NIH Annual Intramural Research Report

Core project

ZIC MH002960-01

Report Title

Data Science and Sharing Team

2017 Fiscal Year

October 01, 2016 - September 30, 2017

Lead Investigator

Adam G Thomas; PhD

Research Organization

Lab Branch Code: DIRP
Functional MRI Core, NIMH

Lab Staff and Collaborators within the *Functional MRI Core*

Nino Migineishvili
Dylan M Nielson; PhD
John Gavin Rodgers-Lee; PhD

Total staff years

2.50

Collaborators from other NIMH organizations

Joyce Chung; MD *(OCD, OSD, DIRP, NIMH)*
Richard Coppola; SB, DSc *(MEG, OSD, DIRP, NIMH)*
Robert Cox; PhD *(SSC, OSD, DIRP, NIMH)*
Allison Carol Nugent; PhD *(NTMD, ETPB, DIRP, NIMH)*
Judith L Rapoport; MD *(CPB, DIRP, NIMH)*
Armin Raznahan *(DNU, CPB, DIRP, NIMH)*
Audrey E Thurm; PhD *(PDNB, DIRP, NIMH)*
Carlos Alberto Zarate; MD *(ETPB, DIRP, NIMH)*

Collaborators from other NIH Institutes/Centers

Dante Picchioni *(AMRI, LFMI, BNP, DIR, NINDS)*
Wallace P Shaw *(NCRS, SBRB, DIR, NHGRI)*

Extramural Collaborators

Oscar Esteban; PhD *(Psychology, Stanford University)*
Satrajit Ghosh; PhD *(Depar McGovern Institute for Brain Research tment, Massachusetts Institute of Technology)*
Chris F Gorgolewski; PhD *(Psychology, Stanford University)*
Anisha Keshavan; PhD *(eScience Institute, University of Washington)*
Jong-Hwan Lee; PhD *(Korea University)*

Human subject research

does not use human cells, human subjects, or human tissues

Keywords

Data Sharing, Data Science, Open Science, Machine Learning, Neuoimaging, MRI, High Performance Computing

Goals and Objectives

The goal of the Data Science and Sharing Team is to support and advance the creation, distribution, and utilization of large, open datasets to accelerate discovery within the NIMH Intramural Research Program. We provide tools and training to help scientists within the IRP embrace open and reproducible science practices. This includes:
- Standardized, community recognized formats and repositories for data storage and dissemination
- Collaborative, version-controlled tools for developing analysis code
- Open distribution of all experimental methods and results to maximize impact and reproducibility

Summary

This is the first annual report for the Data Science and Sharing Team (DSST). In this past year the team has been busy hiring staff and establishing infrastructure. Significant progress has also been made towards the goals and mission of the group.

The first data scientist for the group, John Lee, was hired in June of 2016. The second data scientist, Dylan Nielson, was hired in April of 2017. Nino Migineishvili join the team as a summer student in June of 2017. Adam Thomas has been leading the team since its creation.

Building an Intramural Neuroimaging Data Repository

The Functional Magnetic Resonance Imaging Facility (FMRIF) maintains five MRI scanners that generate over 150 scans per week. All of this data in archived, but it is not currently standardized, organized, or consented such that investigators can aggregate across the participants in the many ongoing studies within the NIH IRP. The DSST is working to change that on three fronts.

First, we have consulted with the scientists responsible for the Alzheimers Disease Neuroimaging Initiative project who have established standard MRI protocols to use on large, diverse groups of participants across different scanner platforms. The protocols are now available on the FMRIF scanners and will soon replace the clinical scans currently collected on all participants on a yearly basis. Second, we have deployed a data sharing repository using the software developed at the Stanford Center for Reproducible Neuroscience. This repository went online in June of 2017 and is currently accessible to intramural researchers on the NIH campus. It will be available to researchers throughout the world before the end of the year. Third, we have worked closely with the office of the clinical director to help develop a protocol for recruiting and scanning healthy volunteers. This protocol will provide a pool of volunteers that can be streamlined into other studies within the IRP while also building a normative database of phenotypic, genetic, and imaging data.

Facilitating Access to Shared Data

The DSST is also working to facilitate and streamline access to existing public data repository for NIMH intramural researchers by creating a local copy on the NIH High Performance Computing System, also known as the Biowulf. Once the local copy is in place, the DSST guides researchers through the necessary data use agreements and then grants them access to the data. This has already been accomplished for the Human Connectome Project which offers comprehensive phenotypic, genetic and imaging data on 1200 participants. We are in the process of making data available from the LIFE Project (2,377 participants), the Adolescent Brain Cognitive Development (ABCD) project (10,000 participants), the Child Mind Institutes Healthy Brain Network, and the UK Biobank (100,000 participants).

Delivering Education and Training

Conducting reproducible science on large datasets requires skillsets that many researchers do not possess. Since May of 2016 the DSST has conducted four hands-on courses and hosted invited talks from internationally recognized experts in the field of open and reproducible science. All slides and materials are available at https://cmn.nimh.nih.gov/dsst

- On May 26 & 27, 2016 the DSST taught a Data Carpentry in the NIH library.
- On June 9 & 10 2016 the DSST taught a course on Software Carpentry for the NIH Graduate Partnership program.
- On Nov 2, 2016 Adam Thomas gave a presentation on reproducible methods to the MEG North America Meeting.
- On Dec 9, 2016 the DSST hosted Dr. Matthew Brett who provided an education lecture on how to improve reproducibility in neuroimaging research.
- On Jan 25, 2017 John Lee taught in the Introduction to R: Data Wrangling & Statistical Analysis course at the NIH Library.
- On March 13-17, 2017 the DSST held a four day course on reproducible neuroscience .
- On April 25th, 2017 John Lee taught a course on using R for reproducible analyses in the NIH Library.
- On May 18, 2017 the DSST held a workshop on reproducible research as part of the NIH Pi Day celebration.
- On August 1-2, 2017 the DSST held a two day course on reproducible neuroscience with guest lectures from Dr. Satra Ghosh of MIT and Dr. Yarik Halchenko of Dartmouth University.
- On August 14th & 17th, 2017 the DSST hosted talks from Dr. Anisha Keshavan on Crowdsourcing in Neuroscience and Dr. Regina Nuzzo on common statistical pitfalls in neuroscience.

Future Directions and Applications

One of the motivations for building large, standardized repositories of data is so that we may more easily employ data-hungry techniques such as machine learning. The Data Science and Sharing Team was conceived of and designed to work closely with the Machine Learning Team which is still forming. In August of 2017, Charles Zheng, the first member of that team arrived and we are actively planning future projects. Two more members of that team are expected to arrive in the next few months.

Nino Migineishvili, a summer student from UCLA who has been working with the DSST since June 2017, has also been applying machine learning to several large datasets: one acquired here by Dr. Philip Shaw from the Social and Behavioral Research Branch at NHGRI (442 participants) and others acquired outside NIH such as the Nathan Kline Institute (888 participants). Nino's project of predicting brain age from structural images has been pre-registered at Open Science Foundation and is currently being prepared for publication (DOI: 10.13140/RG.2.2.25126.63047).

Jong Hwan Lee, a visiting professor from Korea University working with the DSST and Dr. Peter Bandettini, has also been applying machine learning methods to predict fMRI time course using the HCP dataset on the NIH HPC. Several similar, data-intensive projects are planned for the coming year.

Publications Generated during the 2017 Reporting Period

Ordered by author name.

1. Ghosh SS, Poline JB, Keator DB, Halchenko YO, Thomas AG, Kessler DA, Kennedy DN (2017) A very simple, re-executable neuroimaging publication. F1000Res 6:124
   PubMed ID 28781753 Pubmed Central ID 5516225
2. Hodgetts CJ, Voets NL, Thomas AG, Clare S, Lawrence AD, Graham KS (2017) Ultra-High-Field fMRI Reveals a Role for the Subiculum in Scene Perceptual Discrimination. J Neurosci 37:3150-3159
   PubMed ID 28213445 Pubmed Central ID 5373110
3. Nielson D, Sederberg PB (2017) MELD: Mixed Effects for Large Datasets. PLOS One, in press.
4. Thomas AG, Dennis A, Rawlings NB, Stagg CJ, Matthews L, Morris M, Kolind SH, Foxley S, Jenkinson M, Nichols TE, Dawes H, Bandettini PA, Johansen-Berg H (2016) Multi-modal characterization of rapid anterior hippocampal volume increase associated with aerobic exercise. Neuroimage 131:162-70
   PubMed ID 26654786 Pubmed Central ID 4848119
5. Trefler A, Sadeghi N, Thomas AG, Pierpaoli C, Baker CI, Thomas C (2016) Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. Neuroimage 133:41-52
   PubMed ID 26921714
6. Zhou D, Liu S, Dillard-Broadnax DV, Berman RA, Rapoport J, Thomas AG (2016) 7-Tesla MRI reveals regional hippocampal volume deficits of dentate gyrus in childhood-onset schizophrenia. Society for Neuroscience 46th Annual Meeting, in press.