Neuroinformatics Framework
progress toward tools for capturing, sharing and exploring human brain imaging data

Multiple laboratories can share and compare evidence of functional changes in the aging brain (foreground) from MRI imaging with genetic information and results of online cognitive battery tests (which employ simple images such as those at right) using tools developed by this project.

**Selected accomplishments**

- Deployed a robust neuroinformatics framework for capture of human brain imaging data across multiple Harvard sites
- Used the infrastructure to capture 500 human brain imaging data sets linked to genetic information over a period of six months

**Launched** 2007

**Status** Continuing at the Center for Brain Science

**Participants**

- **Randy Buckner**, Psychology/FAS/ MGH
- **Gabriele Fariello**, Center for Brain Science/IIC
- **Timothy O’Keefe**, CBS/IIC
- **Victor Petrov**, CBS/IIC
Several years ago, the Center for Brain Science at Harvard University realized that it needed to create a computer science and informatics infrastructure capable of dealing with a growing problem common to many scientific fields, but perhaps particularly daunting in the field of neuroscience.

Modern neuroscience merges a wide array of scientific fields and a wider array of data acquisition modalities. Advances in neuroscience, genetics, computer science and computing power have increasingly outpaced the ability of individual researchers to manage and effectively analyze the volume and diversity of data captured or derived. The informatics infrastructure to support the ever-accelerating information capture and research has become the limiting agent at the frontiers of neuroscience. Furthermore, the desire and need to acquire data from larger cohorts has been limited by resources and finances. These issues are by no means limited to the field of neuroscience, but are perhaps best represented in it. The Neuroinformatics Framework (NF) has successfully addressed and continues to address these limitations. It is an ongoing endeavor with little likelihood of abating in the foreseeable future.

Project goals

Through the Neuroinformatics Framework Project, the Neuroinformatics Research Group (NRG) is working to facilitate both small- and large-scale high-throughput neuroimaging-based research on the Harvard campus as well as at the teaching hospitals, research centers and the neuroscience community at large. The broad goals of the Neuroinformatics Framework are to:

- Enable large-scale high-throughput data capture in near real-time
- Provide reliable and secure permissions-based storage of all data
- Increase the computational ability of the existing computer science/information technology infrastructure
- Provide automated computational pipelines for processing the data
- Support the ongoing research of dozens of Harvard faculty
- Decrease the cost of acquisition of subjects and running protocols
- Decrease the time and cost to develop and test new protocols
- Allow researchers to ask questions across large-scale multimodal data sets

<table>
<thead>
<tr>
<th>Neuroinformatics Framework: Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
</tr>
<tr>
<td>January 2007</td>
</tr>
<tr>
<td>February 2007</td>
</tr>
<tr>
<td>June 2007</td>
</tr>
<tr>
<td>December 2007</td>
</tr>
<tr>
<td>May 2008</td>
</tr>
<tr>
<td>July 2008</td>
</tr>
<tr>
<td>February 2009</td>
</tr>
</tbody>
</table>
The science
A case illustrative of the need for a cross-domain framework to produce ground-breaking discoveries is the Brain Genomics Superstruct Project (BGSP).

An important application of cost-efficient neuroimaging is in the assessment of genetic risk of psychiatric disease. In most individuals, the risk factors are clinically silent. Brain biomarkers are therefore required to identify which risk factors are diagnostic.

The need is becoming more acute as genome-wide association studies (GWAS) identify growing numbers of associations with unknown impact on the brain. A first question relevant to clinical decision-making is which genetic variants (or combinations of variants) are linked to observable brain dysfunction. Within the framework of the Harvard Catalyst Translational Research Laboratory and its affiliated imaging core, the Center for Brain Science will convert the emerging MRI technologies into a viable cost-efficient clinical neuroimaging platform. We will then supply assistance to investigators to collect 500 data samples using the platform to build an openly available repository of brain and behavioral data for human genetic explorations. The delivery of this openly available resource will be a transformative demonstration that the approach is viable and will also put the technologies in the hands of a consortium of clinical investigators making health care decisions on a wide range of psychiatric and neurologic illnesses.

Proposed work includes:

- A rapid acquisition brain imaging protocol (taking less than 15 minutes) will be made available to a network of collaborating clinical investigators. The protocol will include both structural and functional measures.
- A database will be maintained using the BIRN-enabled XNAT platform to receive, vet, and share MRI data associated with the project. XNAT is a previously developed toolkit for managing and processing large MRI data sets that we have openly released through BIRN (the Biomedical Informatics Research Network).
- Imaging and genetic data will be acquired from 500 participants across sites at Massachusetts General Hospital, Harvard, McLean Hospital, and MIT. Data will include the rapid acquisition sequences, DNA via saliva extraction, and a web-based cognitive and behavioral testing battery. The data will be shared as an open resource.
- The rapid acquisition brain imaging protocol will be made available and tested across three national CTSA (Clinical and Translational Science Award) sites: Harvard, Washington University in St. Louis, and UC San Francisco.

A major goal of the BGSP is to acquire neuroimaging data (including functional connectivity), full genomic data and an array of cognitive data for 2,500 to 5,000 subjects. Already 26 faculty from Harvard and MIT have expressed interest in participating. Without the Neuroinformatics Framework project, such very-large-scale interdisciplinary studies could not take place.

Informatics infrastructure to support ever-accelerating information capture and research has become the limiting agent at the frontiers of neuroscience.

In addition to the BGSP, the Neuroinformatics Framework is currently actively facilitating the research of more than a dozen faculty at HMS and FAS. We expect this number to grow to between 26 and 30 by the end of the year. On the Harvard main campus alone, seven faculty members (Joshua Greene, Daniel Schacter, Randy Buckner, Diego Pizzagalli, Yaoda Xu, Alfonso Caramazza, Jason Mitchell) and their researchers use the NF for their individual research projects. Although we have long lamented that the time scale of research means that there is necessarily a lengthy time delay before significant publications, we are steadfast in our certitude that this effort is paying off and will continue to pay large research dividends in the years to come.
to come at a fraction of previous costs. We are already seeing some publications facilitated by the second-generation framework, NF2, whereas we expect publications relating directly to the NF within the next 24 to 48 months.

**Education**

The Neuroinformatics Framework will be used to teach undergraduates and graduates in upcoming courses to be offered in spring 2010, such as the undergraduate neuroimaging lab for psychology.

By integrating UPPI (and using OpenBT and PPT) into XNAT we are able to allow researchers to perform both automated and ad-hoc analyses. Our proof of concept pipeline was designed to take the raw DICOM images acquired by researchers at the scanners and perform standardized quality control analyses on the images and sessions. Above are some sample screen shots of the results of one such analysis.
concentrators. A “sandbox” version will allow students to download, view, analyze and process brain images, along with anonymized clinical, psychological and genetic information.

This will permit students to actively query “real” data sets and learn how to analyze brain images acquired through standard neuroimaging techniques. Students will also be able to perform complex analyses across very disparate data sets.

**The “database” infrastructure**

In order to accomplish the goal of capturing data in near-real time, we had to deploy a system capable of capturing data directly from the MRI scanners and which would act as a gateway to accessing the data. For this we chose to customize the open-source eXtensible Neuroimage Archiving Toolkit (XNAT). This database has been modified to accept data from a number of different sources including MRI scanners, DNA sequence analysis machines, and cognitive battery tests.

In order to create a highly reliable and available system, the NRG deployed XNAT using virtualization technology which enabled us to rapidly deploy production, test and development images of XNAT quickly and robustly.

**Uniform Pipeline Processing Infrastructure (UPPI)**

One of the greatest challenges in any Neuroimaging endeavor is the ability to process and mine large amounts of data both reasonably and efficiently. To do so, a mechanism to define processing steps and pipelines in a uniform manner and to spawn and manage those computational pipelines is necessary; the NRG has developed the following tools with these goals in mind:

**OpenBT**

As the number of datasets in a project such as BGSP increase, so must available compute cycles. The ability to dispatch jobs onto a compute cluster becomes immediately apparent. Ultimately however, the ability to dispatch jobs to a compute “cloud” must be achievable. The NRG has approached this issue by developing a toolkit for dispatching compute tasks between machines located on the same or disparate networks requiring only one point-of-entry (HTTPS/8080).

In the most basic sense, OpenBT is a web-application for publishing command-line tools over HTTP. One important feature however is the ability to “link” multiple instances of OpenBT on-the-fly. In doing so, several instances of OpenBT are able to share any given task load by distributing jobs precipitously and propagating processed results to the end-user over HTTP.

**Pipeline Processing Toolkit (PPT)**

There exist a wide variety of tools and libraries written for the Neuroimaging community. Experience suggests that researcher generally opt to use bits and pieces from several libraries interchangeably. Unfortunately when doing so—and especially when data fusion is involved—the tendency to run into complex, daunting issues increases. Recent efforts to simplify this problem have not been widely accepted.

The Neuroimaging Toolkit (NT) aims to consolidate several software packages into a simple toolkit by which users may describe an entire analysis pipeline in XML. Each component within NT has multiple execution methods for dispatching jobs locally, through a job dispatcher (LSF) or though OpenBT. A user can perform for example, data retrieval, conversion, motion correction, normalization, and functional connectivity in several lines of XML.

By integrating OpenBT and PPT, UPPI is therefore able to run complex, distributed analysis
pipelines which are either launched from within a system such as XNAT, run on the command line, or spawned via an HTTP request.

**iCognition testing suite and toolkit**

In order to further enable high-volume acquisition of psychological data, it was necessary to develop a software platform capable of assessing subjects online in a manner which produced results similar to those observed in the laboratory. To date, this has been a difficult task, but the NRG NF has successfully implemented iCognition (currently running on http://www.mytimedtest.org), an online cognitive battery test capable of administering a wide array of cognitive battery tests online and minimal cost with results substantially similar to those obtained in the lab. It features a full set of behavioral tests, a database-backed management panel and ability to integrate with other services via XML.

The NRG/NF is working to release iCognition 2.0, which will enable neuroscience and psychology researchers to piece together custom tests on-line with drag-and-drop ease. It will provide an easy to use package for experiment creation and sharing, easy deployment to a web server and the ability to run the tests in an online environment, while promoting Open Source values and ideas for an open and collaborative society.

**Current use**

The Neuroinformatics Framework is already capturing data for the Center for Brain Science neuroimaging faculty, for MGH faculty at the Athinoula A. Martinos Center for Biomedical Imaging, and for cross-institution faculty and researchers for the BGSP, in near-real-time. Data are being captured directly from the scanners, from the surveys and cognitive battery tests. We will be integrating the capture of the genetic and biomarker data shortly (possibly August 2009) and performing automated analyses based on NF created processing pipelines. The NF will ultimately enable the ability to capture, store, share, and analyze genetic data received from microarrays to full genomic sequences, task-based functional MRI, structural MRI, DTI/DWI (diffusion tensor/diffusion-weighted imaging), functional MRI, CT (computed tomography), PiB (Pittsburgh Compound-B imaging, for seeing plaques) and an assortment of other three and four dimensional data sets, and a disparate aggregate of cognitive battery data.
battery data. This is a complicated and involved task, but a fundamental requirement for today’s “bleeding-edge” neuroscience. To then associate such data with subjects of varying demographic qualities and cohort membership, allow searching through the data on many different meta-data dimensions, manage privacy concerns, grant or limit access while ensuring high-availability and data longevity and still be able to provide unfettered access to raw and derived data to researchers with appropriate privileges requires considerable effort.

The informatics framework capable of accomplishing these goals is one which necessarily requires domain knowledge of computer science, hardware and software architecture, information technology support and maintenance, and end-user usability with a fundamental understanding of the underlying science.

The NRG works closely with the Harvard University Research Computing and Information Services at MGH, Life Science Computing resources at Harvard and research faculty at Harvard FAS, Harvard Medical School and MIT.

Future work
The Neuroinformatics Framework project, and the Neuroinformatics Research Group will continue to exist but will be fully funded by the Center for Brain Science.

Publications

