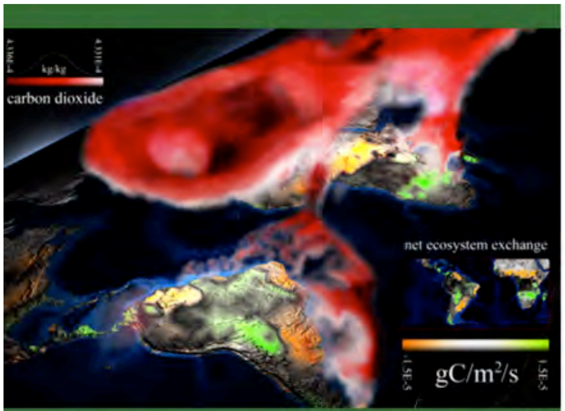
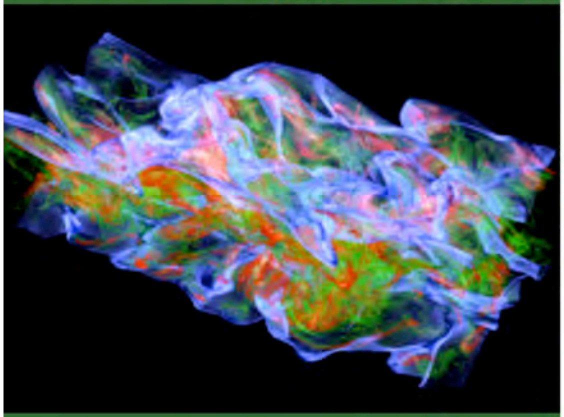
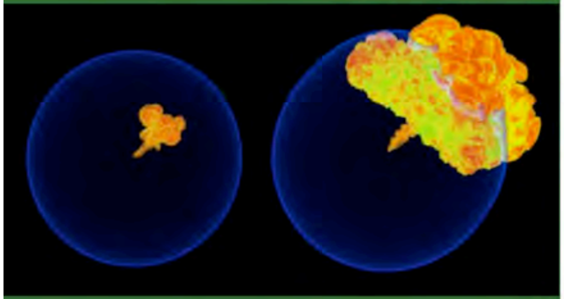






Visualization and Knowledge Discovery:

Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale

October 2007

Workshop Co-Chairs:

Chris Johnson, University of Utah
Rob Ross, Argonne National Laboratory

Workshop Working Group Co-Chairs:

Sean Ahern, Oak Ridge National Laboratory
Jim Ahrens, Los Alamos National Laboratory
Wes Bethel, Lawrence Berkeley National Laboratory
Kwan-Liu Ma, University of California, Davis
Michael Papka, Argonne National Laboratory
John van Rosendale, College of William and Mary
Han-Wei Shen, Ohio State University
Jim Thomas, Pacific Northwest National Laboratory

Executive Summary

The Department of Energy's Office of Science continues to make significant strides in providing for the future of the nation's energy and economic security. It has created unmatched scientific facilities and installations to perform work in areas ranging from high-energy physics to energy-related biological and environmental research built on foundations in genomic science, climate modeling, contamination and transport modeling, and related interdisciplinary sciences. To capitalize on the investments made in these installations, DOE continues to develop the world's pre-eminent computing facilities, planning extreme scale computing capabilities, capable of computing at rates of 100 petaflops per second and greater, that will serve the nation as we move forward in the 21st century. This delivery of forefront computational facilities to scientists has enabled a number of leading scientific advances relevant to the Office of Science. The success of high-performance computing, however, involves not only the construction and effective use of advanced computational facilities but also the development of tools to effectively analyze the flood of data produced by these facilities. Yet the development of such tools has lagged, creating a significant bottleneck to scientific discovery. Thus, in addition to advancing computation and simulation, we must develop modern data analysis techniques and visualization tools. Such developments will be the next important computational contributions to enable scientific discovery.

Earlier this year the Department of Energy (DOE) Office of Advanced Scientific Computing Research (ASCR) convened a workshop to report on the fundamental research in visualization and analysis necessary to enable knowledge discovery from computational science applications at extreme scale. The goal of this report is to identify the most critical issues in visualization and analysis and to suggest future research efforts to address these issues.

Principal Finding: Scientific data analysis, visualization, and data management have evolved over the past few decades as a result of research funding from the DOE, the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), and other agencies. Today's ability to understand and explore spatial-temporal data and nonspatial data is the result of this legacy. However, datasets being produced by experiments and simulations are rapidly outstripping our ability to explore and understand them, and there is, nationwide, comparatively little basic research in scientific data analysis and visualization for knowledge discovery.

Suggested Action: We must restart basic research in scientific data analysis and visualization as a first class citizen within the DOE Office of Advanced Scientific Computing Research. A strong basic research program is vital to our continued success and competitiveness in the international scientific research endeavor. Fundamental advances must be made in visualization to exploit the potential of extreme scale simulations and large datasets derived from experiments. We must also pay much greater attention to human factors; for example, by measuring which visualization techniques are most useful to the end user. We need to treat visualization itself as an experimental science, not just a technology.

Acknowledgments

The DOE/ASCR Visualization and Analytics Workshop was held in Salt Lake City, Utah on June 7-8.

Support for the workshop web site was provided by Erik Jorgensen of the Scientific Computing and Imaging (SCI) Institute.

Gail Pieper, from Argonne National Laboratory helped edit the report.

Nathan Galli from the SCI Institute designed the report cover.

Rachel Smith and Joree O'Neal from the Oak Ridge Institute for Science and Education (ORISE) helped with the workshop organization and provided support throughout the workshop.

Within DOE/ASCR, Yukiko Sekine provided financial support for the workshop and gave valuable advice on the program and on participants to be invited.

The workshop co-chairs extend their thanks to the workshop attendees for their time and active participation and extend a special thanks to working group co-chairs: Sean Ahern, Jim Ahrens, Wes Bethel, Kwan-Liu Ma, Michael Papka, John van Rosendale, Han-Wei Shen, and Jim Thomas.

Chris Johnson
Rob Ross
August 24, 2007

Contents

1 Introduction

Visualization and analysis methods are the principal means of understanding data in many areas of science. Science is increasingly data-driven and multidisciplinary; both experiments and simulations are producing petascale datasets, and larger datasets are on the horizon. But data alone does not suffice; it must be transformed into knowledge to be of any real value. Visual scientific data analysis and representation are central to this transformation—a critical link in the chain of knowledge acquisition.

Humans are innately visual creatures; indeed, half of our brains are devoted to processing visual information. In computational terms, vision is by far our highest-bandwidth data path. Thus, visual data exploration is fundamental to our ability to interpret models and understand complex phenomena. We use our visual perception and cognition to detect patterns, assess situations, and rank tasks. Visual data exploration is one of the most important ways to reduce and refine data streams, enabling us to winnow huge volumes of data—an increasingly critical operation. Visual data exploration has thus become a cornerstone of the scientific enterprise.

Visual data exploration is, however, clearly underappreciated. One reason is the tendency to view computer graphics and visualization mainly as a way to present scientific results. But the field of visual data exploration is much more than "pretty pictures." The real power comes from the integration of interactive visual representation into the end-to-end scientific discovery process, coupling the spectacular visual understanding of the human mind with the scientific problem at hand.

Visual data analysis, facilitated by interactive interfaces, enables the detection and validation of expected results while enabling unexpected discoveries in science. It allows for the validation of new theoretical models, provides comparison between models and datasets, enables quantitative and qualitative querying, improves interpretation of data, and facilitates decision-making. Scientists can use visual data analysis systems to explore "what if" scenarios, define hypotheses, and examine data under multiple perspectives and assumptions. They can identify connections between large numbers of attributes and quantitatively assess the reliability of hypotheses. In essence, visual data analysis is an integral part of scientific problem solving and discovery.

The Department of Energy has been funding visualization research for many years, both in Advanced Scientific Computing Research (ASCR) and in the other Offices in the Office of Science as well as in the National Nuclear Security Administration (NNSA). The knowledge, techniques, and infrastructure enabled by this funding have been key to DOE's success in many areas of science. Thus, visualization, and data exploration more generally, could be seen as a well-worn path, fully integrated into the scientific workflow. Such a view, however, misses the point. The coming of peta- and exascale computing and data acquisition from high-bandwidth experiments across the sciences is creating a phase change. Our ability to produce data is rapidly outstripping our ability to use it. As Herbert Simon, Nobel Laureate in economics, noted:

A wealth of information creates a poverty of attention and a need to allocate it efficiently.

This statement succinctly summarizes the issue with peta- and exascale datasets. We have far more data than we can explore in a lifetime with current tools.

One way of viewing this situation is by analogy with a bicycle. A bicycle is an elegant and refined tool, perfect for exploring a neighborhood. It does not work at all for exploring continents and oceans; for that, one needs a different kind of tool. In a like manner, the visualization and data exploration tools developed over recent decades with funding from DOE, NSF, and other agencies have served us admirably with gigabyte and even terabyte datasets. As we reach the peta- and exascale, however, they will no longer suffice. Yet the number of new ideas in the research pipeline is comparatively meager. A high percentage of current visualization and data exploration funding amounts mainly to direct applications support, as opposed to pioneering the novel approaches that will be needed as we enter the exascale era.

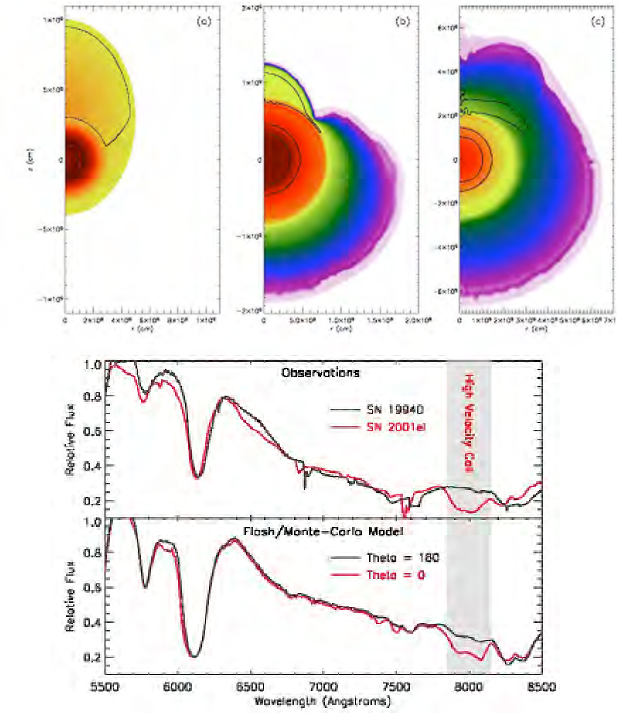
To begin grappling with this broad issue, a group of scientists and researchers met under the auspices of ASCR in Salt Lake City on June 7–8, 2007. The goal was to discuss the coming “data tsunami” and the issues involved in data exploration, data understanding, and data visualization at the petascale and beyond. The Office of Science’s notable success in discovering new science and deploying both experiments and computational simulation to great effect suggests that much of what the Office of Science is already doing is working very well. Yet there was a general feeling that the phase change mentioned above is about to create an unpleasant surprise in the form of our inability to cope with vast amounts of data about to be produced. Charting a roadmap for addressing this problem is a challenging exercise; this document is intended, at most, as a *foundation* for a roadmap. But along with the opportunities for discovering important new science with peta- and exascale data, there is an increasing sense of urgency—we really don’t yet know how to cope with data at this scale. Without better techniques and a new mindset, the data streams so arduously created by researchers in all areas of science will simply fall on the floor.

2 Mission Needs

The Department of Energy has a broad, multifaceted mission. Visualization, data analytics, and data exploration are crosscutting themes powering the research and policy activities within DOE. The paragraphs below highlight some of the many roles that visualization plays as a core computational science technology.

2.1 Computational Astrophysics

Petascale computing on the near horizon and exascale computing over the next decade will enable a quantum leap in complexity in simulating physical systems. Such complexity will motivate a commensurate leap in the priority placed on visualization research and the development of visualization tools that will enable scientific discovery in these systems. *Without visualization, discovery will not be possible.* The challenges are daunting. The datasets are expected to increase in size from hundreds of terabytes to petabytes per simulation. The dimensionality of the datasets is expected to increase well beyond the three dimensions that characterize these systems spatially. Likewise, the datasets are expected to include an ever increasing number (tens to hundreds) of variables of different types. Such challenges can be met only through a dramatic increase in the priority placed on visualization research and funding for it.

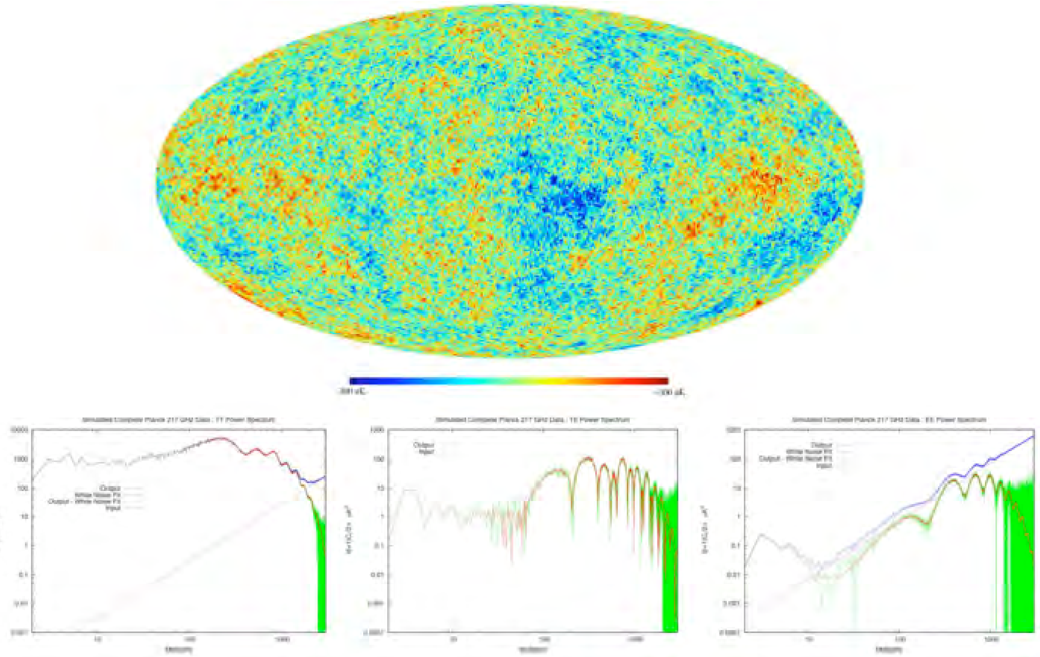


The spectrum synthesis of the Incite model compares favorably to observations and may be the first explosion model that "naturally" explains the transition from deflagration to detonation in thermonuclear supernovae. (Image courtesy T. Plewa, ASC/Alliance Flash Center, University of Chicago, and D. Kasen, Johns Hopkins University and STScI.)

2.1.1 Supernovae Explosion Modeling

We are creating supernovae explosion model datasets comprising dozens of variables per grid cell representing physical, chemical, and radiation attributes on high-resolution, adaptive, time-varying grids. For these datasets, visual data analysis will be a necessity: we need to visualize fluid flow and structures with varying degrees of transparency. We also need movies, not just for presentations and for TV shows, but because the time aspect of the simulations is often even more essential than the spatial one for understanding the results. The enormity of the datasets poses a challenge, and that challenge will become more pronounced as we begin to compare the results of hundreds or thousands of such models, which consist of time-varying 3D models as well as multiband light curves, with the hundreds of thousands of supernovae observed by missions such as the Large Synoptic Survey Telescope and the Joint Dark Energy Mission.

In the future, the ability to perform comparative visual analysis of large collections of observed and simulated data is crucial for validating the correctness of supernovae models. – Stan Woosley, PI of SciDAC Computational Astrophysics Consortium



The 50,000,000-pixel full-sky temperature and Q- and U-mode polarization maps, and the 3000-multipole TT, TE, and EE angular power spectra recovered from them, obtained at NERSC from the first-ever analysis of all of the data from a single Planck frequency, consisting of 75 billion simulated observations. (Images courtesy of J. Borrill, Planck Science Team, LBNL.)

2.1.2 Cosmic Microwave Background Data Analysis

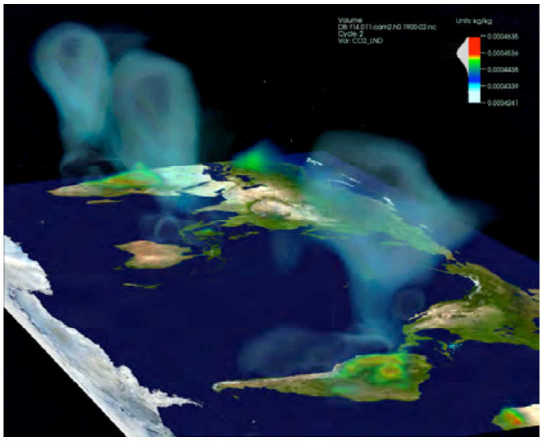
The Cosmic Microwave Background (CMB) radiation offers the earliest possible image of the Universe, as it was only 400,000 years after the big bang. Over the past 40 years observations of the CMB temperature have provided crucial tests of cosmological theories and have constrained many of the fundamental parameters of the preferred inflationary big bang models to very small ranges. In conjunction with supernovae results they have led to the astonishing conclusion that 95% of the universe is composed of forms of matter and energy that we know nothing about.

The new frontier in CMB research is measuring its polarization modes. These signals are orders of magnitude fainter than the temperature and hence require orders of magnitude larger datasets to achieve the necessary signal-to-noise ratio in the data. Experiments such as the joint ESA/NASA Planck satellite mission will gather datasets whose analysis will need peta- to exascale computing.

Visual data exploration will play a key role in providing an easily understandable view of data and data analysis results, and will be a part of a larger set of community-wide capabilities that include both data management and high performance computing. – J. Borrill, LBNL

2.2 Climate Modeling

Climate models provide an integrated understanding of the climate system and provide detailed projections of future climate changes to policy makers. Visualization plays an important role here. For much of this effort, the workhorse visualizations are typically global maps of changes in surface temperature, precipitation, winds, ocean currents, and other relevant fields. Time series of important modes of variability such as El Niño are also used. In coming years, visualization research and development will be important in trying to reduce the dimensionality of the analysis space for Earth system models that simulate the carbon, sulfur, and nitrogen cycles. Moreover, the large datasets produced by new climate models will require high-performance parallel and distance visualization tools that can enable scientists to interact with huge volumes of data and select manageable subsets for further exploration.

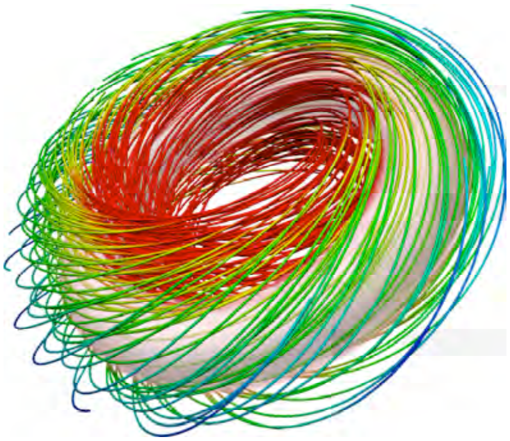


Carbon dioxide plumes from a terrestrial biogeochemistry model in the community climate system model (CCSM). Understanding the relationships between the carbon and nitrogen cycles in coupled climate models is critical to understanding long time-scale climate change. (Image courtesy of Oak Ridge National Laboratory.)

The large datasets produced by new climate models will require high performance parallel and distance visualization tools that can enable scientists to interact with huge volumes of data and select manageable subsets for further exploration. – Phil Jones, Los Alamos National Laboratory

2.3 Magnetically Confined Fusion

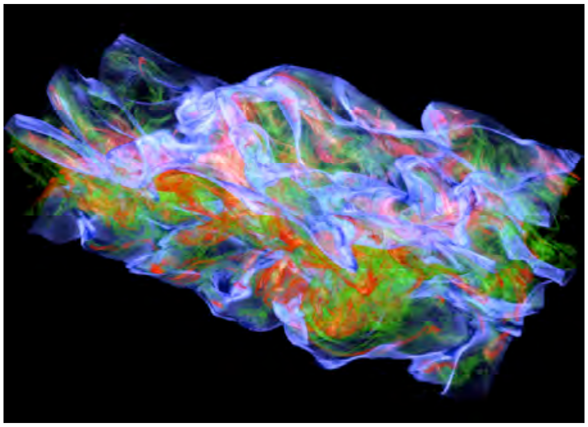
Fusion has the potential to provide a long-term, environmentally acceptable source of energy for the future. While research during the past 20 years indicates that it will likely be possible to design and build a fusion power plant, the major challenge of making fusion energy economical remains. Improved simulation and modeling of fusion systems using peta- and exascale computers are essential to achieving the predictive scientific understanding needed to make fusion practical. Integrated simulation of magnetic fusion systems involves the simultaneous modeling of the core plasma, the edge plasma, and the plasma-wall interactions. Each region of the plasma has anomalous transport driven by turbulence, abrupt rearrangements of the plasma caused by large-scale instabilities, and interactions with neutral atoms and electromagnetic waves. Many of these processes must be computed on short time and space scales, while the results of integrated modeling are needed for the whole device on long time scales. The mix of complexity and widely differing scales in integrated modeling results in a significant computational challenge.



Lines of magnetic flux confining a simulated tokamak plasma within the SIESTA fusion equilibrium code. (Image courtesy of Oak Ridge National Laboratory.)

In fusion, visualization could help us understand the specifics of onset of the H-mode in tokamaks and the formation of hotspots in electromagnetic structures. -- John Cary, Tech-X Corporation

2.4 Combustion Simulation



Multivariate visualization of a turbulent combustion simulation. (Image courtesy of the University of California, Davis and Sandia National Laboratories.)

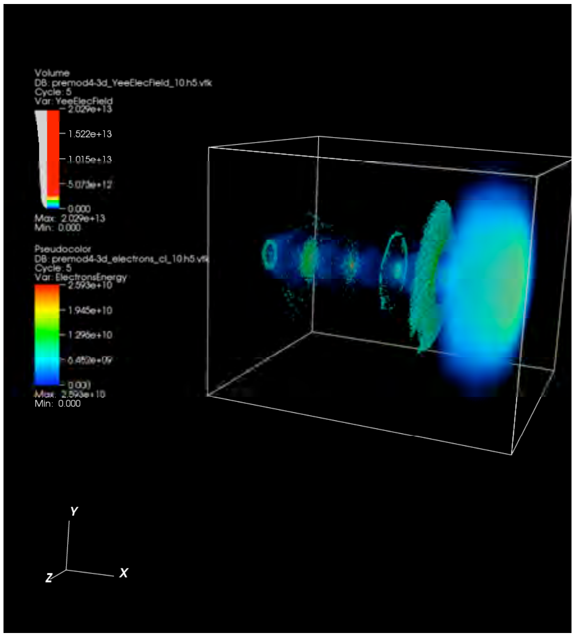
High-fidelity combustion simulations provide benchmark data to develop predictive models used to optimize the design of fuel efficient, clean burning, advanced low-temperature engine concepts using new, diverse fuel sources of the 21st century, such as oil sands, oil shale, and biodiesel fuels that are carbon neutral and hence do not contribute to the greenhouse effect. The data resulting from peta- and exascale combustion simulations are both multiscale and complex, and the sheer volume of raw data defies traditional methods of visualization and analysis for knowledge discovery; moreover, data rates are projected to increase by tenfold over the next few years. A new paradigm for knowledge discovery that integrates the following key components is needed to successfully extract meaning from datasets resulting from upcoming simulations: parallel, efficient, scalable feature-detection, segmentation, and tracking algorithms; parallel volume visualization of time-varying multivariate data and particles; feature-borne analysis software; and efficient parallel collective I/O.

The data resulting from peta- and exascale combustion simulations are both multi-scale and complex, and the sheer volume of raw data defies traditional methods of visualization and analysis. –Jacqueline H. Chen, combustion researcher at Sandia National Laboratories

2.5 Accelerator Design

Particle accelerators are critical to research in many fields, including high energy physics, nuclear physics, materials science, chemistry, and the biosciences. Accelerators have also been proposed that address national needs related to energy, the environment, and national security. The 3D, multiscale, nonlinear, and many-body aspects of accelerator design problems and the complexity and immensity of the associated computations add up to extreme technical difficulty. SciDAC's accelerator modeling project is providing scientists with advanced simulation tools for petascale computers that can perform detailed modeling of plasma accelerator experiments in their full scale for the first time. Furthermore, these tools are building the foundation for codes that will be able to design and test future experiments in high fidelity in advance of their construction.

Visualization helps us better see and understand the trapping and acceleration of particle beams in laser-plasma interactions. – John Cary, Tech-X Corporation



This image shows electrons that are "trapped," or being accelerated, in a Plasma-Wakefield particle accelerator simulation. (Image courtesy of C. Geddes and C. Siegerist, LBNL.)

3 Research Areas

A strong basic visualization and analysis research program is vital to the continued success of the scientific research endeavor. Fundamental advances must be made to extract meaning from large and complex datasets derived from experiments and from upcoming petascale and exascale simulation systems.

3.1 Fundamental Algorithms

Effective data analysis and visualization tools in support of predictive simulations and scientific knowledge discovery must be based on strong algorithmic and mathematical foundations and allow scientists to reliably characterize salient features in their data. New mathematical methods in areas such as topology, high-order tensor analysis, and statistics will constitute the core of feature extraction and uncertainty modeling using formal definition of complex shapes, patterns, and space-time distributions. This will benefit a wide variety of applications ranging from climate modeling to fusion and nuclear physics and will support petascale to exascale scientific simulations.

Findings: *Visualization is more than a "pretty picture." Effective visual data analysis must be based on strong mathematical foundations to reliably characterize salient features and generate new scientific knowledge.*

Suggested Action: *Basic research in developing fundamental mathematical methods such as topology, statistics, high-order tensors, uncertainty, and feature extraction must be established to tackle tomorrow's exascale visualization problems.*

3.1.1 Robust Topological Methods

Topological methods are becoming increasingly important in the development of advanced data analysis because of their expressive power in describing complex shapes at multiple scales. For instance, local and global trends in the flow of CO_2 are crucial to understanding the interaction of ocean models with atmospheric models, the effectiveness of carbon sequestration, and the effects of climate change in general. The recent introduction of robust combinatorial techniques for topological analysis has enabled the use of topology, not only for presentation of known phenomena but for the detection and quantification of new features of fundamental scientific interest.

3.1.2 High-Order Tensor Analysis

Tensors are general representations of scalars and vector quantities used to describe many physical properties, such as fluid flows and strength of materials. The visualization community has focused mainly on 0^{th}-order and 1^{st}-order tensor fields, and only more recently on 2^{nd}-order tensor fields. Higher-order tensors such as 4^{th}-order stiffness tensors found in geomechanics or 6^{th}-order longitudinal structure function in statistical vortex flows found in plasma physics are largely neglected by the visualization community. The challenge of visualizing higher-order tensor fields is similar in some ways to the challenge of visualizing multivariate datasets. Both deal with a high number of interrelated values at each location, where the relationships of the variables need to be highlighted, while mathematical properties and invariants need to be preserved in tensor fields. Novel methods must be developed to help scientists understand such datasets, possibly including glyph-based techniques, topological representations via critical region analyses, or continuous field representations.

3.1.3 Statistical Analysis

Our current data analysis capabilities lag far behind our ability to produce simulation data or record observational data. A particular gap exists in the mathematics needed to bring analysis and estimation methodology into a data-parallel environment. Data parallel solutions that can support, as well as use, exascale resources require new mathematics that consider an entire estimation or analysis problem in a specific application for developing scalable data-parallel algorithms in data analysis. Although scalable parallel analysis methods often will work across specific applications, generalized tools for this purpose are lacking. Browsing or looking at data is no longer possible as we approach a petabyte, so there is an enormous need for methods to dynamically analyze, organize, and present data by variability of interest across all application domains. Solutions to these problems will likely come from dynamically considering high-dimensional probability distributions of quantities of interest. This requires new contributions from mathematics, probability, and statistics.

3.1.4 Feature Detection and Tracking

The scaling of simulations to ever finer granularity and timesteps brings new challenges in visualizing the data that is generated. It is crucial to develop smart, semi-automated visualization algorithms and methodologies to help filter the data or present "summary visualizations" to enable scientists to begin analyzing the immense data following a more top-down methodological path. A key requirement for effective sharing and querying of scientific data is to develop a solid mathematical foundation to define and extract features and track their evolution over time. Also needed are formal semantic schemas, taxonomies, and ontologies for describing, characterizing, and quantifying features and for highlighting areas of interest in massive time-varying data, thereby giving the scientists a handle on where to look or to make more high-level queries. Feature-based techniques are also important for analyzing the results of different simulations and making comparisons between simulations and experimental data. Once features and their evolution are identified and measured, tools are needed to enable researchers to identify interfeature relationships and evolutions or configurations of a set of objects and their interactions.

3.1.5 Uncertainty Management and Mitigation

A significant problem faced by the Office of Science simulation efforts is the robust treatment of uncertainty. Numerical simulations are rife with sources of uncertainty, which can be introduced in the form of numerical imprecision, inaccuracy, or instability. Predictions and forecasting inherently contain uncertainty arising from the variability in the physical processes under study. Scientific experiments and measurements introduce uncertainty in the form of calibration errors, differences in repeated measurements, and the like. Visualization of petascale datasets also can introduce uncertainty during processing, decimation, summarization, and abstraction as an artifact of creating much-condensed representations of the data.

The ability to fully quantify uncertainty in high-performance computational simulations will provide new capabilities for verification and validation of simulation codes. Having a robust mathematical framework for tracing the sources of uncertainty and its propagation throughout the simulation process turns simulation into a strong predictive capability. Handling uncertainty must be an end-to-end process, where the different sources of uncertainty are identified, quantified, represented, tracked, and visualized together with the underlying data. Hence, uncertainty representation and quantification, uncertainty propagation, and uncertainty visualization techniques need to be developed in order to provide scientists with credible and verifiable visualizations.

3.2 Complexity of Scientific Datasets

Scientific simulation codes are producing data at exponentially increasing sizes, but spatial resolution is only one of the axes by which datasets are expanding. As computational scales reach the petascale and extend into the exascale, simulation codes are also increasing in their temporal resolution, degree of code coupling, and extent of parametric exploration. Although some similarity may be leveraged, each of these scales requires research and expansion to enable new scientific discovery.

Findings: *Trends in scientific simulation—which include coupled codes, hierarchical computation and data models, extreme and varying scales of spatial and temporal resolution, and increasing numbers of variables to more faithfully represent physics and chemistry phenomena—present challenges that cannot be met by extrapolating existing approaches, known techniques, and familiar methodologies.*

Suggested Action: *A concerted and long-term visual data understanding and representation research effort is a sound and crucial investment for providing the technologies needed to enable knowledge discovery on the complex, heterogeneous, multiresolution datasets projected to be produced by scientific simulations on peta- and exascale platforms.*

3.2.1 Multimodel Data Understanding

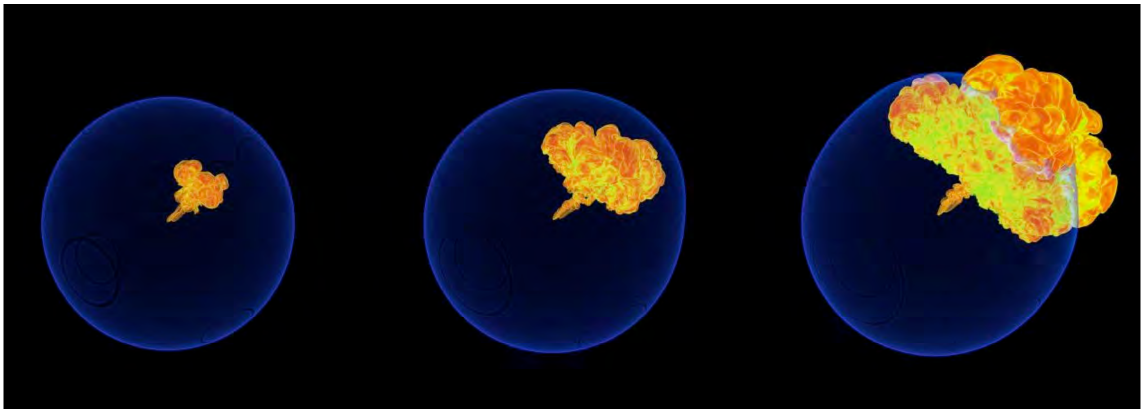
One area of significant advancement in computational science in recent years enabled by more powerful computing platforms is multimodel codes. These codes, which play a significant role in SciDAC projects aiming to model complex facilities, such as fusion tokamaks and particle accelerators, and complex scientific phenomena, such as supernovae explosions and Earth system models, consist of combinations of codes each modeling some individual scientific regime. Data produced by one component is often used as input to another, resulting in an extremely complex and information-rich dataset. In other cases, input from instruments is combined with simulation results. Traditional approaches to visual data analysis have focused on data generated from a single code or code family. These approaches do not lend themselves to use on the complete systems simulated with such multimodel codes.

New approaches to visual data analysis and knowledge discovery are needed to enable researchers to gain insight into this emerging form of scientific data. Such approaches must take into account the multimodel nature of the data; provide the means for a scientist to easily transition views from global to local model data; offer the ability to blend traditional scientific and information visualization; perform hypothesis testing, verification, and validation; and address the challenges posed by vastly different grid types used by the various elements of the multimodel code. Tools that leverage semantic information and hide details of dataset formats will be critical in enabling visualization and analysis experts to concentrate on the design of these approaches rather than becoming mired in the trivialities of particular data representations.

3.2.2 Multifield and Multiscale Analysis

In many scientific fields of study, computational models aim to simulate phenomena that occur over a range of spatial and temporal scales spanning several orders of magnitude. Those models also attempt to capture the interaction of multiple variables, often referred to as multivariate or multifield data. Visualization of multivariate or multiscale datasets is helping scientists discover hidden relationships among the data, as well as transient events (occupying a small fraction of simulation time) that have a profound influence on the outcome of the simulation.

Unfortunately, while current visual data analysis technologies are capable of processing many types of adaptive, multivariate, multiresolution data, these technologies lack the ability to take into account various types of constraints that would improve their usability and applicability to a broader set of fields of study. Multiresolution techniques are needed to support zooming in to regions of interest, generating geometry with high accuracy where needed, and displaying animations that are short enough to match a



A series of 3D volume-rendering of a Type Ia supernova as hot nuclear ash erupts from the surface of the white dwarf progenitor. The blue surface shows an isocontour of the density of the white dwarf at 10 million grams per cubic centimeter and represents a length scale of roughly 2,000 kilometers in radius. (Image courtesy DOE-supported NNSA ASC/Alliance Center for Astrophysical Thermonuclear Flashes at the University of Chicago and Argonne National Laboratory.)

viewer’s desired context while providing sufficient detail for transient important events. For multifield data, visualization cannot simply map different variables to different visual parameters, as one will quickly run out of visual parameters and introduce a visual overload on the user, hampering the task of data understanding. We therefore need to bring in different approaches from visual analytics, projections and dimensionality reduction, database queries, feature detection, and novel visualization techniques.

3.2.3 Time-Varying Datasets

New challenges for scientists have emerged in the past several years as the size of data generated from simulations has experienced an exponential growth. One major factor contributing to the growth of data size is the increasingly widespread ability to perform very large scale time-varying simulations. Although intensive research efforts have been undertaken to enable visualization of very large datasets, most of the existing methods have not specifically targeted time-varying data. New visualization techniques and user interfaces must be developed to assist the user in understanding exascale time-varying multivariate datasets. Scientists must be able to interactively browse through different spatial and temporal scales, visualize and identify scientific phenomena of different temporal lengths, and isolate and track salient features in both time and space. Multiresolution spatial and temporal data management and encoding techniques need to be fully integrated with current and future visualization algorithms so that the scale and location of the time-varying data will be completely transparent to the visualization users.

3.3 Advanced Architectures and Systems

Research into computational methods cannot stand alone without consideration of the computational platforms on which they depend. Emerging peta- and exascale architectures provide both a blessing and a curse: unprecedented computational power, but also the ability to generate results far faster than we can store—much less visualize—them. This change is as disruptive as the shift from vector to distributed memory supercomputers 15 years ago, which took years of effort to address. Software systems are simultaneously growing in complexity, and additional work is needed to enable scientists to integrate visualization and analytics tools into the scientific process.

Findings: *Upcoming system architectures are a significant departure from systems of the past decade. Current approaches for performing visualization and analysis are not well suited to the processing or storage capabilities of petascale and exascale architectures. Likewise, software environments surrounding these algorithms are not adequate for scientific discovery using these resources.*

Suggested Action: *Sustained research in exploiting parallelism, in situ processing, data access, and distance visualization is necessary to adapt visualization and analysis techniques to the rapidly changing computational landscape in order to help scientists gain insight into their problem using advanced systems.*

3.3.1 Pervasive Parallelism

Computer architectures are undergoing revolutionary change. In the near term, all computer architectures will involve parallelism on a single chip. In the longer term, all computer architectures will involve massive parallelism. For example, AMD and Intel have changed their product lines to include dual-core and quad-core processors, with roadmaps for continued increases in the number of cores. The Sony/Toshiba/IBM Cell Processor has eight stream processing cores in addition to a conventional scalar processor. Commodity GPUs now feature hundreds of processors. GPUs and CPUs are also being merged, which will enable tight coupling between applications and graphics. This is likely to be the biggest change to the PC platform in the past 20 years.

We are entering an era of pervasive parallelism. As the number of transistors doubles, the number of cores will also double. This trend means that software of the future will be very different from the sequential programs of today. This revolution in computer architecture will impact the graphics and visualization enormously. The visualization pipeline as we know it today will likely be radically different in order to exploit the new architectures. These new architectures will also enable an entirely new class of interactive visualization applications. Since graphics is the main driving application for such high-performing chips, it is critical that the graphics and visualization community actively participate in the research and development of these technologies. One key focus for near-term research is the integration of the CPU and GPU, and the programming models for each. Future architectures likely will be heterogeneous, with multiple kinds of processors on a single die. Visualization, which can use both multicore-CPU-style thread parallelism and GPU-style data parallelism, will play a major role in understanding the results from such heterogeneous systems.

3.3.2 In Situ Processing

As processing power grows, so does the amount of data processed and generated. Increased computation rates enable simulations of higher fidelity, which in turn yield more data. Unfortunately, storage system bandwidth is not increasing at the rate at which our ability to generate data is growing. The divide between what we are producing and what we are capable of storing is critical. It is already common for simulations to discard over 90 percent of what they compute. With storing data no longer a viable option, output processing and visualization must be performed in situ with the simulation. Collocating certain visualization algorithms with simulation can simultaneously improve the effectiveness of the algorithm and maximize the information stored in the data. For example, saliency analysis can help the simulation make better decisions about what to store and what to discard. Feature extraction becomes much more effective when all variable information is available, and feature tracking is much more reliable when temporal fidelity is high. Features can provide far more information to an analyst and can require far less storage than the original volume. Because these techniques must be integrated into the application and supported by the run-time environment, interaction with designers of programming models and system software for advanced architectures is warranted.

3.3.3 Data Access

In situ processing can mitigate the disparity between data generation rates and storage system capabilities and is an important component in managing petascale and exascale datasets. However, applications on upcoming systems will store an unprecedented amount of simulation data during their run time. The current practice of postprocessing datasets from leadership-computing applications on separate visualization clusters will likely fall short at the petascale and certainly will be impossible at the exascale. Research in alternative mechanisms for processing large datasets is critical for enabling visualization at these scales. These could include out-of-core mechanisms and streaming models of processing, likely used in conjunction with in situ processing.

Data models and formats are an important issue for applications as a whole, because the decisions made when defining these models and formats affect the scientists' ability to describe the results of their work as well as the efficiency with which that data is moved to storage and subsequently processed. The explosion of data formats and models present in the DOE application space is causing significant problems in our ability to generalize tools for visualization and analysis, and this situation is exacerbated by the use of multiple formats and models in applications that combine simulation with other data sources or that leverage coupled codes. The disconnect between the data models used in simulation codes and subsequent postprocessing access patterns, in conjunction with an increase in the complexity of these datasets, is leading to increased overhead in the I/O component of the visualization and analysis process. Attention is needed to ensure that storage organizations are optimal for state-of-the-art visualization algorithms and map well to the systems on which this data will be processed. Achieving this objective will require the combined effort of scientists, visualization experts, and storage researchers.

Mechanisms for reducing data within the storage system provide another avenue for reducing the I/O requirements of analysis. Active storage technologies, under research in the storage domain, could be an important enabler by allowing analysis primitives to execute within the storage system. In cases where scientists prefer to locally view results of remote simulations, minimizing the amount of data that must be transferred is critical. Additional research is necessary to understand how best to integrate data reduction into remote I/O protocols so that reduction can be performed prior to movement of datasets over long-haul networks.

3.3.4 Distance Visualization

For DOE Office of Science application teams, visualizing, analyzing, and understanding their results is key to effective science. These activities are significantly hampered by the fact that scientists and the supercomputing resources they work on are located in geographically different locations. These teams are expecting to generate petabytes of data soon and exabytes of data in the near future, making this problem increasingly challenging. To address this challenge, we need to look beyond application and adaptation of existing technologies. Many orders of magnitude separate the data sizes we need to visualize and the data sizes our current gigabit networks can handle.

A diverse and broad set of interrelated research and development activities is needed to address specific distance visualization challenges. These include development of latency-tolerant techniques for delivering interactive visualization results to remote consumers using distributed and parallel computational platforms; techniques for delivering visualization results that gracefully accommodate the wide variance in network capacity, from multiple OC-192 rings (ESnet) to consumer-grade broadband; resource- and condition-adaptive partitioning of the visualization pipeline to meet performance or capability targets; and data storage and transmission techniques that leverage advances in compression, progressive refinement, subsetting, and feature-based methods to help reduce the I/O bandwidth requirements to a level more appropriate for distance-based visualization.

3.3.5 End-to-End Integration

In order to analyze and understand scientific data, complex computational processes need to be assembled and insightful visualizations need to be generated, often requiring the combination of loosely coupled computational and data resources, specialized libraries, and Grid and Web services. Typically this process involves data management and statistical analysis tasks, such as data extraction from very large datasets, data transformation or transposition, statistical summarization, pattern discovery, and analytical reasoning. Rather than attempting to develop a single, monolithic system with such a wide range of capabilities, technologies and tools from different domains must be integrated in a single framework to provide iterative capabilities of interacting with and visualizing scientific data.

Multiple visualization and data analysis libraries and tools are available today, some of which (e.g., VTK, VisIt, ParaView, and SCIRun) are capable of processing very large data volumes in parallel, and some (e.g., VisTrails) have advanced provenance, comparative and multiview capabilities. Statistical and plotting tools (e.g., R, matplotlib, and IDL) are used routinely by scientists. Integrated environments (e.g., Matlab and Mathematica) are also very popular. For data management, various tools (e.g., such as NetCDF and HDF5) support specialized data formats, and others (e.g., FastBit) support specialized indexing methods for efficiently performing value-based queries and subset extraction. The lack of integration among these tools is a major shortcoming, however, and hampers visualization and data analysis efforts.

A framework is needed that allows multiple tools to interact, permitting the integration of existing and future software modules into end-to-end tasks. Research is needed to have visualization, data management, statistical, and reasoning tools interoperate seamlessly. Further work is needed to develop specialized workflow capabilities for visualization and data analysis. The development of these tools is especially challenging when dealing with expected peta- and exascale datasets and multiple scientific domains.

3.4 Knowledge-Enabling Visualization and Analysis

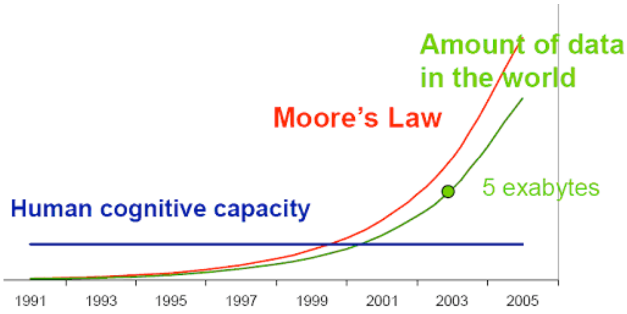
The goal of visualization in science is to support generation of new scientific knowledge. Current visualization systems, however, address knowledge implicitly rather than explicitly. That is, they are not formally integrated with methods and tools that enable capture of knowledge, representation of the knowledge and its provenance, or management and reuse of knowledge gained in support of subsequent visual exploration and discovery. Capturing knowledge as it is generated is important to assessing that knowledge and determining its applicability. Capturing provenance allows the process of reasoning to be reconstructed, in turn enabling other users to evaluate the utility and trustworthiness of knowledge representations. As we develop new methodologies for capturing knowledge about the discovery process, that knowledge can be shared and reused by collaborators or even broader communities, providing increased capability in the area of reproducible scientific results that others can validate and verify.

Findings: *Analysis is about interaction among people working with each other and computational resources to understand results. Little about this process is currently captured for reuse except for anecdotal summaries and final snapshots in the form of images and movies. New capabilities will be required to enable discovery at the exascale, including the ability to reconstruct previous analyses for reuse, leverage previously acquired and related knowledge, and provide guidance and discovery aids to the scientist.*

Suggested Action: *Basic research is needed to develop novel methods for capturing knowledge about the analysis process and providing that knowledge for reuse in collaboration and interaction with other team members and computational resources.*

3.4.1 Interaction, Usability, and Engineering Knowledge Discovery

Even as simulation datasets have been growing at an exponential rate, the capabilities of the natural human visual system have remained unchanged. Furthermore, the bandwidth into the human cognitive machinery remains constant. As a result, we have now reached a stage where the petascale and exascale datasets critical to the DOE ASCR mission can easily overwhelm the limits of human comprehension. Over the past 30 years, computerized techniques for visualizing information have concentrated on incrementally improving techniques for the graphical display of data. While these improvements have extended the field of visualization, they have concentrated on only a small part of the problem that scientists and engineers face. To enable the creation of a visual analysis, reasoning, and discovery environment targeted at peta- and exascale datasets, we need research to develop a better scientist-computer interface— the nexus of cognitive science, effective visual presentation of information and data, usability analysis and optimization, methodologies for exploring and interacting with large and complex, hierarchical, multimodal, and possibly incomplete and conflicting data.



Human cognitive capacity remains flat while our ability to collect and generate data continues to grow at an exponential rate. (Image courtesy Jeffrey Heer, PARC User Interface Research Group.)

Advances in the area of the scientist-computer interface will have a profound, positive impact on our ability to gain knowledge and understanding from data of increasing size and complexity and on our ability to perform hypothesis testing and knowledge discovery in peta- and exascale data, and will fundamentally change our understanding about how humans perceive and gain knowledge from large, complex data. Research directions in this area include formal usability studies and analysis across diverse domains such as code, data, and graphics interfaces; alternative display technologies; quantitative analysis and optimization of workflow; mappings from data to visual representations; and inclusion of cognitive principles into the visualization and data analysis tools.

One approach to improving interaction could be through a common interface across multiple tools. Ideally, this new technology would result in reusable user interfaces that enable intuitive and interactive exploration and discovery—for example: interoperable user-interface libraries that contain widgets having a common look and feel that are specifically intended for large-scale data exploration yet usable by multiple applications. One design objective is interfaces that capture the best interaction methods to support data reduction, feature extraction, querying, and selection. These interfaces should also support synchronous collaborative interaction between multiple users who may be separated by great distances.

3.4.2 Collaboration

Today's scientific research is inherently distributed, with science teams often consisting of researchers at universities and national laboratories around the country or around the globe. A new generation of visualization and data exploration tools are needed to significantly enhance interaction between these distributed scientists, their data, and their computational environments.

Also needed is a collaboration infrastructure that supports both asynchronous and synchronous collaboration. Asynchronous collaboration infrastructure might include large-scale equivalents of wikis, blogs, mashups, and other emerging social networking tools. Synchronous collaboration infrastructure might include context- and location-aware, persistent visualization and collaboration environments. These environments should seamlessly display information from both local and remote sources, while simultaneously providing an environment that fully exploits local capabilities without lowering the experience to the lowest common denominator.

Faculty, researchers, and students in their weekly research progress meetings in front of the 100 Megapixel LambdaVision display at the Electronic Visualization Laboratory (University of Illinois at Chicago). Using EVL's Scalable Adaptive Graphics Environment (SAGE) middleware, a variety of high-resolution information are juxtaposed on the walls to enable group discussion. Participants can be in the room and or communicating over distance mediated via high-definition video conferencing.

It is also necessary to deal concurrently with both distributed human-human and distributed human-computer interactions. The ideal would thus be environments allowing remote and local participants alike to effectively participate in real-time computation, visualization, and data exploration. Unfortunately, little infrastructure is currently available to enable graphics and visualization developers to build tools with such collaborative capabilities. A clear need exists for both "building blocks" to allow these developers to create effective, interoperable, collaboration tools. And the tools themselves are central to the scientific enterprise, enabling distributed teams to make the discoveries of the future.

3.4.3 Quantitative Metrics for Parameter Choices

The fundamental process of visualization involves choices of parameters for queries of different types. Examples are the selection of spatial and temporal scales, transfer functions, and lighting and camera parameters. To glean insight into a scientific dataset, the user often needs to go through a lengthy, sometimes prohibitively expensive, process to obtain a large ensemble of visualization results. Quantitative feedback about the choice of visualization parameters is crucial for streamlining visual analysis. Techniques are needed to help scientists quickly narrow down the immense parameter search space, identify salient features, and decide the right level of detail in the data to perform further investigation. Also needed are metrics to help users understand the tradeoff between the computational cost and the information gain, and the completeness of the visualization results. The users need to be informed not only about what they have seen but also about what they have not yet seen.

4 Supporting a Basic Research Program

Investment in long-term basic research in visualization and knowledge discovery is crucial to ensuring the advancement of science required to meet future national needs. In addition, a successful basic research program must be complemented with adequate infrastructural support and a comprehensive education program.

4.1 Infrastructure for Successful Research

Scientific research programs are necessarily supported by the unheralded but vital resources that form the computing infrastructure. Frequently it is infrastructure—whether hardware, software, or data—that enables researchers to create breakthroughs in the sciences. For example, only after the telescope was refined by Dutch craftsmen was Galileo able to discover the moons orbiting Jupiter. Such circumstances are not atypical; often, advances in basic infrastructure free scientists to take the next step forward in their particular field of study. Thus, both infrastructure and research must be funded at appropriate levels to ensure that technology innovation continues.

The computational sciences presuppose that an adequate computing environment exists to support the envisioned numerical simulations and experiments at a scale appropriate to the data sizes anticipated. While this is certainly a necessary condition, many other facilities can significantly benefit the stability, flexibility, and efficiency of the scientific research process. For example, using standard datasets to compare and contrast the efficacy of algorithms from competing research groups enables useful comparisons. Further, collaborative research projects are the norm as systems become more complex. In addition to sharing data, therefore, infrastructure that supports the development of tools across multiple organizations is mandatory. Another important consideration is to ensure that new technologies are rapidly transitioned to those organizations that can benefit from them, or even form the basis of new business opportunities to benefit the U.S. economy.

The following topics are seen as critical to a successful R&D program in visualization and knowledge discovery:

- **Collaborative software process.** Research teams are becoming distributed and larger. Thus, software processes that support distributed collaboration are essential. Furthermore, such processes must facilitate the development of stable systems that are thoroughly tested and managed.
- **Data repositories.** Visualization and analysis researchers typically are starved for example datasets. Often simplistic data is used to initially develop computational techniques; however, in many cases this data is not representative of the targeted applications. Thus the computational community must be encouraged to gather, distribute, and manage representative datasets to help enable creation of effective computational methods.
- **Toolkits and reusable components.** Many researchers use standard toolkits and applications on which to base their research. These have the benefit of accelerating research because such

foundational elements do not have to be recreated, and researchers can focus on the particular problem at hand. Creation of standard toolkits should be encouraged within the community to help continue this tradition.

- **Open science.** The practice of science requires the ability to recreate the results of experiments. It also implies full disclosure as to the methods and data used to generate the results of an experiment. In the computational sciences this means access to data, source code, and publications. It is imperative that the practice of open science be employed to ensure the full benefit of scientific funding.

4.2 Fostering Education

A strong research program cannot be established without a complementary education component, which is as important as adequate infrastructure support. A continuing supply of first-quality computational scientists available for work at DOE laboratories is critical. For example, the DOE Computational Science Graduate Fellowship (CSGF) program has successfully provided support and guidance to some of the nation's best scientific graduate students, and many of these students are now employed in DOE laboratories, private industry, and educational institutions. However, in order to meet the DOE Office of Science's Advanced Scientific Computing mission, there is also a significant need for a similar program supporting training in large-scale visual data analysis. The DOE High-Performance Computer Science Fellowship formed by Los Alamos National Laboratory, Lawrence Livermore National Laboratory, and Sandia National Laboratories to foster long-range computer science research efforts in support of the challenges of high-performance computing was a step in right direction. Unfortunately, these fellowships have been discontinued. A DOE Graduate Fellowship in High-Performance Computer Science is needed that trains people in large-scale visual data analysis, as well as in scientific data management and high-performance software and hardware. In addition, a successful education program should include the following:

- **Undergraduate research experience (REU).** The goal is to involve undergraduate students in visualization and data analysis research during their sophomore and senior years. REU supplements should be provided to ongoing DOE-funded research projects. In order to further improve the quality of emerging visualization scientists, these undergraduate students should be encouraged to apply for DOE graduate fellowships.
- **Postdoctoral research fellowships.** The need for visualization scientists in DOE laboratories is expected to increase rapidly. Postdoctoral research fellowships can help attract more fresh Ph.D.'s into the field of visualization in support of DOE missions. These fellowships would expose the postdoctoral fellows to the most challenging research problems and advanced computational facilities and would provide a comfortable transition into faculty or DOE scientist positions. The DOE Early Career Principal Investigator Program is a great first step in this direction but covers only a limited number of individuals.
- **Workshops and tutorials.** The diversity of visualization research makes it difficult for a single institution to offer comprehensive training to its students. Furthermore, most academic institutions do not have a balanced visualization curriculum. With adequate funding support, leading researchers and institutions in visualization can help others develop their education program through workshops and tutorials. Also, workshops and tutorials offered regularly at major conferences in other scientific disciplines can help educate application scientists about the latest visualization technologies.
- **Data repositories and benchmarks.** The education program should include the creation of data repositories and benchmarks. The ultimate goal is to create a multi-institutional effort with a coherent intellectual theme and shared education resources, tightly coupled with research activities at DOE laboratories.

4.3 Integrating Basic Research Programs

For maximum effect, a research program in visual analysis and knowledge discovery should be well-integrated into the broader Office of Science portfolio, into other U.S. research programs, and into research programs in this field around the world. To this end, we envision three complementary strategies.

Interagency collaborations. A number of government agencies support research programs in visualization and data exploration. For example, the Department of Homeland Security (DHS), with its National and Regional Visualization and Analytics Centers (NVACs and RVACs), focuses on mission-critical issues of homeland security addressed by information visualization technology. In addition, data exploration and visualization programs are supported by the National Institutes of Health (NIH), Department of Defense, and DARPA. Moreover, the National Science Foundation has for many years funded a number of individual research projects in computer graphics, along with larger-scale graphics/visualization research efforts in the Partnerships for Advanced Computational Infrastructure (PACI) program. In addition, several programs are being launched in computer graphics, visualization, and data exploration. The first of these is a program in the basic science of visual analysis, as described in "Illuminating the Path: The Research and Development Agenda for Visual Analytics." The second is a visualization and data exploration component being added to the Cyberinfrastructure program; this will include telecollaboration and remote visualization and remote data exploration. The third is a new program called Cyber-enabled Discovery and Innovation (CDI), planned to encompass five thrusts all involving visual data exploration: knowledge extraction, complex interactions, computational experimentation, virtual environments, and education.

While DOE has a unique mission and its laboratories provide an essential national resource, DOE's research agenda in data exploration and visualization overlaps with those of these and other agencies. Through shared research investment, the DOE can stretch scarce resources and exploit potential synergies.

International partnerships and collaboration. Through the formation of international partnerships, DOE can leverage advances occurring in other countries and can better participate in the global advance of science. As an example, several countries are starting programs in visualization and data analysis, including Australia, New Zealand, Canada, and several countries within the European Union (EU); and there is a growing interest within the Asian community. Also, the EU countries have surged ahead of the United States in "collaboratories"; this is an area ripe for international partnerships.

Centers of excellence in infrastructure and education. The breath, depth, complexity, and richness of modern scientific investigations can no longer be accomplished within a single discipline or institution. The DOE SciDAC II program supports two visualization projects with potentially broad impacts on DOE research: the Visualization and Analysis Center for Enabling Technologies (VACET) and the Institute for Ultrascale Visualization. VACET focuses on research, development, and deployment of production-quality petascale-capable visualization analysis technology for SciDAC computational sciences at DOE's open computing facilities. The Ultravis Institute plays a leading role in peta- and exascale visual data understanding by combining basic research with intense collaborations with science teams and by reaching out to the visualization and science communities to teach them about these new techniques. Centers such as these play an important role in educating the next generation of visual analysts. Educating the next generation is key to a vibrant, successful community.

5 Conclusions

Visual analysis and knowledge discovery is an indispensable cornerstone of the contemporary scientific discovery process. One of the main messages of this report is that without long-term investment in visual analysis and knowledge discovery research, existing and future scientific investments face a serious risk: there is an alarming divergence in the trajectories of the flood of scientific data computed by simulations and collected by experiments and our ability to gain understanding from such data.

Basic research in visual analysis and knowledge discovery is critical in order to remedy this situation. We consider work in the following technical areas to be of the highest priority:

- **Interaction and Collaboration** – A new generation of visualization and data exploration tools are needed to significantly enhance interaction and collaboration between these distributed scientists, their data, and their computational environments.
- **Pervasive Parallelism and Multiscale Analysis** – New developments in computer architecture will enable the development of visualization applications that are parallel at multiple levels. This capability must be provided to ensure that scientists are able to maximize time analyzing data.
- **Feature Detection and Tracking** – New algorithms will allow the detection and tracking of features that are of interest to the scientist, aiding in the discovery of important regions to investigate further.
- **Multifield and Multimodel Data Understanding** – New approaches will enable comparison and combined analysis of multi-variable data that is becoming increasingly common in extreme scale datasets.
- **Distance Visualization** – Because computational resources, data, and scientists are rarely collocated, new visualization and analysis pipeline architectures must accommodate distance visualization at the extreme scale and remote users with a diverse set of display and processing capabilities.
- **In Situ Processing** – In order to maximize the effectiveness of large computational resources, new visualization efforts will collocate certain visualization algorithms with simulation.
- **Time-Varying Datasets -** New visualization techniques and user interfaces must be developed to assist the user in understanding extreme scale, time-varying, multivariate datasets. Scientists must be able to interactively browse through different spatial and temporal scales, identify scientific phenomena of different temporal length, and track salient features in both time and space.
- **Visual Analysis, Quantification, and Representation of Uncertainty and Error** – New approaches must be developed to quantify the uncertainty and error in the analysis process and present scientists with immediate feedback as they choose different forms of visual analysis.
- **End-to-End Integration** – End-to-end integration strategies, considering the entire simulation/analysis process as an analog to a physical experiment, allow scientists to more easily document and track the entire scientific process.

Further, we envision several key attributes of a successful visual data analysis research program. It must be sufficiently well-funded over an adequate duration of time to allow research results to be conceived, germinate, and reach fruition. The funding profile needs to be commensurate with the time horizon for research so that staff may be attracted and retained for the duration of the project. The research program needs to be complemented by an education program for the creation of a new generation of visualization and analysis researchers. The research program can not exist in isolation, but rather must be part of a larger portfolio of research, development, and deployment activities that have coordination and interaction across programs and, where feasible, other agencies. Such coordination and interaction reduce duplication of effort and help new technological advances make their way into the hands of scientists. The program needs to facilitate a reward structure that accommodates and encourages high-risk projects that may not have payoff for a long period of time. Successful basic research efforts, which target potentially high-risk, high-reward areas having a relatively long time horizon (5–15 years), will be the source of the major technological advances required to meet future data understanding challenges.

Appendix A: Summary of Basic Research Program Focus Areas

Mathematical Foundations

- **Robust Topological Methods.** Provides mathematically-based capability for identifying and quantifying phenomena in multi-resolution, temporally varying, large and complex scientific data to aid in accelerating knowledge discovery.
- **High Order Tensor Analysis.** Provides visual analysis capability responsive to the increasingly complex, multivariate data emerging from many science projects.
- **Statistical Analysis.** Helps bridge the gap between visual and traditional analysis in the extreme scale regime by focusing attention on global characteristics rather than those of individual data points.
- **Feature Detection and Tracking.** Focuses visual and traditional analysis processing at the extreme scale on subsets of data deemed to be scientifically “interesting” for a given line of hypothesis testing.
- **Uncertainty Management and Mitigation.** Helps scientists better understand and analyze uncertainty, which is present in some form in all scientific data.

Data Fusion

- **Multimodel Data Understanding.** Ability to perform visual data analysis for science projects that use multiple codes (tightly or loosely) coupled to model complex, multiregime phenomena.
- **Multifield and Multiscale Analysis.** Comparative visual analysis and data exploration at multiple resolutions on high-resolution, complex data offers a powerful capability for enabling extreme scale knowledge discovery.
- **Time-Varying Datasets.** Fundamental visualization, analysis, and data access capabilities in response to simulations producing more data in the temporal domain and the corresponding need for temporal analysis/access.

Advanced Architectures and Systems

- **In Situ Processing.** Integration of visual data analysis processing with the simulation code itself, enabling leverage of large computational resources for visual data analysis as well as potentially avoid extreme scale data I/O and management problems.
- **Data Access.** Data models, formats, and high-performance access patterns occupy a central role in all visual data analysis endeavors.
- **Distance Visualization.** Technologies that help geographically-distributed teams of researchers visualize and understand their remotely-located scientific results.
- **End-to-End Integration.** Design patterns and engineering practices that simplify combining disparate technologies from many different research and development teams—visualization, statistics, data management, analysis, and so forth—into vertical applications.

Knowledge-Enabling Visualization and Analysis

- **Scientist-Computer Interface.** Increases scientific knowledge discovery through a combination of optimizing the interfaces between data, software algorithms, visual presentation, and users, as well as evolution in fundamental visual presentations of large, complex, and abstract data characteristic of contemporary science.
- **Collaboration.** Provides infrastructure enabling distributed teams of scientific researchers to engage in multi-participant, interactive knowledge discovery.
- **Quantitative Metrics for Parameter Choices.** Accelerates time to discovery by simplifying use of complex visual data analysis tools.

Appendix B: Report Contributors

Workshop Participants
Sean Ahern - Oak Ridge National Laboratory
Jim Ahrens - Los Alamos National Laboratory
David Banks - University of Tennessee
Hank Childs - Lawrence Livermore National Laboratory
Wes Bethel - Lawrence Berkeley National Laboratory
David Ebert - Purdue University
Sam Fulcomer - Brown University
Pat Hanrahan - Stanford University
Chris Johnson - University of Utah
Gary Johnson - DOE Program Management
Ken Joy - University of California, Davis
Jason Leigh - University of Illinois at Chicago
Kwan-Liu Ma - University of California, Davis
Alan MacEachren - Pennsylvania State University
Pat McCormick - Los Alamos National Laboratory
Don Middleton - University Corporation for Atmospheric Research (UCAR)
Ken Moreland - Sandia National Laboratories
Alex Pang - University of California, Santa Cruz
Michael Papka - Argonne National Laboratory
Valerio Pascucci - Lawrence Livermore National Laboratory
David Rogers - Sandia National Laboratories
John van Rosendale - College of William and Mary
Rob Ross - Argonne National Laboratory
Will Schroeder - Kitware Inc.
Yukiko Sekine - DOE Program Management
Han-Wei Shen - Ohio State University
Arie Shoshani - Lawrence Berkeley National Laboratory
Claudio Silva - University of Utah
Deborah Silver - Rutgers University
Jim Thomas - Pacific Northwest National Laboratory
Amitabh Varshney - University of Maryland
Terry Yoo - National Institutes of Health

External Participants
Jacqueline Chen - Sandia National Laboratories
Anthony Mezzacappa - Oak Ridge National Laboratory
George Ostrouchov - Oak Ridge National Laboratory
John Owens - University of California, Davis
Thomas Peterka – Argonne National Laboratory