1. Brief description of Problem to be solved

One of the most important problems in neuroscience today is determining how the brain is wired; which regions are connected? How strong are those connections? What do they imply about how our brains function? With the advent of tissue clearing methods, hundreds of labs around the world have begun examining the connections in the brain with a level of detail previously impossible. Historically, neuroscientists had to physically slice and image each section of a given brain, taking weeks for data acquisition while also severing connections between different regions of the brain. Tissue clearing methods, such as CLARITY, have enabled researchers to image the whole brain intact using light sheet microscopy, preserving connections between regions. We hope to drive breakthrough neuroscience discoveries by enabling multi-terabyte (TB) cleared brain tissue analysis.

Specifically, we will solve three challenges that are collectively the bottleneck in this promising and growing line of research:

1. A lack of robust algorithms: Hundreds of algorithms exist to performing analyses on microscopy data. However, very few, if any, of these algorithms are optimized for cleared tissue data. Furthermore, there is a wide variety of parameters to tune for each of these individual algorithms, hindering robustness and reproducibility of analyses. To reliably and robustly analyze these data, there must be comprehensive, well-documented analysis tools along with meaningful quality assurance at each step of the pipeline.
2. Making the algorithms scale to multi-terabyte data: CLARITY coupled with light sheet microscopy can generate datasets between 1 and 100 TB per brain, rendering many existing analysis tools computationally intractable. As microscopy resolution improves over time, these data will only get bigger. Analysis infrastructure constraints currently prevent data exploration in neuroscience research on datasets this large. To tackle this problem
3. Facilitating subject matter expertise: It can be quite challenging for neuroscientists who are more familiar with data collection than data analysis to set up an environment where they can analyze data. Often, there are useful analysis tools that are too complex to enable use by neuroscientists. A useful tool must be easily accessible and require minimal training to use.

2. Proposed AWS Solution

1. COBALT pipeline: A cleared brain tissue analysis pipeline designed for scalability with robust algorithms. COBALT will consist of 3 main submodules: registration, cell detection, and tractography. This addresses our computational challenge as we will make our pipeline and code open-source via Dockerhub and Github respectively.
2. COBALT cloud deployment: A Docker container that is compliant with AWS Lambda and/or AWS Batch, along with deployment scripts allowing users to automate analysis, given data in S3, with one-click from the command line. Cloud deployment will automate the creation of the AWS Batch compute environment and the associated resources. This addresses our big data challenge as the Batch compute environment will automatically scale up as necessary for additional jobs.
3. COBALT web-service: An AWS-hosted web-service will be created to allow users to automate upload of local data to S3, creation of a compute environment, queue, and job definitions with minimal input from users. To ensure that there are enough resources for

each job, we will use Elastic Beanstalk to scale our web-service. This addresses the subject-area expertise challenge as we will set up a cloud-hosted, scalable web-service accessible to all neuroscientists.

Tools Used:
We will be using the following AWS resources: EC2, ECS, Elastic Beanstalk, Batch, S3, CloudFront

Current Progress:

Our pipeline is in the pre-alpha stage of development. We are developing scalable algorithms for the initial stages of our pipeline. The Neurodata team has published two journal papers and released open-source code on CLARITY-based registration (ndreg). Our registration package has been tested and refined on numerous CLARITY datasets. Preliminary investigation of cell detection algorithms has yielded promising results especially with regards to unsupervised methods. These methods are far more computationally tractable and desirable for scale. Current 3D spot detection methods tested provide reasonably accurate results but are not optimized for CLARITY. However in order to reproducibly assess algorithms we are developing metrics and scripts to automatically analyze their performance.

Timeline/Milestones:

September 30, 2017 *(Completed)*: Understand the relevant neuroscience and optics techniques involved in generating datasets including possible artifacts or other issues with data. Read about and replicate current unsupervised machine learning and image processing cell detection methods on CLARITY data. Understand state-of-the art machine learning methods in biological image analysis and narrow focus to a small subset of potentially useful cell detection algorithms. Deliverables included complete jupyter notebooks in python to run and evaluate existing unsupervised cell detection algorithms and a bibliography to solidify the context in which we are developing our software.

December 31, 2017: Finalize cell detection algorithms and include quality assurance metrics for COBALT pipeline and prototype cloud deployment using EC2 instances through AWS Batch. Verify quality and robustness of algorithms on 5 publicly available CLARITY datasets. Deliverables will include a well-documented python package on PyPi, open-source code on Github, and a simple web-service which can process small (<50GB) datasets.

March 31, 2018: Understand and quantitatively assess current unsupervised tractography methods especially as applied to light microscopy images. Complete cloud deployment. Finalize quality assurance metrics in COBALT pipeline and deploy Docker container using AWS Batch or ECS. Run COBALT on 10 publicly available large (>1 TB) light microscopy CLARITY datasets. Deliverables will include analysis results stored in a public S3 bucket, a public Docker image on Dockerhub, and our open-source code available on Github.

June 30, 2018: Finalize COBALT pipeline web-service deployment. Analyze all publicly available cleared brain tissue light microscopy datasets through final COBALT pipeline and make results publicly available in an S3 bucket. Deliverables will include final web-service with

Elastic Beanstalk back-end to scale with 100 TB+ data and publicly available analysis results in an S3 bucket.

3. Plan for sharing outcomes (tools, data, and/or resources) created during project.

The Neurodata team believes strongly in open source development. Our pipeline and web-service will be actively developed using open-source containerized environments like Docker. All code developed will be publicly available on github. Furthermore, our web-service will be public-facing, so all neuroscientists can use the pipeline without need for technical expertise. Our goal is to provide a cloud-hosted, publicly available, science-as-a-service platform to facilitate research in the neuroscience community. We have already created an open-source registration package (https://github.com/neurodata/ndreg) among others (https://github.com/neurodata).

4. Any potential future use of AWS beyond grant duration by individual research group or broader community

The COBALT pipeline and web-service will be made publicly available through our research website http://neurodata.io to facilitate long-term, reproducible neuroscience research through the AWS cloud infrastructure. Any datasets obtained and processed by our COBALT pipeline will also be made available using one of our public S3 buckets.

5. Names of AWS employees you have been in contact with

Chris Goodson – JHU Lead Account Manager
Angel Pizarro – Technical Life Sciences Manager
Mike Kuentz – Senior Solutions Architect
Dan Sinnott – Account Manager

6. Any AWS Public Data Sets to be used in your research

None right now, but we will be using 12 public CLARITY-Optimized Light Sheet Microscopy (COLM) datasets hosted by Amazon in S3.

7. Keywords

computational neuroscience, docker, pipeline, scalability