Published in final edited form as:

Nat Methods. 2018 November ; 15(11): 846–847. doi:10.1038/s41592-018-0181-1.

A community-developed open-source computational ecosystem for big neuro data

Joshua T. Vogelstein[1,*], **Eric Perlman**[1], **Benjamin Falk**[1], **Alex Baden**[1], **William Gray Roncal**[2], **Vikram Chandrashekhar**[1], **Forrest Collman**[3], **Sharmishtaa Seshamani**[3], **Jesse L. Patsolic**[1], **Kunal Lillaney**[1], **Michael Kazhdan**[1], **Robert Hider Jr.**[2], **Derek Pryor**[2], **Jordan Matelsky**[2], **Timothy Gion**[2], **Priya Manavalan**[2], **Brock Wester**[2], **Mark Chevillet**[4], **Eric T. Trautman**[5], **Khaled Khairy**[5], **Eric Bridgeford**[1], **Dean M. Kleissas**[6], **Daniel J. Tward**[1], **Ailey K. Crow**[7], **Brian Hsueh**[7], **Matthew A. Wright**[7], **Michael I. Miller**[1], **Stephen J. Smith**[3], **R. Jacob Vogelstein**[6], **Karl Deisseroth**[7], and **Randal Burns**[1]

[1]Johns Hopkins University, Baltimore, MD, USA.

[2]Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA.

[3]Allen Institute for Brain Sciences, Seattle, WA, USA.

[4]Facebook, Menlo Park, CA, USA.

[5]Janelia Research Campus, Ashburn, VA, USA.

[6]Gigantum, Washington, DC, USA.

[7]Stanford University, Stanford, CA, USA.

To the Editor — Recent technological developments, such as high-throughput imaging and sequencing, enable experimentalists to collect increasingly large, complex, and heterogeneous ‘big’ data. These studies result in terabytes of data per day, yielding petabytes across experiments and laboratories. These experimental capabilities exceed the scale or feature set of existing software. For example, such data cannot be stored, processed, and visualized on a laptop or workstation. Instead, big data must be stored on data centers and processed on high-performance compute clusters.

In 2011, we launched Open Connectome Project[1], an open-access data repository powered by open-source web-services software applications that store, analyze, and visualize large imaging datasets. However, as technology changed, features were added, and scale increased, our academic development team and resources became overwhelmed. We overhauled our custom stack into a community-built and -maintained software ecosystem deployed in the commercial cloud, integrating multiple open-source projects and extending

* jovo@jhu.edu.

Competing interests

D.M.K. and R.J.V. are employed by Gigantum; this company provided support in the form of salaries for these authors, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

All code is available from https://neurodata.io/tools/ under an Apache 2.0 license unless otherwise specified. All publicly available data are accessible at https://neurodata.io/data/ under an ODC-By v1.0 license, unless otherwise specified.

them for our needs (https://neurodata.io). The ecosystem makes it possible to analyze disparate datasets by reusing components originally designed for other applications.

All datasets use NeuroGlancer for visualization in a browser, bossDB—a cloud spatial database—for data management, NDWeb to store visualization links, NDeX to exchange data into and out of the system, and Jupyter notebooks for data analyses. Spatial data derivatives are uploaded and coregistered for quality control. The other software tools are for subsets of the data types, and make the ecosystem flexible and extensible. Figure 1 shows our reusable tools and web visualizations of two exemplary datasets described below.

Conjugate array tomography images ≥ 20 channels of microscale protein expression in three dimensions, and can be followed by nanoscale serial electron microscopy[2]. Our workflow for these data aligns and stitches a collection of 2D images with custom scripts based on TrakEM2's nonlinear registration. After data have been ingested and visualized, probabilistic synapse detection[3] is run on the entire dataset. We study the statistics of the synapses using multiscale graph correlation[4].

CLARITY—a brain-clearing method[5]— collects data from multiple channels, including background for registration, and a 'signal' channel that exhibits protein expression localized to a particular subpopulation of cells. Our workflow uses TeraStitcher to align and stitch the collection of raw 2D image tiles, and registration to the Allen Brain Atlas with NDReg[6], a multiscale, multichannel implementation of large deformation diffeomorphic metric mapping.

To date, NeuroData holds ≥ 100 public and private datasets, with ≥ 200 teravoxels from 30 collaborators, making it the world's largest and most diverse public neuroscience data repository (the up-to-date listing can be found at https://neurodata.io/). We add new data based on our collaborations and community input (data incur ongoing costs for cloud storage and computation). Without running any code or downloading any data, anyone with internet access can visualize image data from different technologies to generate hypotheses or plan new experiments. If investigators choose to download data and/or code, they can access and analyze disparate data with the same functionality and syntax, which allows for faster comparisons and scientific discoveries. Because all of the code is open source, anybody can download, set up, and modify this ecosystem.

Acknowledgements

R.B., E.P., B.F., A.B., V.C., K.L., M.K., E.B., J.L.P., D.J.T., M.I.M., and J.T.V. were supported by the Defense Advanced, Research Projects Agency (DARPA) SIMPLEX program, SPAWAR contract N66001-15-C-4041, NIH-NINDS TRA, 1R01NS092474, "Synaptomes of Mouse and Man," and NSF, 16–569 NeuroNex contract 1707298. W.G.R., B.W., R.H., D.P., T.G., P.M., and J.M. were supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA contract no. 2017-17032700004-005 under the MICrONS program and APL Internal Research and Development Funds. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation.

References

1. Burns R. Szalay A. The open connectome project data cluster: scalable analysis and vision for high-throughput neuroscience; Proc. 25th International Conference on Scientific and Statistical Database Management; New York: ACM; 2013. Article 27
2. Collman F et al. J. Neurosci 35, 5792–5807 (2015). [PubMed: 25855189]
3. Simhal AK et al. PLoS Comput. Biol 13, e1005493 (2017). [PubMed: 28414801]
4. Shen C et al. arXiv Preprint at https://arxiv.org/abs/1609.05148 (2016).
5. Chung K et al. Nature 497, 332–337 (2013). [PubMed: 23575631]
6. Kutten KS et al. A large deformation diffeomorphic approach to registration of CLARITY images via mutual information In Medical Image Computing and Computer Assisted Intervention—MICCAI 2017 Vol. 10433 (eds Descoteaux M et al.) 275–282 (Springer, Cham, 2017).

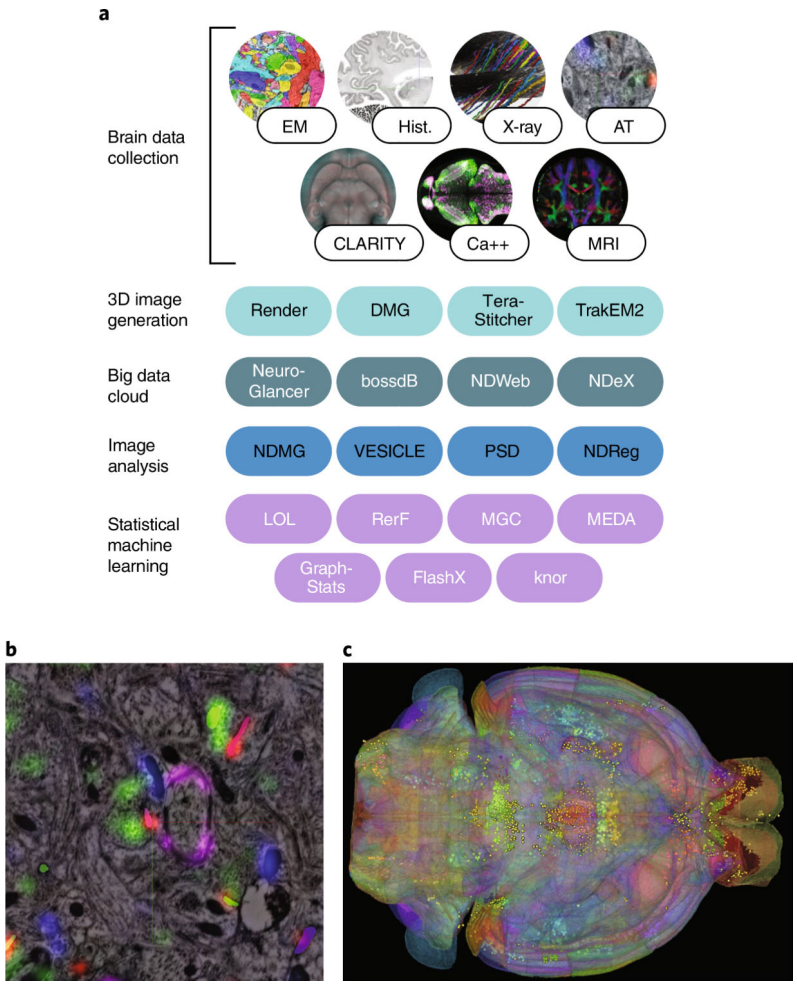

Fig. 1 |. NeuroData's open-source software ecosystem and example visualizations.

a, NeuroData's open-source software ecosystem. EM, electron microscopy; Hist., histology; AT, array tomography. **b**, Image of mouse brain tissue processed with array tomography including multi-spectral light microscopy with electron microscopy. Synapses were labeled manually. **c**, A section of the Allen Reference Atlas registered to an image obtained via whole-brain CLARITY of a mouse. All cell bodies were detected by machine vision.