I. Brief description of Problem being solved

Computational neuroscience has led to many novel discoveries about how our brains work, allowing us to better evaluate personality, intelligence, and psychological disorders. However, there is a bottleneck in computational neuroscience discoveries due to the difficulties in handling big data. Many datasets, such as conjugate array tomography (cAT) synapse data, and CLARITY brain images, can be terabytes in size. Therefore, current practices for running analysis and storing/uploading the data simply do not work as they were not made to handle large volumes.

Two independent leading laboratories have built complementary tools that each partially solve the problem. Specifically, the Johns Hopkins University's NeuroData Lab excels at storing, viewing, and pulling large spatial data using AWS, but does not provide a service for researchers to run different analysis jobs. The OpenNeuro platform provides an easy-to-use interface where users can launch and monitor analysis pipelines on AWS at scale, but is limited to small, public datasets and only a few specific data tools. There is therefore a gap: *researchers are unable to run computationally expensive pipelines on large spatial data in the cloud*.

II. Proposed AWS Solution:

Integrate Neurodata data tools and the OpenNeuro framework to provide a common, intuitive interface for visualization and analysis of big neurodata that run at scale in the cloud.

AVATR Data ingest service:

This will allow users to go to our AWS-hosted web service and automate the upload procedure to S3. This service will also allow users to pull big data using Neurodata tools into a format ready to be run on dockerized pipelines. Our service will ideally require minimal user input to further simplify the cloud deployment process.

AVATR Cloud deployments:

The docker containers we generate will be compliant with AWS Batch. The cloud deployment will automatically create the AWS Batch compute environment, queue, and jobs required for the specific data, plugin, and user input.

AVATR Plugin-based architecture:

Each data set has its own intricacies, and ongoing support for each modality will be individually supported through docker container plugins that can run and scale easily through AWS Batch.

Tools Used:

AWS Batch, AWS S3, AWS EC2

Timeline and Key Milestones:

December 15 2017:

AVATR core infrastructure will be completed and able to be run in the cloud. For the beginning of the project, we will use cAT data. Core infrastructure includes data ingest, analysis, and visualization, as well as a job metadata archiver.

Data ingest infrastructure will be composed of subroutines pulling and uploading data to S3 buckets. Data analysis infrastructure will consist of functionality to distribute computation on large volumes of data, and stitch results together for meaningful output. Visualization infrastructure will operate on the output of analysis, and all jobs will be archived with meaningful metadata.

Graphic design of UI to indicate data and algorithms to use will be completed. UI for a "history" of archived job metadata will also be completed.

February 15 2018:

We will work with select computational neuroscience groups to integrate their pipelines into our infrastructure. Thorough documentation will be made and used to introduce groups to the infrastructure. The UI will be linked to the infrastructure to create a visually elegant cloud-based web service. These select groups will be able to use our infrastructure effectively with their own applications.

April 15 2018:

We will initiate collaboration with http://openneuro.org. Our infrastructure will be able to be run alongside other pipelines on their website such as NDMG, which we previously developed and is currently amongst the most frequently used pipelines in OpenNeuro.

III. Plans for Sharing Outcomes:

Our pipeline and web-service are actively developed using open-source docker containers. All code is publicly available on github, and docker containers will be in dockerhub.

IV. Any potential future use of AWS beyond grant duration:

The tools we develop will be incorporated into our lab website, http://neurodata.io, and http://openneuro.org for long-term research. Researchers investigating supported datasets will have the continued ability to analyze their data using AWS cloud infrastructure. All the data and derivatives will continue to be stored in S3, and when anybody runs a pipeline using this infrastructure in the future, they will be utilizing EC2 and Batch.

V. Name 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

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

None

VII. Keywords to facilitate proposal review:

Neuroscience, Docker, S3, Infrastructure, Neurodata, OpenNeuro

In its current state, the HBNB dataset is about 2 TB compressed. With derivatives and visualizations stored, we expect to have around 10 TB of data. This would amount to $235.58 of storage costs per month according to the Simple Monthly Calculator. TIn order to process the entire HBNB dataset, preliminary results suggest we expect to require one1 hour of compute per data point on a m4.4.xlarge instance. To account for testing and debugging, we budget 1000 hours on a m4.4x.large instance to process the HBNB dataset. We budget and an additional 500 hours per month during and after the completion of the project to ingest additional datasets uploaded by the broader community. This would require $804.72 for the HBNB dataset, and an additional $404.70 per month to ingest new datasets according to the Simple Monthly Calculator. Finally, running a dedicated web service on a t2.large instance would require $64.81 per month. In total, this project would cost ($235.58 * 12) + ($64.81 * 12) + ($404.70 * 12) + $804.72 = $9265.80, which is the value in AWS credits we request.