

Harnessing cloud computing for high capacity analysis of neuroimaging data from NDAR



Daniel Clark¹, Christian Haselgrove², David Kennedy², Zhizhong Liu³,
Michael Milham¹, Petros Petrosyan⁴, Carinna Torgerson³, John Van Horn³, Cameron Craddock¹

¹Child Mind Institute, New York, NY, ² University of Massachussetts Medical School, Worcester, MA, ³University of Southern California, Los Angeles, CA, ⁴UCLA, Los Angeles, CA, ⁵Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY

Introduction

- ▶ The National Database for Autism Research (NDAR) hosts a vast collection of neuroimaging datasets that can be processed and utilized to yield significant scientific discoveries.
- ▶ This amount of resources necessitates a high-performance computing (HPC) infrastructure, which is not always readily available for researchers in-house.
- ▶ Amazon Web Services (AWS) Elastic Compute Cloud (EC2) computing service offers a “pay as you go” model that allows researchers to utilize HPC performance without the up-front capital costs and maintenance of an in-house solution.
- ▶ The developers of the Laboratory of Neuro Imaging (LONI) Pipeline, the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) Computational Environment (CE) and the Configurable Pipeline for the Analysis of Connectomes (C-PAC) have implemented pipelines in EC2 that interface with NDAR

Methods

LONI Pipeline

- ▶ The LONI Pipeline software was extended to include new pipeline modules to access data from the NDAR database, transfer input data out of Amazon S3 (Simple Storage Service), and to load results back into S3¹
- ▶ A pipeline was constructed to extract cortical thickness and subcortical region volume data from structural MRI images in the NDAR database, which included:
 1. Reorient images to standard orientation using FSL’s reorient2std module
 2. Extract cortical thickness using FreeSurfer recon-all
 3. Calculate volumes of subcortical regions using FSL’s BET and FIRST all
- ▶ The resulting pipeline was used to process 780 T1-weighted structural images and return the results to NDAR

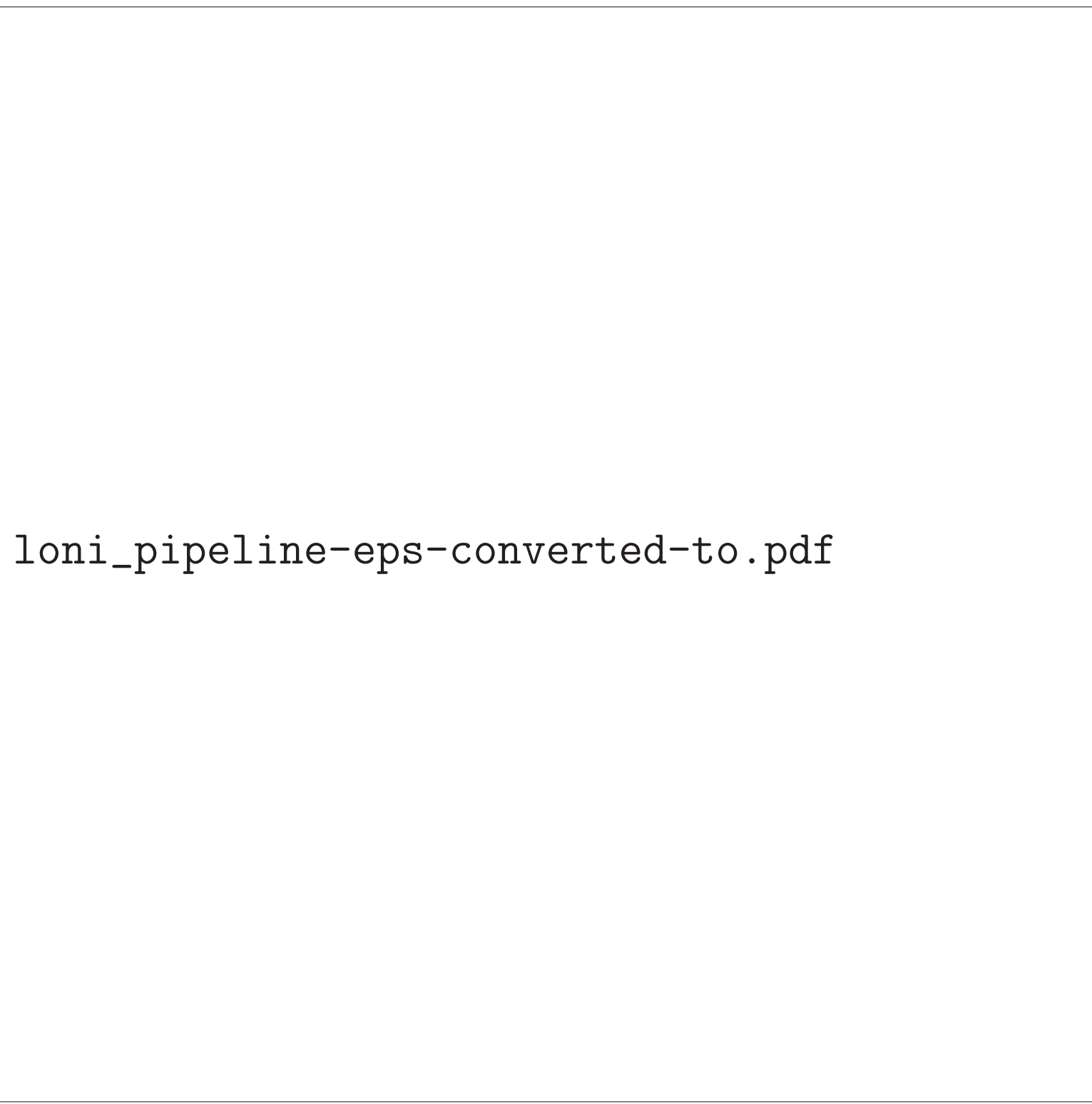


Figure 1 : Graphical layout of the constructed pipeline

Configurable Pipeline For the Analysis of Connectomes (C-PAC)

- ▶ C-PAC modules were written in Python to build input data lists by querying NDAR, read input data from S3, write processed results to S3 and write values back to the NDAR database²
- ▶ New pipelines were created to perform the ANTS cortical thickness³ procedure and the Preprocessed Connectomes Project’s Quality Assessment Protocol (<http://preprocessed-connectomes-project.github.io/quality-assessment-pipeline>)⁴
 1. Reorient images to standard orientation using FSL’s reorient2std module
 2. Extract cortical thickness using FreeSurfer recon-all
 3. Calculate volumes of subcortical regions using FSL’s BET and FIRST all
- ▶ The resulting pipeline was used to process 780 T1-weighted structural images and return the results to NDAR

Connections	Columns	Data Constraints	Grants	Statistics	Triggers	Flashback	Dependencies	Details	Partitions	Indexes	SQL
Columns	DATA	TYPE	NULLABLE	DATA	DEFAULT	Column ID	Comments				
1	IMAGE_ID	NUMBER(38, 0)	Yes	(null)		1	(null)				
2	OUTSET_ID	NUMBER(38, 0)	Yes	(null)		2	(null)				
3	SUBJECT_ID	VARCHAR2(255 BYTE)	Yes	(null)		3	(null)				
4	SPC_SUBJECT_ID	VARCHAR2(255 BYTE)	Yes	(null)		4	(null)				
5	INTERVIEW_DATE	DATE	Yes	(null)		5	(null)				
6	INTERVIEW_AGE	NUMBER(38, 0)	Yes	(null)		6	(null)				
7	USER	VARCHAR2(255 BYTE)	Yes	(null)		7	(null)				
8	COMMENTS_MISC	VARCHAR2(1000 BYTE)	Yes	(null)		8	(null)				
9	IMAGE_FILE	VARCHAR2(255 BYTE)	Yes	(null)		9	(null)				
10	IMAGE_THUMBAILL_FILE	VARCHAR2(255 BYTE)	Yes	(null)		10	(null)				
11	IMAGE_DESCRIPTION	VARCHAR2(512 BYTE)	Yes	(null)		11	(null)				
12	IMAGE_FILE_FORMAT	VARCHAR2(50 BYTE)	Yes	(null)		12	(null)				

- ▶ Launched an AWS-hosted miNDAR database by querying NDAR website for the data of interest (e.g. from a particular study)

Results

Table 1 : Processing completed as a part of the effort. Nodes corresponds to the number of hosts used in the calculation. PF is parallelization factor and corresponds to the number of jobs ran in parallel on each node. On demand instances were used for the master node and spot instances were used for all computation nodes in the cluster. CPU Time is the total amount of time required to perform the computation and Wall Time is the amount of time that passed. # DS: Number of datasets. CPD: Cost Per Dataset. C-PAC: Configurable Pipeline for the Analysis of Connectomes. NITRC-CE: NITRC Computational Environment

Processing	# DS	Platform	Nodes	PF	CPU Time	Wall Time	Cost	CPD
ANTS Cortical Thickness	3197		20	8	23,018	147	\$760.24	\$0.24
Resting state fMRI processing w/ 4 strategies	1112	C-PAC	20	3	834	22	\$80.54	\$0.07
Quality Assessment Protocol	1112		20	4	380	14	\$19.02	\$0.02
Freesurfer recon-all	986		4	32	23,664	193	\$211.44	\$0.21
FSL FIRST	1247	NITRC-CE	4	32	208	3	\$2.19	> \$0.01
Temporal QA	1349		4	32	450	13	\$4.69	> \$0.01
Freesurfer recon-all and FSL FIRST	780	LONI Pipeline	20	32	18,720	49	\$252.36	\$0.32

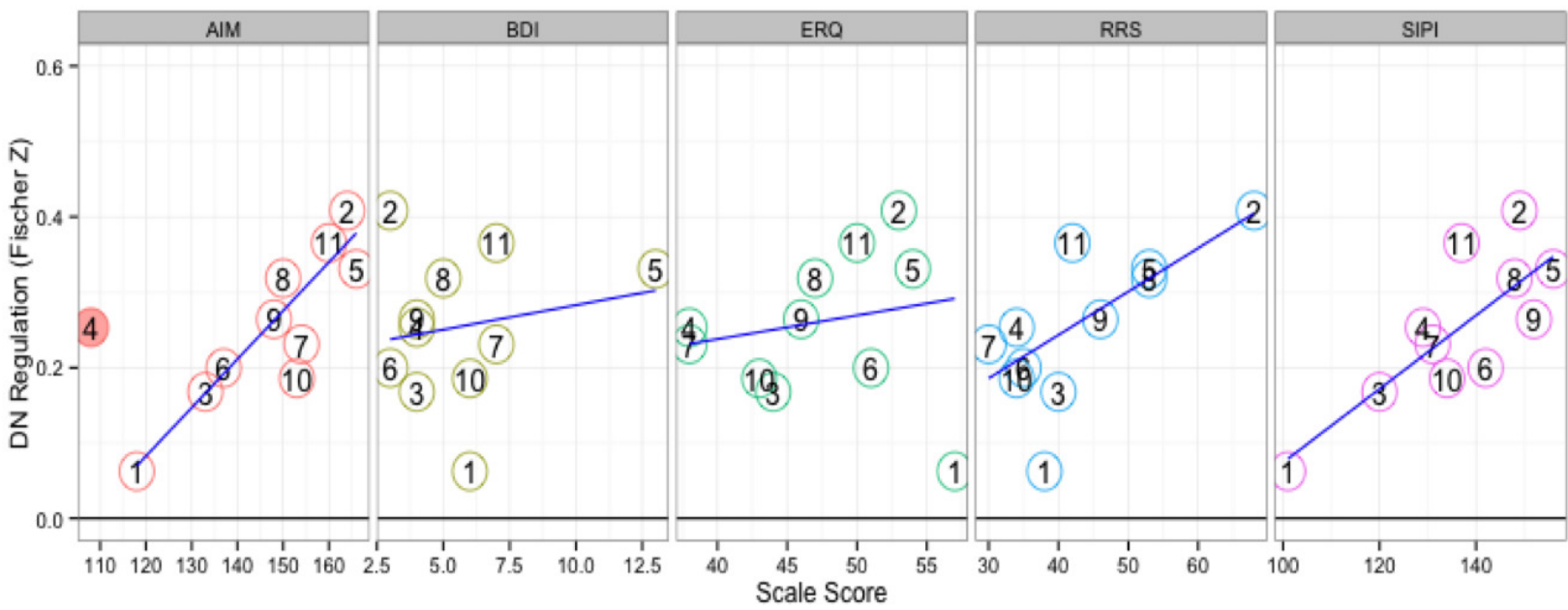


Figure 4 : Inter-individual variation in performance correlates with behavioral measures.

- ▶ As shown in figure 4 models learned for the best and worst performing participants match with the canonical pattern of the default network.
- ▶ The best participant was able to follow the instructions very well 4, the worst seems to have been corrupted by noise.
- ▶ Figure 5 shows 12 of the subjects were able to modulate the DN at above chance levels, performance on feedback runs is consistent independent of order, but performance on nonfeedback runs improves if they occur after feedback runs.
- ▶ Measures that were significantly associated with DN regulation include ($p < 0.05$, FDR corrected): the affect intensity measure (AIM), ruminative responses scale (RRS), and the imaginal processes inventory.

Conclusion

- ▶ We developed a system for measuring DN regulation using realtime neurofeedback.
- ▶ Participants were able to modulate their DN activity and their ability to do so was correlated with phenotype.
- ▶ This system provides a new experimental paradigm for understanding network dysregulation and how it maps to disease states and phenotype.

References and Acknowledgements