Multi-Modal Brain Visualizations

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

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 create data-driven hypotheses and create a better understanding and treatment of brain disorders.

In any Exploratory Data Analysis (EDA) task, the ubiquitous first step is to visualize and look at the data. This informs the explorer, in the context of his/her wealth of domain knowledge, of visual trends, geometric structure, and groupings of data points which inspire more rigorous numerical analysis and concrete results. When data are small, this step is easy. When data are a multi-modal mess of high dimensional spatio-tempotal time series and high dimensional ordinal/categorical tables, this step is impossible for somebody without extensive statistical and computational knowledge.

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 are planning on creating a web application which will enable brain scientists to produce high quality, meaningful visualizations of neuroimaging datasets at scale by leveraging cloud computing services from Amazon Web Services. This service will be made public and advertised as a utility to brain scientists, who can use it freely if their (non sensitive) data is then made available to the world.

The Developmental Dataset

For development purposes, we will make use of the Healthy Brain Network Biobank 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)” (http://biorxiv.org/content/early/2017/06/13/149369). 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.

Components of the Full Application

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.

Iterations and Deliverables

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 HBNB dataset, and create visualization functions and a pipeline to generate all of these plots automatically. 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 webservice. 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 MVP web page 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, user outreach – 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.

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 according to the Simple Monthly Calculator. In order to process the entire HBNB dataset, we expect to require 1 hour of compute on a m4.4xlarge instance. To account for testing and debugging, we budget 1000 hours on a m4.4x large instance to process the HBNB dataset, and an additional 500 hours per month on a m4.4x large instance to ingest new datasets. 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.

If this budget cannot be met, we can develop and deploy on a subset of the HBNB dataset. Costs dependent on the data-set size are everything but the dedicated web server which would cost in total $777.72 for the year. Thus, for example, if we were awarded ($9265.80 - $777.72) / 2 + $777.72 = $5021.76, we would be able to operate on half of the data, which would allow us still be of benefit to the communities we are trying to impact.