Enabling Discovery Through Visual Exploration

An Introduction to Data Visualization & Its Applications

By Harsh Bhatia

Visual metaphors have assisted human understanding since early days of mankind; the modern scientific and social-scientific evolution especially benefits greatly from the visual medium. With increasing size and complexity of contemporary data, visualization research has focused on developing techniques for novel mathematical and visual representations to assist data exploration for intended as well as fortuitous discovery. This article introduces the reader to the field of visualization, and discusses some of its important applications.

Keywords: *Visualization, Computer & Society*
Categories: *Human-centered computing —Visualization*
Corresponding Author: *Harsh Bhatia*
Email: *hbhatia@llnl.gov*

What is Visualization?

The human brain is capable of processing images and visual cues better and faster than reading textual information. It has a tendency to find and retain patterns in everything that the eyes see; *visualization* makes it easier for the brain to absorb information — *"a picture is worth a thousand words!"* Visualization is a way of telling stories and finding patterns in the data; it existed before computers, before technology, and even before language. The prehistoric humans would look up at the sky and connect the stars by invisible lines to imagine various constellations. Instead of the names of the stars, it was easier to remember which constellations they belonged to — *the dots make more sense when they are connected.* Among countless success stories of visualization, we discuss one.

John Snow's Cholera Map. In 1854, "the most terrible outbreak of cholera in the UK" had spread, and within a month, about 500 people had died. The scientific hypothesis of the day was that cholera spreads through air, leading the health administration to align their efforts correspondingly, until a local physician John Snow published a map (see Figure 1) clearly showing the cases of cholera centering around a certain water pump. Since the contemporary technology could not adequately test the pump's water for cholera, the patterns identified by Snow's visualization played a crucial role in convincing the administration to investigate the water pump, which enabled them to identify the source of contamination, establishing the true nature of cholera. This work is considered a major event in the history of public health and geography, and gave birth to a new field — epidemiology.

Visualization in the Modern Context. The goal of the contemporary visualization research is to leverage existing scientific methods by providing new insights through

Figure 1: John Snow's 1854 cholera map spatially visualized the distribution of cholera cases in London around a particular pump (zoomed in), affecting the administration to act correspondingly. Image source: Wikipedia.

visual medium. Recent advances in computing, sciences, and social sciences produce more-sophisticated data, which requires more-advanced visualization techniques in order to explore natural and man-made phenomena, and derive new knowledge. In the last 30 years, visualization has become increasingly important by demonstrating its utility in enabling scientists to explore their increasingly complex experiments and simulations in more detail, allowing them to ask questions of much greater import. A National Science Foundation Panel initiative in 1985[1] provides a formal definition: *"Visualization is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights."*

Visualization researchers work at the intersection of many disciplines: computer science and computer engineering, computer graphics, mathematics and statistics, cognition and perception, etc., as well as a wide variety of application domains. Although the goal of pure visualization research is to find novel visual representations of complex data to make it easy to understand, visualization, often, goes in hand with data analysis. In particular, visualization plays an important role in applying different mathematical techniques to analyze the data. For instance, *topological visualization* focuses on applying rigorous topological techniques to understand the shape of the data enabling concise mathematical and visual representations, and define physical and/or intuitive features in mathematical terms.

The goal of this article is to give an introduction and overview of visualization in the context of some modern scientific and non-scientific application areas.

[1]"Visualization in Scientific Computing". In: *SIGGRAPH Computer Graphics* 21.6 (1987). Ed. by B. H. McCormick, T. A. DeFanti, and M. D. Brown. ISSN: 0097-8930.

Scientific vs. Information Visualization

The visualization community broadly categorizes itself into two: scientific vs. information visualization. While there exist cases at the intersection of the two categorizes, the classification is primarily based upon the application areas and types of data involved. As the name suggests, *scientific visualization* typically addresses the use-cases in scientific applications: climate modeling, combustion, aerodynamical designs, medical imaging, molecular dynamics, etc., the end users of which are also mostly scientists and experts. On the other hand, *information visualization* is a more general category that encompasses all sorts of non-scientific data, e.g., social scientific data such as social networks, epidemic outbreaks, sports, etc., the audience of which may include specialists as well as general public.

The type of data usually targeted by scientific visualization is *spatiotemporal* — a quantity whose variation is studied over space and time, and has up to 3 spatial and 1 time dimensions. Examples include temperature or atmospheric pressure variation at the surface of the earth, i.e., *scalar fields*, velocity of wind flows around an automobile and flow of oceanic currents, i.e., *vector fields*, and the distribution of stress and shear in an elastic material, i.e., *tensor fields*. On the other hand, information visualization addresses data that does not fit the above description, e.g., graph-type data such as social networks, high-dimensional tabular data such as housing market, which depends upon a number of variables, sports or music data targeted at identifying patterns to improve a game or a composition, etc.

Together, the two branches of visualization aim to assist both specialists and masses to explore vast amounts of data and highlight hidden patterns and behaviors in the data. Nevertheless, with its increasing impact on making informed scientific and/or policy decisions, the concerns of uncertainty or errors in visualization have become more palpable.

Uncertainty Visualization

Imagine a surgeon using modern medical visualization tools to analyze the CT (computed tomography) scans of a patient to detect cancerous cells, and decide how and where to cut out the tumor. The CT scan captures different types of tissues as objects of different densities, and the visualization system typically computes the "decision" surfaces between cancerous and healthy cells using the difference in their densities. Needless to say that the stakes are high, and the surgeon needs to be absolutely sure that the visualization can be trusted, and can be used to plan the surgery. Even the slightest errors due to numerical precision, or uncertainties in visualization and computational algorithms can cause the surface to enclose a healthy tissue, or leave a cancerous tissue out. As a result, the ideal visualization would acknowledge this uncertainty, and show the cancerous cells it is absolutely certain about, along with the regions it may have compiled an error for — *honest and agnostic visualization is better than an oblivious one.* The role of a visualization scientist is not to decide what a cancerous cell is, but to assist the domain scientist, the surgeon in this case, make the correct decision. Uncertainty visualization provides the confidence required to make the right decisions, by displaying the data honestly and as

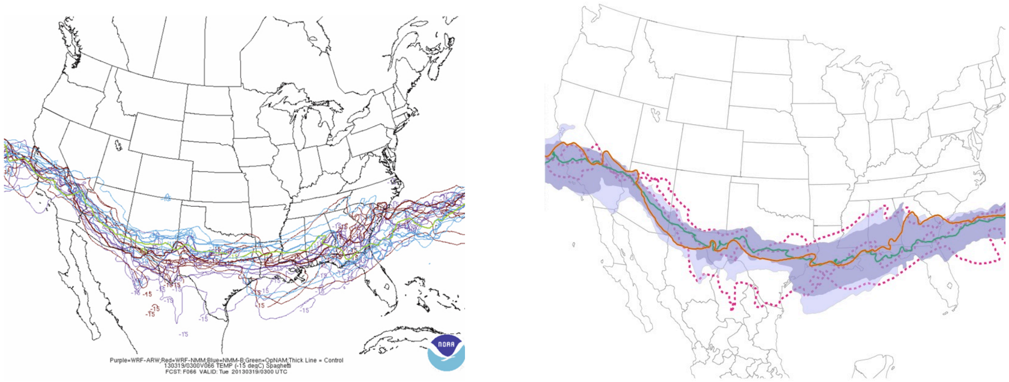

Figure 2: Contour Boxplot (right) provides an easy-to-understand visualization of an ensemble of simulation runs of a weather model, showing median, order statistics, and three outliers. Compare it with standard visualization of multiple contours (left) which obscures the statistical behavior. Image source: Wikipedia.

accurately as possible.

Uncertainty visualization is also useful to study *ensembles* — collections of measurements or simulations run with different parameters, where statistical behavior needs to be studied. For example, Figure 2 shows an ensemble data created through different runs of a simulation model with different perturbations of the initial conditions to account for the errors in the initial conditions and/or model parameterizations. The ensemble consists of *isocontours* of the temperature field (isovalue -15 °C) at 500mb in altitude. As illustrated in the figure, the conventional visualization obscures the information, whereas Contour Boxplots,[2] an uncertainty visualization technique, displays the statistical variation cleanly making it easier to understand the data and its implications.

Applications

In this section, we discuss a representative set of applications where visualization has been integral in gaining new insights; we focus on application areas for both scientific and information visualization. Almost all of the research referenced in this article is chosen from recent years, to give the reader an idea of the current advancements of the field. This choice is a very small subset of visualization research; it is neither extensive nor merit-based representation of the field.

Visualization for Climate

Climate monitoring and modeling is an application area that uses visualization and analysis extensively. Visualizations for climate data are as often created for public as they are for

[2] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. "Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2713–2722. ISSN: 1077-2626. DOI: `10.1109/TVCG.2013.143`.

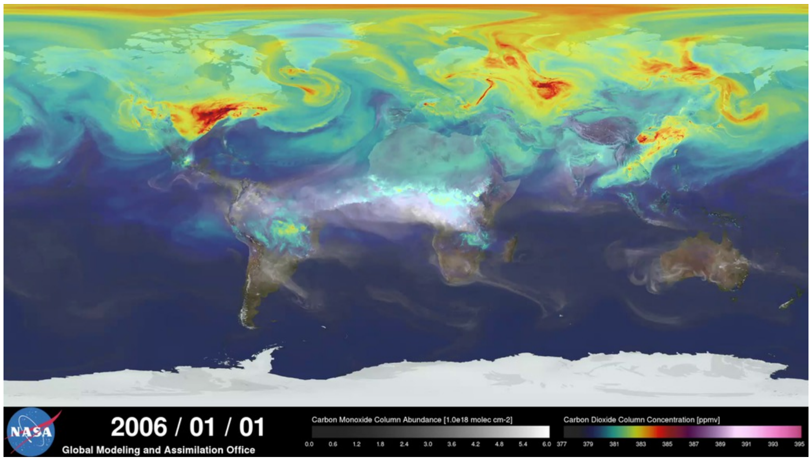

Figure 3: Visualization of carbon emission levels. Image source: NASA.

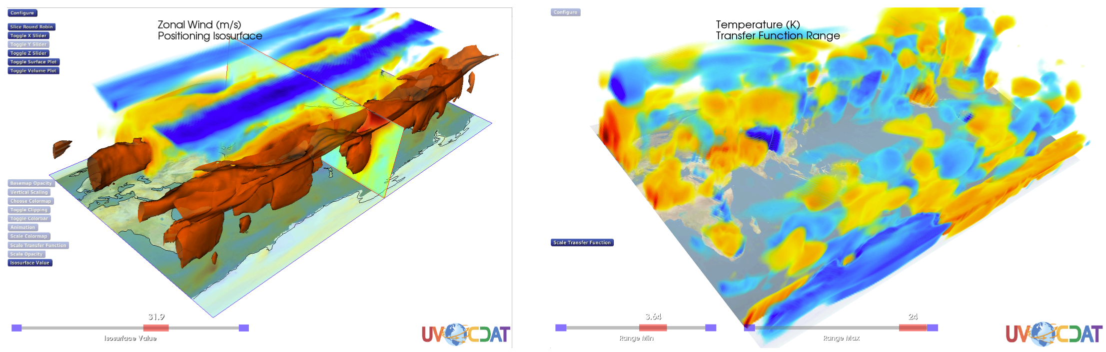

Figure 4: Visualization created using UV-CDAT to explore climate data. Left: Superposition of volume rendering, slicing, and isosurface visualization of zonal wind. Right: Understanding temperature anamolies through volume rendering. Image source: UV-CDAT.

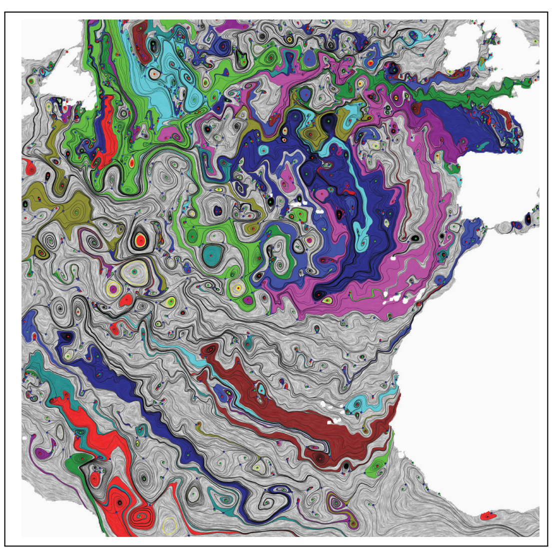

Figure 5: Visualization of patterns in oceanic currents in North Atlantic Ocean extracted using robust computational techniques. Image source: Harsh Bhatia.

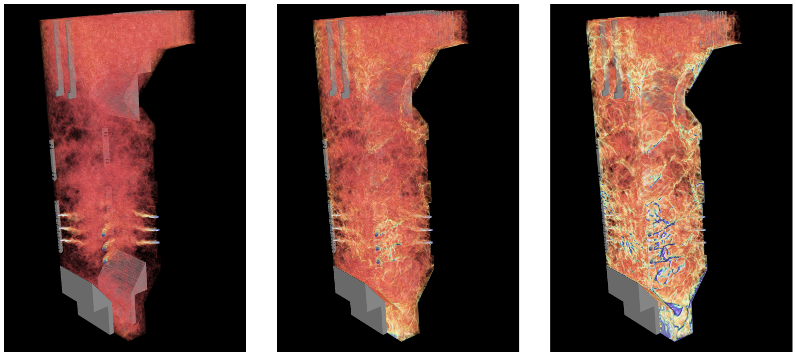

Figure 6: Visualization of dispersion pattern of carbon particles in coal-boilers allow scientists to understand and improve combustion and performance of such engines. Image source: Uintah: `http://uintah.utah.edu/gallery.html`

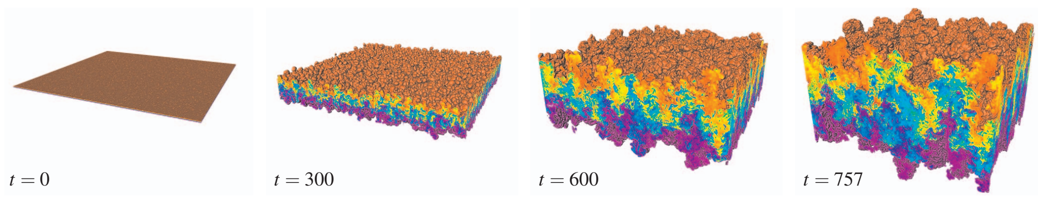

Figure 7: Visualization and topological of Rayleigh-Taylor instability (turbulent mixing of two fluids of different densities) allows explaining the evolution of "bubble"-like structures in the data. Image source: IEEE TVCG.

climate scientists. For example, NASA[3] publishes visualizations about climate change (see Figure 3), weather predictions, geological and geographical data, terrestrial and planetary information, etc., to convey the information to general public. On the other hand, climate scientists utilize advance visualization and analysis techniques to understand and improve their models and interpret the data. UV-CDAT[4], for instance, is a set of visualization tools for exploration and analysis of large-scale climate data (see Figure 4).

Visualization research challenges in such cases not only include addressing large-scale data, but how to efficiently (often interactively) sift through the data to make scientific queries, and how to develop visual representations to best identify and represent multiple parameters affecting the system. Scientists are typically interested in understanding how climatic factors, such as oceanic currents, wind flow, rainfall, etc., behave in certain regions. Computationally robust techniques to identify features of interest form an important research direction for visualization. In this context, Figure 5 shows a visualization of oceanic flow with different isolated regions of currents in different colors. Robust identification of such complex structures hinge upon delicate inaccuracies in data, numerics, and visualization; therefore, techniques that provide explicit control over how data is represented and processed are highly desirable.[5]

[3]NASA Scientific Visualization Studio: `https://svs.gsfc.nasa.gov/`

[4]Ultrascale Visualization – Climate Data Analysis Tools: `http://uv-cdat.llnl.gov/index.html`

[5]Harsh Bhatia. "Consistent Feature Extraction From Vector Fields: Combinatorial Representations and Analysis Under Local Reference Frames." PhD thesis. The University of Utah, 2015.

Visualization for Combustion and Mixing

Simulation of complex chemical and physical reactions has been an integral component of scientific advancement in the last half century. With increasing need for cleaner energy, a huge amount of effort is directed towards developing low-emission combustion engines; scientists regularly perform large-scale simulations of mixing and combustion of turbulent fluids, fire, explosions, etc., and require innovative solutions to not only visualize but also analyze the resulting data. As an example, Figure 6 visualizes simulations of coal-fired boilers with different sizes of carbon particles. Understanding the difference between them — smaller particles disperse evenly whereas large particles tend to form clusters — allows improving the performance of the boiler. For applications ranging from astrophysics to nuclear fusion, Laney et al.[6] developed a topological analysis and visualization techniques for exploring the turbulent mixing of two fluids. Using visualizations such as Figure 7, they were able to characterize the evolution of "bubble"-like structures in the process, which are a defining property of such turbulent mixing.

Visualization for Molecular Dynamics

Molecular dynamics (MD) is another important application area benefited by visualization. In general, MD simulations create large-scale data containing complex behavior at microscopic scale where the goal is to identify how atoms interact with each other. Examples include understanding protein folding and unfolding, diffusion and solvation of ions in certain electrolytes, effects of impurities on crystalline structures, etc. Topological analysis and visualization has proven useful in this area as well. Gyulassy et al.[7] developed a technique to identify interstitial and interlayer sites through which lithium ions can diffuse in batteries with graphitic electrolytes. In particular, by applying topological analysis, they were able to identify the regions in the battery that allow the lithium ion to travel through, directly allowing the scientists to understand the diffusion behavior, and hence, the performance of the battery (see Figure 8). Günther et al.[8] devised a new way of characterizing molecular interactions by using a composite of two quantities derived from the electron density. By applying topological analysis on these fields, they were able to identify different types of bonds in the system, and presented a visualization system for the same.

In some cases, tailored visualization systems can offer interactive exploration of statistics, and therefore, reduce the effort required in sifting through large amounts of data. For example, the interactive statistical exploration presented by Bhatia et al.[9] allows

[6] D. Laney et al. "Understanding the Structure of the Turbulent Mixing Layer in Hydrodynamic Instabilities". In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006), pp. 1053–1060. ISSN: 1077-2626. DOI: 10.1109/TVCG.2006.186.

[7] A. Gyulassy et al. "Interstitial and Interlayer Ion Diffusion Geometry Extraction in Graphitic Nanosphere Battery Materials". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 916–925. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467432.

[8] D. Günther et al. "Characterizing Molecular Interactions in Chemical Systems". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2476–2485. ISSN: 1077-2626. DOI: 10.1109/TVCG.2014.2346403.

[9] H. Bhatia et al. "Interactive exploration of atomic trajectories through relative-angle distribution and

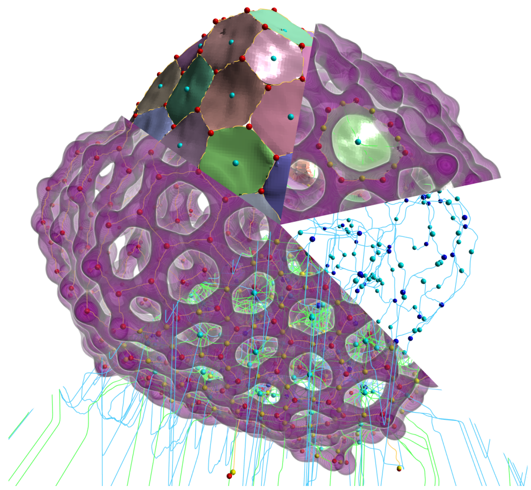

Figure 8: Visualization of carbon nanosphere including volume rendering, surface visualization, and topological components to understand the defects in the material and where lithium ions can diffuse to. Image source: Attila Gyulassy.

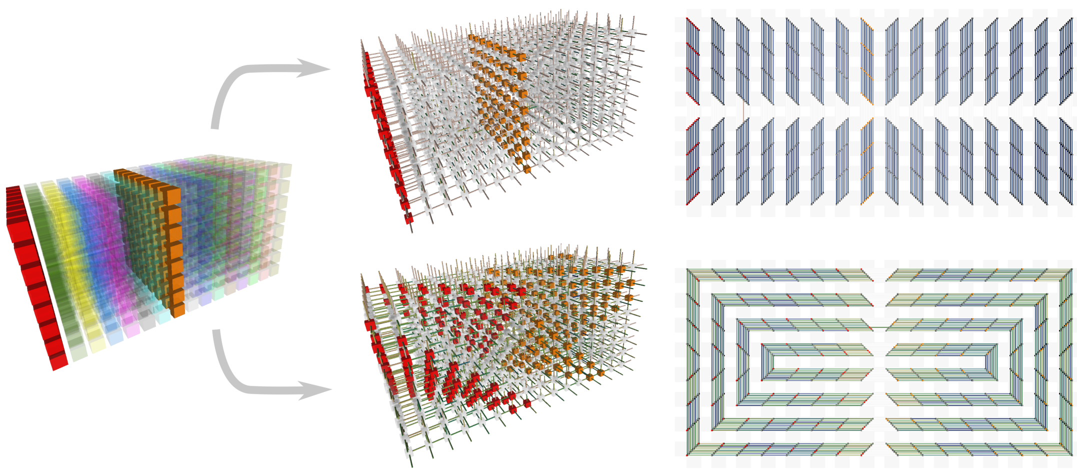

Figure 9: Visualization of 3-dimensional torus networks allow interacting with complex topology and understand the communication behavior. Image source: IEEE TVCG

visualizing 2-dimensional statistics of atomic trajectories, and detect certain time-scales where trajectories show characteristic behaviors. As compared to generic visualization tools, the custom design allows visualizing uncertainty and focusing on desired properties.

Visualization for High-Performance Computing

Understanding the communication and utilization patterns for HPC resources is an important contemporary challenge. In order to make the most efficient use of the

associated uncertainties". In: *2016 IEEE Pacific Visualization Symposium (PacificVis)*. 2016, pp. 120–127. DOI: 10.1109/PACIFICVIS.2016.7465259.

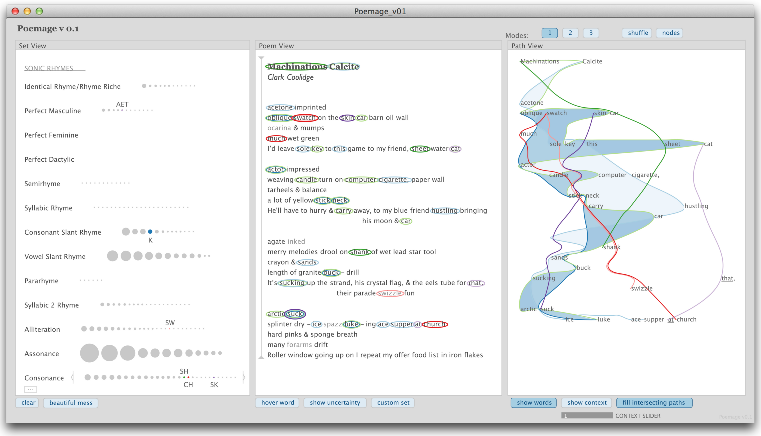

Figure 10: Poemage is a visualization system for exploring the sonic topology of a poem. Image source: Nina McCurdy, `http://www.sci.utah.edu/~nmccurdy/Poemage/`

investments in supercomputers, it is important to optimize their usage for large-scale applications. A lot of research has been done in this direction; the primary challenges are high-dimensionality of the data, e.g., complex network topology, and large amounts of data, e.g., trace for thousands to millions of processors for tens of thousands of time-steps, with an important constraint that the visualizations must maintain spatial, temporal, as well as application-specific contexts.

To this end, Landge et al.[10] developed a novel visualization scheme for network layouts with 3-dimensional torus topology (see Figure 9), extended by McCarthy et al.[11] for 5-dimensional torus topology. Such visualizations have been useful in identifying important communication bottlenecks for large-scale simulations over distributed computing resources, and can help scientists redesign their simulations to make more-effective use of the supercomputers. Isaacs et al.[12] developed a visually intuitive scheme for displaying traces of large-scale parallel codes using *logical time*. By considering the ordering of message-passing (instead of actual time), they provide a simpler overview and allows identifying bottlenecks in the code, ultimately enabling the scientists to improve the simulation codes and/or improve job scheduling and node mapping on supercomputers.

[10] A. G. Landge et al. "Visualizing Network Traffic to Understand the Performance of Massively Parallel Simulations". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2467–2476. ISSN: 1077-2626. DOI: `10.1109/TVCG.2012.286`.

[11] C. M. McCarthy et al. "Visualizing the Five-dimensional Torus Network of the IBM Blue Gene/Q". in: *First Workshop on Visual Performance Analysis (VPA), 2014*. 2014, pp. 24–27. DOI: `10.1109/VPA.2014.10`.

[12] K. E. Isaacs et al. "Combing the Communication Hairball: Visualizing Parallel Execution Traces using Logical Time". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2349–2358. ISSN: 1077-2626. DOI: `10.1109/TVCG.2014.2346456`.

Visualization for Music, Poetry, and Sports

McCurdy et al.[13] developed a visualization system (see Figure 10) for poetic compositions by analyzing sound and sonic devices — the literary devices involving sound that are used to convey meaning or to influence the experience of the listener or the reader, such as rhyme, rhythm, meter, etc. By formalizing the linguistic semantics of the words, they consider the *topology of a poem*, which reduces the complex structures formed by the interactions of words into simpler visual representations. When fully mature, this type of visualization research could potentially quantify the quality of a poem in an objective manner by identifying some defining constructs that are appealing to the human mind and make a certain composition delightful.

Polk et al.[14] designed a visualization system to analyze tennis matches to assist non-professional tennis coaches and players. Utilizing easily collectible data such as scores and videos, and combining a variety of simple visualizations based upon real-life metaphors, the system provides insights into match performance by giving a holistic visualization of the match progress as conveyed by the data. The authors plan to work towards extending the system to include multiple matches, which could make it a powerful tool for non-professional players to assess and improve their performance.

Discussion

Visualization is, in general, a multidisciplinary field, aiming to assist in deriving insights from data. It is an umbrella term that encompasses data and applications of a wide variety. This article introduces the reader to the field of visualization research — both scientific and information. It describes some key aspects of visualization, and discusses few important application areas benefited by visualization.

Nevertheless, a small overview article cannot do justice to the entire field. We invite the interested reader to consider attending/following important visualization conferences: IEEE Visualization (`http://www.ieeevis.org/`), European Conference on Visualization (`http://www.cs.rug.nl/jbi/eurovis2016/`), IEEE Pacific Visualization Symposium (`http://www.pvis.org/`), Topology-based Methods in Visualization (`http://vis.uni-kl.de/topoinvis/`), IEEE Symposium on Large Data Analysis and Visualization (`http://www.ldav.org/`), etc., and journals: IEEE Transactions on Visualization and Computer Graphics, Computer Graphics Forum, Information Visualization, etc. Readers may also find useful publicly-available visualization tools and softwares for a variety of scientific and non-scientific data: Paraview (`http://www.paraview.org/`) and VisIt (`https://visit.llnl.gov/`).

[13] N. McCurdy et al. "Poemage: Visualizing the Sonic Topology of a Poem". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 439–448. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467811.

[14] T. Polk et al. "TenniVis: Visualization for Tennis Match Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2339–2348. ISSN: 1077-2626. DOI: 10.1109/TVCG.2014.2346445.

Visualization research has made substantial progress in the past two decades. However, going forward, there exist important challenges that must be met. With increasing size of data, the conventional visualization approaches are often reaching their bottlenecks, and novel techniques and frameworks are required to address large- and extreme-scale visualization. Similarly, as mathematical models and simulations become highly-sophisticated, the resulting data is becoming more complex, and new ideas are required to address the additional complexity. Finally, uncertainty visualization, which has been identified as one of the top research problems in visualization, requires much work to encompass all kinds of data and visualization techniques.

About the author. Harsh Bhatia is a postdoctoral researcher at the Lawrence Livermore National Laboratory, CA, USA. His research interests include scientific visualization, data analysis, scientific computing, computer graphics, mathematical modeling, high-performance computing, computational topology and more, but he has been actively engaged primarily in visualization research.

Acknowledgements. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-MI-704740.