# Interactive Visualization Solutions for Large-scale Complex Image Data

Kyle Harrington[1], Jacob Hinkle[1], Tiago Ferreira[2], Matthew Wolf[1], Wei Xu[3], Berk Geveci[4], and Yunha Lee[5]
[1]Oak Ridge National Laboratory, {harringtonki, hinklejd, wolfmd}@ornl.gov
[2]Howard Hughes Medical Institute, Janelia Research Campus, ferreirat@janelia.hhmi.org
[3]Brookhaven National Laboratory, xuw@bnl.gov
[4]Kitware, berk.geveci@kitware.com
[5]Center for Advanced Systems Understanding, HZDR, y.lee@hzdr.de

**Challenge:** Large-scale image data requires significant computational power and data storage, yet also requires interactive data visualization, curation, and processing. The need for HPC resources makes this line of research well suited for DOE HPC facilities; however, methods for interactive visualization must be extended to enable users to interact with the increased volume of image data. While increased requirements for data processing, algorithm evaluation, and machine learning model training motivate large projects to use centralized HPC resources, the need to use interactive visualization to inspect and analyze data introduces a motivation to use data locally for reduced latency and data transfer. There are challenges that arise in common approaches for addressing this tradeoff where each approach has limitations that can impact research quality and efficiency. First, algorithm developers often work with small subsets of data locally, but other data subsets may be important for inspection and testing. Second, browsing entire datasets is often achieved by incrementally streaming data of increasing resolution for the region of data currently being visualized, but the user may miss important details by interacting with the visualization before they have seen the data at sufficient resolution. Third, when it is critical to interactively work with the entire dataset it is common to simply work on-site with a high-speed connection directly to the compute resources, but this clearly limits the user base. Although it would be ideal for algorithm developers, data curators, and other researchers to be able to work with and inspect entire datasets at the real-time speeds required for interactive visualization, compromises are required in practice.

Some of the most challenging examples of large-scale imaging problems are emerging in neuroscience, biology, and geophysics, where datasets can include large image volumes, 3D geometries, data annotations, and data with spatiotemporal overlaps and discontinuities. In biology, projects like OpenOrganelle at Janelia Research Campus freely share large amounts of image data on cloud resources; however, with most internet connections it is not possible to explore the high-resolution raw image data at interactive speeds due to the time it takes for the images to load. This makes it difficult to inspect the images when searching for data to use for developing algorithms or testing new hypotheses. In geophysics, digital twins of earth systems are popular tools that integrate large amounts of observation and simulation data in the 3D domain. Digital twin research depends heavily on interactive 3D visualization to improve public engagement and provide decision support for policy makers. In neuroscience, there are multiple brain mapping efforts that have a significant component dedicated to interactive labeling of anatomical structures. Interactive neuron tracing and data labeling can be enhanced by software tools that use intelligent algorithms and graphical structures [1]. While improving interactive visualizations of large-scale imaging data can have a broad impact across fields, a critical challenge is knowing how much impact a particular improvement has.

**Opportunity:** This paper is focused on the challenges associated with visualizing large imaging datasets efficiently and supporting user interaction for these datasets. To address these challenges, we propose to investigate the following directions: **(1)** use machine learning models to guide visualization at the edge, **(2)** leverage information in images to restructure data for improved data streaming, and **(3)** enhance interactive machine learning efficient sparse representations. Finally, we suggest high-level ways to quantitatively measure impact in these directions.

**(1)** Sometimes it is clear that it is not necessary to visualize or access an entire dataset to evaluate the performance of an algorithm or model. In these cases, it can be possible to perform the computation and visualization using an edge computing paradigm. We propose to investigate solutions that use machine learning-based guides and supplemental data from the relevant datasets to help select which regions should be transferred to edge visualization nodes. Consider the case of refining an image segmentation algorithm, it is often important to focus on evaluating performance in regions of high uncertainty as opposed to evaluating performance uniformly across the entire image. By applying Bayesian methods to estimate uncertainty in an algorithm's predicted segmentation it becomes possible to discover regions of high uncertainty. Regions with high uncertainty, or

regions with domain-specific relevance, can be processed and visualized on user-facing machines, while the bulk of the data is processed on central HPC resources.

**(2)** The majority of approaches that support streaming large volumes of data utilize pyramids defined over regular grids of decreasing resolution. However, recent work in level sets and particle methods have led to the introduction of data structures that are defined over irregular structures which take advantage of features and information contained in the underlying image data. These methods can be extended to account for supplemental information, such as segmentation labelings, polygons, and meshes, which can be used to inform tesselations that structure data favorably for task relevant visualizations. We propose to investigate solutions that use context-aware particle methods to accelerate the streaming of multiresolution image data to users.

**(3)** Interactive machine learning is becoming increasingly important when new datasets are obtained in the absence of ground truth. Interactive machine learning requires users and curators to incrementally annotate data and validate predictions. This is often achieved with software tools that allow the users to paint labels or features onto images in a pixel-wise fashion. Pixel-wise tools can become cumbersome in situations where it takes significant amounts of time to stream the data at high resolution. On the other hand, graphical objects, like polygons, lines, splines, points, and meshes, are efficient to transmit. We propose to provide enhanced support for interactive tools to create sparse graphical annotations for image data.

Furthermore, we suggest that it is critical to develop strategies for evaluating the impact of interactive visualization solutions. **(1)** The adoption of machine learning-guided edge visualization in HPC projects should be measured across the user base. **(2)** The performance of streaming particle methods should be quantified in terms of latency and bandwidth usage for both biological and geophysical domains. **(3)** The efficiency of sparse graphical annotations should be compared to pixel-wise labeling across multiple image data domains and projects. By quantitatively assessing interactive visualization solutions we can ensure that our efforts are driven by impact.

**Timeliness:** Image-based atlas projects, like Janelia Research Campus' MouseLight project, have produced large amounts of data to share, including ground truth annotations. These datasets represent open problems for machine learning researchers to address, especially with respect to automated image segmentation. The MouseLight project is preparing to make additional imaging data and annotations publicly available. This presents an ideal opportunity for using machine learning-guided visualization at the edge to support the development of new algorithms for discovering neural circuits. A key impact of this approach is that it will expand the pool of computational scientists that can contribute to developing algorithms for large-scale image data.

Digital twin efforts are expanding from earth systems to other parts of science, including biology. It is an ideal time to leverage the advanced state of digital twin research in geophysics to discover the relevant practical limitations of interactive visualization. Geophysical digital twins are composed of many layers of information, from satellite imagery to 3D buildings, that make them an ideal candidate for evaluating the impact of context-aware particle methods on data streaming. Algorithms and tools that successfully improve streaming for interactive visualizations will enable users to work with significantly larger and higher resolution data.

The number of large imaging datasets that are available is continuing to grow as institutes, such as Janelia Research Campus, the Allen Institutes, and the Chan Zuckerberg Initiative Biohub, undertake major mapping projects. However, it is often the case that these datasets are not fully analyzed due to project scope and goals. Interactive machine learning is making it easier for researchers quickly begin working with datasets that lack relevant ground truth data. By improving support for interacting with data using sparse graphical annotations, the power of researchers to develop new studies based on large-scale imaging efforts will be expanded.

The ORNL Frontier HPC is becoming established and is well-suited for large-scale image analysis projects. While compute power has increased, the human capacity to explore data has not increased at the same rate. Therefore, there is a great need to assist users in visualizing data and supporting processing on edge nodes, accelerate streaming visualizations to support exploration of large datasets, and improve capabilities for interacting with data using bandwidth efficient sparse representations. Narrowing the gap between centralized HPC resources and the respective users is a major challenge for large-scale image data. A key step in this direction is facilitating interactive visualization and making it possible for a broader audience to develop new research questions with these valuable resources.

[1] Arshadi, C., Günther, U., Eddison, M., Harrington, K. I., Ferreira, T. A. SNT: a unifying toolbox for quantification of neuronal anatomy. Nature Methods, 18(4), 2021. 374-377.