Department: Visualization Viewpoints
Editor: Theresa-Marie Rhyne, theresamarierhyne@gmail.com

Enabling Domain Expertise in Scientific Visualization With CinemaScience

Terece L. Turton
Los Alamos National Laboratory

Divya Banesh
University of California, Davis; Los Alamos National Laboratory

Trinity Overmyer
Purdue University

Benjamin H. Sims
Los Alamos National Laboratory

David H. Rogers
Los Alamos National Laboratory

***Abstract*—Scientific users present unique challenges to visualization researchers. Their high-level tasks require them to apply domain-specific expertise. We introduce a broader audience to the CinemaScience project and demonstrate how CinemaScience enables efficient visualization workflows that can bring in scientist expertise and drive scientific insight.**

Science is a creative enterprise and users of scientific visualization need flexible, creative ways to visualize and derive insight from their data. Scientists often have unique challenges in the visualization process. Their data can span many scales and can include experimental, observational, and simulation data. Current science simulations can produce terabytes to exabytes of data, challenging the ability of visualization applications to quickly render or efficiently process the data. Observational data sources may be orders of magnitude smaller. Yet the scientist will need to distill down both simulation and observational data to a compatible representation in order to validate the model.

Scientists also need the ability to apply their domain expertise in a way that facilitates data

Digital Object Identifier 10.1109/MCG.2019.2954171
Date of current version 6 January 2020.

Published by the IEEE Computer Society

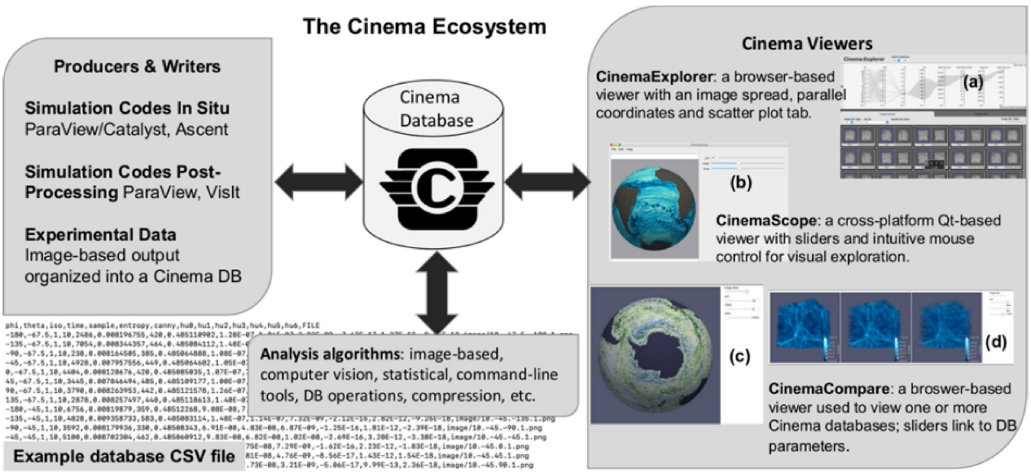

Figure 1. Cinema Ecosystem consists of functionality to generate Cinema databases; analysis algorithms and frameworks; and flexible viewers that can be tailored to meet the needs of a scientist. An example database CSV file is shown. The viewers display (a) a Nyx[3] cosmology simulation in CinemaExplorer, where various statistical and image-based properties have been extracted for each image. An MPAS-Ocean[4] database is viewed in (b) CinemaScope and (c) CinemaCompare viewer. The sliders are linked to the database parameters. Cinema databases often include time and viewing angles, enabling the user to explore the data temporally and spatially. (d) CinemaCompare can be used to explore multiple datasets; three Nyx runs with different parameters can be compared for verification and validation of the simulation.

exploration, analysis, and knowledge discovery. One such visualization approach to enabling scientific expertise is the CinemaScience project. The use of CinemaScience enables domain scientists to explore their data through a domain-science perspective rather than through the more computer-graphics approach that traditional visualization applications provide.

In this article, we describe the CinemaScience project and then provide a survey of examples of its use in scientific visualization to illustrate how it enables domain expertise.

CINEMASCIENCE ECOSYSTEM

The CinemaScience project provides a flexible visualization ecosystem, shown in Figure 1, for extreme scale scientific datasets. Cinema[1,2] is a database approach that encompasses *in situ* (occurring while the simulation is running) and postprocessing visualization workflows for both simulation and experimental datasets. The original Cinema concept was based on the idea that one could render a visualization *in situ*, saving the visualization images rather than writing out the full output data thus drastically decreasing the output size. The user decides what to visualize and save, downsampling from the high-dimensional simulation parameter space to, for example, specific variables, camera angles, or slices. The resultant images, when viewed with a Cinema viewer, provide a similar postprocessing experience to exploring the data with a typical visualization application, but without the rendering overhead. In addition to saving images, the user can instead save floating-point data values within the PNG output file, allowing access to that data for analysis rather than a value possibly distorted by an applied color map. Due to these space saving advantages, the use of CinemaScience as an *in situ* tool can save the results of a simulation at much higher temporal and spatial resolutions than when saving the full output data.

That original Cinema concept has since expanded to include the following:

- A database (DB) specification that is agnostic to the source of data.
- A comma separated value (CSV) based relational database structure that maps a set of parameters to a set of data artifacts. An example CSV file can be seen in Figure 1.
- A range of data abstracts including metadata such as run parameters, output variables,

and other possible data forms such as CSV files, text files, or meshes.
- A database/query approach to visual analytics leveraging common visualization techniques such as parallel coordinates.
- A set of image and data-based algorithms for postprocessing analysis.

At the core of the Cinema concept is the database specification that allows data to be written and read by any application. This makes it straightforward, for example, to create a Cinema database from a set of experimental images using a script to write out a CSV file with run and image properties mapped to each image. Or to read in a Cinema database, apply analysis routines and write out an augmented database with derived output quantities associated with the input data abstracts. This innate flexibility of the CinemaScience project—the ability to explore any collection of images regardless of the source of these images—is part of what makes Cinema especially useful in the area of image-based *in situ* techniques. Cinema database generation tools are available, but they are not required to make full use of CinemaScience's exploratory capabilities.

CINEMA VIEWERS

Working in tandem with the Cinema database is a set of flexible viewers and viewer components that can be combined to produce a visual analysis tool tailored to a scientist's workflow. The same Cinema database can be loaded into any viewer without modifications to the data. These open source viewers provide a foundation on which more specific components can be built to suit the scientists' data exploration needs. There are three basic Cinema viewers as shown in Figure 1. Each viewer facilitates visual exploration and analysis in different ways:

- *CinemaScope:* A cross-platform Qt viewer that enables users to explore an image database both with sliders and with intuitive mouse control for zooming and rotating. Useful for collaborations and integration with other tools such as OpenCV, Python, and R.
- *CinemaCompare:* A browser-based viewer that enables users to compare one or more image databases through parametric exploration.
- *CinemaExplorer:* A browser-based ensemble viewer that combines parallel coordinates, scatter plots, and image artifacts for in-depth analysis. Implemented using D3, the Cinema Explorer parallel coordinates can be used to select ranges on the parallel coordinate axes to explore relationships between parameters, find outliers, and view specific data. Using linked views, this ensemble viewer enables scientists to correlate parametric information to resulting images, making it easier to find convergences, divergences, and correspondences in large datasets.

Browser-based viewers benefit analysis workflows for two main reasons: 1) Collaboration and sharing of information within the scientific community becomes easier when lengthy software installations are not required, and 2) Exploring data results at remote locations or experimental labs can be made possible and expedited when software installation permissions are restricted.

Producing Cinema Databases *In Situ* and Postprocessing

In a traditional visualization workflow for a large-scale simulation, the code may output petabytes of data, which is then visualized with a standard application such as ParaView[5] or VisIt.[6] The time it takes to render each timestep or variable is often a factor limiting exploration of the data. I/O and storage considerations preclude saving every timestep and there is usually some intentional or unintentional downsampling of data by only saving every *n*th timestep. Cinema provides a complementary approach in which downsampling takes place by saving the simulation at the resolution of the data artifacts. However, the space saving capability of Cinema usually means higher temporal resolution.

Cinema database export capabilities are part of standard scientific visualization applications. ParaView and VisIt both output Cinema databases, and ParaView's *in situ* framework Catalyst can be used to instrument a simulation code for *in situ* generation of Cinema databases. Ascent,[7] an *in situ* infrastructure under development as part of the Exascale Computing Project (www.exascaleproject.org/), also includes functionality for *in situ* production of Cinema databases.

Due to the universal nature of a CSV file, Cinema is uniquely suited to experimental image data. Scientists naturally save run parameters and other metadata in spreadsheet form. Linking that information to images or other data abstracts can be done via scripts running during data collection or via a postprocessing script to organize experimental data into a Cinema database.

CINEMA-ENABLED SCIENTIFIC WORKFLOWS

In order to spur introspection and progress in scientific visualization, Johnson[8] proposed a list of the top scientific visualization research problems. CinemaScience provides practical, user-focused solutions that address many of these issues (the numbers refer to the list in the work of Johnson).[8]

- (1) *Think about the science:* The main goal of CinemaScience is to enable scientists to bring domain expertise to visualization and facilitate discovery.
- (5) *Efficiently utilizing novel hardware architectures:* Cinema works with many leading applications (Paraview, VisIT) and infrastructures (Ascent), providing a straightforward path for integration into *in situ* workflows on a variety of hardware architectures (such as NERSC's Cori, ORNL's Summit, and LLNL's Sierra).
- (6) *Human–computer interaction:* With its flexible suite of components, CinemaScience makes it possible to create interfaces adaptable for many types of analyses.
- (7) *Global/local visualization (details within context):* Cinema can display a global overview and enable close examination of local features through use of views and call-outs, as well as through the parallel coordinate interface.
- (8) *Integrated problem-solving environments (PSEs):* CinemaScience's CSV-based data structure is simple and flexible, making it easy to update the database and make decisions in real-time.
- (9) *Multi-field visualization:* The ability to compare multiple databases and analyze ensembles of data both address this issue.
- (10) *Integrating scientific and information visualization:* The parallel coordinates and scatter plot views provide simple but effective ways to find clusters and correlations.
- (11) *Feature Detection:* CinemaScience's database approach has been combined with various image processing algorithms to support a wide range of feature detection needs.
- (12) *Time-dependent visualization:* Time is a common parameter included in a Cinema database file. Standard feature matching and tracking techniques have also been applied to spatial and temporal sequences of Cinema database images.

Drawing on some of the tasks and challenges from Johnson, we survey specific analyses and use cases that demonstrate how CinemaScience can enable scientists to bring in their domain specific expertise, foster scientific insight and discovery, and create more efficient scientific workflows.

Feature Detection and Feature Tracking Over Time

Feature detection, matching, and tracking are separate but interconnected tasks. Identifying changes in features as a function of time is a corollary of this task—when does a feature appear, disappear, undergo a physically significant change in shape or size. The process must consider multiple features and the complex interactions among them.

A Cinema-enabled tracking system can be seen in Banesh *et al.*,[9] a multistage workflow that enables the user to identify ocean eddies, view images to follow eddy movement, and count and track eddies through time. The flexibility of Cinema is leveraged to create a highly specialized tool that enables a scientist to apply their domain specific knowledge. Leveraging the ability to save data values at the resolution of an image (within a Cinema database), analysis algorithms are applied directly on the data through floating-point images that record a projection of the simulation values. The combination of these low-cost Cinema database images and optimized computer vision algorithms creates an interface for scientists, where exploration and application of their domain knowledge to their data is made easier and possible in real time.

Detecting the Gulf Stream Western Boundary In a similar approach, the Cinema-based framework can be used to define the

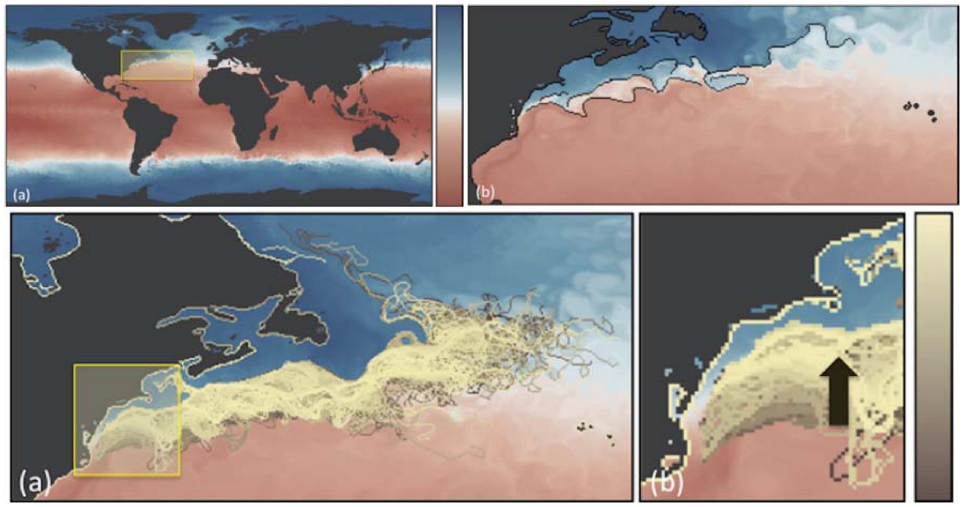

Figure 2. *Top:* (a) Two-dimensional projection of MPAS-Ocean sea surface temperature dataset with a hot–cold colormap. (b) The inset yellow region in (a) is magnified. Canny edge detection is applied to extract the edges (black polylines) indicating the large temperature gradient changes between the warmer Equatorial waters and the cooler Polar waters. These edges correspond to the northern boundary of the Gulf Stream. *Bottom:* Mapping 173 time steps of the simulation. (a) Enables scientists to view shifts in the temperature front of the Gulf Stream over time. For example, focusing on the region highlighted in yellow in (a) and enlarged in (b), the leftmost region of the temperature front of the Gulf Stream shows a gradual shift northward (as indicated by the arrow). This shift corresponds to seasonal activity. Note that lighter colors refer to later time steps.

extent of ocean currents through the application of edge detection. Understanding the path and fluctuations of ocean currents is important for understanding ocean climate and their impact on coastal cities. Analyzing current locations is a major component of validating the accuracy of an ocean simulation such as MPAS-O. By using edge detection techniques within an analysis framework applied to a Cinema database of sea surface temperature, the Gulf Stream can be extracted from the image database. Again, the direct representation of the simulation output as floating-point values within the saved PNG images enables access to the actual data for analysis.

The Gulf Stream is defined as the location where the warmer waters from the Equator meet the cooler waters from the North Pole. This intermingling of waters creates a high gradient in sea surface temperature. This type of phenomena can be extracted by an edge detection technique based on gradients in the image intensity. Figure 2 (top) shows a single timestep of MPAS-O sea surface temperature with the fine black line in (b) delineating the northern boundary of the Gulf Stream. Extracting the highest gradient change in the region corresponds to the northern boundary of the Gulf Stream.

Using a temporal Cinema database, the aggregate movement of the Gulf Stream northern boundary can be mapped. In Figure 2 (bottom), the individual boundaries from each timestep of the two and a half year MPAS-O simulation are aggregated, with the lighter edge color indicating the later timesteps. Together, this mapping shows the temperature front closer to the coast line experiencing directed, seasonal movement, while the temperature front further out into the Atlantic Ocean exhibits more chaotic behavior. This Cinema-based workflow can be compared to the previous manual workflow of the scientist in which many timesteps were printed out and Gulf Stream boundaries were hand drawn—a process fraught with error and non-reproducibility.

The results from the Cinema workflow on simulation data can be compared to observational data to see if model predictions conform

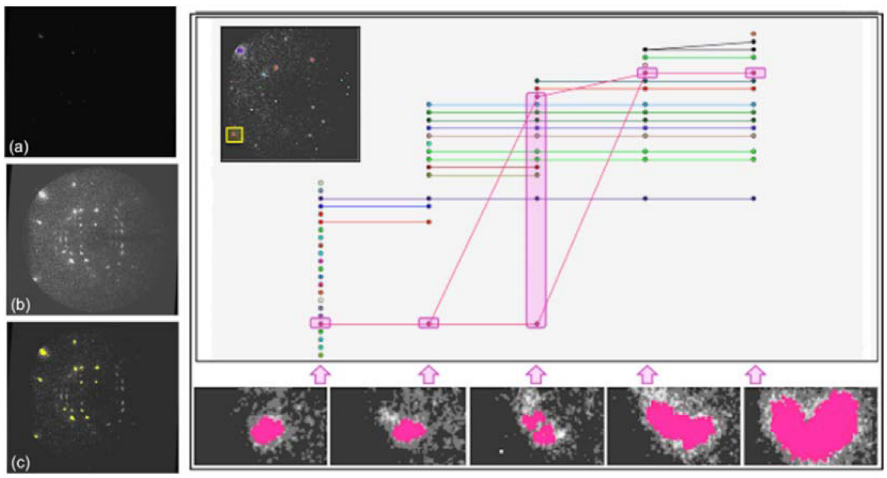

Figure 3. *Left:* (a) Experimental data TIFF image. (b) Brightness- and contrast-enhanced reference image; bright regions correspond to the Bragg reflections, and (c) enhanced image with Bragg reflections extracted (in yellow) through contour detection. *Right:* Given the extraction of Bragg reflection and thresholding of features based on size to remove noise and experimental artifacts, the remaining Bragg reflections can be tracked temporally to understand the effect of shock compression on a crystalline material. The feature tracking graph shows the creation, annihilation, and persistence of features across the discrete time steps. The highlighted region in pink and close-up analysis depict the feature split and merge back of a specific Bragg reflection.

to observation. Using this Cinema-enabled approach enables scientists to quickly and efficiently extract accurate information about the Gulf Stream and apply domain specific knowledge to not only make scientific advancements, but to validate ocean models. Since CinemaScience is agnostic to the origin of the images, ocean scientists can apply the same techniques to observational data as well as a range of ocean simulation models, enabling a direct benchmarking of theory and observations.

Extracting Bragg Reflections From Noisy Data Shock compression experiments to determine crystalline structure and properties are challenging due to the very short time in which measurements can be made and because the physical sample is irreparably damaged during the experiment. The short experimental time scale requires a high flux of X-rays to make accurate measurements. The physicists enhance the X-ray diffraction signal through electronic amplification to improve the data quality. This amplification produces high-frequency noise in the data images. Key tasks of the scientists include 1) extracting the Bragg reflections in the experimental images; 2) removing the noise in the data; and 3) tracking the Bragg reflections over the timesteps of each compression experiment. In Figure 3 (left), one can see a) the original experimental image, b) a brightness- and contrast-enhanced representation of the original image to highlight the locations of the Bragg reflections, and c) extracted Bragg reflections as determined by a workflow where the images were arranged into a Cinema database and analyzed by the scientist using the computer vision analysis framework. In Figure 3(c), scientists can vary the isovalue to identify closed contours that correlate with Bragg reflections of interest. Moment invariants are then used to set an upper and lower threshold to remove noise.

Changes in Bragg reflection position and size can help scientists identify the type of transformation the crystalline material undergoes during shock compression. For example, in Figure 3 (right), during the transition from the first step to the second, a majority of the existing reflections have died, and new reflections are formed. However, in steps two to five most of the features are constant or identify splits and merges. The appearance or disappearance of reflections

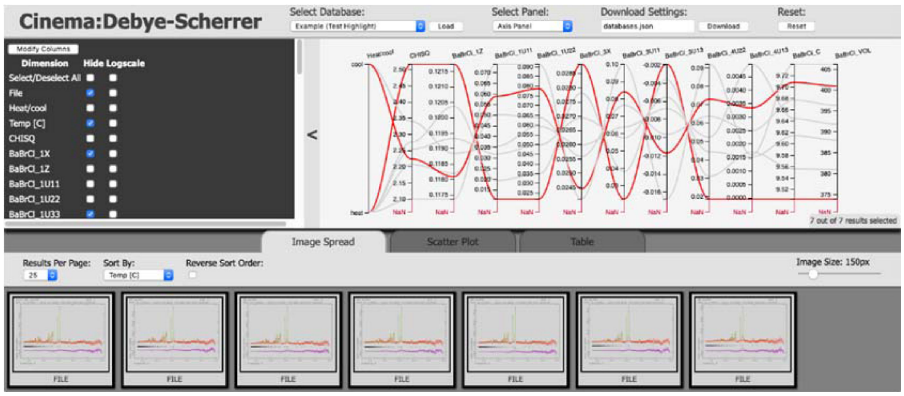

Figure 4. Example of the CinemaDebye-Scherrer user interface. The view in the upper left can be used to select specific variables which the parallel coordinates can be used to threshold and select ranges to aid the scientist in finding outliers in the data. (Sample dataset available from github.com/cinemascience.)

indicate a transition to a different crystal structure, whereas splitting or combining reflections is indicative of changes in strain or grain size of the crystals without a change in the structure.

The previous scientist workflow consisted of manually identifying the Bragg peaks and visually scanning the image sequences to try to track changes. The scientists found it much easier to visualize the temporal evolution through CinemaScience-based feature tracking, specifically for certain crystalline materials, where the ability to categorize Bragg reflections had proven to be especially difficult in past research. Results such as these can be used to validate simulations or compare across experiments. For shock compression science, CinemaScience enables scientists to gain insight into the underlying physics.

Ensemble Visualization and Analysis

Ensemble visualization is a common task in scientific visualization, both for simulation and observational data. Ensembles can be used to explore the parameter space for models, for understanding and quantifying uncertainty, or to analyze a range of experimental data.

Powder Diffraction Analysis In Figure 4, a specialized Cinema viewer is used to visualize a high-dimensional ensemble of Rietveld analysis plots from powder diffraction experiments.[10] The CinemaDebye-Scherrer viewer combines the parallel coordinates view with a view to the left that can be used for analysis related tasks. In this figure, the view panel is used to select axes for the analysis. Other views are available to run queries or set thresholds. In addition to the usual image spread and scatter plot, a Table tab enables the user to drill down into the actual numbers in the analysis plots that are the data abstracts in this Cinema database. The plots can link to tens to hundreds of input and output parameters associated with the diffraction run and Rietveld analysis plot.

The CinemaDebye-Scherrer viewer enables the scientist to manipulate, threshold, and select ensemble members. Important goals for the scientists in this analysis were to find outliers, explore the parameter space to look for coverage gaps, and determine the next sets of runs. The previous scientist workflow was not able to take into account the many input and output parameters. The parallel coordinates and query views enable the scientist to apply range and thresholding selections, identifying trends, relationships, and runs that are problematic or meaningful. From the information derived, the scientist brings in their knowledge of powder diffraction to decide the next steps in the analysis to optimize time spent at the experimental facilities.

Beamline Science for XFEL Shock Physics Experiments Figure 5 shows CinemaBandit,[11] a specialized Cinema viewer integrated with the workflow at a beam-time facility to enable fast-paced, real-time decision-making using multiple data types, and an ensemble of experimental results. Using CinemaBandit, scientists can correlate experimental

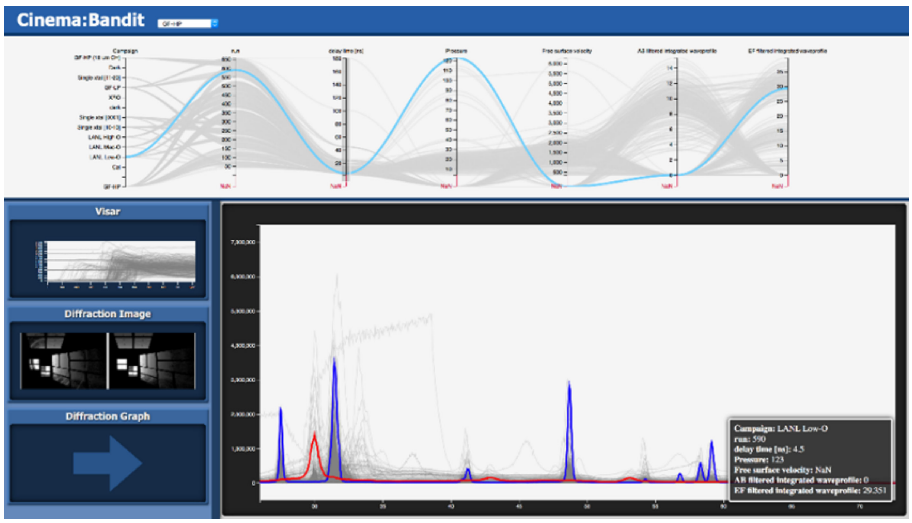

Figure 5. Overview of the CinemaBandit interface.[11] The parallel coordinates plot at the top shows parametric (i.e., inputs and outputs) information about each run in an experimental dataset. At the bottom, the user can alternate between the VISAR, Diffraction Image, and Diffraction Graph views (Source: Orban *et al.*;[11] used with permission.)

input parameters with resulting values, velocity interferometer system for any reflector (VISAR) and diffraction graphs, and reconstructed diffraction images at a beam-time facility. This enables them to make analytical decisions on upcoming experiments in tandem to the experimental process, creating an integrated visualization environment that provides real-time feedback. This contrasts with their previous workflow in which they would analyze the results offline to plan how to use the next beam time allotment. CinemaBandit can be integrated into a continuous workflow that automatically adds new experimental data to the CSV file and database. The browser-based viewer is automatically refreshed with each new addition ensuring a seamless workflow. Users can also narrow the data of interest in the parallel coordinates plot for focused analysis and to determine if there are data ranges that need more experimental exploration. VISAR, diffraction image, and diffraction graph views are linked to the parallel coordinates view so that selection of a run in one view highlights the appropriate data in the others. Being browser based, CinemaBandit can be run on any laptop or network without software installation requirements.

As X-ray free electron laser facilities (XFEL) produce data at higher frequencies, automated, and integrated visualization tools such as CinemaBandit are becoming invaluable to catalog experimental data and perform scientific analysis on the fly. The limited time and high stress environment at a beam-time facility does not foster manual and in-depth examination of each experimental run. CinemaBandit is especially useful for such scenarios because high level, global generalizations can be inferred through the parallel coordinates view, but refinements and selections allow for more localized study if and when required.

CONCLUSION

Science is always a collaborative enterprise and this work brought together Cinema developers, a student in rhetoric and technical communication, and a sociologist. The diverse viewpoints in such a group gave us a new perspective on how Cinema enables scientific insight. Scientific visualization requires support for applying domain expertise, theory, and models to enable insight to be derived from the data and advance scientific understanding. CinemaScience is a flexible system of both *in situ* and postprocessing software components that easily and efficiently achieve this across a wide range of scientific domains and datasets.

The open source CinemaScience project is available at: github.com/cinemascience. Documentation and more information can be found at: www.cinemascience.org.

ACKNOWLEDGMENTS

The CinemaScience Project is led by D. H. Rogers. D. H. Rogers, D. Banesh, and T. L. Turton would like to acknowledge many Cinema collaborators and developers whose work is summarized in this article, in particular, J. Ahrens, A. Biswas, C. Biwer, S. Dutta, D. Orban, and C. Tauxe; they thank members of the LANL Data Science at Scale Team; their colleague F. Samsel. They would like to thank the many scientists whose challenges helped to drive the development of Cinema and would like to thank: J. Schoonover, C. A. Bolme, K. Ramos, S. Vogel, R. Sandberg, M. Petersen. T. Overmyer and B. H. Sims thank the many scientists whose discussions on scientific tasks and the role of visualization in driving scientific insight helped form their contributions. This work was supported by the NNSA and LANL LDRD 20170029DR. This work was released under LA-UR-19-29339.

REFERENCES

1. J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, and M. Petersen, "An image-based approach to extreme scale in situ visualization and analysis," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Piscataway, NJ, USA, 2014, pp. 424–434.
2. J. Woodring, J. P. Ahrens, J. Patchett, C. Tauxe, and D. H. Rogers, "High-dimensional scientific data exploration via cinema," in *Proc. IEEE Workshop Data Syst. Interactive Anal.*, 2017, pp. 1–5.
3. A. S. Almgren, J. B. Bell, M. J. Lijewski, Z. Lukić, and E. Van Andel, "Nyx: A massively parallel AMR code for computational cosmology," *Astrophys. J.*, vol. 765, no. 1 , Feb. 2013, Art. no. 39.
4. T. Ringler, M. Petersen, R. L. Higdon, D. Jacobsen, P. W. Jones, and M. Maltrud, "A multi-resolution approach to global ocean modeling," *Ocean Modelling*, vol. 69, pp. 211–232, 2013.
5. J. Ahrens, B. Geveci, and C. Law, "ParaView: An end-user tool for large data visualization," in *Visualization Handbook*. Amsterdam, The Netherlands: Elsevier, 2005
6. H. Childs *et al.*, "VisIt: An end-user tool for visualizing and analyzing very large data," in *High Performance Visualization–Enabling Extreme-Scale Scientific Insight*, New York, NY, USA: Taylor & Francis, 2012, pp. 357–372.
7. M. Larsen *et al.*, "The ALPINE in situ infrastructure: Ascending from the ashes of Strawman," in *Proc. In Situ Infrastructures Enabling Extreme-Scale Analysis Visualization*, 2017, pp. 42–46.
8. C. Johnson, "Top scientific visualization research problems," *IEEE Comput. Graph. Appl.*, vol. 24, no. 4, pp. 13–17, Jul./Aug. 2004.
9. D. Banesh, J. A. Schoonover, J. P. Ahrens, and B. Hamann, "Extracting, visualizing and tracking mesoscale ocean eddies in two-dimensional image sequences using contours and moments," in *Proc. Workshop Visualisation Environ. Sci.*, 2017, pp. 43–47.
10. S. C. Vogel *et al.*, "Interactive visualization of multi-dataset Rietveld analyses using *Cinema:Debye-Scherrer*," *J. Appl. Crystallograph.*, vol. 51, pp. 943–951, 2018.
11. D. Orban *et al.*, "Cinema:Bandit: A visualization application for beamline science demonstrated on XFEL shock physics experiments," *J. Synchrotron Radiation*, vol. 24, no. 1, 2020.

Terece L. Turton is currently a staff scientist in the Data Science at Scale team, Los Alamos National Laboratory. She received the Ph.D. degree in physics from the University of Michigan. She is a member of IEEE. Contact her at tlturton@lanl.gov.

Divya Banesh is currently working toward the Ph.D. degree at the Computer Science Department, University of California, Davis and a Graduate Student Researcher at Los Alamos National Laboratory. She is a member of IEEE. Contact her at dbanesh@lanl.gov.

Trinity Overmyer received the M.A. degree in rhetoric and composition from Purdue University. She is currently working toward the Ph.D. degree in rhetoric and technical communication, Purdue University. Contact her at trinity@purdue.edu.

Benjamin H. Sims is currently a staff scientist in the Statistical Sciences Group, Los Alamos National Laboratory. He received the Ph.D. degree in sociology and science studies from the University of California San Diego. Contact him at bsims@lanl.gov.

David H. Rogers is a staff scientist in the Data Science at Scale team at Los Alamos National Laboratory. He received the M.S. degree in computer science from the University of New Mexico and the M.F.A. degree in literature from Vermont College of Fine Arts. Contact him at dhr@lanl.gov.

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.