Processing and Analyzing Terascale Conjugate Array Tomography Data

Alexander Baden[1], Eric Perlman, Forrest Collman[2], Stephen Smith[2], Joshua T. Vogelstein[1], Randal Burns[1]

[1] Johns Hopkins University, Baltimore, MD USA

[2] Allen Institute for Brain Science, Seattle, WA USA

Challenge

- Routine collection of millions of micrographs.
- Multi-stage software reconstruction pipelines require detailed record keeping.
- Tera- and peta-voxel data volumes required specialized data management.

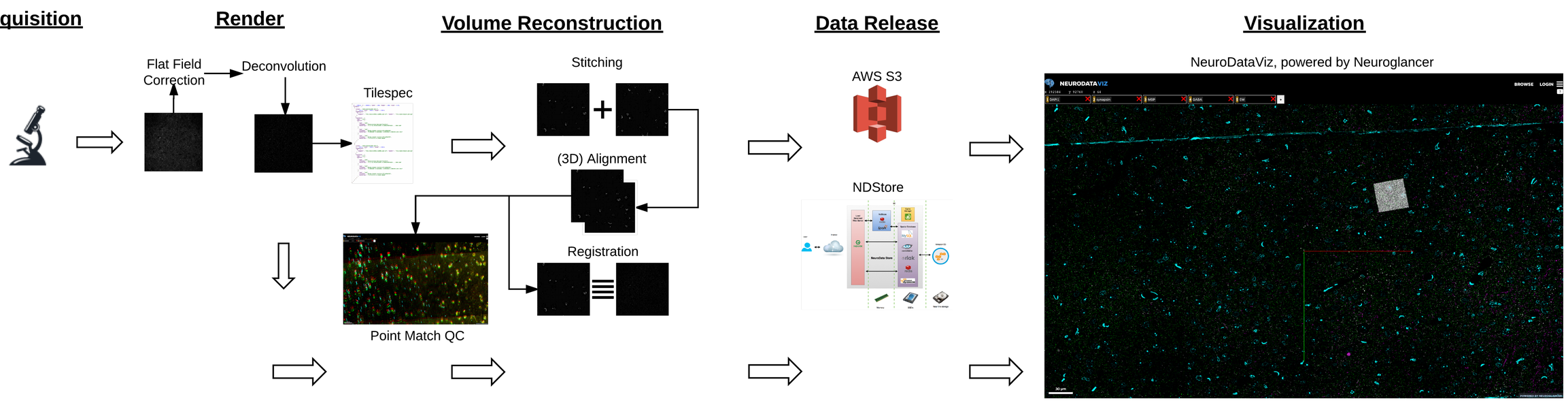
Action

- Designed a pipeline for processing and releasing AT data.
- Provide visualization capabilities at key stages of the pipeline.
- Leveraging the cloud for scale out.

Resolution

- Open source. Anyone can benefit from our work or join in the development process.
- Resource sharing. Software developed for AT is relevant for EM and vice versa.
- Community focus. Collaboration benefits everyone.

Pipeline Overview



Neuroglancer Integration into NDViz

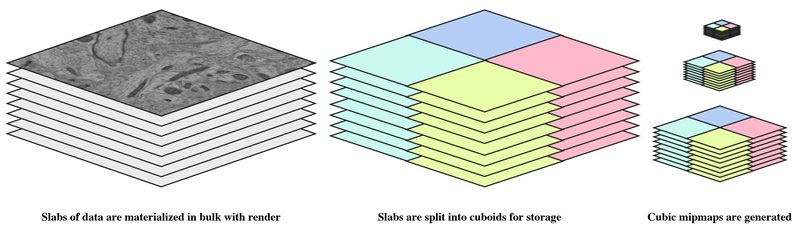
Imaging data is inherently visual. As such, there is a strong desire to convey information with rich, dynamic visualization tools. The neuroscience community has produced several state-of-the-art tools (TrakEM2, Reconstruct, Vaa3D) that enable interactive data exploration and analysis. As data sizes increase, key features of these tools have been migrated to the Web. In addition to the basic graphics rendering required of a Desktop application, a Web application must handle the process of retrieving and caching remote data.

The open source Web viewer Neuroglancer [4] handles the data retrieval and data rendering for 3D microscopy data and volumetric annotations. By building on top of Neuroglancer, we can immediately begin to develop interesting data visualizations specific to the problems we want to solve without having to re-engineer the fundamentals (see Point Match Visualization). Additionally, since Neuroglancer is open source, we can contribute features of interest back to the neuroscience community and integrate features contributed by the community as well.

Render Integration

A single array tomography dataset can encompass millions of data tiles corresponding to different (x, y, z) locations and different stains (imaging channels), resulting in a large volume of metadata. During 3D Volume Reconstruction, we wish to apply filters and transforms to specific image tiles, further complicating matters. Render [1] provides a framework for storing tile metadata and transformations in *tilespecs* and materializing transformed tiles using a rich suite of web services. These services enable us to dynamically interact with the data, enabling rapid visualization to assess both imaging and alignment quality.

NDStore Materialization

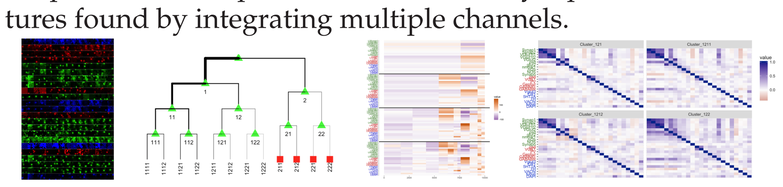


Fast access to data is essential for sharing among collaborators and with the public at large. We materialize image stacks from render into 3-dimensional cuboids, the basic building block of the NeuroData storage infrastructure. In this format we can then support fast, multi-planar access to data for random-access visualization or computation.

We render the data at a resolution as close as possible to the input images. For the conjugate EM and array tomography, EM is rendered at a voxel size of $3 \times 3 \times 50$ *nm* and the AT channels at $96 \times 96 \times 50$ *nm*, both being stored in the same coordinate space with different base resolutions. Each type of data are stored as a unique channel in the same project.

NDStore Analysis

Once materialized, algorithmic exploration of the data becomes simple . One example is classification of synapses based on features found by integrating multiple channels.



(a) (b) (c) (d)

(a) Synaptogram: Visualization of the identified synapses across all imaged channels.

(b) Dendrogram: Shows the Hierarchical GMM (Gaussian mixture model) structure, branch size denotes size of cluster.

(c) Stacked Means Plot: The block structure corresponds with the dendrogram. Rows show the feature means within the cluster and columns represent the size of each cluster.

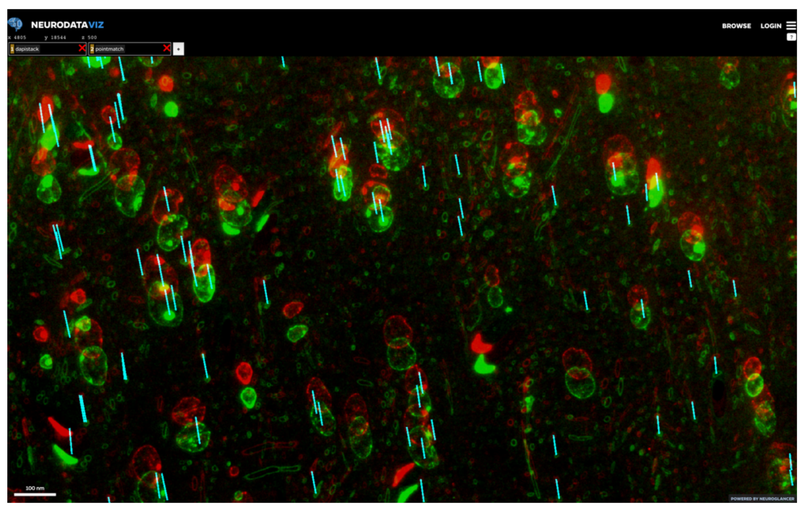
(d) Correlation Matrices: A selected few correlation matrices to highlight clusters 1211 and 1222. Both clusters have highest average excitatory expression and show an interesting anti-correlation in nr2b and nmdar.

To The Cloud

We have a prototype of the software stack running in the cloud (on Amazon Web Services). This deployment consists of:

- TIFF images are stored directly in Amazon S3
- Render and MongoDB both running in Docker containers
- ndstore (with S3 for cuboid storage, redis for caching, and DynamoDB for key storage)

Point Match Visualization



NeuroDataViz displaying DAPI (green) and DAPI transformed one voxel in z (red) with point matches overlayed as lines (cyan). Data from Forrest Collman (unpublished).

We are able to visualize point matches to aid in diagnosing alignment problems. First, we dynamically request sets of points from the Render Point Match database. Then, we render lines between each pair of points overlaid on top of imaging channels. We are able to render the same imaging channel with an offset – e.g. an increase of 1 voxel in the z-direction. By rendering two offset layers and the point matches, we can determine whether the point matches are correct or whether we have enough point matches. Using that information, we can regenerate point matches or re-run the solver with different parameters to improve alignment.

Future Work

The current workflow is a functional prototype. All components are being designed to be used by others, either on local hardware or in the cloud. We plan to take advantage of the elasticity of the cloud and price benefits of "spot" instances to improve performance and reduce computational costs for the tasks which can be trivially parallelized (e.g., materialization).

References

[1] Render Tools and Services (Renderer). `https://github.com/saalfeldlab/render/`.

[2] Janelia EM Aligner. `https://github.com/khaledkhairy/EM_aligner/`.

[3] Saalfeld et al. *Elastic volume reconstruction from series of ultra-thin microscopy sections.* Nature Methods 9, 717-720 (2012) doi:10.1038/nmeth.2072

[4] Neuroglancer. `https://github.com/google/neuroglancer/`.

[5] Burns et al. *The Open Connectome Project data cluster: Scalable analysis and vision for high-throughput neuroscience.* SSDBM 2013

Acknowledgements

Kunal Lillaney (JHU/ndstore), Jesse Patsolic (JHU/synapse clustering) and Sharmishtaa Seshamani (AIBS/alignment) are invaluable colleagues. Eric Trautman (HHMI/Render) Jeremy Maitin-Shepard (Google/Neyroglancer) provided invaluable assistance (and patience). NeuroData is supported by awards from NIH, NSF, IARPA, DARPA, Johns Hopkins University, and the Kavli Foundation. Specific award information can be found at https://neurodata.io/about/.

Johns Hopkins Whiting School of Engineering

Kavli Neuroscience Discovery Institute — Bridging Biology, Engineering and Data Science.

Allen Institute for Brain Science