Justification for Eligibility of Proposed Research

As summarized in my project proposal, during my PhD in Biomedical Engineering and tenure holding this award I will be developing infrastructure for enabling, communicating, and lowering the barrier for reliable and reproducible big-data neuroinformatics. While direct applications and demonstrations of my work will be in neuroscience, the core achievements and technical challenges I will address lie in computer science and the application of network statistics.

My work will focus on the creation of a platform to enable pipeline development, evaluation, deployment, and aid in the communication of scientific results. The platform I will develop will expose three modes of operation: *prototype*, *parallelize*, and *publish*. While the *publish* mode will expose a web-interface and RESTful Application Programming Interface (API) for users to query, the *prototype* and *parallelize* modes will operate as a "wrapper" around tools which will record provenance and execution information about the pipeline being executed, and enable the evaluation of pipelines though graph theoretical measures. Tools and environments which I will leverage to this end, include: Docker and Singularity (container-platforms for creating lightweight virtualized computational environments), Boutiques (a descriptive command-line framework), Reprozip (provenance tracing tool for pipeline record-keeping), Amazon Web Service (high powered computing cloud and data storage solution), and Compute Canada.

While existing platforms which manage high performance deployment of pipelines (e.g. CBRAIN) and data management (e.g. LORIS) in the neuroinformatics domain provide exceptional functionality to users, there is a significant barrier to using or managing these resources for researchers on their own. The tool developed through my project will drastically lower the barrier to entry for deploying pipelines and distributing datasets at scale, as well as providing an interface with these existing platforms (and others) so that researchers may leverage existing datasets and tools in their analyses. Through the development of an extensible common API, my tool will be able to leverage with core functionality of these systems and be flexible enough to easily connected with systems using significantly different APIs but aim to solve similar challenges, such as XNAT, Datalad, or COINS.

Another essential component of my project is the evaluation of the stability and reproducibility of pipelines through the analysis of provenance graphs created at runtime. While previous literature and my background in statistical connectomics provide a starting-point for graph measures which may highlight relevant differences in pipeline executions, thorough exploratory analysis of commonly used and tested pipelines, both considered "reproducible" and "non-reproducible," will be necessary to determine effective measures in this context. My experience in neuroimaging as both a tool developer and user will be invaluable towards identifying a variety of pipelines representative of a wide range of data and computational properties in neuroinformatics throughout this analysis.

As described above, the work undertaken in this project is technical in nature, and my achievements will be of benefit to NSE. During my tenure at McGill University I will be working with Dr. Alan Evans and Dr. Tristan Glatard, who have appointments in the Biomedical Engineering and Computer Science departments, respectively. Both of my supervisors have considerable experience developing scalable infrastructure for informatics applications, including Alan's experience developing the CBRAIN platform for cluster management, and Tristan's experience developing the CARMIN API for launching pipelines and managing data and the Boutiques framework. Their collective experience will be invaluable to me in this project.