextual cues to help the user rapidly interpret what he or she is seeing. Interaction techniques strive to enable users to go beyond data exploration to achieve a dialogue with their information space to detect trends and anomalies, evaluate hypotheses, and uncover unexpected connections.

Computer scientists wish to develop effective interfaces to computers that facilitate communication and interaction between the human and the information in the machine. In the past, the importance of interface design has not always been fully recognized, or it may have been even ignored completely. Today, good design is increasingly recognized as being a key requirement for a user interface to be usable, flexible, and successful. With the current proliferation of computing devices, including mobile phones, PDAs, and other handheld devices, design is even more important in order to enable the user to manage the complexity that this introduces. With the intelligence in these devices, they can communicate with each other and reduce the cognitive load they place on the user. However, if information is filtered before it is presented to the user, how do we ensure that it is filtered appropriately and that key information that subsequently turns out to be important is not omitted or deleted?

Studies have focused on the ways that users interact with different kinds of devices. For example, the human perception of information on a mobile phone is different from that on a wall-size display. We need to be aware of these differences and the opportunities and constraints that they present both for the display of information and also the user's interaction with it.

Published by the IEEE Computer Society 0272-1716/07/$20.00 © 2007 IEEE

Visualization Time Line

Visualization in a presentation sense has been used for at least a thousand years.

- 1952 – A.S. Douglas receives a PhD at the University of Cambridge on human-computer interaction using the CRT display on Edsac 1 computer (http://www.pong-story.com/1952.htm).
- 1962 – Jack E. Bresenham develops one of the first graphics algorithms.
- 1963 – Ivan E. Sutherland's sketchpad system uses graphical techniques for human-machine communication (http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-574.pdf).
- 1963 – General Motors Research develops the DAC-I system, the first computer-aided design system (paper published at the 1963 American Federation of Information Processing Societies Fall Joint Computer Conference).
- 1967 – ACM Special Interest Group on Graphics and Interactive Techniques is founded.
- 1973 – William Newman and Robert Sproull publish *Principles of Interactive Computer Graphics* (McGraw-Hill).
- 1974 – Ted Nelson publishes *Computer Lib/Dream Machines*.
- 1974 – ACM Siggraph hosts first annual ACM Siggraph Conference.
- 1977 – Apple releases Apple II micro with color graphics.
- 1977 – ACM Siggraph hosts first annual ACM Siggraph Workshop on User-oriented Design of Interactive Graphics Systems.
- 1978 – ACM Special Interest Group on Social and Behavioral Computing hosts a conference on "People-oriented Systems: When and How?"
- 1982 – Silicon Graphics is founded, developing 3D graphics terminals and workstations.
- 1982 – ACM Special Interest Group on Computer-Human Interaction is founded.
- 1983 – *The Psychology of Human-Computer Interaction*, edited by Stuart K. Card, Thomas P. Moran, and Allen Newell, is published (Lawrence Erlbaum Assoc.).
- 1983 – Edward R. Tufte publishes *The Visual Display of Quantitative Information*.
- 1983 – ACM Sigchi hosts first annual conference.
- 1987 – NSF panel publishes a report on *Visualization in Scientific Computing* (McCormick, DeFanti, and Brown). Principal recommendations: National funding should be granted for short- and long-term provision of tools and environments to support scientific visualization, and these should be made available to the community at large.
- 1987 – Volume data. William Lorensen and Harvey Cline publish their Marching Cubes algorithm (http://en.wikipedia.org/wiki/Marching_cubes).
- 1989 – IEEE hosts first annual IEEE Visualization Conference.
- 1990 – The World Wide Web debuts.
- 1990 – Gregory Nielson and Bruce Shriver publish *Visualization in Scientific Computing*, the first book on visualization (IEEE Computer Society Press).
- 1992 – OpenGL real-time 3D graphics standard is published.
- 1992 – Modular visualization environments. Users create their application interactively by connecting modules using a point-and-click interface (for example, AVS, Khoros, Iris Explorer).
- 1995 – IEEE hosts first annual Information Visualization Conference.
- 1998 – Collaborative visualization. Networked multiuser environments developed.
- 1999 – *Readings in Information Visualization*, edited by Stuart Card, Jock Mackinlay, and Ben Shneiderman, is published (Morgan Kaufmann).
- 2000 – Robert Spence publishes *Information Visualization* (Addison-Wesley).
- 2004 – National Visualization and Analytics Center is founded.
- 2005 – *Illuminating the Path: the R&D Agenda for Visual Analytics*, edited by James Thomas and Kristin Cook, is published (IEEE Computer Society Press).
- 2006 – NIF/NSF publishes a report on *Visualization Research Challenges* (http://tab.computer.org/vgtc/vrc/index.html)
- 2006 – IEEE hosts first annual IEEE Visual Analytics Symposium.

A more detailed version of this time line, which you can add to, is available at http://www.inf.brad.ac.uk/home/Visualization Time Line Long v2.doc.

Models and data

Data complexity inherently complicates the analytic process. Some analytic challenges require the understanding of massive volumes of data, such as simulation data or network data. In other cases, the complexity results not from the data's large scale but from the diversity of the data types required for analysis. In still other situations, data that is readily interpretable by humans, such as text, is much more difficult for a computer to interpret. In cases where well-formed models can be reliably constructed for identifying information and situations of interest, they can form the basis for automated data analysis. However, in situations that are not well understood, or in which the purpose of the analysis is to detect surprising information, traditional models alone will not suffice. These models must be augmented with feature extraction techniques that draw on statistical and mathematical approaches, as well as mathematical representations that simplify data in ways that are appropriate to the task at hand.

Cognitive loading

A human can observe information being displayed in real time or explore an information space using interactive techniques. However, what the human brain can receive is limited in terms of information that it must process and make judgments about, often in the context of adjacent information either in time or space in the display environment. More specifically, it commonly refers to the load on the human's working memory during problem solving, reasoning, and thinking. Cognitive load theory, as defined by Sweller,[3] states that optimum learning occurs in humans when the load on working memory is kept to a minimum to

best facilitate the changes in long-term memory. Displaying information in visual form may circumvent this to some degree, but not all visual representations may be appropriate to searching for new pieces of information or anomalies in the data. This suggests the cognitive and human-computer interface aspects of visualization are extremely important, and until these issues are addressed effectively, information and knowledge will remain undiscovered, at least where computers are being used.

Further challenges

One of the difficulties with discovering new information is that it often lies outside the boundaries of the current investigation, or it may be transitory—only present for a particular period of time. In certain circumstances these time constraints can be external. In security investigations, for example, we may be given a time limit within which an investigation must be conducted. If no new information is uncovered within a specified time interval, the investigation must be concluded. What methods might be used to optimally home in on areas where new information may be found? Our investigations therefore could be subject to internal and external time constraints. If we knew what we were looking for, we would open up the relevant part of the boundary or time window to ensure we could investigate it.

In addition, if we don't know what to expect, how do we know if we have found it? According to the principle of falsifiability, defined by Karl Popper in the 1960s, progress in scientific discovery and understanding is made through the iterative refinement of existing theories by discovering new information that is inconsistent with the theory.[4] Thomas Kuhn has found little evidence of this and has argued that scientists work more in a series of paradigms[5] —hence, the use of the term *paradigm shift*.

Given the increasing size and complexity of data sets produced by laboratory experiments or the observations of natural phenomena, the volume of data to be analyzed is a major challenge. Even with interactive visual tools and sophisticated data analysis algorithms, this is still difficult and time-consuming.

Sense-making

Once data has been gathered and organized into forms to facilitate further inquiry, analysts perform a variety of sense-making activities on and with them. One would hope to be so fortunate that the threads of evidence and discovery fit together seamlessly to expose the greater insights embedded in the data, but this is rarely the case in practice. Analysts must make connections between disparate pieces of data and begin to construct plausible scenarios of the bigger picture.

Visual analytic systems like those the articles in this issue describe can help analysts examine the data under new perspectives or simply in a fashion that makes it easier to understand the trends, themes, and relationships the data suggests. Essentially, analysts construct schema that map the facts and data being examined into higher-order plans and activities. Visual analytic tools assist in the evidence gathering and information foraging aspects of this process as well as the integration and construction of new knowledge phases.

Analytical reasoning

Discovery of the unexpected is a critical part of the analytical reasoning process. When people are trying to make sense of their data to understand situations and decide on an action, they develop various scenarios for actions and their outcomes then evaluate data against their mental models of these scenarios to determine how to maximize the outcome. People must be able to identify unexpected information and have support for incorporating that information into their thought processes to determine not only how it affects the potential outcomes that they envision, but also whether it invalidates the potential scenarios themselves.

Visual analytic systems can help analysts examine data under new perspectives or simply in a fashion that makes it easier to understand the trends, themes, and relationships the data suggests.

However, this can be a challenging process. To take a simple example, when using computer systems, users often stick with the particular settings they have always used (often the defaults), even though other settings might be better. In more complex analytic situations, cognitive biases can prevent us from seeing and interpreting information accurately. Tools and techniques are needed to help overcome users' inherent human limitations to be able to see and truly understand their data.

Discovering knowledge

Using data analysis algorithms to search large volumes of data might uncover new items of information. However, their significance may be related to other items of information that are not discovered. Thus, only a partial, and perhaps erroneous, picture is obtained. Is it possible to adopt a more holistic approach that uncovers all the new and unexpected pieces of information in a data set—and the relationships between them?

More traditional information discovery approaches have relied on search engines to find significant pieces of data. This is appropriate for some problems. However, the significance of one piece of data may lie more in its relationship to another piece of data so that the total is more than the sum of the discrete parts. Furthermore, the importance of new information may be apparent only in the context of the user's understanding of the problem at hand. Although information discovery can be supported by learning techniques such as hybrid neural networks and genetic algorithms, the human user's understanding of the situation plays a

key role in discovering knowledge and developing insight.

Stephen H. Muggleton says:

> During the twenty-first century, it is clear that computers will continue to play an increasingly central role in supporting the testing, and even formulation, of scientific hypotheses. This traditionally human activity has already become unsustainable in many sciences without the aid of computers. This is not only because of the scale of the data involved but also because scientists are unable to conceptualize the breadth and depth of the relationships between relevant databases without computational support. The potential benefits to science of such computerization are high—knowledge derived from large-scale scientific data could well pave the way to new technologies, ranging from personalized medicines to methods for dealing with and avoiding climate change [*Towards 2020 Science* (Microsoft, 2006); http://research.microsoft.com/towards2020science]. … Meanwhile, machine-learning techniques from computer science (including neural nets and genetic algorithms) are being used to automate the generation of scientific hypotheses from data. Some of the more advanced forms of machine learning enable new hypotheses, in the form of logical rules and principles, to be extracted relative to predefined background knowledge. … One exciting development that we might expect in the next ten years is the construction of the first microfluidic robot scientist, which would combine active learning and autonomous experimentation with microfluidic technology.[6]

The articles in this special issue

"Visual Discovery in Computer Network Defense," by D'Amico et al., explores using visual tools to assist in locating patterns of network activity in large volumes of data. It also aims to provide a framework that synchronizes with the cognitive and operational requirements of analysts who work in this field. Since this approach is designed to uncover both known and currently unknown forms of user activity, it is designed to assist with the discovery of new forms of activity—that is, unexpected within the current framework of activity. Thus, the approach seeks to extend the boundaries of current systems.

"Insights Gained through Visualization for Large Earthquake Simulations," by Chourasia et al., applies visualization techniques to simulations using massive data sets. The objective is to make the simulation as close to the physical situation as possible, so that the simulation can be predictive of the future. The visualizations have enabled instabilities in the simulation process to be uncovered and also delivered new results in the end points of the simulations that the seismologists did not expect.

In "Visualizing Diversity and Depth over a Set of Objects," by Pearlman et al., the authors have developed tools to understand the attributes of a set's members. They used two parameters: depth, which refers to the prevalence of the distribution of attribute values in the set, and diversity, which refers to the distribution of these values across a range. Composite representations capture these values in the set and communicate this information to the user. The technique has been studied in three application domains and delivered some unexpected results.

In "nSpace and GeoTime: A VAST 2006 Case Study," Proulx and his colleagues discuss the use of their visual analytics systems in working on the 2006 IEEE Symposium on Visual Analytics Science and Technology Contest. The authors describe the analytic processes undertaken in working on this challenge and how these systems assisted their exploration and sense-making activities. Their suite of tools combines data analysis algorithms and techniques with flexible visualizations and user interfaces, resulting in an environment that allows analysts to pose and research hypotheses about the data.

"Bridging the Semantic Gap: Visualizing Transition Graphs with User-Defined Diagrams," by Pretorius and van Wijk, presents a method for assisting with the sense-making of data. Custom diagrams convey the semantics associated with the data. Two applications of the technique to large real-world applications show how new properties were discovered and an unknown error was identified. ■

References

1. J.H. Clark, "Foreword," *An Introductory Guide to Scientific Visualization*, R.A. Earnshaw and N. Wiseman, Springer Verlag, 1992.
2. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, J.J. Thomas and K.A. Cook (eds.), IEEE CS Press, 2005, p. 186.
3. J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science*, vol. 12, no. 1, 1988, pp. 257-285.
4. K. Popper, *The Logic of Scientific Discovery*, Routledge, 2002 (originally published 1959).
5. T.S. Kuhn, *The Structure of Scientific Revolutions*, Univ. Chicago Press, 1996 (originally published 1962).
6. S.H. Muggleton, "Exceeding Human Limits," *Nature*, vol. 440, 2006, pp. 409-410.

Kris Cook *is a project manager at Pacific Northwest National Laboratory, where she has led R&D efforts in information visualization and visual analytics projects for the past 11 years. As part of the leadership team for the National Visualization and Analytics Center, she coordinates the work of five Regional Visualization and Analytics Centers at universities throughout the United States. She has a BS in chemical engineering from The Ohio State University. Contact her at kris.cook@pnl.gov.*

Rae Earnshaw *is pro vice-chancellor (Strategic Systems Development) at the University of Bradford, UK, and professor of electronic imaging and media communications in the School of Informatics. He has authored and edited 33 books on computer graphics, visualization, multimedia, design, and virtual reality, and published 140 papers in these areas. He has a PhD in computer science from the University of Leeds. He is a member of ACM, IEEE, Computer Graphics Society, Eurographics, a Fellow of the British Computer Society, a Fellow of the Institute of Physics, and a Fellow of Royal Society of Arts. Contact him at r.a.earnshaw@bradford.ac.uk.*

John Stasko *is a professor in the School of Interactive Computing and the Graphics, Visualization, and Usability Center at the Georgia Institute of Technology, where he is director of the Information Interfaces Research Group. His research is in the area of human-computer interaction with a specific focus on information visualization, visual analytics, and the peripheral awareness of information. He has a PhD in computer science from Brown University. Stasko is on the editorial staff of five journals focusing on the topics of visualization and HCI, and is on the steering committee for the IEEE Information Visualization Conference. Contact him at stasko@cc.gatech.edu.*

Coming Next Issue: Real-Time Interaction with Complex Models

Interacting with 3D models of almost unlimited size and complexity is a key challenge in computer graphics and scientific visualization. Such models can contain millions, even billions of 3D primitives, such as point sets, surfaces, voxels, and higher dimensional data sets, and each data set is often associated with a complex set of parameters. Many techniques accelerate the management and interaction of large data sets based on sample-based representation and rendering, polygon rasterization hardware, and ray tracing.

IEEE Computer Graphics and Applications