Ph.D. Thesis Proposal: Gregory Kiar

December 10th, 2017

Characterizing Study Performance for Increasing Power in Big Data Neuroscience

With an explosion in data collection across nearly all scientific disciplines, we are entering the big-data era of scientific exploration. From the Sloan Digital Sky Survey [1] in cosmology, to the UK Bio-Bank [2], Human Connectome Project [3], and the Consortium of Reproducibility and Reliability (CoRR) [4] in Neuroscience, initiatives are being launched to federate this data-tsunami, many with particular emphasis and commitment towards open data-sharing. As access to data has increased drastically, it has become apparent that a lack of reproducibility in data analysis methods or collection paradigms is becoming a plague in many current disciplines of science [5]. Unless addressed, our analyses will continue to fail to do justice to the extraordinary datasets we possess. Computational sciences in particular have an opportunity to develop more rigorous practices and provide tools which lower the barrier for producing reproducible analyses.

The aim of my PhD is to study the reproducibility, stability, and generalizability of pipelines in structural neuroimaging. The main hypothesis of my work is that optimizing pipelines for accuracy on small, largely homogenous cohorts of data (as is currently commonplace in neuroscience), results in tools which are prone to overfitting, and resultantly perform inadequately on large, heterogeneous datasets. I am planning to 1) demonstrate this issue in T1w MRI processing pipelines (i.e. FreeSurfer, CIVET), 2) develop a framework and tools for reporting pipeline performance beyond quality control, and 3) will demonstrate that optimizing tools over these values results in more stable, reliable, generalizable, and thus impactful pipelines. The successful completion of my Ph.D. has potential to transform the process by which scientists and tool developers quantify their analyses, and will lead to the production of more richly described, trustworthy scientific studies. I will describe the core efforts for each of these chapters below.

Chapter 1: Evaluate generalizability in T1w MRI cortical surface estimating processing pipelines

The first component of my project is to evaluate the quality and turn-key generalizability of structural neuroimaging pipelines. If a tool optimized on a small dataset fails to generalize to large heterogeneous datasets (i.e. derivatives are of significantly lower quality) statistical tests performed on this data may be untrustworthy. Thus, quantifying the stability and generalizability of tools is essential when transitioning to big data neuroscience. I will demonstrate the need for such a quantification for structural MRI processing. I will run FreeSurfer [6] and CIVET [7] cortical surface estimation pipelines on a small homogeneous dataset, the Nathan-Kline Institute Rockland Sample (NKI-RS), and optimize them based on accuracy as described through quality control (QC) figures and measures intrinsic to this pipeline, and quantify the variability of outputs over one or more features of the cortical surfaces, such as cortical thickness. This dataset was selected because of its high quality, ease of access, and membership of the CoRR initiative. I will apply the optimized pipeline to the remainder of the CoRR datasets, and evaluate the performance of the pipeline similarly to the above. While the CoRR datasets have been collected similarly, I expect this analysis will demonstrate that the remaining datasets, when run with parameters tuned for NKI-RS, will produce lower quality and more highly variable results. This will motivate the

following chapters by demonstrating that optimizing tools for accuracy on small homogenous samples is ineffective when scaling to large heterogeneous datasets.

Chapter 2: Develop a framework for reporting pipeline performance beyond quality control

While disciplines such as machine learning and computer vision expect the performance of tools to be quantified with sensitivity (portion of "true" entities identified) and specificity (portion of "false" entities rejected), such statistics are uncommonly computed in the context of neuroimaging. These statistics can be useful for describing not only the accuracy and appearance of derivatives as is done through quality control, but describes them in the context of hypothesis-testing when expected performance is known. The recent "Cluster Failure" [8] manuscript empirically computing false-positive rates – required for the computation of sensitivity and specificity – and reported alarmingly high rates for several tools. Extending this approach, I will develop a method for empirically computing sensitivity/specificity, and develop a combined feature of the two, tool-alpha values, through iterative bootstrapping and permutation testing. This technique will enable scientists to not only quantify the quality of their derivatives and results through quality control, but to summarize the trustworthiness of their processing. This tool will also be useful for tool developers who may run this analysis on phantom datasets which will provide a general intuition of the tools performance to users. I will demonstrate this approach for analysis and tool evaluation using the study conducted in Chapter 1, and expect to show that the higher variability in the heterogeneous analysis is captured in these measures and suggests caution in statistical analysis.

Chapter 3: Joint optimization as an effective approach for neuroinformatics tools

While neuroimaging pipelines aim to accurately capture variability within the brain, processing paradigms must be resilient to various sources of noise that are common across data. However, ensuring that tools are stable and generalizable comes at the cost of noise reduction or smoothing, which may negatively impact the accuracy of the results. To ensure that tools are successful both in terms of producing high fidelity results and generalizability, I propose the joint-optimization of accuracy (as determined by intrinsic quality control measures) and the tool-alpha value as generated by the paradigm put in place through Chapter 2. I will optimize the pipelines used in Chapter 1 considering both of these properties. I will deploy both the original pipelines used in Chapter 1 and the newly optimized pipelines on data from the UK Biobank, and will attempt to discriminate between populations based on the derivatives. I expect to demonstrate that the power of the study is increased in the jointly-optimized tool as compared to the original, and propose the extension of this principle to other heterogeneous datasets and tool, ultimately leading to more trustworthy and generalizable findings in big data neuroinformatics.

Additional Tools Developed

I will develop several additional tools in order to enable the work described above. *Clowdr* is platform which will enable the deployment of pipelines efficiently and quickly in the cloud, dramatically shortening the feedback loop between tool development and evaluation at scale. This will be a necessity for performing the analyses described above, as hyper-parameter optimization and deployment on large datasets may be computationally expensive. *Provenance Graph Analysis* is a method which I may develop for identifying sources of variability and instability inside pipelines, providing a higher-resolution evaluation of tool performance. Development of this tool is initially being undertaken by another student of Dr. T. Glatard, so it is currently unknown how much this will ultimately fit into my contributions.

Conclusion

The contribution of my Ph.D. will be the development of several tools and methods for deploying and evaluating neuroinformatics pipelines on big data. Working with FreeSurfer and CIVET, I will be in particular improve the quality of cortical surface estimation at scale and provide a principled way for other tools to be improved or deployed to large heterogeneous datasets. I believe the contributions of my Ph.D. will have the potential to improve the quality and scalability of neuroinformatics research as a whole.

Justification of Academic Background

Throughout my Master's degree I developed a structural connectome estimation pipeline from multimodal MRI (M3RI) scans of the human brain. I packaged this tool using containerization engines such as Docker and Singularity, documented it with the Boutiques descriptive command-line framework, and deployed it across a variety of computational infrastructures including Amazon Web Services, CBRAIN, and OpenNeuro. The packaging and deployment of this pipeline afforded me considerable experience in a variety of high performance computing environments, and required that I learn to evaluate pipeline performance across infrastructures and datasets and compare the quality of produced derivatives.

The pipeline I developed was optimized with respect to the reliability and stability of resulting connectomes, and provided the user with a variety of intermediate quality control figures to enable proof-reading of the results and ensuring their accuracy qualitatively (as no ground-truth exists, quantitative accuracy is not strictly-speaking achievable). The connectomes produced were also evaluated through a variety of graph statistical measures that operated on either an entire graph, its nodes, weighted edges, binary edges, or the largest connected component of the connectome.

During my tenure at Johns Hopkins I also taught several courses on applications of graph statistics, including Introduction to Connectomics, Statistical Connectomics, The Art of Data Science, and Neuro Data Design which was a design class for undergraduate engineering and computer science students who were tasked with developing and validating properties of their own neuroimaging processing pipelines that operated on Electron Microscopy, CLARITY, M3RI, or EEG data. I have significant experience in applying graph theoretical measures to biologically derived networks.

I have taken courses at a graduate level in computer vision, machine learning, data science, statistics, medical image processing, magnetic resonance imaging, and neural systems. I believe that my experience and education in image collection, processing, and pipeline development, evaluation, execution on HPC environments, as well as applications of graph statistics to real-world graphs will be invaluable assets towards me completing the project described and outlined above.

References

[1] York. Astronomical J. 2000 Sep;120(3):1579.
[2] Sudlow. PLoS med. 2015 Mar;12(3):e1001779.
[3] Van Essen. Neuroimage 80 (2013): 62-79.
[4] Zuo. Scientific data 1 (2014): 140049.
[5] Baker. Nature News. 2016;533(7604):452.
[6] Fischl. Neuroim. 2012 Aug 15;62(2):774-81.
[7] Ad-Dab'bagh, Y. OHBM. Florence, Italy, 2006.
[8] Eklund. PNAS (2016): 201602413.