Billions for Big Brains Proposal

Sunny Thodupunuri

Denis Jered McInerney

Hannah Cowley

1 Scientific Question

Neurodegenerative diseases have an irrefutable impact on our society through causing human suffering and high monetary cost. Two well-known diseases, Alzheimer's and Parkinson's, impact more than 5.4 million and 500,000 individuals, respectively. Although efforts are underway to understand the neurodegenerative processes underlying these diseases, more information is needed regarding the impact they have on the brain's macro and microstructure. Although it is known that different neurodegenerative diseases impact brain functioning in different ways, few studies have executed a large-sample approach to comparing and contrasting the connectomes of such diseases. As such, we ask, how do connectivity patterns differ between neurodegenerative diseases, and how do individual differences influence connectivity patterns within a class of neurodegenerative diseases? Answering this question will help clinicians make more accurate diagnoses, further categorize possible subcategories of neurodegenerative diseases, aid researchers in developing targeted therapies, and illuminate further the relationship between brain structure and behavior.

Although there are many well-documented neurodegenerative diseases, we will investigate seven disorders and and one age-matched control group per disease. Individuals in the mid to late stages of their respective diseases will be recruited so they may be studied as the damage to their connectome becomes more severe; with stark damage to the brain, it may be easier to determine differences in the connectome. We will acquire the connectomes of individuals with mixed dementia and dementia of unknown pathology, Alzheimer's disease, Parkinson's disease, Pick's disease, Huntington's Disease, Posterior Cortical Atrophy, and Dementia with Lewey Bodies. While these dieseases share some aspects of symptomology, their symptomology and pathology differs significantly from each other, allowing for meaningful comparisons. Control groups will be acquired to create a "model" connectome for healthy participants. Through databases such as the one developed by NeuroData, these healthy connectomes may be available with little additional scanning needed by the present project.

2 Hardware, Facilities, & Data Collection Processes

We propose using structural MRI, resting state fMRI, task-based fMRI, DSI, Clarity, and Electron Microscopy to collect information about both the macroscale and microscale connectome of different neurodegenerative diseases. DSI is chosen over DTI as it has the ability to produce better images of crossing white matter connections. MR methods will be used to develop macroscale connectomes for each individual in the study. These macroscale connectomes will be compared against each other between and within neurodegenerative diseases to investigate the connectome differences between neurodegenerative diseases and whether individual differences play a role in determining an individual's macroscale connectome within a specific class of neurodegenerative disease. To visualize and organize data,

tools such as the ones developed by NeuroData will be used.

To develop the macroscale connectomes, participants will be asked to provide any and all MR imaging studies from their time of diagnosis to the present. It is anticipated that individuals will be able to provide structural MRI scans, as this is a common metric used for diagnosis and monitoring of neurodegeneration. Individuals will be asked to come in twice a year for 5 years, and at each of these times, resting state fMRI, task-based fMRI, and DSI scans will be completed. Structural scans will be completed once a year. If the participant has had a structural scan within the past 4 months, they will not be re-scanned.

Patients in the study will be asked whether, upon death, they would like to donate their brain tissue to our study. If the patient consents to such, the tissue will be randomly assigned to either CLARITY or EM scanning after collection. Additional tissue samples will be accrued from individuals whose cause of death was a neurodegenerative disease, who did not complete the longitudinal portion of the study. CLARITY and EM imaging will be used to develop a microscale connectome of brain regions most affected by each specific neurodegenerative disease. These regions will be determined from past studies of the pathology of these diseases, and from significant differences found, if any, in the macroscale connectomes across individuals suffering from the same neurodegenerative disease. These microscale connectomes will be used to further investigate the differences betweeen different neurodegenerative diseases and differences between individuals.

A total of 9 institutions will be involved in this project. These institutions will be geographically diverse to acquire a wide range of patients and accumulate sound data. Groups will be centered at 3 east coast institutions (including Johns Hopkins University), 3 central institutions, and 3 west coast institutions. Having this geographic spread is especially important for EM and CLARITY tissue collection, as once the individual is deceased, there is a short time window before the tissue degrades too much for data collection.

3 Information Extraction Processes

The part of the analysis of the data from structural MRI, resting state fMRI, task-based fMRI, and DTI, once it is acquired will be split into two parts: graph extraction and graph classification. The results of this will determine on which part of the brain we focus EM and Clarity.

The graph extraction will be done using techniques discussed in "Meta-connectomics: human brain network and connectivity meta-analyses". Once the graph extraction is done, we will create a connectome based off of each of the separate graphs from each of the images. For the EM and Clarity data analysis, we will use NeuroData's analysis tools in order to analyze the enormous amount of data efficiently and in a standard format.

In the graph classification step, there are two different ways in which we will use the data:

1. We will use the disease labels to train a Gaussian Mixture Model (supervised learning),

and we will use this to predict the diseases of new connectomes.

2. We will use the graphs of the connectomes but not their labels to train a Gaussian Mixture Model using a clustering algorithm called mclust (unsupervised learning), which is implemented in the package from MCLUST: Software for Model-based Cluster Analysis (1999).

We will use this algorithm to figure out the optimal number of clusters, which can each correspond to a disease or a disease subgroup. To do this, we will use some metric, for example Bayesian Information Criterion, to evaluate how well the final clustering is for various different numbers of clusters. We may also try to use other similar algorithms such as k-means clustering, which also to determine the optimal number of clusters. This type of analysis will help determine if we need to split a disease up into two or more different classifications or (less likely) merge two diseases together.

In order to perform both of these algorithms, we need to come up with features for the connectome graphs. The simplest set of features is simply a number for each edge in the graph, and the number would indicate if the edge exists and how strong the weight is for it. Because this may be a lot of edges, it may be a good idea to reduce the number of features by one of the following ways in order to be able to run the above algorithms faster:

1. We could research possible aspects of the graphs that may be very important for certain cognitive functions which one suspects to be different for different diseases as well as in a healthy brain. This would be one way of doing manual feature development.
2. We could spend time researching and developing an algorithm that would reduce the number of edges in a graph.
3. We could simply only take the strongest connections in graph or look at the images in a lower resolution so as to get a less complicated connectome with less edges.

It would be useful to explore all three of these options. None of these would cost much money other than paying for researchers.

4 Data Upload

Due to the large amounts of data generated by EM and CLARITY, simply uploading to the internet is not feasible, even with gigabit internet connections. Instead, we propose to make use of the Amazon Snowmobile service, which can be used to upload 100 PB per

snowmobile. According to Amazon's estimates, the entire Snowmobile can be loaded up in a few weeks using the high-speed data connections that come with the Snowmobile service.

For the structural MRI, resting state fMRI, task based fMRI, and DSI data, we propose to upload them to a local machine. Since these techniques generate much smaller data sets than EM and CLARITY, they can be uploaded to AWS from the local machine.

5 Data Storage

After data is uploaded using an Amazon Snowmobile, it can then be transferred to Amazon's S3 service. We propose to use this service since it integrates well with Snowmobile and can be used in conjunction with Amazon EC2 for data processing. Additionally, once data processing is completed, results and data sets can be shared with a website hosted on S3.

We propose to upload the data on the local machine to S3 as well. However, the processing of this data can be done of the local machine. Doing the processing on a local machine may be more convenient because it is easier to manipulate and use data on a local machine than it is to use on an online machine.

6 Cost

TO BUY: NO mri scanners (most institutions have that) CLARITY scan equipment JERED AND SUNNY DO THIS PART BECAUSE YOUR STUFF (MICRO STUFF) ARE THE MOST COST-INTENSIVE STUFF AND I'M NOT SO SURE ABOUT HOW TO DO THIS

HANNAH SAYS: MRI scanning is about $1000 per scan session. I propose 2 scan sessions a year for a max of 10 years per participant. Let's plan on 50 participants per disease. An equivalent amount of EM/CLARITY data should be collected, so y'all crunch numbers on that

Table 1: Table detailing costs

Item	**Cost per Item**	**Number**	**Cost**
MRI Scans	$1,000	7000	$7,000,000
Light Sheet Microscope	$400,000	18	$7,200,000
Researcher Salaries and Benefits	$200,000	180	$36,000,000
Local Computer	$10000	45	$450,000
AWS Snowmobile	$2,000,000	15	$30,000,000
S3 Storage	$2.52	2,000,000,000	$5,040,000,000
EC2	$2	100,000,000	$200,000,000
		Total Cost:	$5,320,450,000

7 Feasibility

Because we are using previously existing technology for all of the scans we do, and each scanning session before death doesn't take very long, the data acquisition is very feasible for the scans other than EM and Clarity. EM and Clarity would only have to be performed on small parts of the brain, and would also only be done on a few patients per scan per disease. This would make the number of EM scans, which take the longest, about 42 if we do two per patient. Over the course of 5 years, that would be a little over 8 scans per year, which means that if we do a one per every 3 months, we should have 3 Electron Microscopes working in parallel. Because we are using over 9 different institutions, this should definitely be feasible.

Another consideration is the number of people it would take to run the data acquisition. Because it is across so many institutions, it would take on the order of 100 people to get all of this data. Right now, the field of connectomics isn't that big, so it might be hard to get 100 people to join this project. However, because it is a 10 billion dollar project, it would attract many in the field and even those that were previously outside the field.

With regards to data acquisition, it is well within both our budget and the technology of the current day to store 100s of petabytes of data and transport it to the cloud. Once there, we will need a great deal of processors working in parallel to analyze the data, but we have methods for scaling down the amount of computing necessary in order to make the time scale on which we will need to run programs on the order of months (i.e. via dimensionality reduction which is discussed in the information extraction processes section). This is also well within our budget. Hopefully, once we finish processing the 5 years of data and even before, we will have results that can be used to direct further research on different therapy and possibly drug treatment options. There is room in our budget to do research on this as well.

References

[1] Liu AK, Hurry ME, Ng OT, DeFelice J, Lai HM, Pearce RK, Wong GT, Chang RC, and Gentleman SM. Bringing clarity to the human brain: visualization of lewy pathology in three dimensions. *Neuropathology and Applied Neurobiology*, 2015.

[2] Alzheimer's Association. Types of dementia.

[3] David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner. Graph partitioning and graph clustering. *American Mathematical Society Center for Discrete Mathematics and Theoretical Computer Science*, 2012.

[4] Kwanghun Chung and Karl Deisseroth. Clarity for mapping the nervous system. *Nature Methods*, 2013.

[5] McGovern Institute for Brain Research at MIT. Brain disorders: By the numbers, January 2014.

[6] C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysimeta-connectomics: human brain network and connectivity meta-analyses. *Journal of Classification*, 1999.

[7] Stephan Gerhard, Alessandro Daducci, Alia Lemkaddem, Reto Meuli, Jean-Philippe Thiran, and Patric Hagmann. The connectome viewer toolkit: An open source framework to manage, analyze, and visualize connectomes. *Frontiers in Neuroinformatics*, 2011.

[8] Patric Hagmann. From diffusion mri to brain connectomics. *PRESENTEE A LA FACULTE SCIENCES ET TECHNIQUES DE LINGENIEU*, 2005.

[9] Xiao-Bo Liu1 and Cynthia M Schumann. Optimization of electron microscopy for human brains with long-term fixation and fixed-frozen sections. *ACTA Neuropathologica Communications*, 2014.

[10] Crossley NA, Fox PT, and Bullmore ET. Meta-connectomics: human brain network and connectivity meta-analyses. *Psychological Medicine*, 2016.

[11] Human Connectome Project. Gallery.

[12] Joshua T. Vogelstein, Brett Mensh, Michael Häusser, Nelson Spruston, Alan C. Evans, Konrad Kording, Katrin Amunts, Christoph Ebell, Jeff Muller, Martin Telefont, Sean Hill, Sandhya P. Koushika, Corrado Calì, Pedro Antonio Valdés-Sosa, Peter B. Littlewood, Christof Koch, Stephan Saalfeld, Adam Kepecs, Hanchuan Peng, Yaroslav O. Halchenko, Gregory Kiar, Mu-Ming Poo, Jean-Baptiste Poline, Michael P. Milham, Alyssa Picchini Schaffer, Rafi Gidron, Hideyuki Okano, Vince D. Calhoun, Miyoung Chun, Dean M. Kleissas, R. Jacob Vogelstein, Eric Perlman, Randal Burns, Richard Huganir, and Michael I. Miller. To the cloud! a grassroots proposal to accelerate brain science discovery. *Neuron*, 2016.

[13] Joshua T. Vogelstein, William Gray Roncal, R. Jacob Vogelstein, and Carey E. Priebe. Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[14] V.J. Wedeena, R.P. Wanga, J.D. Schmahmannb, T. Bennera, W.Y.I. Tsengc, G. Daia, D.N. Pandyad, P. Hagmanne, H. D'Arceuila, and A.J. de Crespignya. Diffusion spectrum magnetic resonance imaging (dsi) tractography of crossing fibers. *NeuroImage*, 2008.