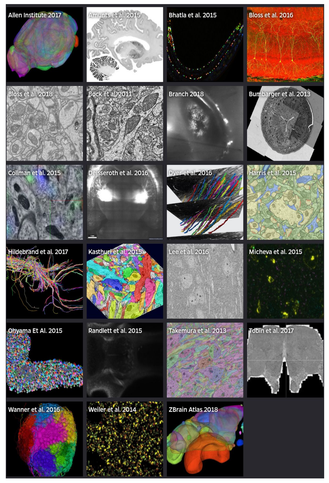
Benjamin Falk

NeuroData's Open Data Cloud Ecosystem

Benjamin Falk, Joshua T. Vogelstein | Johns Hopkins University, Kavli Neuroscience Discovery Institute

Introduction

- Neuroscience imaging produces datasets too large for traditional approaches
- We have deployed a comprehensive suite of tools to support
  - Hosting hundreds of datasets
  - Fast visualization
  - Scale to petabytes
- Difficult for end users and maintenance
- We explore different methods to manage data
  - AWS Open Data bucket for hosting public data
  - Data stored in flat file system
  - Flywheel for managing derivatives and metadata
- Flat file system is faster and less expensive
- Need to build additional missing services



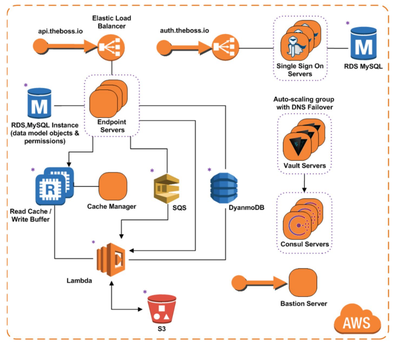
https://neurodata.io/ndcloud/

BOSS

Design

- Built for MICrONS project[1]
- Traffic goes through endpoint(s)
- Mix of microservices and monolith, making it complex

1: https://www.iarpa.gov/index.php/research-programs/microns



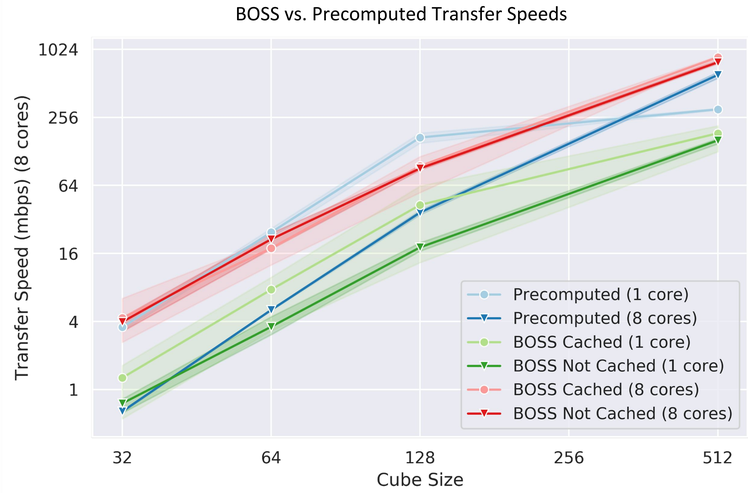
Cons

- $$$
- Central point of failure
- Requires cache to be performant
- (Re)deployments are *tricky*
- Downsampling could be faster
- Limited LIMS system
- No place to put derived data (meshes, skeletons, computations)
- Difficult for end users to use
- Bugs are difficult to track down due to interconnected services
- Requires login
- Coordinates have to be positive

Pros

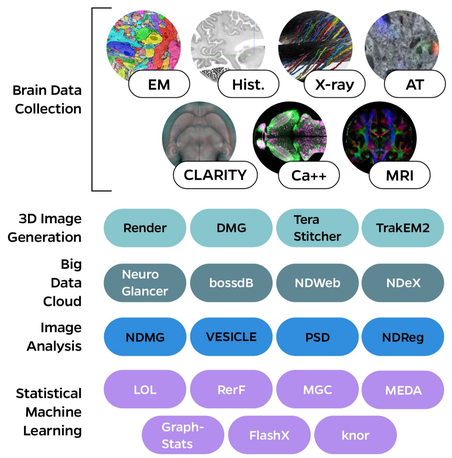
- Provides everything
- Authentication
- Scales reasonably well
- Downsample inexpensive
- Timeseries support
- Database indexing of segmentation locations for easy lookup

Benchmarks



- We do not observe a large benefit of cache for BOSS in this test
- Precomputed can be as fast as BOSS
- Cube size has a strong effect on transfer speed for both BOSS and precomputed

Neurodata tool ecosystem



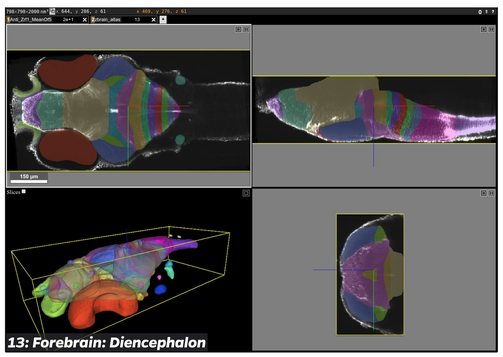
1: https://viz.neurodata.io
2: https://api.boss.neurodata.io
3: https://ndwebtools.neurodata.io

Neuroglancer deployment with zbrain atlas: https://zbrain.viz.neurodata.io

- Neuroglancer: visualization of image volumes, annotations, segmentations, skeletons, meshes[1]
- BOSS: hosting datasets using cloud infrastructure[2]
- ndweb: web application which allows custom interactions with BOSS[3]
  - channel selection, TIFF cutouts, custom neuroglancer links
- Ndex: exchange data w/ neurodata BOSS programmatically

Static storage for image volumes

Neuroglancer precomputed format (created by Jeremy Maitin-Shepard) is a flat file storage specification that has several advantages, and some disadvantages, over a centralized system such as BOSS. Specification is open source and available at the following link: https://github.com/google/neuroglancer/blob/master/src/neuroglancer/datasource/precomputed/README.md



Pros

- Flat file format is flexible (hosting)
- Custom chunk size per dataset
- Direct access to data
- No endpoint
- Fewer things can go wrong
- $

Lacking

- Authentication
- Cache
- Database (bounding box lookups)
- LIMS
- Infrastructure
- No timeseries support

Conclusions and Future Directions

- Neuroglancer precomputed format allows hosting of large image volumes using distributed storage provided by cloud providers, without the complications of running the BOSS.
- Availability of AWS Open Data[1] and cheap storage pricing, hosting these datasets requires minimal infrastructure.
- Pipelines for downsampling, registration, statistical analysis and computer vision can operate on these precomputed volumes, but they take additional resources and infrastructure, which has yet to be finalized.
- Cloud-Volume[2] and Igneous[3] provide excellent command line utilities for working with precomputed volumes.
- Commercial solutions (e.g. Flywheel) could provide the needed LIMS frontend, organization, and pipeline orchestration.

1: https://aws.amazon.com/opendata/
2: https://github.com/seung-lab/cloud-volume
3. https://github.com/seung-lab/igneous/

NIH MSCZ U19 Multi-Scale Circuits in Zebrafish