Multi-Modal Brain Visualizations

Team Red Lemur – NeuroData Design 2017-2018 – Johns Hopkins University

Brief description of the problem to be solved

With a dataset of *n* subjects containing *m* measurements, where each measurement can be demographic, functional neuroimaging, structural neuroimaging, genetic, and phenotypic data, brain scientists could use Exploratory Data Analysis (EDA) techniques to discover new hypotheses and treatments of brain disorders.

In any EDA task, the first step is to visualize the data. Given his/her domain knowledge, an explorer can use visualizations to find of trends, geometric structure, and clusters which inspire more rigorous numerical analysis and concrete scientific results. When data are small, this step is easy. When data are a multi-modal mess of high dimensional spatio temporal and ordinal measurements, this step is impossible for somebody without an extensive statistical and computational background.

Brain scientists are not data scientists, and do not have the technical expertise to create complicated visualizations of neuroimaging datasets. Data scientists are not brain scientists, and do not have the wealth of domain knowledge required to map visual curiosities to clinically meaningful results. This problem is currently holding back scientific breakthroughs which can be leveraged to benefit human-kind, and we want to build the technology to solve it.

We will create a science-as-a-service web application that will enable the brain science community to produce high quality, meaningful visualizations of current and future neuroimaging datasets at scale by leveraging cloud computing services from Amazon Web Services. This web application will be made public and advertised as a utility to brain scientists, who can use it freely to facilitate their own research.

The Developmental Dataset

For development purposes, we will make use of the dataset described in the manuscript "The Healthy Brain Network Biobank: An open resource for transdiagnostic research in pediatric mental health and learning disorders. bioRxiv (2017)". This is a novel dataset with 664 subjects and measurements spanning multiple modalities including cognitive, psychiatric, EEG, MRI, and demographic data. This is the first release of a planned 10,000 subject sample, and is an ideal example dataset for this application because of its scale, wealth of modalities, and public availability.

Proposed AWS solution

User Experience – Brain scientists will interact with the service via a web page. By navigating to a Visualizations page, they will be able to browse and download visualizations from publicly available neuroimaging datasets. In order to process their own data, brain scientists can view a detailed data format page which will closely follow the standard Brain Imaging Data Structure (BIDS). If not comfortable with data formatting, they can generate an issue on our Github page which will resolved by a one of us or a volunteer.

Backend Processing – Once raw data is uploaded, a batch processing job will be triggered which automatically runs preprocessing pipelines on data from each modality and saves preprocessed data, derivatives, and quality assurance information. There are frequently new open source pipelines which can preprocess each distinct modality, so these will be applied in a modular fashion. Once preprocessed data is available, a visualization pipeline for each modality will be applied to the data, and the results of this will also be saved.

Cloud Infrastructure – Raw neuroimaging data, preprocessed neuroimaging data, derivatives, quality assurance information, and visualizations will exist in a S3 bucket, with one folder per distinct dataset available. Processing will be done using AWS Batch. Finally, the website will be hosted with an Amazon EC2 instance.

Timeline:

- *Iteration 0 (one month) - Background Research and Planning (Completed):* In this phase, we will scope the feasibility of building out web service and the impact that it will have on the Brain Science community. This includes background reading on psychopathological research to determine whether analysis capability can be improved, background reading on the current state of visualization of neuroimaging data to determine what we can build upon and what we will need to build from scratch, background reading on high dimensional statistics and multimodal data analysis to determine robust and scalable algorithms for visualizing neuroimaging datasets, and finally the creation of a project proposal and statement of work to formalize our scope and create a plan of action. Deliverables for this iteration will include a bibliography, notes from background reading, a powerpoint proposal, and a statement of work.

- *Iteration 1 (1.5 months) – Top 10 visualizations:* In this phase, we will determine the most descriptive visualizations for each modality in the Healthy Brain Network Biobank (HBNB) dataset, and create visualization functions and a pipeline to generate all of these plots automatically. These visualizations may be one-to-one (one plot per data point) or many-to-one (aggregate data summarizing all data points), and will be chosen based on questions we would like to answer or biomarkers from literature we would like to visualize. Deliverables for this iteration will include a well documented pypi package which could be used to generate each of these plots individually.

- *Iteration 2 (1.5 months) – Cloud deployment of a Minimum Viable Product:* In this phase, we will develop the cloud infrastructure required to run our Top 10 Visualizations pipeline in an automated one click web-service, in addition to continually improving our Top 10 Visualizations based on peer feedback. Deliverables will include a Docker image which takes as input a specifier for a data point living in a S3 bucket, pre-processes the data, creates visualizations of the data using our Top 10 Visualizations package, and uploads all derivatives and plots to a S3 bucket. We will also build a Minimum Viable Product (MVP) web service which can be used to automatically launch a set of these Docker images on EC2 using AWS Batch and process each data point in a user defined data set. The results will then be browse-able with simple bucket folder listings in a different tab on the web page.

- *Iteration 3 (2 months) – Cross-modal visualization:* Up until this point, we will have only generated visualizations that rely on only one data modality. Here, we will use methods borrowed from the statistical community to create visualizations which rely on >1 modality. E.g., visualizations where two different neuroimaging modalities are embedded into the same latent space, or visualizations describing relationships between two modalities. Deliverables will included an update to our visualizations package, and extensive documentation and scientific notebooks explaining new findings.

- *Iteration 4 (2 months) – Web service update, usability, and promotion*– In this final phase, we will add work from Iteration 3 into the web service and make it production ready (e.g, ready for an alpha launch to the Brain Science community). This will involve updating web service aesthetic and security, and curating publicly available neuroimaging data sets to populate the visualizations section of the service. Deliverables will include a functioning production web service which brain scientists anywhere can access, and the creation of demos and documentation to show off functionality and promote use by the brain science community.

Plan for sharing outcomes
All code and will be available on GitHub, Docker images on DockerHub, and a link to our production web service will be public and available to all.

Any potential future use of AWS beyond grant duration by individual research group or broader community
This publicly available science-as-a-service application will be available to and used by the scientific community for the foreseeable future depending on popularity of the product. If adopted, this application would likely encourage further development and use by the brain science community.

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

Any AWS Public Data Sets to be used in your research
No AWS Public Data Sets will be used. We will make use of the dataset described in the manuscript “The Healthy Brain Network Biobank: An open resource for transdiagnostic research in pediatric mental health and learning disorders. bioRxiv (2017)”. This is a novel dataset with 664 subjects and measurements spanning multiple modalities including cognitive, psychiatric, EEG, MRI, and demographic data. We would be interested in seeing this dataset become an AWS Public Data Set, as it is the first release of a planned 10,000 subject sample, and is an ideal example dataset for this application because of its scale, volume and variety of modalities, and public availability.

Budget
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. To process the entire HBNB dataset, preliminary results suggest we require one 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. 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.

Keywords to facilitate proposal review

Visualizations, multi-modal, neuroscience, psychology, biomarkers, open science, science in the cloud