UC Davis
IDAV Publications

Title
Visualizing Large-Scale Data Sets: Challenges and Opportunities, Panel Presentation

Permalink
https://escholarship.org/uc/item/4pc035j8

Authors
Eick, S.
Hamann, Bernd
Heermann, P.
et al.

Publication Date
1999

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Visualizing Large-Scale Datasets: Challenges and Opportunities

Organizer
Kwan-Liu Ma
University of California, Davis

Moderator
John Van Rosendale
US Department of Energy

Panelists
Stephen Eick
Visual Insights/Lucent Technologies

Bernd Hamann
University of California, Davis

Philip Heermann
Sandia National Laboratory

Christopher Johnson
University of Utah

Mike Krogh
Computational Engineering International Inc.

While we are seeing an unprecedented growth in the amount of data from both computational simulations and instrument/sensor sources, our ability to manipulate, explore, and understand large datasets is growing less rapidly. Visualization transforms raw data into vivid 2D or 3D images that help scientists reveal important features and trends in the data, convey ideas, and communicate their findings. However, massive data volumes create new challenges and makes previous visualization approaches impractical. The new generation of visualization methods must scale well with the growing data volumes and cope with other parts of the data analysis pipeline, such as storage devices and display devices.

To accelerate the development of new data manipulation and visualization methods for massive datasets, the National Science Foundation and the US Department of Energy have sponsored a series of workshops on relevant topics. The workshops have generated a concept, Data and Visualization Corridors, that represents a combination of innovations on data handling, representations, telepresence, and visualization. In the next few years, we expect to see more human and financial resources invested to solve the problem of visualizing large-scale datasets as more demanding applications emerge.

In this panel, the findings and results of the workshop on Large-Scale Visualization and Data Management (Salt Lake City, Utah USA) are reported. The panel and the audience collaborate to answer the following questions:

- How large is large?
- Where do large datasets come from?
- Can current graphics and visualization technology cope with the volume and complexity of data produced by tera-scale calculations or high-resolution/high-volume data collection devices?
- How much of the data do we need to see, and how do we find what we need to see?
- What are ideal data representations that can enable more efficient visualization?
- How much processing power, storage space, bandwidth, and display resolution do we need?
- How much visualization computing should we do at runtime, when the data are being created, vs. at postprocessing time?
- Is computational steering a reality?
- Are there common visualization solutions for scientific, engineering, medical, and business data?
- What can the visualization software industry offer now and in the near future?

Stephen Eick

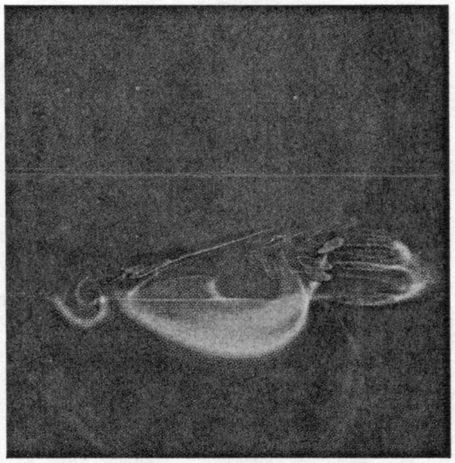
The amount of data collected and stored electronically is doubling every three years. With the widespread deployment of DBMS systems, penetration of networks, and adaptation of standard data interface protocols, the data access problems are being solved. The newly emerging problem is how to make sense of all this information. The essential problem is that the data volumes are overwhelming existing analysis tools. Our approach to solving this problem involves computer graphics. Exploiting newly available PC graphics capabilities, our visualization technology:

1. Provides better, more effective, data presentation.

2. Shows significantly more information on each screen.

3. Includes visual analysis capability.

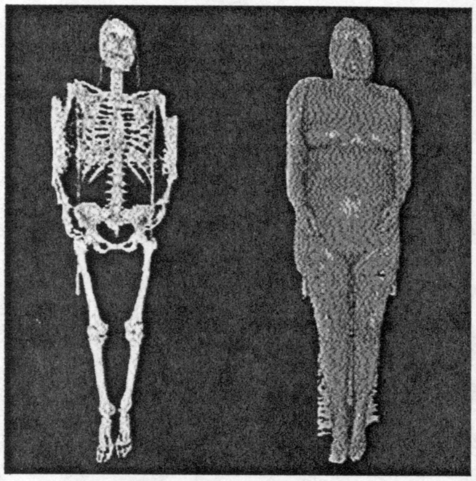
Visualization approaches, such as ours, have significant value for problems involving change, high dimensionality, and scale. In this visualized space, the insights gained enable decisions to be made faster and more accurately.



Volume visualization of data from high-lift analysis revealing flow structures surrounding an aircraft wing. The data contains over 18 million tetrahedra, and the visualization was generated using a parallel computer. Image provided by Kwan-Liu Ma, University of California, Davis.

Bernd Hamann

We are now reaching the limits of interactive visualization of large-scale datasets. This is to be interpreted in two ways. First, the sheer amount of data to be analyzed is overwhelming, and researchers do not have enough time available to "browse" and visually inspect an extremely high-resolution dataset. Second, the rendering-resolution capabilities of current rendering and projection devices are too "coarse" to visually capture the important, small-scale features in which a researcher is interested.

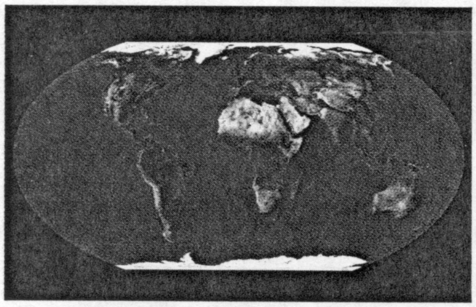


Ray tracings of the bone and skin isosurfaces of the Visible Woman. Image provided by Christopher Johnson, University of Utah.

Two active areas of research can help in this context: multiresolution methods used to represent and visualize large datasets at multiple levels of resolution, and automatic methods for extracting features defined a priori, and identifying regions characterized by "unusual" behavior. Multiresolution methods help in reducing the amount of time it takes a researcher to "browse" the domain over which a physical phenomenon has been measured or simulated, while automatic feature extraction methods assist in steering the visualization process to those regions in space where a certain interesting or unusual behavior has been identified.

Using a full system design approach, visualization requirements are compared with technology trends to quantify the visualization system requirements. Researchers are exploring data reduction and selection, parallel data streaming, and run-time visualization techniques. The system performance goals require each technique to consider a balanced combination of hardware and software.

In summary, multiresolution and automatic feature extraction methods both serve the same purpose. They reduce the amount of time required to visually inspect a large dataset. We should investigate in more depth the synergy that exists between these two approaches. For example, one could envision a coupling of these two methodologies by applying feature extraction methods to the various levels in a pre-computed multiresolution data hierarchy, which would lead naturally to extraction and representation of qualitatively relevant information at multiple scales.



Earth surface image mapped onto a 9-km digital elevation model and rendered (in parallel) at 3000x2400 resolution using a Wagner IV equi-arial projection. Image provided by Tom Crockett, ICASE.

Philip Heermann

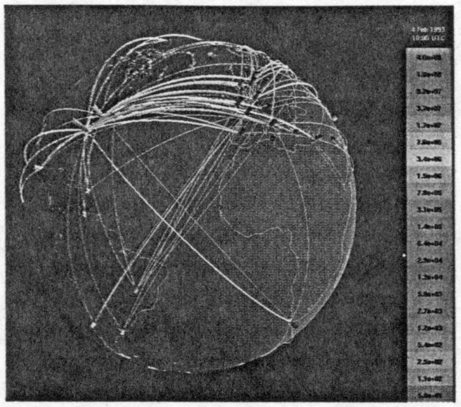
The push for 100 teraflops, by the US Department of Energy's Accelerated Strategic Computing Initiative (ASCI) Program, has driven researchers to consider new paradigms for scientific visualization. ASCI's goal of physics calculations running on 100 teraflops computers by 2004 generates demands that severely challenge current and future visualization software and hardware. To address the challenge, researchers at Lawrence Livermore, Los Alamos, and Sandia National Laboratories are investigating new techniques for exploring the massive data.

The leap forward in computing technology has impacted all aspects of visualizing simulation results. The datasets produced by ASCI machines can greatly overwhelm common networks and storage systems. Data file formats, networks, processing software, and rendering software and hardware must be improved. A systems engineering approach is necessary to achieve improved performance. The common approach of improving a single component or algorithm can actually decrease performance of the overall system.

Christopher Johnson

Interaction with complex, multidimensional data is now recognized as a critical analysis component in many areas, including computational fluid dynamics, computational combustion, and computational mechanics. The new generation of massively parallel computers will have speeds measured in teraflops and will handle dataset sizes measured in terabytes to petabytes. Although these machines offer enormous potential for solving very large-scale realistic modeling, simulation, and optimization problems, their effectiveness will hinge upon the ability of human experts to interact with their computations and extract useful information. Since humans interact most naturally in a 3D world, and since much of the data in important computational problems have a fundamental 3D spatial component, I believe the greatest potential for this human/machine partnership will come through the use of 3D interactive technologies.

Within the Center for Scientific Computing and Imaging at the University of Utah, we have developed a problem-solving environment for steering large-scale simulations with integrated interactive visualization called SCIRun. SCIRun is a scientific programming environment that allows interactive construction, debugging, and steering of large-scale scientific computations. SCIRun can be envisioned as a "computational workbench," in which a scientist can design and modify simulations interactively via a dataflow programming model. It enables scientists to modify geometric models and interactively change numerical parameters and boundary conditions, as well as to modify the level of mesh adaptation needed for an accurate numerical solution. As opposed to the typical offline simulation mode, in which the scientist manually sets input parameters, computes results, visualizes the results via a separate visualization package, then starts again at the beginning, SCIRun "closes the loop" and allows interactive steering of the design, computation, and visualization phases of a simulation.

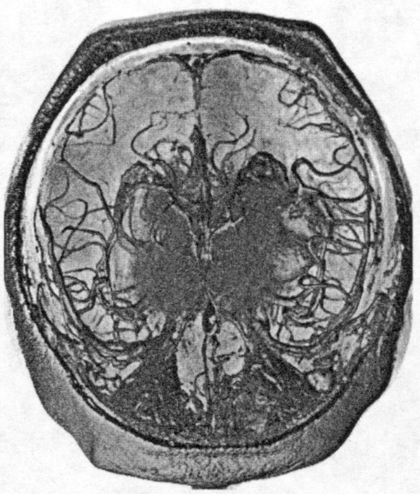


One frame from an animation showing world-wide internet traffic over a two-hour period. Image provided by Stephen Eick, Visual Insights/Lucent Technologies,

Mike Krogh

The DOE's ASCI program brought about the recent advent of terascale supercomputers. These machines, and the calculations that they perform, are magnitudes larger than what is typically found in industry. A reasonable question is: Is commercial visualization software a viable option for large data visualization?

Visualization was identified in the landmark 1987 NSF report "Visualization in Scientific Computing" as essential for analysis, understanding, and communication of scientific data. In a second report issued in 1998 by the DOE and NSF ("Data and Visualization Corridors"), visualization was once again identified as an essential technology without which "we risk flying blind if visualization does not keep pace with simulations."



Volume visualization of cerebral arteries in the brain to isolate a large aneurism on the right side of the image. Image provided by Christopher Johnson, University of Utah.

Visualization is still correctly identified as essential to computational science. However, even after 12 years, the 1998 report indicates that further research is required for scientists to understand the data they are generating. So in this context, how does a commercial package measure up to the demands of terascale computing? In particular, is such software a viable option for supercomputer users and their management? What features do users want? What, if anything, do they have to be willing to sacrifice? For the software provider, what hurdles must be faced? Where are the overlaps between mainstream and bleeding-edge requirements? And where does the real research lie?