Report from 2015 Brainhack at OHBM

Sharing Data in the Cloud

Project URL: https://github.com/DaveOC90/INDI-Organization-Scripts

David O'Connor[1,2*], Daniel J. Clark[2], Michael P. Milham[1,2] and R. Cameron Craddock[1,2]

1 Introduction

Cloud computing resources, such as Amazon Web Services[1] (AWS), provide pay-as-you-go access to high-performance computer resources and dependable data storage solutions for performing large scale analyses of neuroimaging data[1]. These are particularly attractive for researchers at small universities and in developing countries who lack the wherewithal to maintain their own high performance computing systems. The objective of this project is to upload data from the 1000 Functional Connectomes Project (FCP)[2] and International Neuroimaging Datasharing Initiatives (INDI) [3] grass-roots data sharing initiatives into a Public S3 Bucket that has been generously provided by AWS. This will make the data more quickly accessible for AWS-based analysis of these data, but will also improve the speed and availability of access to this data for analyses performed outside of the cloud. To begin with, we focused on the following collections:

- The *Autism Brain Imaging Data Exchange (ABIDE)* consists of structural MRI and resting state functional MRI from 1113 individuals (164 F, 948 M, 6-64 years old, 539 with autism spectrum disorders, 573 typical controls) aggregated from 20 different studies [4].
- The *ADHD-200* contains structural MRI and resting state functional MRI from 973 individuals (352 F, 594 M, 7-21 years old, 362 with attention deficit hyperactivity disorder (ADHD), 585 typically developing controls) collected from 8 sites [5].
- The *Consortium for Reliability and Reproducibility (CoRR)* consists of 3,357 structural MRI, 5,093 resting state fMRI, 1,302 diffusion MRI, and 300 cerebral blood flow scans from 1629 subjects (673 F, 956 M, 6-84 years old, all typical controls) acquired in a variety of test-retest designs at 35 sites [6].
- The *Enhanced Nathan Kline Institute - Rockland Sample (ENKI-RS)* consists of structural MRI, resting state functional MRI, diffusion MRI, cerebral blood flow, and a variety of task functional MRI scans and deep phenotyping on over 700 participants from across the lifespan and a variety of phenotypes acquired at a single site [7]. The acquisition of this collection is ongoing.
- The *Addiction Connectome Preprocessed Initiative (ACPI)*[2] consists of 216 structural MRI and 252 functional MRI from 192 subjects (44 F, 148 M, 18-50 years old) from three datasets generated by NIDA investigators.

2 Approach

Data for the ADHD-200, ABIDE, CoRR, and Rockland Sample data collections are currently downloadable from NITRC[3] as a series of large (>2GB) tar files. The process of uploading the data involved downloading and extracting the data from these tar files, organizing the individual images to the standardized INDI format [4], and then uploading the data to S3. We developed a S3 upload script in python using the Boto AWS software development kit[5] to facilitate this process. We also developed a download script in python that provides basic query functionality for selecting the data to download from a spreadsheet describing the data.

*Correspondence: david.oconnor@childmind.org
[1]Center for Biomedical Imaging and Neuromodulation, Nathan Kline Institute for Psychiatric Research, Orangeburg, 140, 10962, New York, USA
Full list of author information is available at the end of the article

[1] http://aws.amazon.com

[2] http://fcon_1000.projects.nitrc.org/indi/ACPI/html/index.html

[3] http://fcon_1000.projects.nitrc.org/

[4] http://fcon_1000.projects.nitrc.org/indi/indi_data_contribution_guide.pdf

[5] https://aws.amazon.com/sdk-for-python/

3 Results

The entirety of the CoRR, ABIDE, ACPI, and ADHD-200 data collections and ENKIRS data for 427 individuals were uploaded during the OHBM Hackathon event. The data are available as individual files to make it easily indexable by database infrastructures such as COINs [8], LORIS [9], and others. Additionally, this makes it easy for the users to download just the data that they want. The data in the bucket can be browsed and downloaded using a GUI based S3 file transfer software such as Cyberduck[6] (see Fig. 1), or using the Boto python library[7]. One can connect to the bucket using the configuration shown in Figure 1. From there one can specify the path to data of interest; the data is structured as follows: bucketname/data/Projects/ProjectName/DataType. So if one wihsed to access raw data from the ENKIRS, the following path would bring them directly to it: fcp-indi/data/Projects/RocklandSample/RawData. It is also possible to just connect to the Projects folder and browse all available data.

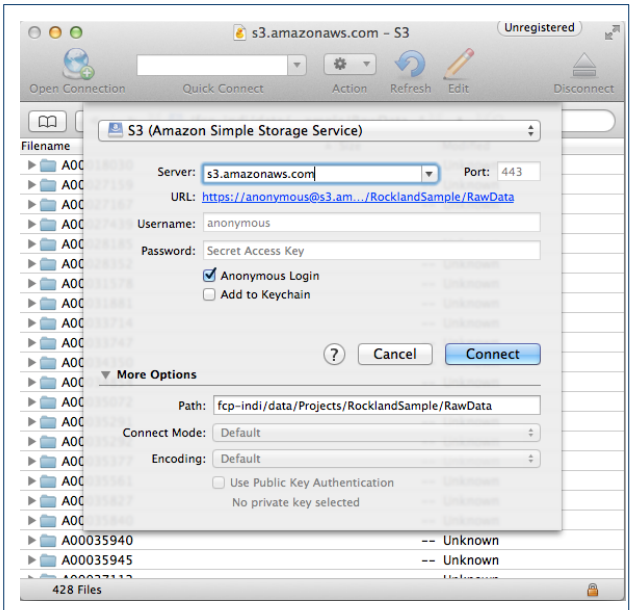

Figure 1 Connecting to the data repository.

4 Conclusions

Uploading data shared through the FCP and INDI initiatives improves its accessibility for cloud-based and local computation. Future efforts for this project will include uploading the remainder of the FCP and INDI data and organizing the data in the new brain imaging data structure (BIDS) format [10].

[6] http://cyberduck.org

[7] https://github.com/FCP-INDI/INDI-Tools

Availability of Supporting Data

More information about this project can be found at: https://github.com/DaveOC90/INDI-Organization-Scripts. Further data and files supporting this project are hosted in the *GigaScience* repository REFXXX.

Competing interests

None

Author's contributions

DO performed quality control, and uploaded the data. DJC wrote code to interact with AWS, preprocessed and uploaded data. MPM and RCC lead the data collection and sharing projects. All of the authors contributed to writing the project report.

Acknowledgements

The authors would like to thank the organizers and attendees of the OHBM Brainhack in Hawaii. This project was made possible by the S3 public bucket generously provided by Amazon Web Services.

Author details

[1]Center for Biomedical Imaging and Neuromodulation, Nathan Kline Institute for Psychiatric Research, Orangeburg, 140, 10962, New York, USA. [2]Center for the Developing Brain, Child Mind Institute, New York, 445, 10022, New York, USA.

References

1. Clark, D., Haselgrove, C., Kennedy, D.N., Liu, Z., Milham, M., Petrosyan, P., Torgerson, C., Van Horn, J., Craddock, C.: Harnessing cloud computing for high capacity analysis of neuroimaging data from ndar. Frontiers in Neuroscience (21). doi:10.3389/conf.fnins.2015.91.00021
2. Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kotter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P.: Toward discovery science of human brain function. Proc. Natl. Acad. Sci. U.S.A. **107**(10), 4734–4739 (2010)
3. Mennes, M., Biswal, B.B., Castellanos, F.X., Milham, M.P.: Making data sharing work: the FCP/INDI experience. Neuroimage **82**, 683–691 (2013)
4. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**(6), 659–667 (2014)
5. Milham, M.P., Fair, D., Mennes, M., Mostofsky, S.H.: The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in Systems Neuroscience **6**(62) (2012). doi:10.3389/fnsys.2012.00062
6. Zuo, X.N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S.J., Courtney, W., Craddock, R.C., Di Martino, A., Dong, H.M., Fu, X., Gong, Q., Gorgolewski, K.J., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S.M., Lainhart, J.E., Lei, X., Li, H.J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D.S., Mayer, A.R., Meindl, T., Meyerand, M.E., Nan, W., Nielsen, J.A., O'Connor, D., Paulsen, D.,

Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.X., Weng, X.C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.T., Milham, M.P.: An open science resource for establishing reliability and reproducibility in functional connectomics. Sci Data **1**, 140049 (2014)

7. Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedict, M.M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz, S.T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R.T., Kamiel, S.M., Anwar, A.R., Hinz, C.M., Kaplan, M.S., Rachlin, A.B., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Craddock, C.C., Calhoun, V., Courtney, W., King, M., Wood, D., Cox, C.L., Kelly, A.M., Di Martino, A., Petkova, E., Reiss, P.T., Duan, N., Thomsen, D., Biswal, B., Coffey, B., Hoptman, M.J., Javitt, D.C., Pomara, N., Sidtis, J.J., Koplewicz, H.S., Castellanos, F.X., Leventhal, B.L., Milham, M.P.: The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. Front Neurosci **6**, 152 (2012)
8. Landis, D., Courtney, W., Dieringer, C., Kelly, R., King, M., Miller, B., Wang, R., Wood, D., Turner, J.A., Calhoun, V.D.: COINS Data Exchange: An open platform for compiling, curating, and disseminating neuroimaging data. Neuroimage **124**(Pt B), 1084–1088 (2016)
9. S., D.: LORIS: a web-based data management system for multi-center studies. Front.Neuroinform **5** (2011)
10. Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., Handwerker, D.A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B.N., Nichols, T.E., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Turner, J.A., Varoquaux, G., Poldrack, R.A.: The brain imaging data structure: a standard for organizing and describing outputs of neuroimaging experiments. bioRxiv (2015). doi:10.1101/034561. http://biorxiv.org/content/early/2015/12/16/034561.full.pdf