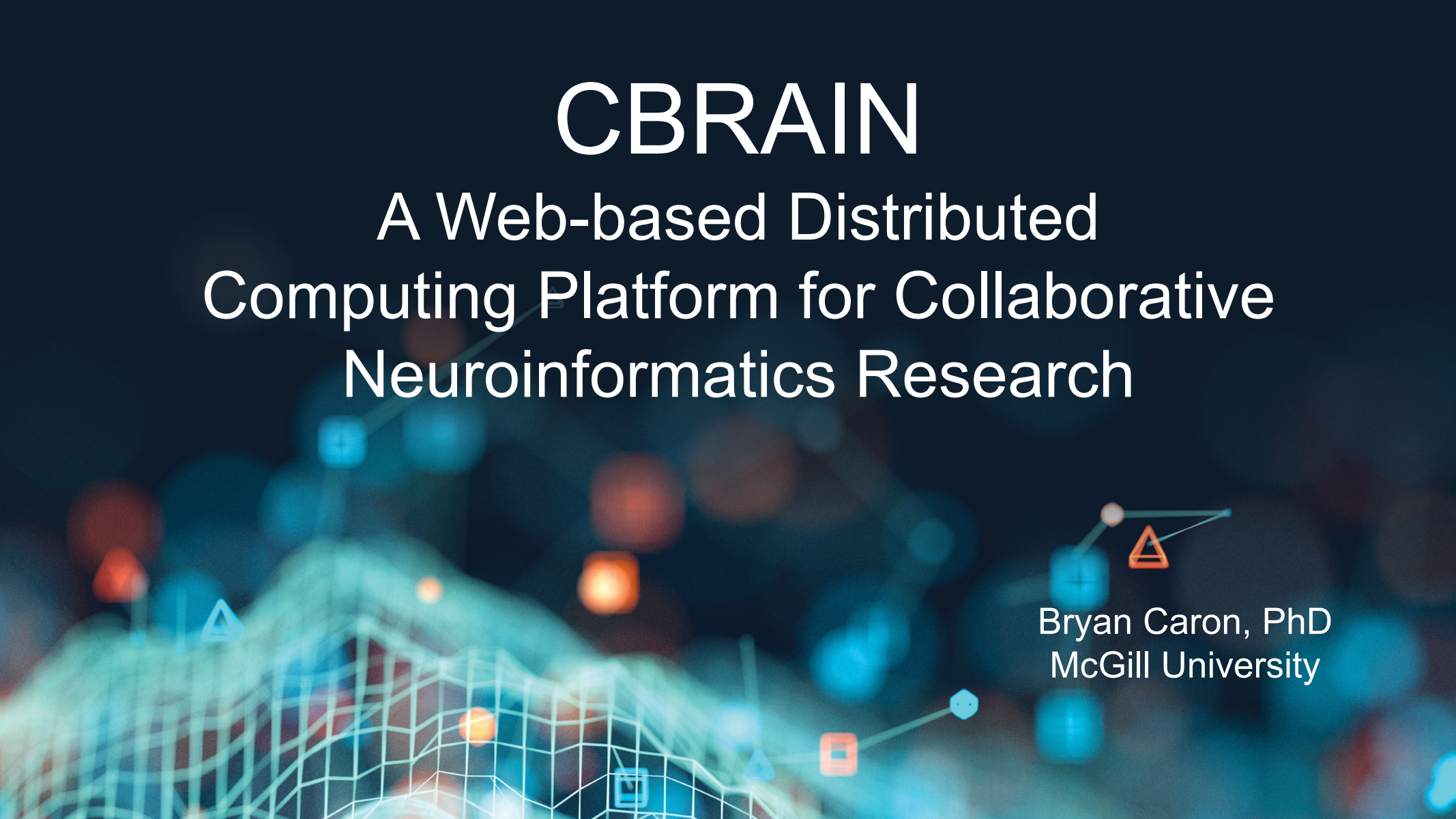
CBRAIN

A Web-based Distributed Computing Platform for Collaborative Neuroinformatics Research



Bryan Caron, PhD
McGill University

Outline

- The motivating challenges of neuroinformatics and big data science
- CBRAIN for Big Data Processing and Workflow Management
- The software ecosystem for reproducibility and interoperability

The Challenges in Neuroinformatics


Big Data


Big Computation


International Collaboration


Interdisciplinary Health Research

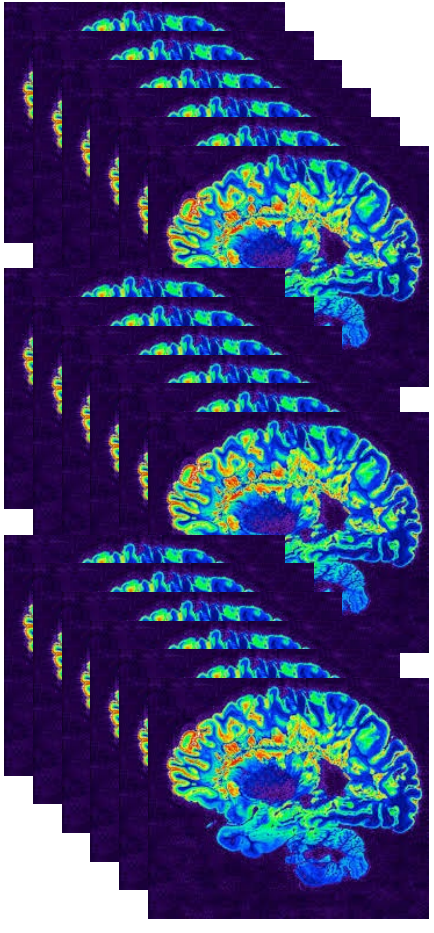
What does it take?
Infrastructure and services for accomplishing science

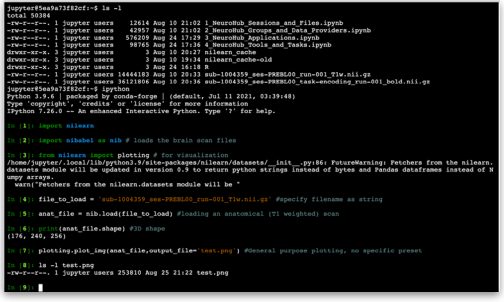
Longitudinal Acquisition, Storage and Curation, Interoperability, Reproducibility, Transfer, Anonymization, Security, Privacy, Ethics, APIs, Validation, Quality Control, Protocol Checking, Preprocessing, Analysis, HPC, Cloud,

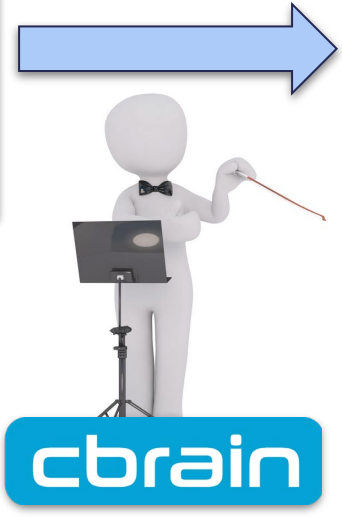
Provenance, Ontological Standardization, Data Harmonization, Upgrades, Maintenance, Bug Fixes, User Interfaces, Bootstrap, Tracking, Extensibility, Data Management, Summary Statistics, Workflows,

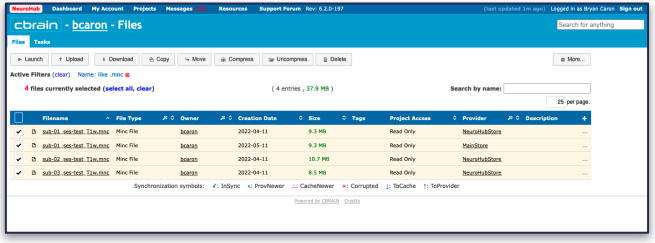
Development, Tool Integration, Data Sharing, Download, Multi-Modal Linking, Querying, Image Processing, Visualization, Networking, System Administration, Partnerships, Funding, HR

Big Data Processing









CBRAIN orchestrates large amounts of data and computation running on remote High Performance Computing and Cloud resources.

Makes big science easier, manageable, reliable and reproducible!

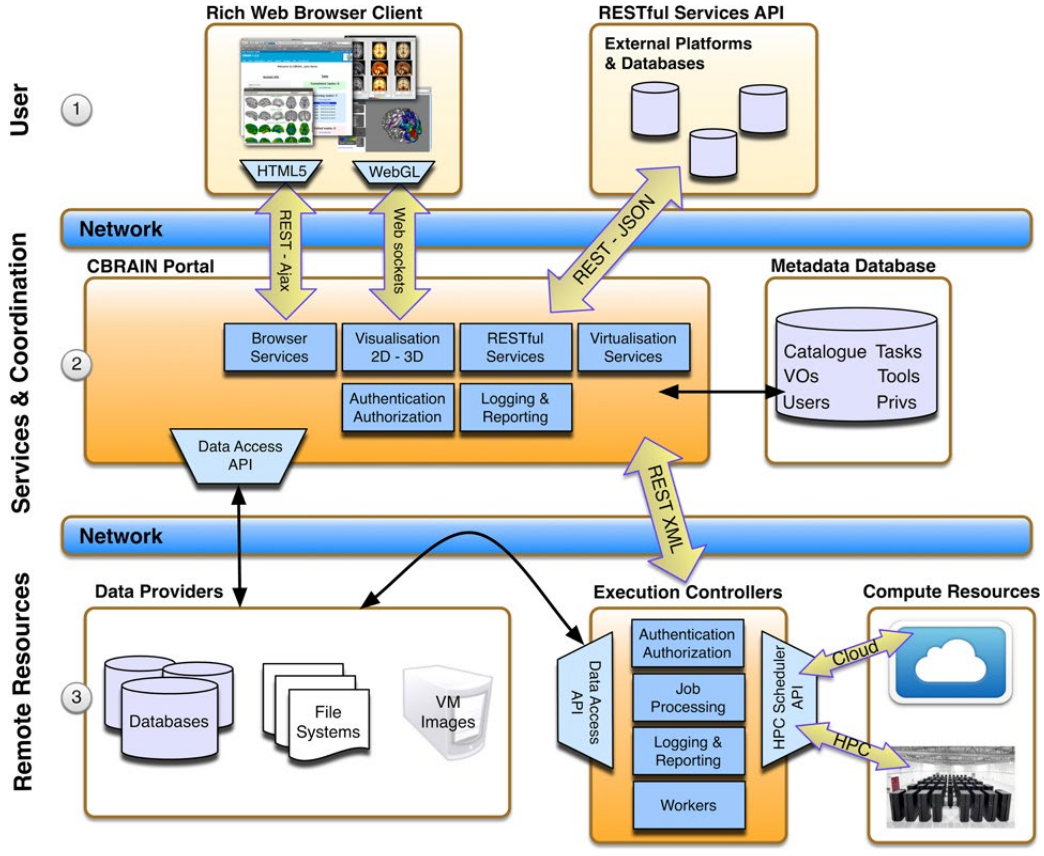
Big Data Processing

cbrain

orchestration

containerization

standardization



Sherif T, Rioux P, Rousseau M-E, Kassis N, Beck N, Adalat R, Das S, Glatard T and Evans AC (2014) *CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research.* 5 Front. Neuroinform. 8:54. doi: 10.3389/fninf.2014.00054

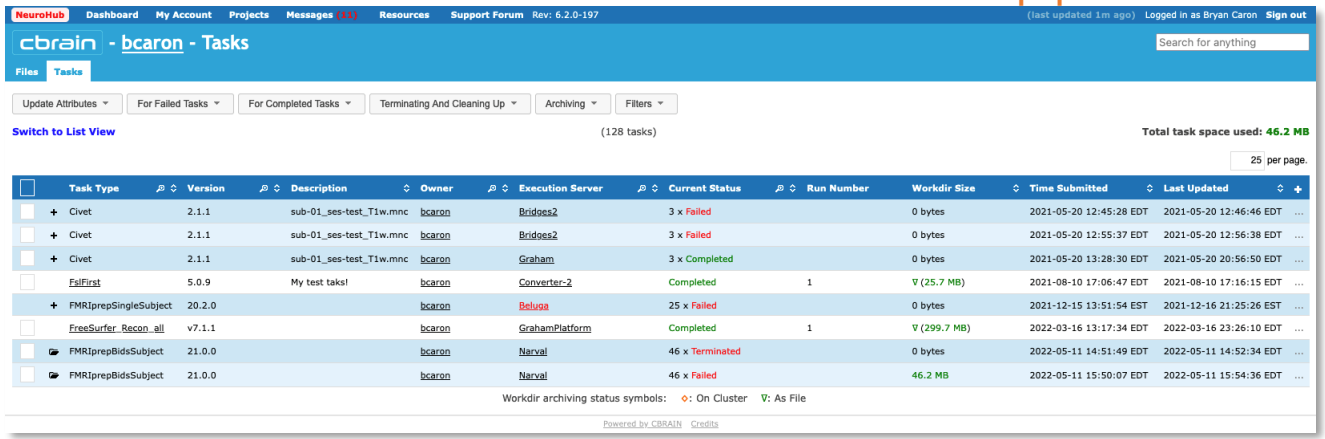
Features of CBRAIN

- Convenient, secure web access and API
- Lightweight core components, low requirements for deployment and operation
- Distributed storage with automated, multipoint data movement, and cataloging
- Transparent access to research tools and computing (HPC and cloud)
- Flexibility to adapt to extremely heterogeneous computing and data sites
- Full audit trail (data provenance) and logs across all user actions
- Scalability (no architectural bottlenecks)
- Full ecosystem security and monitoring

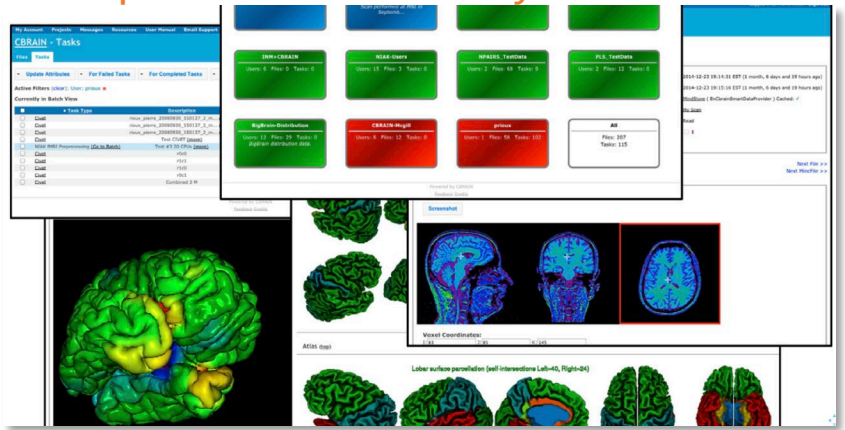
https://github.com/aces/cbrain

Features of CBRAIN

Enables distributed execution of software pipelines



Aggregates multiple distributed file systems into uniform view

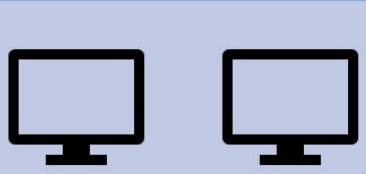


Why use CBRAIN?

cbrain

Access to Big Data and Big Computing

- Easy access to a full range of lab-based through to large-scale ARC resources including Alliance systems
- Access to a diverse range and scale of datasets and storage for high-speed, high throughput computing



Reproducible science

- Deployed, standardized, and tested over 115+ software tools
- Software runs exactly the same, no matter where it executes

Scale up your science

- Web-based platform makes it easy to execute large computation easily
- High degree of parallelism can accelerate your science

Collaborate and Share your science

- Work in a virtual organization sharing data with your team anywhere in the world
- Make your pipelines and results available to others

What is Boutiques?

- Fully-automated integration of applications
- Deployment on heterogeneous computing resources through containers
- Comprehensive input validation through a strict JSON schema
- Flexible application description through a rich JSON schema

https://github.com/boutiques

OXFORD (GIGA)n SCIENCE

Giga Science, 7, 2018, 1–11

doi: 10.1093/gigascience/giy016
Advance Access Publication Date: 23 March 2018
Technical Note

TECHNICAL NOTE

Boutiques: a flexible framework to integrate command-line applications in computing platforms

Tristan Glatard[1,*], Gregory Kiar[2,3], Tristan Aumentado-Armstrong[2,3], Natacha Beck[2,3], Pierre Bellec[4], Rémi Bernard[2,3], Axel Bonnet[5], Shawn T Brown[2,3], Sorina Camarasu-Pop[5], Frédéric Cervenansky[5], Samir Das[2,3], Rafael Ferreira da Silva[6], Guillaume Flandin[7], Pascal Girard[5], Krzysztof J. Gorgolewski[8], Charles R.G. Guttmann[9], Valérie Hayot-Sasson[1], Pierre-Olivier Quirion[4], Pierre Rioux[2,3], Marc-Étienne Rousseau[10] and Alan C. Evans[2,3]

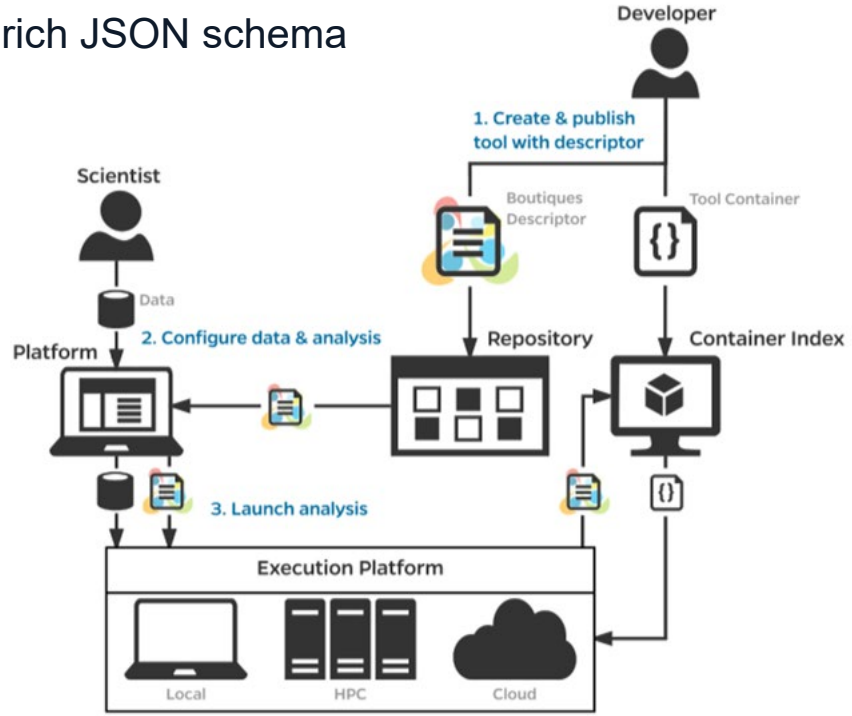


Figure 1. Publication, integration, and execution of applications with Boutiques.

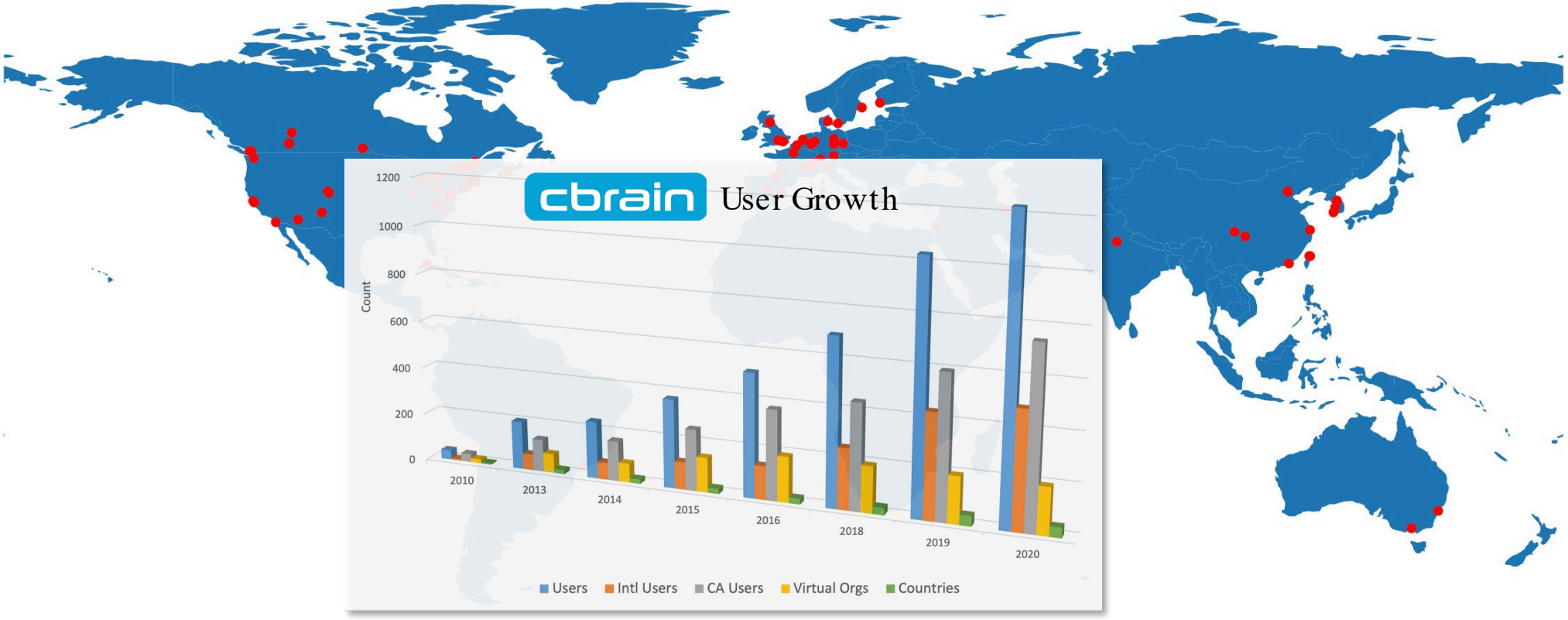
Software Tools

cbrain

cbrain 115+ Tools

- fMRI Analysis
- fMRI Preprocessing
- Structural MRI
- Genetics
- Tractography
- PET
- Arterial Spin Labeling
- EEG
- Data Conversion Tools
- Statistics

Global User Community



1400+ users, 193 sites in 32 countries

Connected Resources

cbrain

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

JÜLICH Forschungszentrum

XSEDE

SURF SARA

PSC

SDSC SAN DIEGO SUPERCOMPUTER CENTER

TACC TEXAS ADVANCED COMPUTING CENTER
Powering Discoveries That Change The World

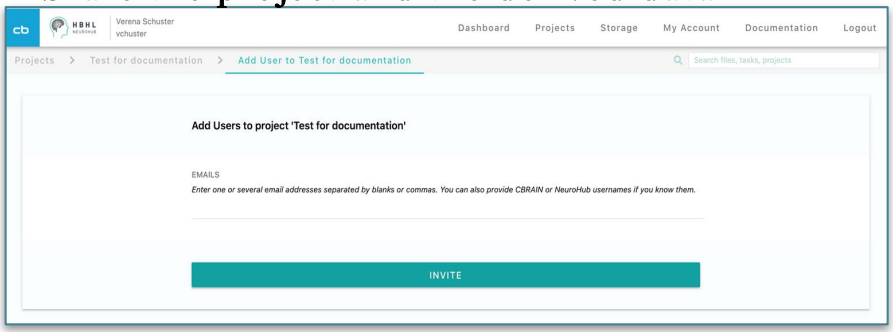
UNIVERSITY OF MINNESOTA
Driven to Discover®

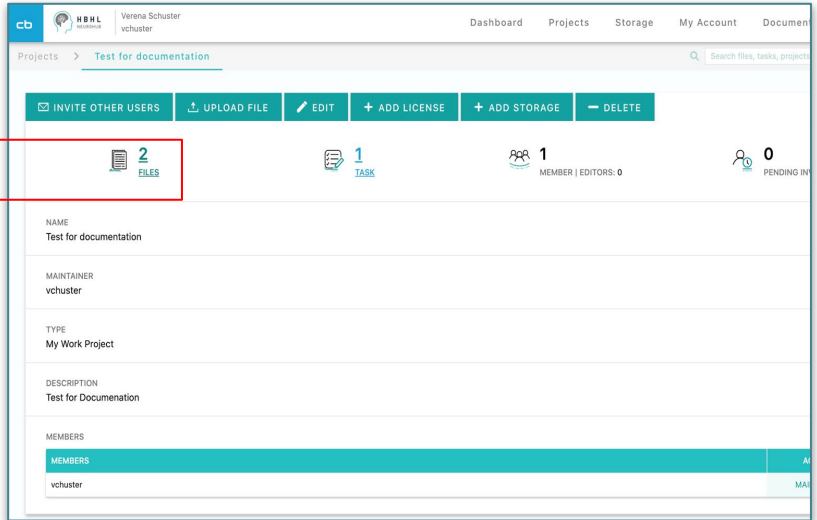
200k+ tasks launched for >5M+ CPU hours processed since 2020
44M+ files, 415 TB registered files currently in CBRAIN

aws

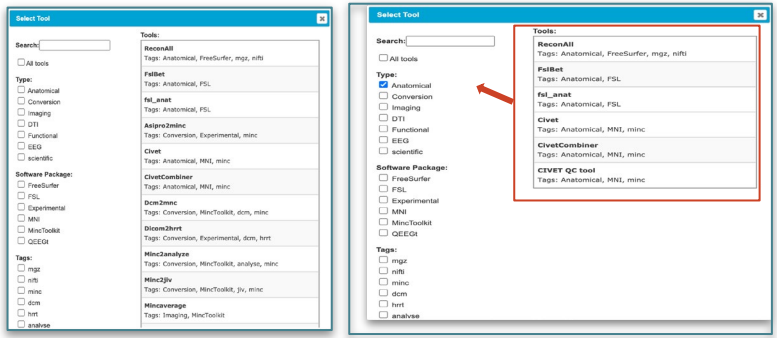
Process, Store and Share Data

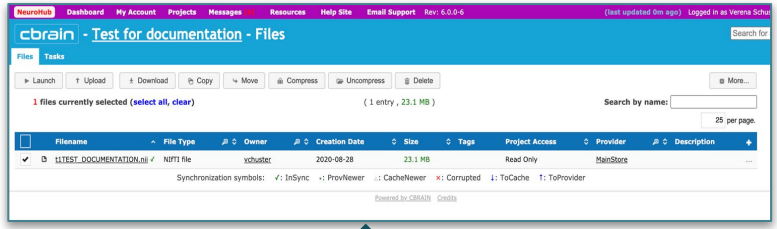
Share the project and the derived data

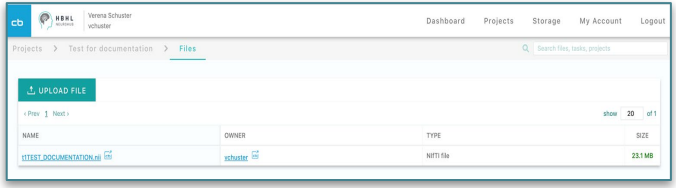




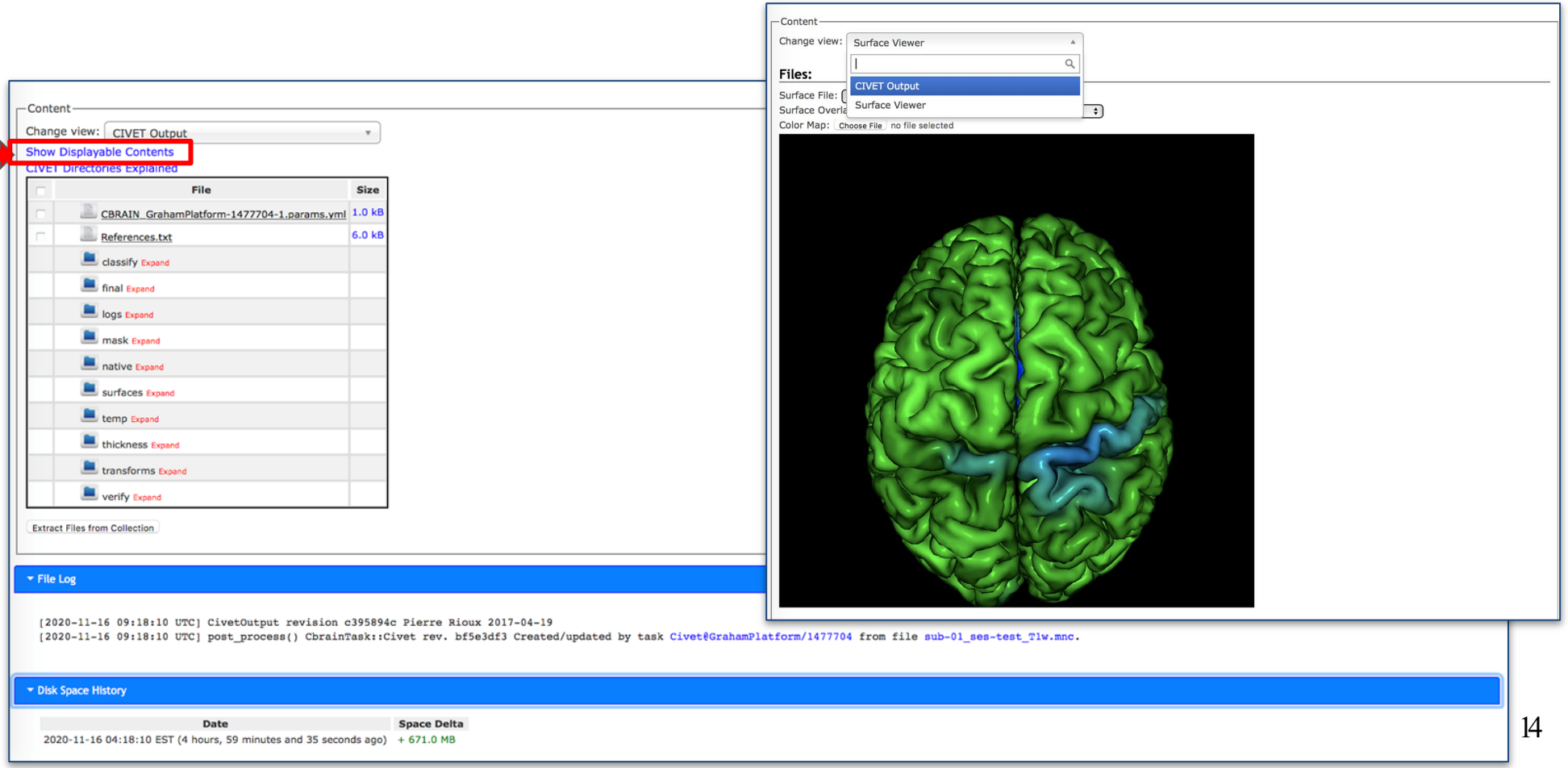
Process data using a variety of tools





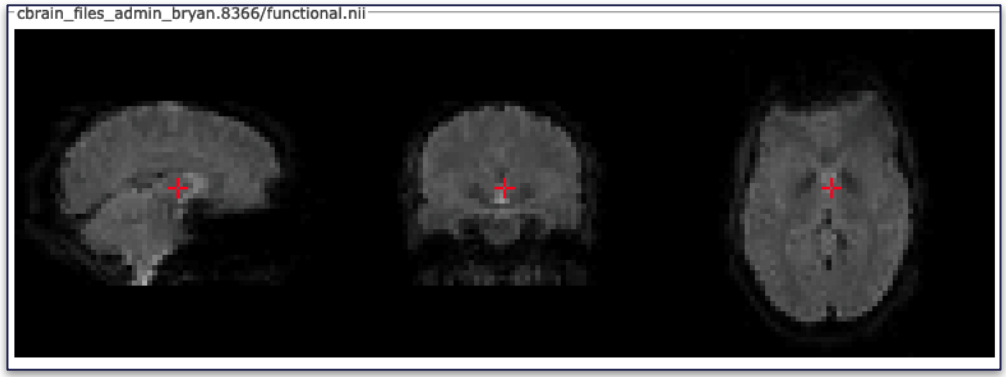


Data Processing and Visualization

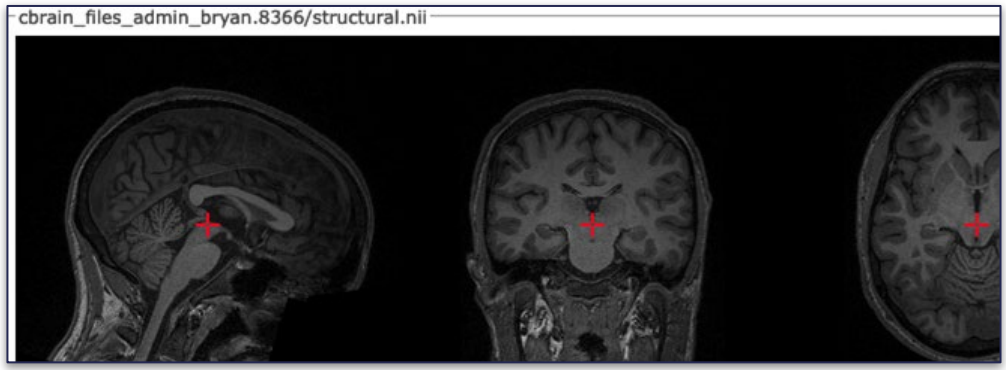


Data Processing and Visualization

Functional



Structural

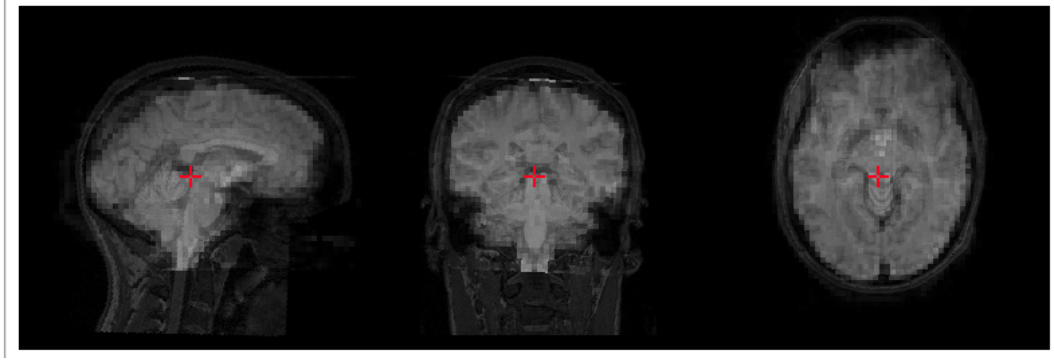


BrainBrowser

cbrain

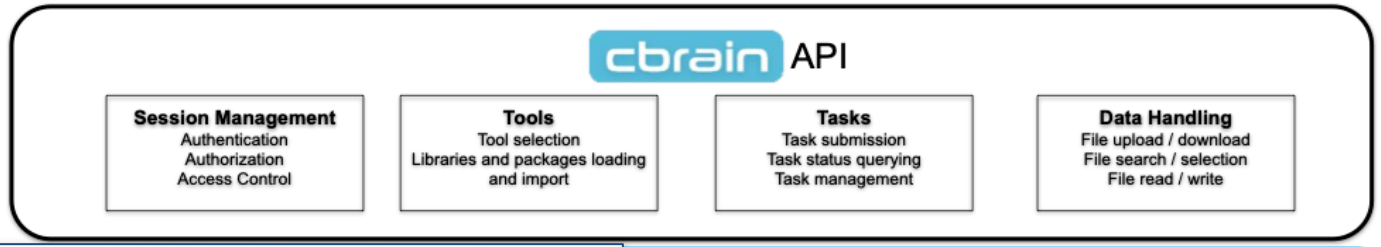
	File	DL	Size
☐	functional.nii	⬇	40.4 MB
☐	structural.nii	⬇	46.1 MB

Overlay



Interactive Data Exploration

- CBRAIN API - users can interact programmatically with data and tools



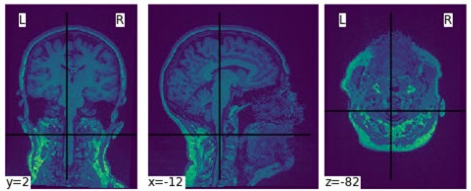

```
In [24]: file_to_load = 'sub-1004359_ses-PREBL00_run-001_T1w.nii.gz' #specify filename as string
         anat_file = nib.load(file_to_load) #loading an anatomical (T1 weighted) scan
         print(anat_file.shape) #3D shape

         file_to_load2 = 'sub-1004359_ses-PREBL00_task-encoding_run-001_bold.nii.gz'
         bold_file = nib.load(file_to_load2) #loading an fMRI (BOLD) scan
         print(bold_file.shape)#4D shape

         plotting.plot_img(anat_file) #General purpose plotting, no specific preset
         plotting.plot_anat(anat_file) #Anatomical image plotting

         (176, 240, 256)
         (64, 64, 32, 183)

Out[24]: <nilearn.plotting.displays.OrthoSlicer at 0x7f352eae9b70>
```



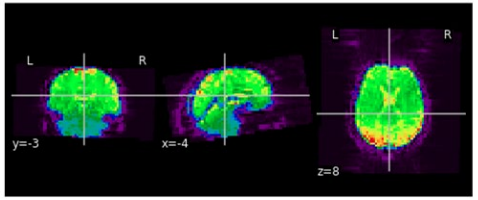
5. fMRI visualization

```
In [26]: from nilearn.image.image import mean_img

         mean_fmri = mean_img(bold_file) #Compute the mean of the images over time or the 4th dimension

         from nilearn.plotting import plot_epi, show
         plot_epi(mean_fmri) #plot cuts of the image (by default 3 cuts: Frontal, Axial, and Lateral)

Out[26]: <nilearn.plotting.displays.OrthoSlicer at 0x7f352ecce908>
```



Access to Large Datasets

biobank uk
Improving the health of future generations

- Centralized and efficient handling and processing
- 40000 participants with imaging data + tools to load into PyTorch
- 200000+ participants with genetics and phenotypic data

C-BIG Repository

+ tools for genetics

More than 3000 patients including those with Parkinson's and ALS, with genetic data accessible via NeuroHub

Human Connectome Project

CONP PCNO

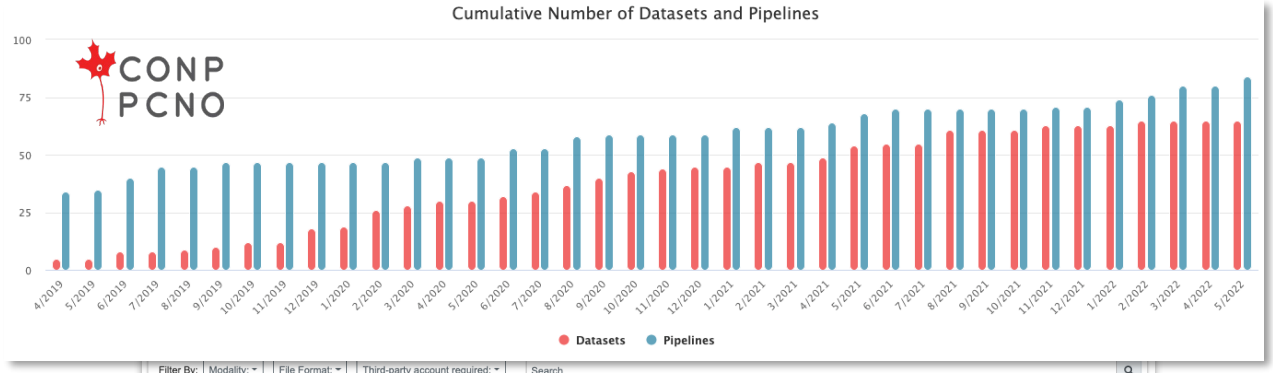
(50+ datasets)

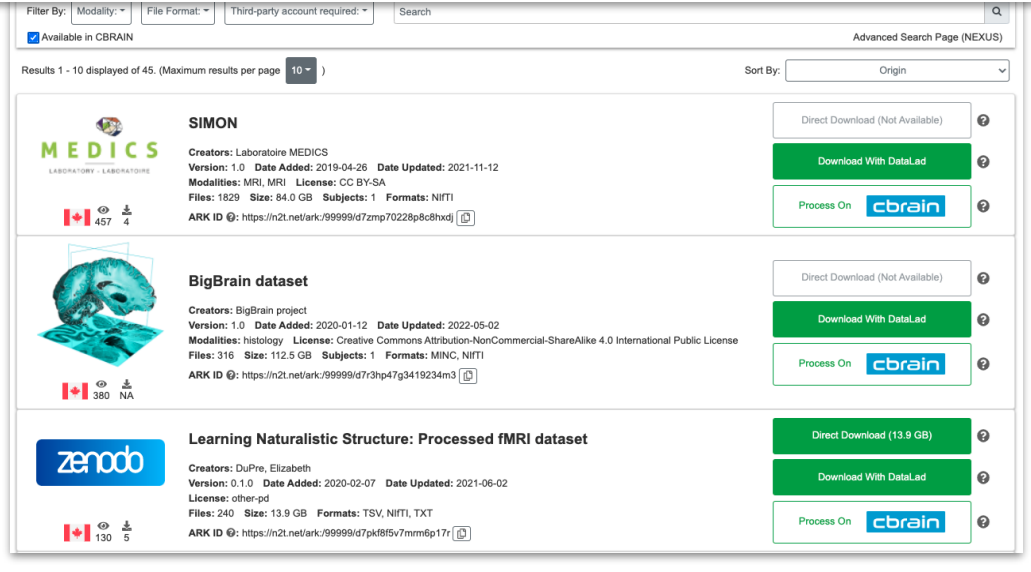
HBCD (HEALthy Brain and Child Development) Study

- 7500 participants study on the effects of opiate addiction

- 12000 participants
- NIH wants major NeuroHub components within its data curation, dissemination plan

Canadian Open Neuroscience Platform (CONP)





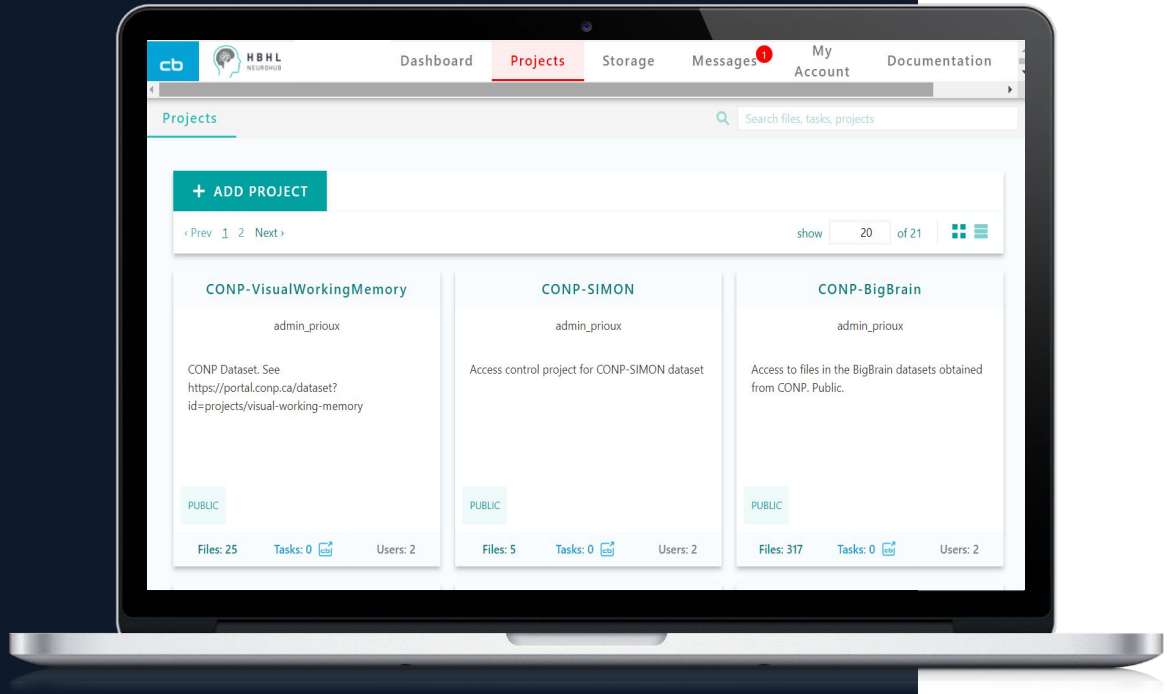
Recent Developments

- Some recent highlight features:
  - New Boutiques integrator - more modular, easier maintenance and tool integration
  - Users can link their account using Globus Auth for Federated Identity Management
  - Data Providers access files in SquashFS / ext3 overlays with singularity bindmount
    - large datasets are then friendly to HPC file systems
  - Fully supported S3 Data Provider configurable with distinct regions and endpoints
  - User can create API tokens for connections from external applications
  - User definable networked storage resources managed through SSH keys
  - User storage resource quotas on storage provided centrally by CBRAIN

Platform interoperability is key to building researcher communities and their success

NeuroHub

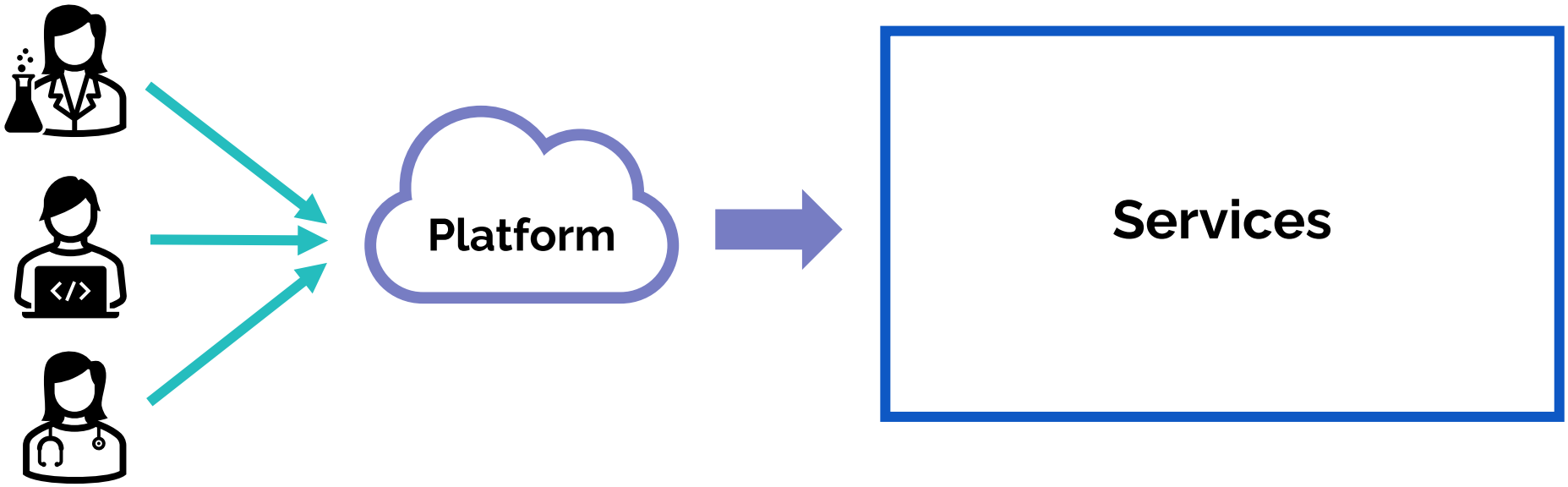
What is NeuroHub?



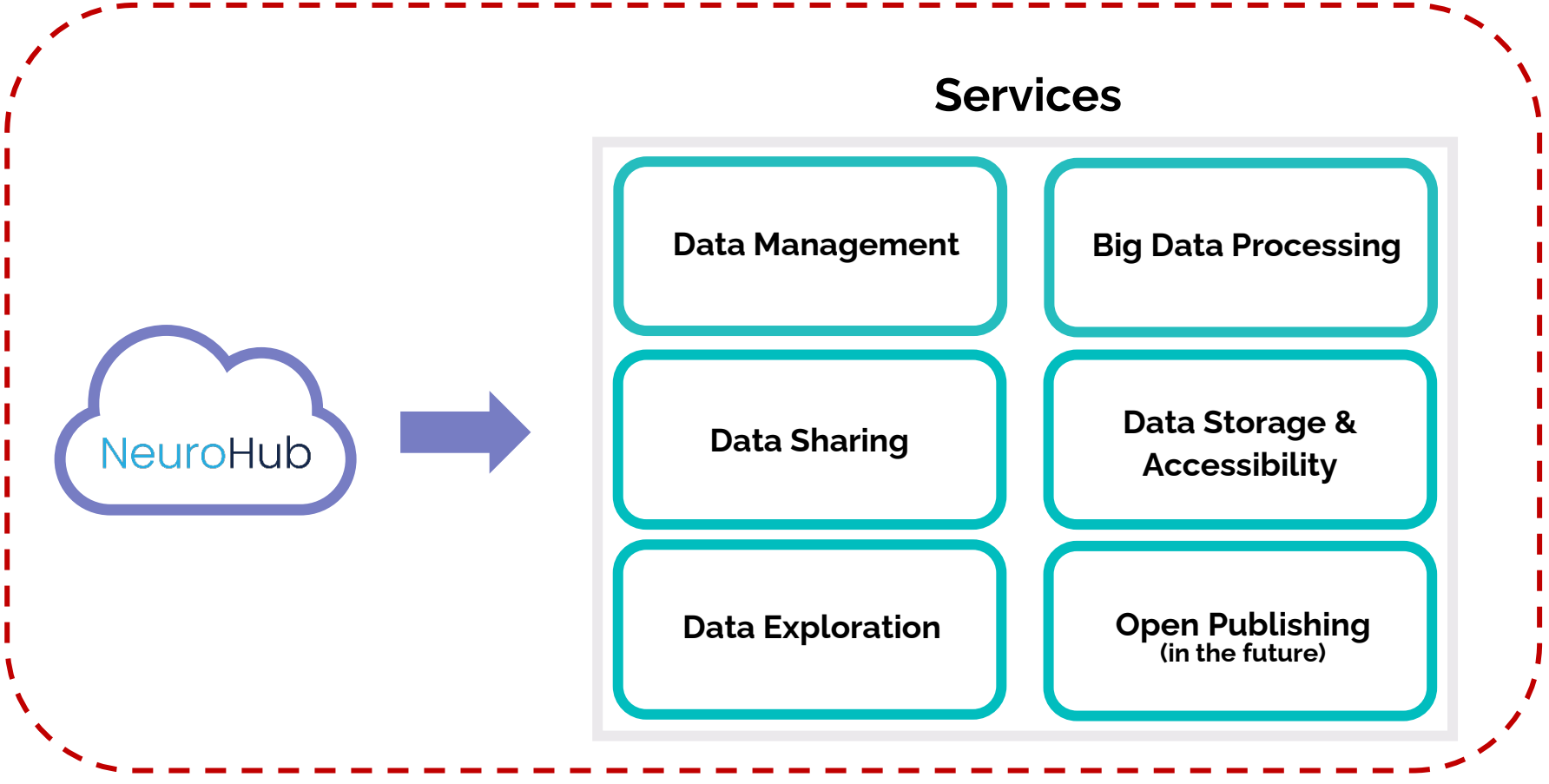
Multi-modal data management

Data analysis

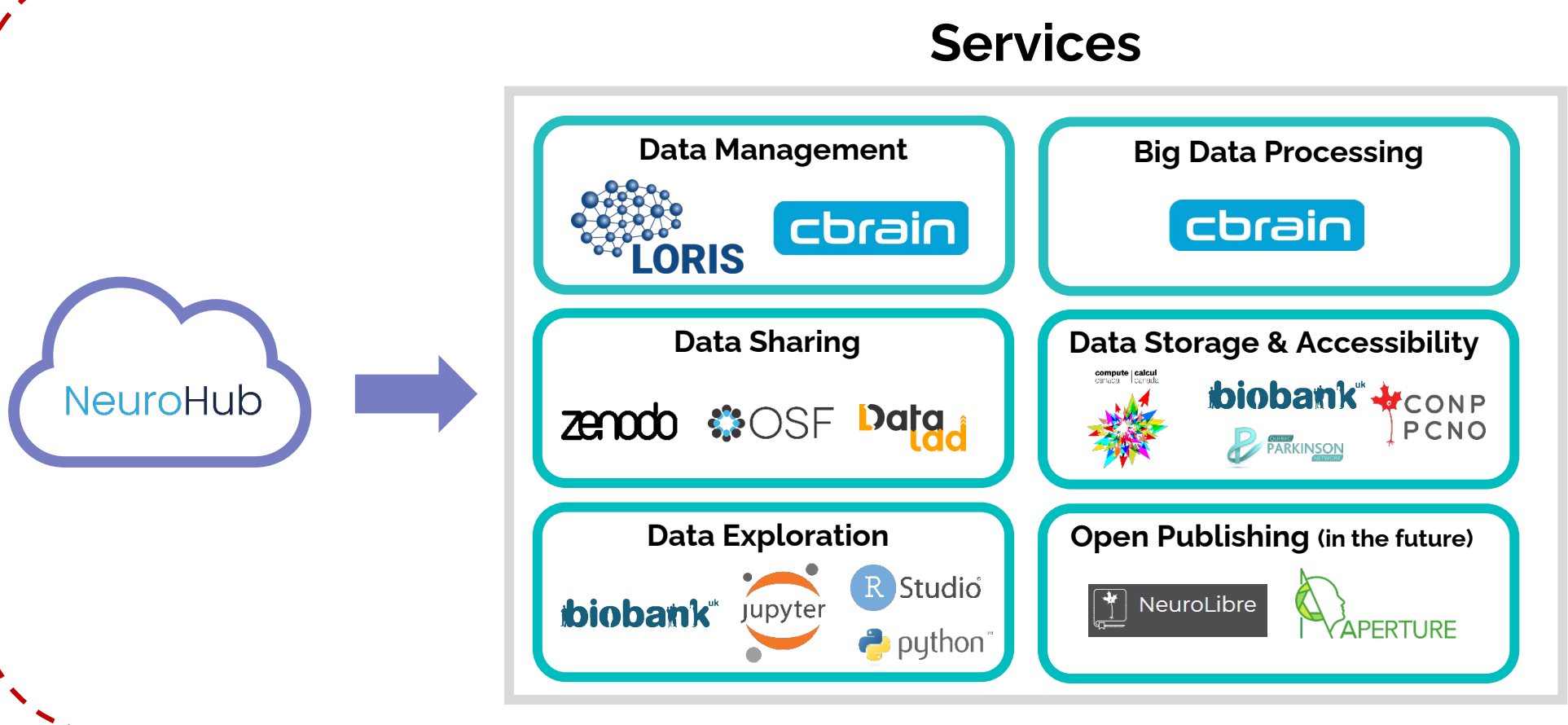
Data Storage & Sharing


Platform
Services

NeuroHub Ecosystem



NeuroHub Ecosystem



NeuroHub Ecosystem Features

Easy-to-use interface and rich API

Securely store and share data

Access to high performance computing

Access large international datasets

Capture and curate multi-modal data

Datasets suitable for machine learning

NeuroHub for Reproducibility and Interoperability

Which dataset version you worked on

Know which software version you have used

Be able to relaunch on new datasets : Replicability

Cost and maintenance aspects

Leverage expertise that cannot be built locally

Build the bridges across communities and across platforms

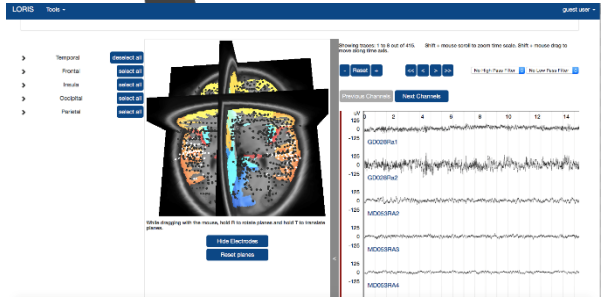
Large Scale Projects

C-BIG Repository

Canadian Consortium on Neurodegeneration in Aging

IBIS

EEGNet



CONP PCNO

Adolescent Brain Cognitive Development®

Teen Brains. Today's Science. Brighter Future.

THE HEALTHY BRAIN & CHILD DEVELOPMENT STUDY

Future Directions

- Additional CBRAIN Data Provider types to support cloud-based storage
  - (e.g., Google Drive / Cloud, Microsoft SharePoint, ..)
  - Update the CBRAIN Portal web interface to improve and simplify the user experience
    - Cleaner and more intuitive workflows
    - Enhanced self-help functionality
    - Updated look and feel based upon modern UI design
- Adapt CBRAIN for generalized use for scientific workflows across HPC and Cloud infrastructures

- Update the NeuroHub Portal web interface to be a clean entry point into the ecosystem

Conclusion


Big Data


Big Computation


International Collaboration


Interdisciplinary Health Research

Infrastructure and services for accomplishing science


CBRAIN and NeuroHub are core components for reproducibility and interoperability

- Interoperate or integrate – not reimplement


Standards are game changers


Community and use-case based development and governance

Acknowledgements

cbrain

Alan Evans
Bryan Caron
Pierre Rioux
Natacha Beck
Serge Boroday
Darcy Quesnel
Xavier Lecours-Boucher
Najmeh Kalili-Manhani
Darius Valevicius
Safa Sanami
Reza Adalat
Tristan Glatard

NeuroHub

Bryan Caron
J-B Poline
Diana Le
Pierre Rioux
Natacha Beck
Serge Boroday
James Mehta
Samir Das
Xavier Lecours-Boucher
Rida Abou-Haidar
Darcy Quesnel
Emmet O'Brien
Sebastian Urchs
Derek Lo
Ksenia Zaytseva

Past Contributors (alphabetical order)

Adam Trefonides, Andrew Doyle, Angela McCloskey, Anton Zoubarev, Armin Taheri, Candice Czech, Ehsan Afkhami, Greg Kiar, Liam O'Callaghan, Marc Rousseau, Mathieu Desrosiers, Melanie Legault, Nicolas Kassis, Rémi Bernard, Shawn Brown, Sonam Tamang, Tarek Sherif, Tristan Aumentado-Armstrong, Verena Schuster

Acknowledgements

NeuroHub is funded by the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains, Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada.