Visualization for Data-Intensive Science

CHARLES HANSEN
CHRIS R. JOHNSON
VALERIO PASCUCCI
CLAUDIO T. SILVA
University of Utah

Since the advent of computing, the world has experienced an information "big bang": an explosion of data. The amount of information being created is increasing at an exponential rate. Since 2003, digital information has accounted for 90 percent of all information produced [1], vastly exceeding the amount of information on paper and on film. One of the greatest scientific and engineering challenges of the 21st century will be to understand and make effective use of this growing body of information. Visual data analysis, facilitated by interactive interfaces, enables the detection and validation of expected results while also 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 using multiple perspectives and assumptions. They can identify connections among large numbers of attributes and quantitatively assess the reliability of hypotheses. In essence, visual data analysis is an integral part of scientific discovery and is far from a solved problem. Many avenues for future research remain open. In this article, we describe visual data analysis topics that will receive attention in the next decade [2, 3].

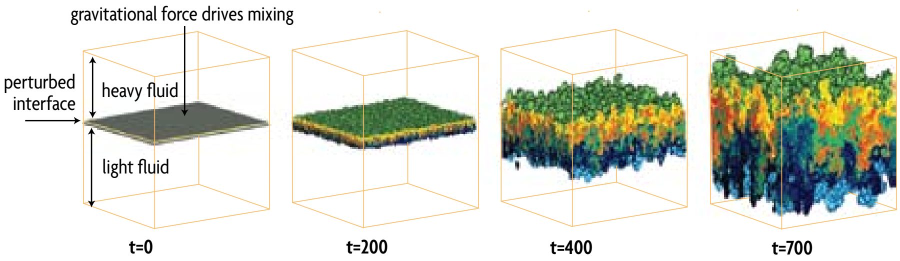


FIGURE 1.
Interactive visualization of four timesteps of the [EQUATION] simulation of a Rayleigh-Taylor instability. Gravity drives the mixing of a heavy fluid on top of a lighter one. Two envelope surfaces capture the mixing region.

VISUS: PROGRESSIVE STREAMING FOR SCALABLE DATA EXPLORATION

In recent years, computational scientists with access to the world's largest supercomputers have successfully simulated a number of natural and man-made phenomena with unprecedented levels of detail. Such simulations routinely produce massive amounts of data. For example, hydrodynamic instability simulations performed at Lawrence Livermore National Laboratory (LLNL) in early 2002 produced several tens of terabytes of data, as shown in Figure 1. This data must be visualized and analyzed to verify and validate the underlying model, understand the phenomenon in detail, and develop new insights into its fundamental physics. Therefore, both visualization and data analysis algorithms require new, advanced designs that enable high performance when dealing with large amounts of data.

Date-streaming techniques and out-of-core computing specifically address the issues of algorithm redesign and data layout restructuring, which are necessary to enable scalable processing of massive amounts of data. For example, space-filling curves have been used to develop a static indexing scheme called ViSUS,[1] which produces a data layout that enables the hierarchical traversal of [EQUATION]-dimensional regular grids. Three features make this approach particularly attractive: (1) the order of the data is independent of the parameters of the physical hardware (a cache-oblivious approach), (2) conversion from Z-order used in classical database approaches is achieved using a simple sequence of bit-string manipulations, and (3) it does not introduce any data replication. This approach has

[1] www.pascucci.org/visus

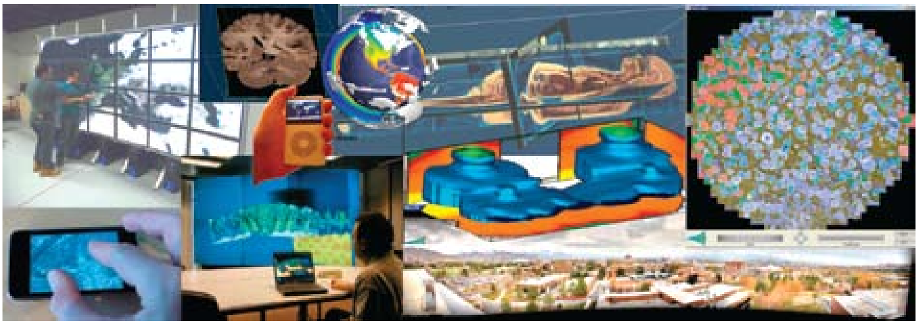


FIGURE 2.
Scalability of the ViSUS infrastructure, which is used for visualization in a variety of applications (such as medical imaging, subsurface modeling, climate modeling, microscopy, satellite imaging, digital photography, and large-scale scientific simulations) and with a wide range of devices (from the iPhone to the powerwall).

been used for direct streaming and real-time monitoring of large-scale simulations during execution [10].

Figure 2 shows the ViSUS streaming infrastructure streaming LLNL simulation codes and visualizing them in real time on the Blue Gene/L installation at the Supercomputing 2004 exhibit (where Blue Gene/L was introduced as the new fastest supercomputer in the world). The extreme scalability of this approach allows the use of the same code base for a large set of applications while exploiting a wide range of devices, from large powerwall displays to workstations, laptop computers, and handheld devices such as the iPhone.

Generalization of this class of techniques to the case of unstructured meshes remains a major problem. More generally, the fast evolution and growing diversity of hardware pose a major challenge in the design of software infrastructures that are intrinsically scalable and adaptable to a variety of computing resources and running conditions. This poses theoretical and practical questions that future researchers in visualization and analysis for data-intensive applications will need to address.

VISTRAILS: PROVENANCE AND DATA EXPLORATION

Data exploration is an inherently creative process that requires the researcher to locate relevant data, visualize the data and discover relationships, collaborate with peers while exploring solutions, and disseminate results. Given the volume of data and complexity of analyses that are common in scientific exploration, new tools are needed and existing tools should be extended to better support creativity.

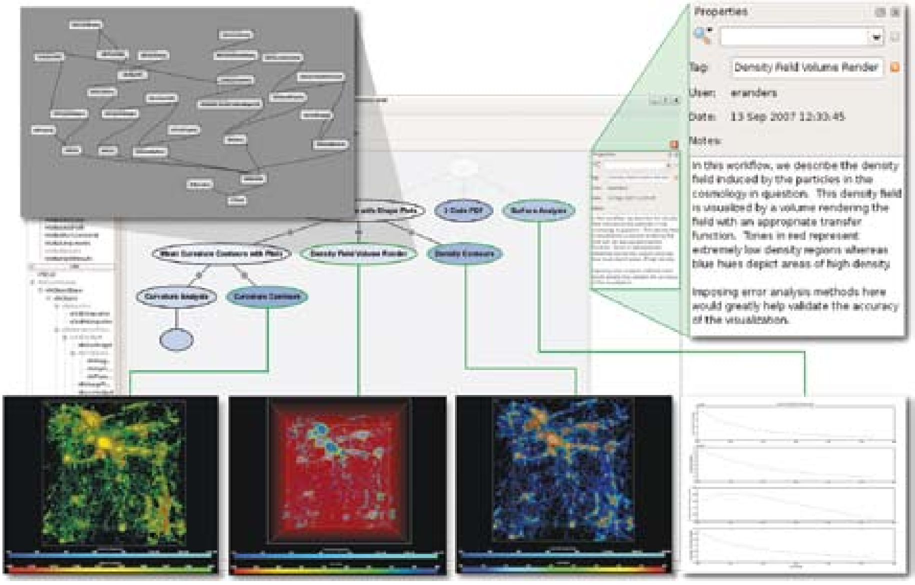
The ability to systematically capture provenance is a key requirement for these tools. The provenance (also referred to as the audit trail, lineage, or pedigree) of a data product contains information about the process and data used to derive the data product. The importance of keeping provenance for data products is well recognized in the scientific community [4, 5]. It provides important documentation that is key to preserving the data, determining its quality and authorship, and reproducing and validating the results. The availability of provenance also supports reflective reasoning, allowing users to store temporary results, make inferences from stored knowledge, and follow chains of reasoning backward and forward.

VisTrails[2] is an open source system that we designed to support exploratory computational tasks such as visualization, data mining, and integration. VisTrails provides a comprehensive provenance management infrastructure and can be easily combined with existing tools and libraries. A new concept we introduced with VisTrails is the notion of *provenance of workflow evolution* [6]. In contrast to previous workflow and visualization systems, which maintain provenance only for derived data products, VisTrails treats the workflows (or pipelines) as first-class data items and keeps their provenance. VisTrails is an extensible system. Like workflow systems, it allows pipelines to be created that combine multiple libraries. In addition, the VisTrails provenance infrastructure can be integrated with interactive tools, which cannot be easily wrapped in a workflow system [7].

Figure 3 shows an example of an exploratory visualization using VisTrails. In the center, the visual trail, or *vistrail*, captures all modifications that users apply to the visualizations. Each node in the vistrail tree corresponds to a pipeline, and the edges between two nodes correspond to changes applied to transform the parent pipeline into the child (e.g., through the addition of a module or a change to a parameter value). The tree-based representation allows a scientist to return to a previous version in an intuitive way, undo bad changes, compare workflows, and be reminded of the actions that led to a particular result.

Ad hoc approaches to data exploration, which are widely used in the scientific community, have serious limitations. In particular, scientists and engineers need

[2] http://vistrails.sci.utah.edu



Tag: Density Field Volume Render

User: eranders

Date: 13 Sep 2007 12:33:45

Notes:

In this workflow, we describe the density field induced by the particles in the cosmology in question. This density field is visualized by a volume rendering the field with an appropriate transfer function. Tones in red represent extremely low density regions whereas blue hues depict areas of high density.

Imposing error analysis methods here would greatly help validate the accuracy of the visualization.

FIGURE 3.
An example of an exploratory visualization for studying celestial structures derived from cosmological simulations using VisTrails. Complete provenance of the exploration process is displayed as a "vistrail." Detailed metadata are also stored, including free-text notes made by the scientist, the date and time the workflow was created or modified, optional descriptive tags, and the name of the person who created it.

to expend substantial effort managing data (e.g., scripts that encode computational tasks, raw data, data products, images, and notes) and need to record provenance so that basic questions can be answered, such as: Who created the data product and when? When was it modified, and by whom? What process was used to create it? Were two data products derived from the same raw data? This process is not only time consuming but error prone. The absence of provenance makes it hard (and sometimes impossible) to reproduce and share results, solve problems collaboratively, validate results with different input data, understand the process used to solve a particular problem, and reuse the knowledge involved in the data analysis process. It also greatly limits the longevity of the data product. Without precise and sufficient information about how it was generated, its value is greatly diminished. Visualization systems aimed at the scientific domain need to provide a flexible

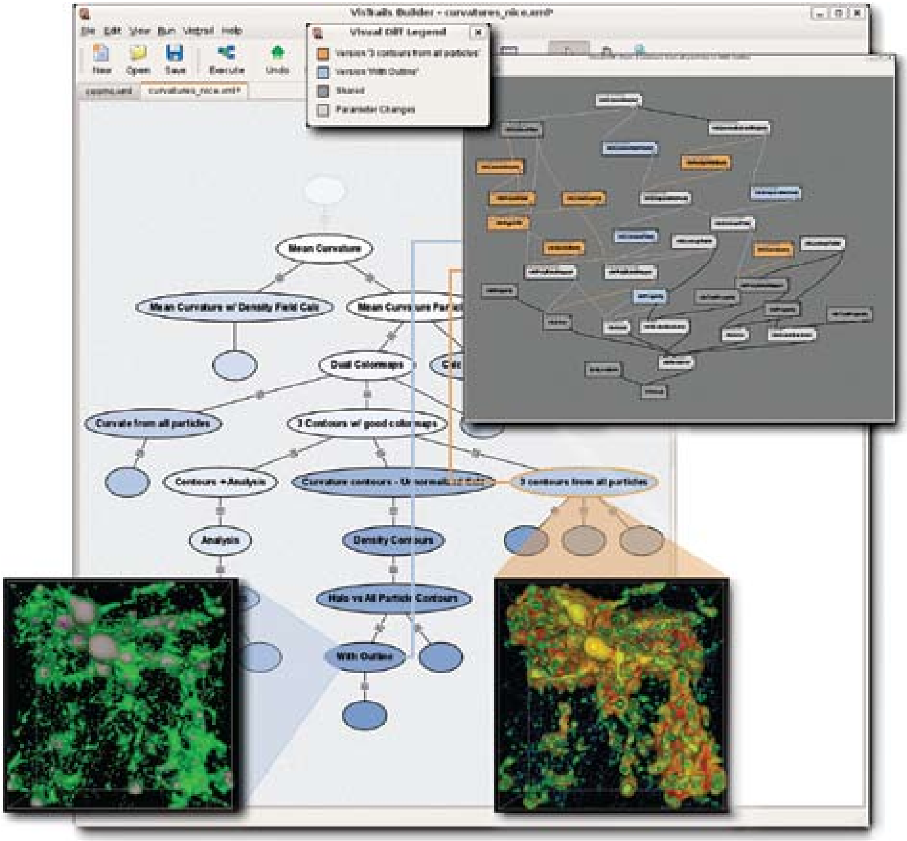

VisTrails Builder - curvatures_nice.xml*
File Edit View Run Vistrail Help
New Open Save Execute Undo
cosmo.xml curvatures_nice.xml*
Visual Diff Legend
Version '3 contours from all particles'
Version 'With Outline'
Shared
Parameter Changes
Mean Curvature
Mean Curvature w/ Density Field Calc
Mean Curvature Partic
Dual Colormaps
Curvate from all particles
3 Contours w/ good colormaps
Contours +Analysis
Curvature contours - Unnormalized data
3 contours from all particles
Analysis
Density Contours
Halo vs All Particle Contours
With Outline

FIGURE 4.

Representing provenance as a series of actions that modify a pipeline makes visualizing the differences between two workflows possible. The difference between two workflows is represented in a meaningful way, as an aggregation of the two. This is both informative and intuitive, reducing the time it takes to understand how two workflows are functionally different.

framework that not only enables scientists to perform complex analyses over large datasets but also captures detailed provenance of the analysis process.

Figure 4 shows ParaView[3] (a data analysis and visualization tool for extreme-

[3] www.paraview.org

ly large datasets) and the VisTrails Provenance Explorer transparently capturing a complete exploration process. The provenance capture mechanism was implemented by inserting monitoring code in ParaView's undo/redo mechanism, which captures changes to the underlying pipeline specification. Essentially, the action on top of the undo stack is added to the vistrail in the appropriate place, and undo is reinterpreted to mean "move up the version tree." Note that the change-based representation is both simple and compact—it uses substantially less space than the alternative approach of storing multiple instances, or versions, of the state.

FLOW VISUALIZATION TECHNIQUES

A precise qualitative and quantitative assessment of three-dimensional transient flow phenomena is required in a broad range of scientific, engineering, and medical applications. Fortunately, in many cases the analysis of a 3-D vector field can be reduced to the investigation of the two-dimensional structures produced by its interaction with the boundary of the object under consideration. Typical examples of such analysis for fluid flows include airfoils and reactors in aeronautics, engine walls and exhaust pipes in the automotive industry, and rotor blades in turbomachinery.

Other applications in biomedicine focus on the interplay between bioelectric fields and the surface of an organ. In each case, numerical simulations of increasing size and sophistication are becoming instrumental in helping scientists and engineers reach a deeper understanding of the flow properties that are relevant to their task. The scientific visualization community has concentrated a significant part of its research efforts on the design of visualization methods that convey local and global structures that occur at various spatial and temporal scales in transient flow simulations. In particular, emphasis has been placed on the interactivity of the corresponding visual analysis, which has been identified as a critical aspect of the effectiveness of the proposed algorithms.

A recent trend in flow visualization research is to use GPUs to compute image space methods to tackle the computational complexity of visualization techniques that support flows defined over curved surfaces. The key feature of this approach is the ability to efficiently produce a dense texture representation of the flow without explicitly computing a surface parameterization. This is achieved by projecting onto the image plane the flow corresponding to the visible part of the surface, allowing subsequent texture generation in the image space through backward integration and iterative blending. Although the use of partial surface parameterization obtained by projection results in an impressive performance gain, texture patterns

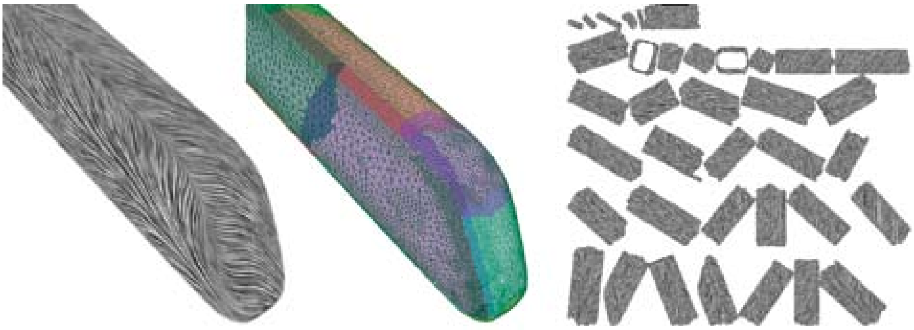


FIGURE 5.
Simulation of a high-speed ICE train. Left: The GPUFLIC result. Middle: Patch configurations. Right: Charts in texture space.

stretching beyond the visible part of the self-occluded surface become incoherent due to the lack of full surface parameterization.

To address this problem, we have introduced a novel scheme that fully supports the creation of high-quality texture-based visualizations of flows defined over arbitrary curved surfaces [8]. Called Flow Charts, our scheme addresses the issue mentioned previously by segmenting the surface into overlapping patches, which are then individually parameterized into charts and packed in the texture domain. The overlapped region provides each local chart with a smooth representation of its direct vicinity in the flow domain as well as with the inter-chart adjacency information, both of which are required for accurate and non-disrupted particle advection. The vector field and the patch adjacency relation are naturally represented as textures, enabling efficient GPU implementation of state-of-the-art flow texture synthesis algorithms such as GPUFLIC and UFAC.

Figure 5 shows the result of a simulation of a high-speed German Intercity-Express (ICE) train traveling at a velocity of about 250 km/h with wind blowing from the side at an incidence angle of 30 degrees. The wind causes vortices to form on the lee side of the train, creating a drop in pressure that adversely affects the train's ability to stay on the track. These flow structures induce separation and attachment flow patterns on the train surface. They can be clearly seen in the proposed images close to the salient edges of the geometry.

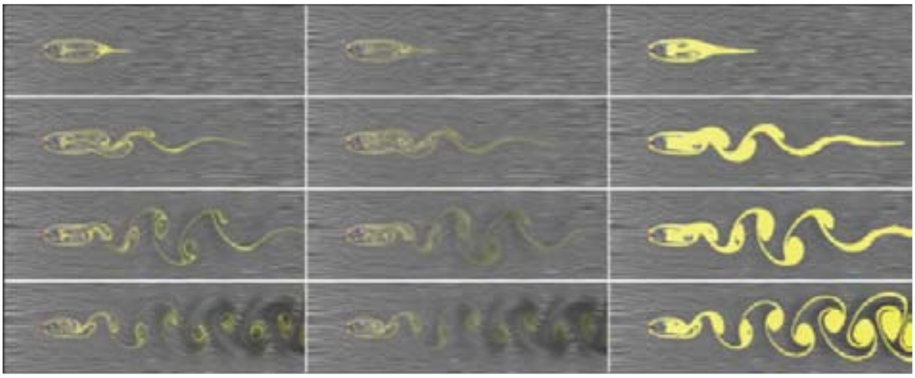


FIGURE 6.
Visualization of the =Karman= dataset using dye advection. Left column: Physically based dye advection. Middle column: Texture advection method. Right column: Level-set method. The time sequence is from top to bottom.

The effectiveness of a physically based formulation can be seen with the =Karman= dataset (Figure 6), a numerical simulation of the classical Von Kármán vortex street phenomenon, in which a repeating pattern of swirling vortices is caused by the separation of flow passing over a circular-shaped obstacle. The visualization of dye advection is overlaid on dense texture visualization that shows instantaneous flow structures generated by GPUFLIC. The patterns generated by the texture-advection method are hazy due to numerical diffusion and loss of mass. In a level-set method, intricate structures are lost because of the binary dye/background threshold. Thanks to the physically based formulation [9], the visualization is capable of accurately conveying detailed structures not shown using the traditional texture-advection method.

FUTURE DATA-INTENSIVE VISUALIZATION CHALLENGES

Fundamental advances in visualization techniques and systems must be made to extract meaning from large and complex datasets derived from experiments and from upcoming petascale and exascale simulation systems. 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 must 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. 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. 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.

Our current data-analysis capabilities lag far behind our ability to produce simulation data or record observational data. New visual data analysis techniques will need to dynamically consider high-dimensional probability distributions of quantities of interest. This will require new contributions from mathematics, probability, and statistics. The scaling of simulations to ever-finer granularity and timesteps brings new challenges in visualizing the data that is generated. It will be 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 datasets using a more top-down methodological path. The ability to fully quantify uncertainty in high-performance computational simulations will provide new capabilities for verification and validation of simulation codes. Hence, uncertainty representation and quantification, uncertainty propagation, and uncertainty visualization techniques need to be developed to provide scientists with credible and verifiable visualizations.

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 multi-model nature of the data; provide the means for scientists to easily transition views from global to local model data; allow blending of traditional scientific visualization and information visualization; perform hypothesis testing, verification, and validation; and address the challenges posed by the use of vastly different grid types and by the various elements of the multi-model code. Tools that leverage semantic information and hide details of dataset formats will be critical to enabling visualization and analysis experts to concentrate on the design of these approaches rather than becoming mired in the trivialities of particular data representations [11].

ACKNOWLEDGMENTS

This work was funded in part by DOE: VACET, DOE SDM, DOE C-SAFE Alliance Center, KAUST GPR: KUS-C1-016-04, NSF CNS-0615194, NSF CNS-0551724, NSF CCF-0541113, NSF IIS-0513212. Our research has been funded by the National Science Foundation (grants IIS-0746500, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0534628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, and CCF-0528201), IBM Faculty Awards (2005, 2006, and 2007), and NIH NCRR Grant No. 5P41RR012553-10. The authors would like to thank Juliana Freire and the VisTrails team for help with the third section of this article.

REFERENCES

[1] C. R. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, Eds., *NIH-NSF Visualization Research Challenges Report*, IEEE Press, ISBN 0-7695-2733-7, 2006, http://vgtc.org/wpmu/techcom/national-initiatives/nihnsf-visualization-research-challenges-report-january-2006, doi: 10.1109/MCG.2006.44.

[2] NSF Blue Ribbon Panel Report on Simulation-Based Engineering Science (J. T. Oden, T. Belytschko, J. Fish, T. Hughes, C.R. Johnson, D. Keyes, A. Laub, L. Petzold, D. Srolovitz, and S. Yip), "Simulation-Based Engineering Science," 2006, www.nd.edu/~dddas/References/SBES_Final_Report.pdf.

[3] NIH-NSF Visualization Research Challenges, http://erie.nlm.nih.gov/evc/meetings/vrc2004.

[4] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," Proc. *ACM SIGMOD*, pp. 1345–1350, 2008, doi: 10.1.1.140.3264.

[5] J. Freire, D. Koop, E. Santos, and C. Silva, "Provenance for computational tasks: A survey," *Comput. Sci. Eng*, vol. 10, no. 3. pp. 11–21, 2008, doi: 10.1109/MCSE.2008.79.

[6] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo, "Managing rapidly-evolving scientific workflows," International Provenance and Annotation Workshop (IPAW), LNCS 4145, pp. 10–18, 2006, doi: 10.1.1.117.5530.

[7] C. Silva, J. Freire, and S. P. Callahan, "Provenance for visualizations: Reproducibility and beyond," *IEEE Comput. Sci. Eng.*, 2007, doi: 10.1109/MCSE.2007.106.

[8] G.-S. Li, X. Tricoche, D. Weiskopf, and C. Hansen, "Flow charts: Visualization of vector fields on arbitrary surfaces," *IEEE Trans. Visual. Comput. Graphics*, vol. 14, no. 5, pp. 1067–1080, 2008, doi: 10.1109/TVCG.2008.58.

[9] G.-S. Li, C. Hansen, and X. Tricoche, "Physically-based dye advection for flow visualization. *Comp. Graphics Forum J.*, vol. 27, no. 3, pp. 727–735, 2008, doi: 10.1111/j.1467-8659.2008.01201.x.

[10] V. Pascucci, D. E. Laney, R. J. Frank, F. Gygi, G. Scorzelli, L. Linsen, and B. Hamann, "Real-time monitoring of large scientific simulations," *SAC*, pp. 194–198, ACM, 2003, doi: 10.1.1.66.9717.

[11] "Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale," C. R. Johnson, R. Ross, S. Ahern, J. Ahrens, W. Bethel, K. L. Ma, M. Papka, J. van Rosendale, H. W. Shen, and J. Thomas, www.sci.utah.edu/vaw2007/DOE-Visualization-Report-2007.pdf, 2007.