Visualization of Structured Data

Thomas A Caswell (Brookhaven National Laboratory) tcaswell@bnl.gov

Challenge

The data collected at DOE facilities, both experimental and simulated, is becoming more complex, higher-dimensional, and multimodal. There is frequently no single obviously correct visualization and several views maybe required to fully understand the data and extract scientific value. For example, at NSLS-II we routinely conduct mapping scans where at each point of an x-y scan over the sample, diffraction data is collected on a 2D imaging detector along with energy-resolved florescence measurements with a multi-element detector. Thus, the total data set is at least 6D: x and y position on the sample, the pixel on the 2D detector, the element of the fluorescence detector and the photon energy. Additional parameters, such as the incoming photon energy or in-situ conditions may also be scanned, further increasing the dimensionality. Current visualization tools are either naive to this structure, requiring significant customization and development to use, or make deep assumptions about the expected structure and cannot be easily moved between applications.

Scientific data has an inherent connectivity structure, but current visualization tools do not fully take advantage of this property. There is a qualitative difference between a time series which, even if sampled discretely, represents a fundamentally continuous phenomenon, and observations about individuals in a population, which are fundamentally discrete. Similarly, the values measured can have a range of mathematical structures from sets to fields. Finally, we need to know how components of the data relate to each other, for example an image has x and y coordinates and an intensity for each pixel. Considering the continuity of the data allows us to differentiate among, for example, volumetric data, a movie, a multi-spectral image, a stack of unrelated images, and a vector field. Despite each of them being representable as a 3D array in memory, they have different topologies and interpretations. During visualization we need to ensure that we preserve the continuity the data and that quantitative changes in the data are yield equivalent changes in the visualization, a property known as equivariance.

Each visualization tool implicitly expresses and exploits only a subset of the full underlying structure in its API and internal data structures. This restriction may be driven by specialization to a particular domain, for example image stacks in ImageJ and Napari, by performance, such as the data structure families in the VTK data model, or to provide a tractable user API, such as Matplotlib. Because each of these tools simplifies the underlying structure differently it is hard for scientists to seamlessly move between visualization tools. Additionally, it is difficult to share tools between domains that have fundamentally equivalent data, such as multi-channel optical florescence microscopy and 2D x-ray florescence mapping, due to domain specific assumptions in most visualization tools.

Opportunity

By rigorously defining the structure of the data and the transformation from data to final visualization we can develop a consistent language, in the spirit of functional programing, to express how the continuity and equivariance are preserved in the visualization process. Others have noted the importance of both continuity [1] and equivariance [2], however accurately enforcing both has been left to the implementation of each library in an ad-hoc way. A consistent conceptual framework will make it easier for scientists and developers to ensure that both properties are preserved in all visualizations.

This is an opportunity to generalize how we think about scientific visualization. Historically it has been centered either on the computer representation (e.g., array or mesh) or on the final domain applications (e.g., multi-spectral imaging or statistical plotting). By building the underlying libraries around the connectivity and logical structure of multimodal and high-dimensional data we can make it easier to write powerful domain specific applications.

Timeliness or maturity

Within the DOE complex we have increasing amounts of complex multimodal and high-dimensional data. We need to provide our scientists and users with the tools they need to efficiently extract scientific value from these data sets.

Just as strided arrays have been the basis of scientific computing over the past 30 years, I expected structured data to be the basis for scientific computing in the next 30 years. Many groups are currently working to specify a standard container for structured data. Active projects include arrow, pandas, xarray, and Awkward Array, all of which can be accessed from Python. Additionally, there is on-going work at NSLS-II to provide access to experimental data [3] in standard data containers. The time is ripe to develop structure-aware visualization libraries.

On-going work by Michael Grossberg at CUNY, in collaboration with Caswell, is formalizing the mathematical description of the structure of scientific data and the transformation from data to visualization. This research is the basis for a planned revision of Matplotlib over the next three years.

References

1. D.M. Butler and M.H. Pendly. Visualization model based on the mathematics of fiber bundles. *Computers in Physics*, 3(5):45, 1989. Doi:10.1063/1.168345
2. W.A. Lea (1969). A formalization of measurement scale form. Electronics Research Center. Cambridge, Mass.
3. https://www.bnl.gov/newsroom/news.php?a=119260
4. https://github.com/matplotlib/2020-NASA-ROSES-E7/blob/main/proposal/matplotlib.pdf