Report from 2015 OHBM Hackathon (HI)

Advancing Open Science through NiData

Project URL: http://github.com/nidata/nidata

B Cipollini[1][*] and A. Rokem[2]

1 Introduction

The goal of this project is to improve accessibility of open datasets by curating them. "NiData" aims to provide a common interface for documentation, downloads, and examples to all open neuroimaging datasets, making data usable for experts and non-experts alike.

2 Approach

Open datasets promise to allow more thorough analysis of hard-to-collect data and re-analysis using state-of-the-art analysis methods. However, open datasets are not truly open unless they are easy to find, simple to access, and have sufficient documentation for use. Currently, publicly available data in neuroscience are scattered across a number of websites and databases, without a common data format no common method for data access, and varying levels of documentation. Datasets are being uploaded to public databases through a number of initiatives, including OpenFMRI and NITRC. In addition, there are funded efforts for collecting data explicitly for the purpose of public sharing – most visibly in the Human Connectome Project (HCP) - but also in the Pediatric Imaging, Neurocognition and Genetics (PING) study. There are a number of funded efforts to collect old data and re-release as public databases, notably the INDI[1] efforts (which include the popular ABIDE and F1000 datasets). The BRAIN initiative aims to collect data that will be a challenge to store, let alone analyze. There are even online journals focused on publishing datasets (e.g. Nature Scientific Data), or with options to release data (e.g. F1000 "Data Notes").

NiData is a Python package that provides a single interface accessing data from a variety of open data sources. The software framework makes it easy to add new data sources, simple to define and to provide access to multiple datasets from a single data source. Software dependencies are managed on a per-dataset basis, allowing downloads and examples to use any public packages without requiring installation of packages required by unused datasets. The interface also allows selective download of data (by subject or type) and caches files locally, allowing easy management of big datasets.

3 Results

We focused on exposing new methods for downloading data from the HCP. We were able to provide a downloader that accepts login credentials and downloads files locally. We created an example that interacts with dipy to produce diffusion imaging results on a single subject from the HCP. We also worked at collecting common data sources, as well as individual datasets stored at each data source, into NiData's "data sources" wiki page. We incorporated downloads, documentation, and examples from the nilearn package and began discussion of making a more extensible object model.

Since the hackathon, we have created such an object model and migrated all code to use it, and a Sphinx-based website is under development.

4 Conclusions

Projects like NiData improve curated data access and increases the effectivity of big data projects with open source data.

[*]Correspondence: bcipollini@ucsd.edu
[1]University of California, San Diego, La Jolla, 9500 Gilman Dr., 92093, California, USA
Full list of author information is available at the end of the article

Author's contributions

BC and AR wrote the software and the report.

Author details

[1]University of California, San Diego, La Jolla, 9500 Gilman Dr., 92093, California, USA. [2]University of Washington, Seattle, 3910 15th Ave NE, Seattle, 98195, Washington, USA.

References

1. Mennes, M., Biswal, B.B., Castellanos, F.X., Milham, M.P.: Making data sharing work: the FCP/INDI experience. Neuroimage **82**, 683–691 (2013)