NeuroHub vision document

Vision Statement: Neurohub will be a robust, functional, flexible and well-supported technical and social platform for students, researchers, and data scientists from the McGill community. It will allow them and their collaborators to:

- find, manage, share, store, annotate, document and curate multi-modal, multi-scale, and multi-disciplinary data
- find, work with, execute and get training for analytics tools
- manage and publish their research
- empower laboratories and researchers for accomplishing highly collaborative and reproducible computational science where results are captured like primary data and can be used for further research and publication

Purpose:

Neurohub was primarily envisioned as the overarching data and computational for the Healthy Brains, Healthy Lives program to support researchers in their work, collaborations, and infrastructure needs. The platform will be a place where McGill labs can manage their projects, data, computational work, and collaborations to build a strong data-sharing research community.

Neurohub will provide both a technological and community-based solution with two major foci:

1. Fostering highly multidisciplinary computational research between traditionally siloed communities for creating solutions in the challenging problems surrounding the mental health and well-being of Canadians.
2. Providing an environment that increases confidence in computational research by enabling reproducible scientific research, provenance information, data-sharing and publication.

Neurohub will serve as the place for HBHL researchers (and eventually external researchers) to place their own data, get access to a bevy of other research data managed, and share data with others in a completely provinenced platform. Neurohub will provide computational facilities for both interactive, small-scale computations as well as large, managed computational workflows, with all of the derived results being imported back into Neurohub for provenance, further research, and sharing with collaborators and in publication.

NeuroHub community

NeuroHub is a community project, acknowledging that without a strong integration with the neuroscience communities a platform risks to not be adapted to the needs. NeuroHub therefore requires a strong interface layer with the scientific community (students, researchers, PIs), communication and training components. NeuroHub also considers that it cannot (and should not) live in isolation from other large international neuroinformatics projects. NeuroHub will work on the interoperability and the development of standards and best practices to foster the emergence of global solutions across the neuroscience and life sciences communities.

HBHL User's Needs

Our team has conducted an in-depth user needs and requirements assessment across HBHL researchers and multiple disciplines. Some of the gaps that we have found through this assessment are:

- Neurohub needs to provide a place where laboratories can place and manage their data so that critical research and knowledge is not lost between generations of researchers and students.
- Neurohub researchers are using a wide diversity of computational tools including Jupyter Notebooks, Python, Matlab, R, Excel, and predetermined pipelines. Neurohub researchers need access to computational and data resources (generally more than is available to individual laboratories) that are connected to these tools in an easy to use manner that allows them to manage their research and collaborate with others.
- Neurohub researchers have a diversity of data security and data sharing requirements and need for the platform to support a spectrum from completely closed to completely open data sharing policies.
- The Neurohub community has need for a web-based portal that allows them to manage their data and computation as individuals, with various organizations (e.g. their laboratories), and with the external community.
- The Neurohub community needs to have an ability to find and access datasets that are currently unavailable to them whether it is due to technical barriers or just not knowing what data is out there.
- The Neurohub community needs help managing the complexity of highly-multidisciplinary computational research with methodologies that allow the work to be done in a reproducible manner with full provenance of the data generation and derived work.
- The Neurohub community needs help in creating standards around their data to enable it to be shared and used in novel analysis such as Artificial Intelligence / Machine Learning techniques.
- The Neurohub community needs training on the latest in data management and computational techniques.

NeuroHub development philosophy

NeuroHub development philosophy is to integrate a set of existing, and community-driven whenever possible, technologies guided by the goals of the platform and the needs of the HBHL community. Anchored by a strong set of existing core platforms such as CBrain and LORIS, our development will focus on how to connect existing components to form an ecosystem targeting the use cases of the McGill neuroscience community. As much as possible, NeuroHub is trying to re-use and not re-invent existing technologies, establishing partnerships when this is the most efficient solution. Our team has vetted several existing technologies based on design and needs for robustness, appropriateness, and ease of customization. NeuroHub developments are goal and use case oriented, bringing solutions to specific problems while developing generic components. Finally, Neurohub is not a hardware-centric project in that, we will be utilizing third-party computing and data storage services to reduce deployment costs as well as to ensure sustainability,

NeuroHub Architectural Components

The Neurohub infrastructure will be comprised of four primary conceptual components.

- The Project Portal Space: This will be the web-based user-facing component of the platform and will provide users a place to login, manage their collaborations, data, and computations. As the primary interface with Neurohub, uses will be able to upload their own data as well as access other shared datasets stored in the *DataSpace*. Each Neurohub user will have a user account and a set of organizations that give them the ability to define how others access their data as well as group data together in one place that they need for their own research. Not only will the Project Portal Space be the access point for users to all other areas of Neurohub, it will also serve to inform them of other research going on throughout the HBHL community and connect them to datasets that will potentially enhance their own research. In addition to data, users will be able to store important documents, add notation to their work, and communicate with other researchers, as well as easily export their research to interactive publication platforms
  Candidate Technology: Open Science Foundation Portal

- The Data Space: This will be the canonical location for cataloguing, accessing, and storing all manners of data in the Neurohub architecture. The Data Space will be primarily populated with metadata providing as standard as possible representations of all of the different data elements that are accessible to users of the platform. Data itself can and will be stored in a multitude of different actual locations (e.g. Compute Canada, GoogleDrive, local hard disk, etc.) and the metadata will include the methodology to actually get the data if needed and move it around the Neurohub ecosystem. The Data Space will also be able to connect with data stored in existing data basing systems, such as LORIS. All data that is created by NeuroHub through analytics will also be placed

back in the Data Space with appropriate standardized metadata that will allow it to be reproduced making these data Findable Accessible Interoperable and Reusable (FAIR); used in other research; and exported to publication. Finally, the Data Space will be protected by an external security layer that will allow users to define specifically the access rules to their data and computation within the platform.
Candidate Technology: Datalad, LORIS

- The Workshop Space (aka the "kitchen"): This is the place where data from the Data Space can be accessed to perform interactive, exploratory computing and analytics. Through both the Project Space and native means, users will have access to spin up interactive processing tools both locally and on remote resources to work intimately and interactively with data in the dataspace. This would include, but not be limited to, tools such as Jupyter Notebooks, RStudio, and command line tools. This space will provide a clear catalogue of available tools as well as a well documented API to access data in the Data Space.
  Candidate Technologies: JupyterHub, BinderHub, RStudio

- The Factory Space: This is the place in Neurohub that will facilitate managed, big-data "analytics". This space is intended for computation that involves a large number of very well-defined data types and a well-defined computational pipeline or workflow that a user would like to perform on them. The Factory Space will provide federated access to a number of computational resources within Compute Canada, at McGill, and in the cloud to provide the necessary computing needed by HBHL researchers. Users will be able to log into the Project Space to define, execute and retrieve computation. All results computed in the Factory will automatically be annotated and placed back into the Data Space for the users to continue further analysis, share results, and potentially use in publication.
  Candidate Technologies: CBRAIN, Boutiques

GAP analysis

NeuroHub acknowledges a set of gaps in the neuroscience community and plans to eliminate or mitigate these. First, irreproducibility has plagued the literature in particular in the biomedical and life sciences fields. NeuroHub will implement provenance and versioning both for data and processing pipelines.
Second, collaborations and building community based tools have been difficult or slow, partly because the incentives to document and share research objects are not necessarily in place. NeuroHub will work on creating tools and processes to ease collaborations across neuroscience fields.
Third, the reuse of research objects (such as data, protocols, software, pipelines) is hampered by lack of findability, accessibility and interoperability. One critical aspect is the capacity of the community to document data and pipelines with standard terminology or ontologies. NeuroHub will work to make neuroscience research objects FAIR.

Fourth, the incentivisation of sharing and making research objects FAIR is still poor. By helping the development of publication platforms that will give credit to the dissemination of research objects beyond the article pdf, NeuroHub will foster a culture in which data or software are FAIR and publishable research objects.
Last -but not least- the computations needed to process complex or large datasets in a machine learning framework are still a significant barrier to many neuroscientists. NeuroHub will strive to streamline and simplify the process by which computations are scaled while keeping provenance information.

Responsibilities

Jean-Baptiste

- **Interface with the scientific community (within and beyond McGill)**
- **Metadata for datasets, tools for FAIR data, data documentation and catalogue**
- **Use cases priorities, user testing (shared responsibilities)**
- **Communication, outreach and link with publishing**
- **Training (neuroinformatics, platform tools)**
- **Liaison with other international projects and initiatives and interoperability**
- **Community building**

Shawn

- **Manage and supervise the Neurohub development team**
- **Staff and hire Neurohub Developers**
- **Develop overall Roadmap and priorities for technical development**
- **Technical platform decision making (i.e. technologies for individual components of the platform)**
- **Manage progress toward completion of development**

Shared

- **Manage User Testing**
- **Prioritize use cases and develop user stories**
- **Represent Neurohub to the RMC and HBHL governance**
- **Manage budget priorities and adjustments**

Platform Goals:

Reproducible Science - The platform should remove data sharing barriers to facilitate reproducibility studies.

- Matchmaking for reproducibility studies
- Capture the provenance of all studies
- Improve interoperability of data between different analytic tools
- Facilitate registration of study protocols (“preregistration”)

Collaborative, Cross-disciplinary Science - The platform should reduce cross disciplinary barriers of data sharing and analytics to synthesize new hypothesis.

- Share data in machine readable, standardized representations.
- Standardize analytic tools allow them to use the data in combined manners.
- Provide a collaborative environment that allows scientists to share findings and data.
- Facilitate neuroimaging and neuroinformatics, as well as other disciplines.
- Proliferate computing/analytics beyond the big-data, HPC-centric competent.

Organize and store data - Provide a sphere for researcher to house data.

- Ensure that data is properly captured (standardization and organization)
- Ensure that data is properly findable and usable for analytics

Open-Science Publication - The platform should enhance automation thereby enabling researchers to share data openly for data publishing.

- Provide full gambit of study data (primary, derived, provenance & analytics).
- Modality-based focus on usability to make it easy for scientists to publish.

Overarching philosophy:

✪ **Integration:** Look first to existing technology to meet requirements.
✪ **Development:** Develop core feature, while also engaging users to support a community.
✪ **Design:** The platform to be built from use cases, not an imposed top down platform.
✪ **Iterative development:** Each phase to build on prior work, and engage more users.
✪ **Open-by-Design**: Use open source, publicly available software.
✪ **Version control**: Use Github to maintain software (in some cases temporarily private).
✪ **User support:** A key idea to be prioritized throughout the development process.
✪ **Strong collaborative environment:** Engage with the international community,
✪ **Leverage assets:** i) awarded projects, ii) other synergistic projects at McGill, iii) well established infrastructure projects, iv) expertise in technical infrastructure and coordinating.