COINSTAC: Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation

Harshvardhan Gazula[1], **Ross Kelly**[1], **Javier Romero**[1], **Eric Verner**[1], **Bradley T. Baker**[1], **Rogers F. Silva**[1], **Hafiz Imtiaz**[2], **Debbrata Kumar Saha**[1], **Rajikha Raja**[1], **Jessica A. Turner**[1], **Anand D. Sarwate**[2], **Sergey M. Plis**[1], **and Vince D. Calhoun**[1]

1 Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA **2** Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

Summary

Central to the field of neuroimaging is the development of techniques for making sense of complex brain data. However, rapid technological advancements are pushing the spatial and temporal resolution of imaging in different modalities to an unprecedented level, leading to large datasets which cannot be analyzed in the traditional desktop computing paradigm. This has led to a paradigm shift in scientific research with an increasing emphasis on collaborative data sharing. However, current approaches to data sharing, such as negotiating multiple data sharing agreements, can be cumbersome. In addition, there are also significant data transfer, organizational, and computational challenges, the result being that collaborative group research requires a great deal of coordination. Human and business factors can hamper research from happening at a constructive pace, maybe even forbidding group research to occur at all.

Software

The Collaborative Informatics and Neuroimaging Suite Toolkit for Anonymous Computation (`COINSTAC` (Plis et al., 2016)) is a web-based framework that addresses the aforementioned issues. It provides a platform to analyze data stored locally across multiple organizations without the need for pooling the data at any point during the analysis, enabling decentralization. It is intended to be the ultimate hub by which researchers can build statistical (Ming et al., 2017) or machine learning models (Gazula et al., 2018) collaboratively in a decentralized fashion. This framework implements a message-passing infrastructure that allows large-scale analysis of decentralized data with results on par with those that would have been obtained if the data were centralized. Because there is no pooling of data, COINSTAC protects the privacy of individual datasets. In addition, COINSTAC also offers differentially private algorithms for enhanced protection from reidentification attacks. Differential privacy is a framework to control the risk that individual data points can be inferred from the output of the algorithm (Dwork, Roth, & others, 2014). Computations can be local or decentralized and are deployed using a containerized model. COINSTAC simulator, a simulation environment for algorithm developers to build COINSTAC computations, is also available. After developers ensure compatibility of their computations with COINSTAC through the simulator, these computations can then be made available within the COINSTAC platform.
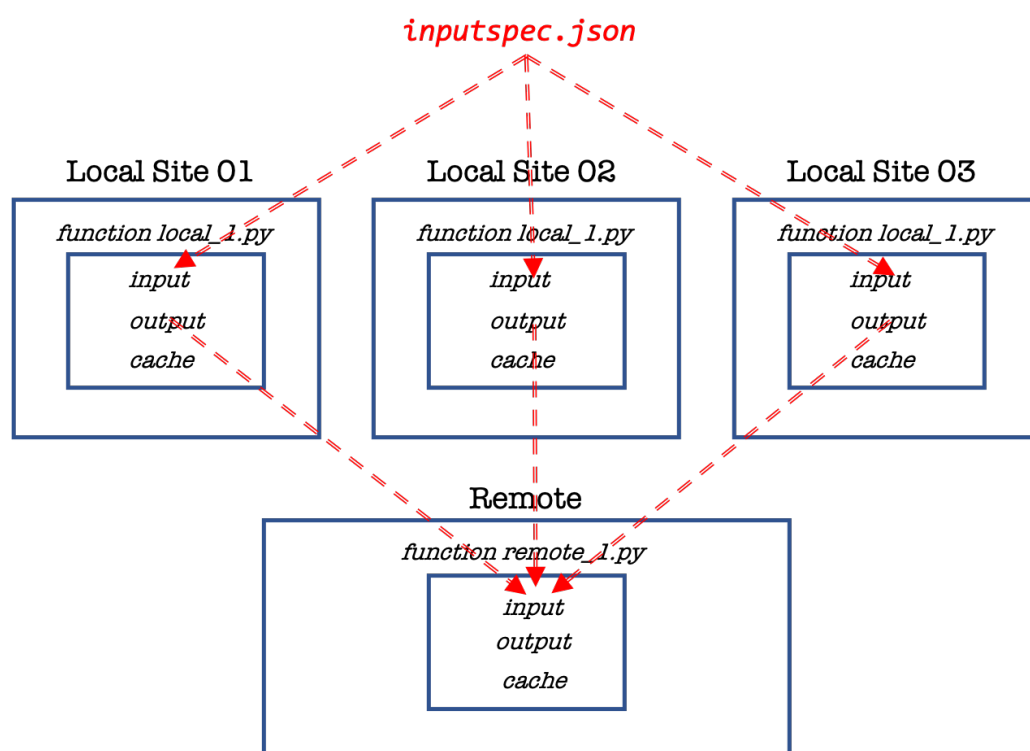
Figure 1: A graphical representation of decentralization in COINSTAC

Related Work

In the past, data-specific collaborative efforts have included either aggregating the data via a centralized data sharing repository or sharing data via agreement-based collaborations. Frameworks such as ENIGMA (Thompson et al., 2014) to some extent bypass the need for data agreements by performing a centrally coordinated analysis at each local site. Another framework called ViPAR (Carter et al., 2016) tries to go one step further by completely isolating the data at the local site but only pooling it via transfer to perform automated statistical analyses. However, the heterogeneity among the local analyses caused by adopting various data collection mechanisms or preprocessing methods can lead to inaccurate meta-analysis findings. Other tools such as CBRAIN (Sherif et al., 2014), Loris (Das, Zijdenbos, Vins, Harlap, & Evans, 2012), XNAT (Herrick et al., 2016), and OpenNeuro (Gorgolewski, Esteban, Schaefer, Wandell, & Poldrack, 2017) exist, but a more detailed comparison is not provided herein, as COINSTAC is, to our knowledge, the first application platform enabling decentralized analysis of brain imaging data.

Features

COINSTAC removes the barriers to collaborative analysis by:

1. Decentralizing analyses and computation.

   - Analysis pipelines only interact with site data on their own computers. Aggregate derived data, such as model weights, from each user may be sent to a central compute node,

---

but never raw data. Over a dozen local and decentralized computations/algorithms have been developed already, with more coming.

- A central compute node performs a complimentary component of the group analysis by coordinating between the various data nodes participating in a consortium. This node may trigger adjusted computations on user machines to improve a model.

2. Not synchronizing full datasets. Instead, synchronizing only aggregate-level analysis metrics (e.g., the gradients of a machine learning algorithm)

- As previously discussed, central compute nodes aggregate these metrics and help the user draw conclusions from the contributors.

- Because machine learning algorithms can be designed to model outcomes via artifacts of analysis pipelines, COINSTAC keeps data safely and conveniently on site computers.

3. Applying differential privacy strategies to further enhance anonymization of private data while still permitting collaboration.

- To protect against malicious embedded attacks inside computations, all computations made available in COINSTAC are manually vetted.

Example Usage

In a recent study (Gazula et al., 2019), the Global Imaging Genetics of Adolescents consortium performed a decentralized voxel-based morphometry analysis of structural magnetic resonance imaging data across two sites to examine the structural changes in the brain related to age, body mass index, and smoking. The code used for the analysis (Gazula, 2020a) and sample data (Gazula, 2020b) are publicly available on GitHub.

Community Guidelines

We encourage collaboration from the community, including creating a computation to run in COINSTAC, conducting a study using COINSTAC, or contributing to our open-source code base.

Acknowledgements

References

Carter, K. W., Francis, R. W., Carter, K., Francis, R., Bresnahan, M., Gissler, M., Grønborg, T., et al. (2016). ViPAR: A software platform for the virtual pooling and analysis of research data. *International journal of epidemiology*, *45*(2), 408–416. doi:10.1093/ije/dyv193

Das, S., Zijdenbos, A. P., Vins, D., Harlap, J., & Evans, A. C. (2012). LORIS: A web-based data management system for multi-center studies. *Frontiers in neuroinformatics*, *5*, 37. doi:10.3389/fninf.2011.00037

Dwork, C., Roth, A., & others. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407. doi:10.1561/9781601988195

Gazula, H. (2020a). Coinstac code for decentralized regression with normal equation on MCIC data. *GitHub repository*. https://github.com/trendscenter/coinstac-regression-vbm; GitHub.

Gazula, H. (2020b). Contains MCIC data for coinstac-regression-vbm. *GitHub repository*. https://github.com/trendscenter/coinstac-regression-vbm-data; GitHub.

Gazula, H., Baker, B., Damaraju, E., Plis, S. M., Panta, S. R., Silva, R. F., & Calhoun, V. D. (2018). Decentralized analysis of brain imaging data: Voxel-based morphometry and dynamic functional network connectivity. *Frontiers in neuroinformatics*, *12*, 55. doi:10.3389/fninf.2018.00055

Gazula, H., Holla, B., Zhang, Z., Xu, J., Verner, E., Kelly, R., Schumann, G., et al. (2019). Decentralized multi-site VBM analysis during adolescence shows structural changes linked to age, body mass index, and smoking: A coinstac analysis. *bioRxiv*. doi:10.1101/846386

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., & Poldrack, R. (2017). OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for human brain mapping. Vancouver, Canada*, *1677*(2). doi:10.7490/f1000research.1114354.1

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). XNAT central: Open sourcing imaging research data. *NeuroImage*, *124*, 1093–1096. doi:10.1016/j.neuroimage.2015.06.076

Ming, J., Verner, E., Sarwate, A., Kelly, R., Reed, C., Kahleck, T., Silva, R., et al. (2017). COINSTAC: Decentralizing the future of brain imaging analysis. *F1000Research*, *6*. doi:10.12688/f1000research.12353.1

Plis, S. M., Sarwate, A. D., Wood, D., Dieringer, C., Landis, D., Reed, C., Panta, S. R., et al. (2016). COINSTAC: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Frontiers in neuroscience*, *10*, 365. doi:10.3389/fnins.2016.00365

Sherif, T., Rioux, P., Rousseau, M.-E., Kassis, N., Beck, N., Adalat, R., Das, S., et al. (2014). CBRAIN: A web-based, distributed computing platform for collaborative neuroimaging research. *Frontiers in neuroinformatics*, *8*, 54. doi:10.3389/fninf.2014.00054

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., Toro, R., et al. (2014). The ENIGMA consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, *8*(2), 153–182. doi:10.1007/s11682-013-9269-5