Computational Needs for Multimodal Explorations in Differential Autism Spectrum Disorder Phenotypes

Zachary Jacokes (zj6nw@virginia.edu) and John Darrell Van Horn (jdv7g@virginia.edu) for the Autism Centers of Excellence Consortium, Department of Psychology and School of Data Science, University of Virginia, Charlottesville, VA 22903

Introduction: Human neuroscience researchers have become increasingly interested in collecting large data sets due to the complex and multi-faceted phenotypic presentation of many disorders. Paramount to successful modern brain research is the ability to synthesize complementary data modalities to paint a circumspect picture of how disorders affect individuals holistically. Such data sets, therefore, include human neuroimaging, full-genome sequencing, microbiome, as well as multivariate phenomic assessments. As computing power availability and efficiency improves, so too does the quality and robustness of research utilizing such next-generation hardware and software. Research on Autism Spectrum Disorder (ASD) is a particularly salient use-case; indeed, ASD research is among the most exemplary cases of revelatory multimodal data sets (Hull et al 2017).

The Autism Centers of Excellence (ACE) Network at UVA: The initial research question the ACE network examined revolved around the vast diagnostic discrepancy between boys and girls with ASD. Boys are more than four times more likely to be diagnosed with ASD than girls (Fombonne 2009), and most research to this point has yielded little evidence indicating why this sex difference exists. Some have posited this could be a result of the female protective effect (Lei et al 2019), or the notion that a greater genetic mutational load is required for the ASD phenotype to manifest in girls than in boys (Chen et al 2017). In order to better understand the disorder, the UVA network initially sought to collect a vast array of data and data types in Wave 1, and then to analyze this data longitudinally after collecting a second set of data in Wave 2. The modalities we collected include imaging data (structural MRI [sMRI], functional MRI [fMRI], diffusion tensor imaging [DTI]), electroencephalography (EEG), genomics, and phenotype data. Data have been obtained under two major collection efforts, described briefly here:

ACE Wave 1: Wave 1 of the project involved four data collection sites and a centralized data coordination center (DCC). The data collection sites (Harvard University, Seattle Children's Hospital, UCLA, and Yale Child Study Center) each recruited and collected data from approximately 125 participants (523 participants in total). Of these participants, 245 were ASD (46.8%), 200 were neurotypical controls (38.2%), and 78 were neurotypical siblings of ASD participants (14.9%). Even numbers of male and female participants were recruited, most importantly in the ASD group (107 female; 43.7%), as the skewed ratio of males-to-females diagnosed with ASD is apparent in the cohort breakdown in many prior studies. Each participant went through the entire protocol, which included two structural imaging scans, four task-based functional scans, a resting-state functional scan, a diffusion tensor imaging scan, an electroencephalography scan, genetic testing, and various phenotypic assessments. This multimodal dataset is uniquely complex and presented significant data transfer, storage, and analysis challenges. The imaging data was aggregated at the DCC in a large-scale imaging database where raw data was stored locally and collaborators could access the data via either web interface or by directly interacting with the compute cluster via secure shell programs. The phenotype and genomics data was sent to the DCC stored on hard drives and disks.

ACE Wave 2: The second data collection phase of the project, beginning five years after the completion of Wave 1, will encompass the same participants and data modalities, with some improvements. The first and most important addition will be the incorporation of REDCap, an online database and survey distribution hub that we will be using to collect and aggregate the phenotype data. REDCap will eliminate the need to send hard copies of the phenotype data to a brick-and-mortar location by ensuring that all data will be collected and stored securely online. The intention is to re-recruit approximately half of the prior cohort, while recruiting new participants that will be age-matched with the Wave 1 cohort in order to better compare results since the participants from the Wave 1 cohort will have aged at least five years. Additionally, imaging data will be collected from our participants at two separate instances to increase the fidelity of the results.

Data Migration, Synthesis, and Security: The network's funding is contingent on consistent reporting of the data collection progress to the NIH as well as bi-annual data deposits into the National Database for Autism Research (NDAR). NDAR is a massive government database comprised of de-identified ASD data from any study that even tangentially involves examining the ASD phenotype. While it is important for the network to maintain a local database, NDAR is an invaluable resource for autism researchers around the world to analyze different cohorts and experimental designs within the ASD research environment.

Because the network is collecting human data, investigators are ethically bound by both Health Insurance Portability and Accountability Act (HIPAA) standards and Institutional Review Boards (IRB). While cloud computing and other cutting-edge data storage solutions are becoming more and more prevalent for huge datasets, the added layer of sensitivity encompassing personal health information (PHI) precludes us from utilizing such technologies. Additionally, data migration from one technology to another carries the risk of data loss and transcription errors, and such technologies tend to be prohibitively expensive at the outset. Longitudinal research typically spans many years, and cloud computing in particular is frequently charged as an ongoing expense, the cost of which compounds rapidly. As a result, local data storage and dissemination is the preferred methodology, which requires vast computing power and resources. This too is expensive, but as a one-time cost it becomes worthwhile as studies extend across years and potentially decades. At the

University of Virginia, we soon expect to house the nation's premier compute cluster designed for human subject research. This state of the art cluster will boast the highest level of security while maintaining inter-connectivity between sites sending us an enormous array of multimodal data. Collaboration is a major factor in successful research, and we intend to leverage our compute cluster to facilitate collaboration and eventually, scientific success, for years to come.

Recommendations: The ACE Network, based at UVA, is only one such example of modern human neuroscientific research having needs for not only HIPAA-compliant security but also for high-performance computing. Current UVA systems are efficient and robust; however, they have been designed to compartmentalize data from various sources and lack ability to centrally coordinate and synthesize data between projects, networks, and institutions. The ACE Network at UVA will likely necessitate an augmented system architecture model: vastly upgraded local computing cluster in terms of hardware as well as the implementation of more sophisticated neuroscience data transfer protocols, specifically utilizing transfer services such as Globus. When handling the estimated petabyte of subject data the ACE Network expects to collect, transfer speed and security become imperative to consortium-wide research efforts. Additionally, front-end containerization solutions are necessary not only to standardize the processing workflow involved in these data types, but also to simplify the human-compute cluster interaction. Research at a granular level is often carried out by transient research assistants, and due to the high turnover rate among research assistants, the on-boarding process needs to be simple enough that anyone with access to an internet browser can effectively produce research-quality results. Stable, robust, and reliable data storage and high-performance computation are, broadly speaking, needed to ensure a solid foundation for neuroscience research conducted 'at scale'.

Conclusion: The UVA network has encountered and addressed unique needs for high performance computing in many ways. Existing UVA computing is excellent though room to expand the middle ground between heavily secure and multi-CPU processing provides a unique opportunity. While this represents the cutting-edge of research computing, soon many more research institutions and universities will likely follow our lead in an attempt to examine disorders and, indeed, the human experience more generally, by collecting multimodal datasets that require large computing resources. With such a system deployed, a large-scale network effect is likely to emerge, where collaboration among compute cluster hubs is commonplace and data are shared as quickly as the internet allows. Moving forward, UVA neuroscience computing resources will represent the next generation of research computing as the community looks to further enhance ways to address the challenges presented by ever-larger brain research data sets. We envision UVA being the standard-bearer for neuroscience research computing and to lead the way toward understanding, treating, and perhaps eventually curing, disorders associated with brain development, dysfunction, and cognitive function across the lifespan.

References

Chen, C., & Van Horn, J. (2017). Developmental neurogenetics and multimodal neuroimaging of sex differences in autism. *Brain Imaging and Behavior, 11*, 38-61.

Fombonne, E. (2009). Epidemiology of Pervasive Developmental Disorders. *Pediatric Research, 65*, 591-598. doi:https://doi.org/10.1203/PDR.0b013e31819e7203

Hull, J., Dokovna, L., Jacokes, Z., Torgerson, C., Irimia, A., & Van Horn, J. (2017). Resting-State Functional Connectivity in Autism Spectrum Disorders: A Review. *Frontiers in Psychiatry, 7*(205). doi:https://doi.org/10.3389/fpsyt.2016.00205

Lei, J., Lecarie, E., Jurayj, J., Boland, S., Sukhodolsky, D., Ventola, P., . . . Jou, R. (2019). Altered Neural Connectivity in Females, But Not Males with Autism: Preliminary Evidence for the Female Protective Effect from a Quality-Controlled Diffusion Tensor Imaging Study. *Autism Research, 12*(10), 1472-1483. doi:https://doi.org/10.1002/aur.2180