I. Brief description of Problem being 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 terabyte-scale, cleared brain tissue analysis through the use of a comprehensive and well-documented software package and web interface.

Challenges:

1. Computational challenge: 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. In order 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. Big Data challenge: CLARITY coupled with light sheet microscopy can generate datasets over 1.5 TB, rendering many existing analysis tools useless. 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.
3. Subject-area expertise challenge: It can often 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.

II. Proposed AWS Solution

1. COBALT pipeline: a cleared brain tissue analysis pipeline designed for scalability with algorithms that can perform robust analyses. This addresses our computational challenge as we will make our pipeline and code open-source via Docker and Github respectively.
2. COBALT cloud deployment: we will create a Docker container that is compliant with AWS Lambda and/or AWS Batch and develop deployment scripts allowing users to automate analysis, given data is in S3, with one-click from the command line. Cloud deployment will automate the creation of the AWS Batch/Lambda compute environment and the associated resources.
3. COBALT web-service: To address the subject-area expertise challenge, 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. In order to address the big data challenge, we will use AWS Lambda/Batch to create the required compute resources in parallel.

Tools Used:
AWS EC2
AWS Lambda
AWS Batch
AWS S3

Current Progress:

Solution 1: Our pipeline is under active development and is currently in the pre-alpha stage of development. We are developing scalable analysis algorithms for the initial stages of our pipeline.
Solution 2: Our cloud deployment will rely on a more finalized analysis pipeline. Once solution 1 is in it's beta stage, we will begin automating the creation of a compute environment using AWS cloud infrastructure.
Solution 3: We will have a working beta of our web-service by February 2018 but the completion of this process requires automation scripts from solution 2.

Timeline/Milestones:

End of December 2017: Finalize analysis algorithms for COBALT pipeline and prototype cloud deployment/web-service using AWS EC2 instances or AWS Lambda. Verify quality and robustness of algorithms on 5 publicly available CLARITY datasets

End of February 2018: Complete COBALT pipeline and cloud deployment. Finalize algorithms and quality assurance metrics in COBALT pipeline and deploy Docker container using AWS Batch/Lambda. Run COBALT on 10 publicly available large (>1 TB) light microscopy CLARITY datasets.

End of April 2018: Complete COBALT 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.

III. Plans for sharing outcomes

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.
See here for open-source pipelines: https://github.com/neurodata.

IV. Any potential future use of AWS beyond grant duration

The COBALT pipeline and web-service will be made publicly available through our research website neurodata.io for long-term, reproducible research through the AWS cloud infrastructure.

Any datasets obtained and processed by our COBALT pipeline will also be made available using a public S3 bucket.

V. Name of AWS employees you have been in contact with

None.

VI. AWS Public Data Sets to be used in your research

None right now, but hope to have our S3-hosted CLARITY data ultimately become a public dataset.

VII. Keywords

computational neuroscience, docker, pipeline, scalability