

NeuroCI: Continuous Integration of Neuroimaging Results Across Software Pipelines and Datasets

Jacob Sanz-Robinson, Arman Jahanpour, Natalie Phillips,
Tristan Glatard, Jean-Baptiste Poline.



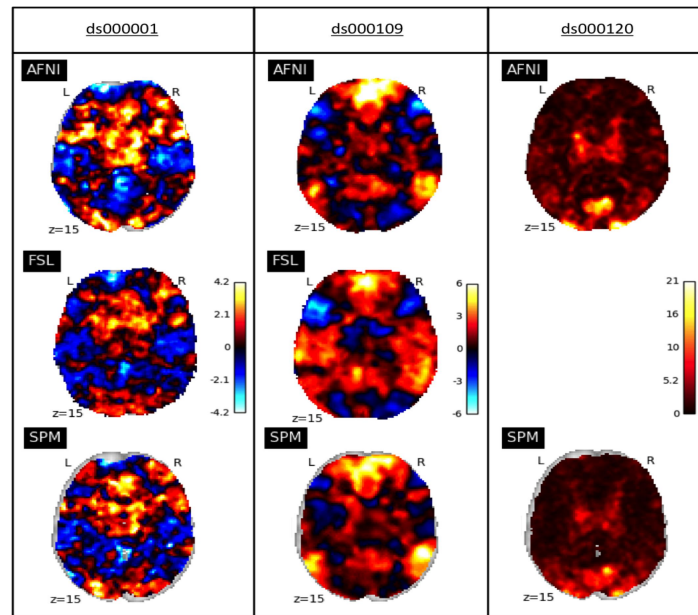
McGill
UNIVERSITY



UNIVERSITÉ
Concordia
UNIVERSITY

Context - Neuroimaging Reproducibility Crisis

- Neuroimaging results are sensitive to variations in processing pipelines, contributing to scientific result reproducibility issues. [\[Kennedy et al., 2019\]](#)[\[Botvinik-Nezer et al., 2020\]](#)[\[Bowring et al., 2019\]](#)
- Researchers often faced with multiple choices of pipelines featuring similar capabilities, and which may yield different results when applied to same data.
- No ground truth: unclear which pipeline to use when they yield different results.



Comparison of unthresholded statistic maps of brain activations [\[Bowring et al., 2019\]](#)

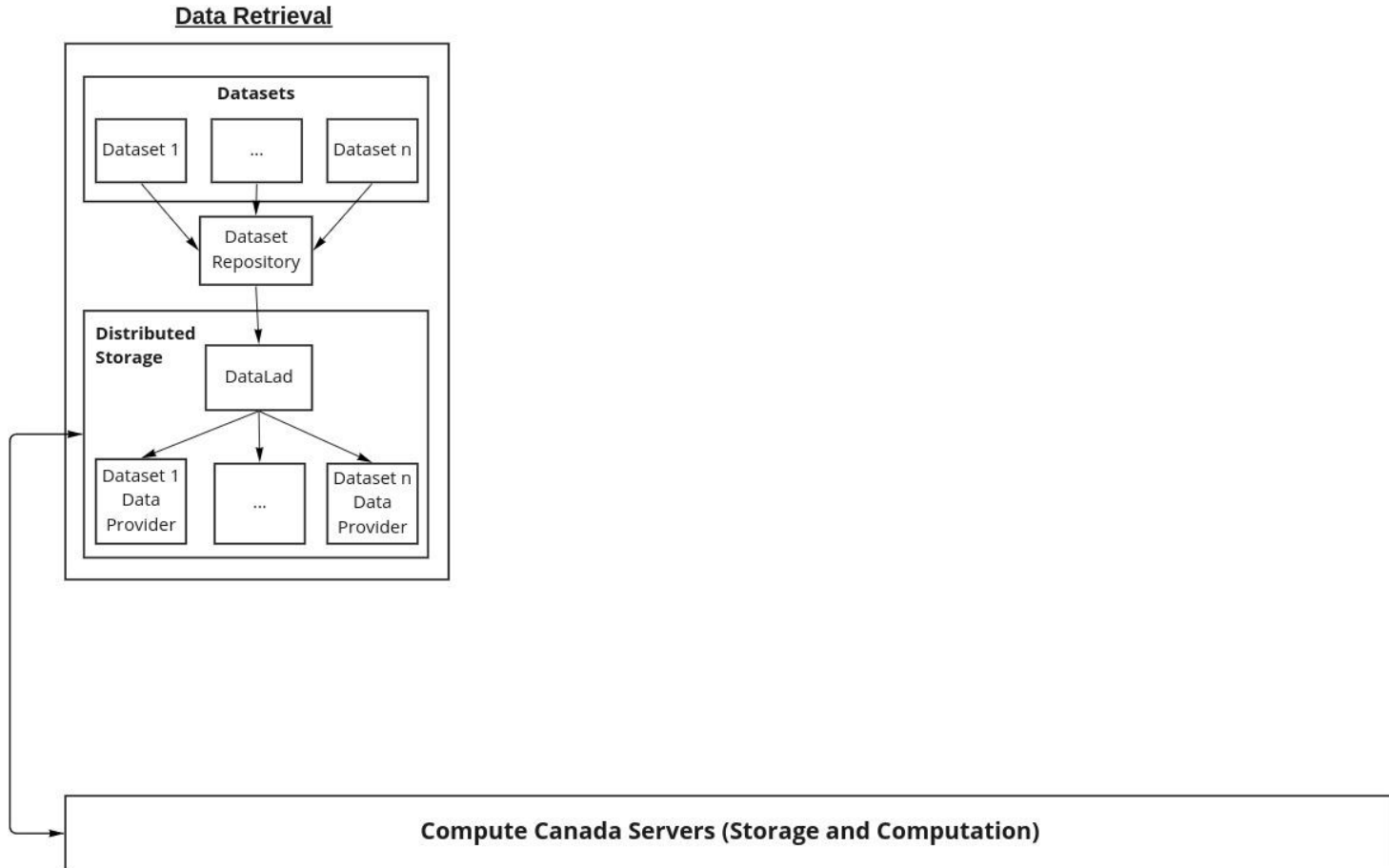
Context - Continuous Integration (CI)

- CI has been proposed as a method to evaluate and facilitate reproducibility [\[Krafczyk et al., 2019\]](#)[\[Beaulieu-Jones et al., 2017\]](#).
- Software engineering technique: series of tests launched as soon as new code has been committed to repository to check if expected code functionality is obtained.
- Applying CI to neuroimaging: challenging due to the large storage & memory requirements, and privacy conditions of datasets.
- To solve these limitations, we built a capable distributed computation infrastructure!

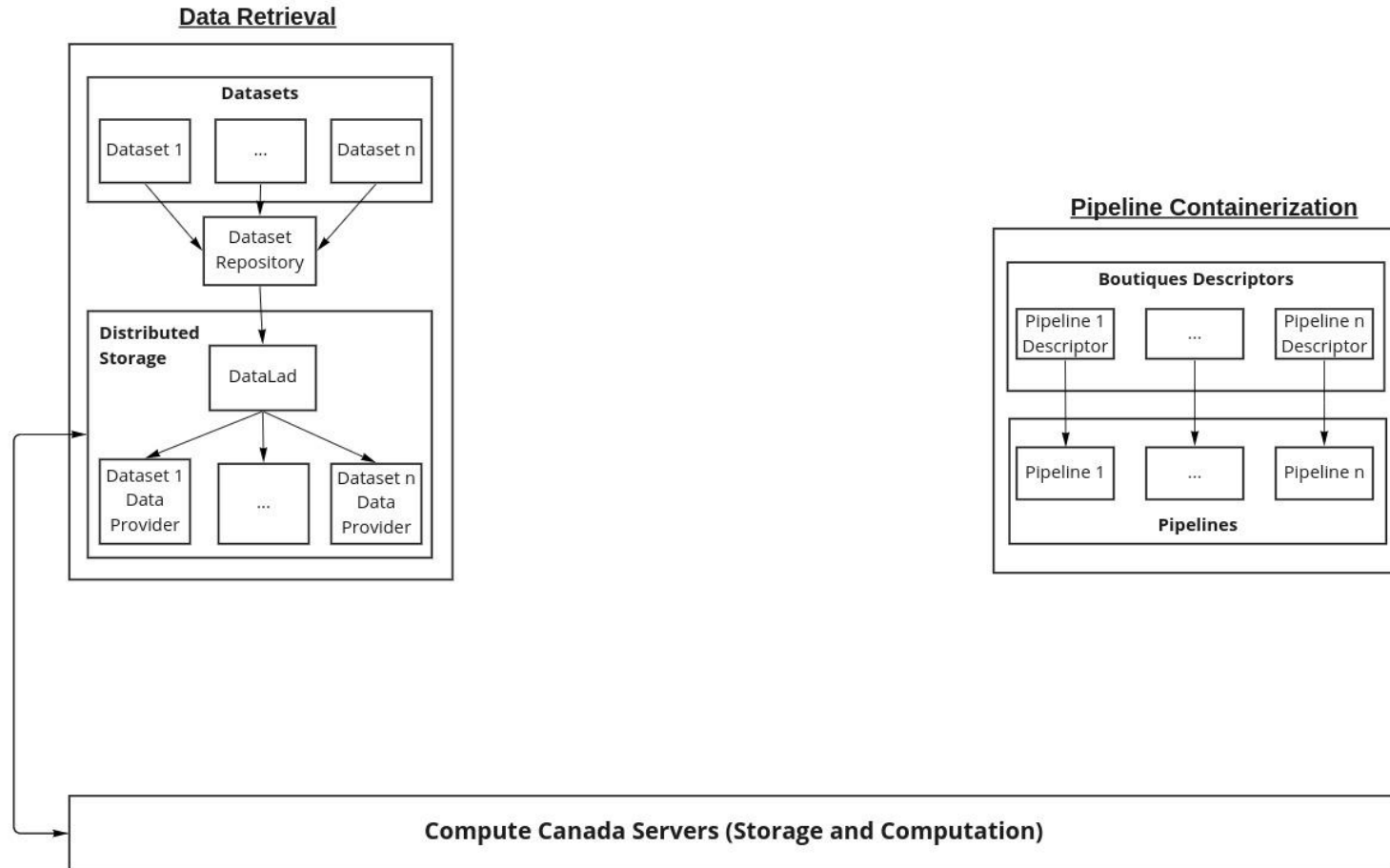
Goal

- Use CI Framework we built to **investigate associations between hearing loss and brain structure** - robustness and reliability of result unclear.
- Some studies agree on hearing loss reliably predicting gray & white matter atrophy and being associated with dementia [\[Lin et al., 2014\]](#)[\[Alfandari et al., 2018\]](#)[\[Armstrong et al., 2019\]](#).
- Null results have also been found [\[Profant et al., 2014\]](#), and other null results may not have been published (cf. the “file drawer effect” [\[Rosenthal, 1979\]](#)).

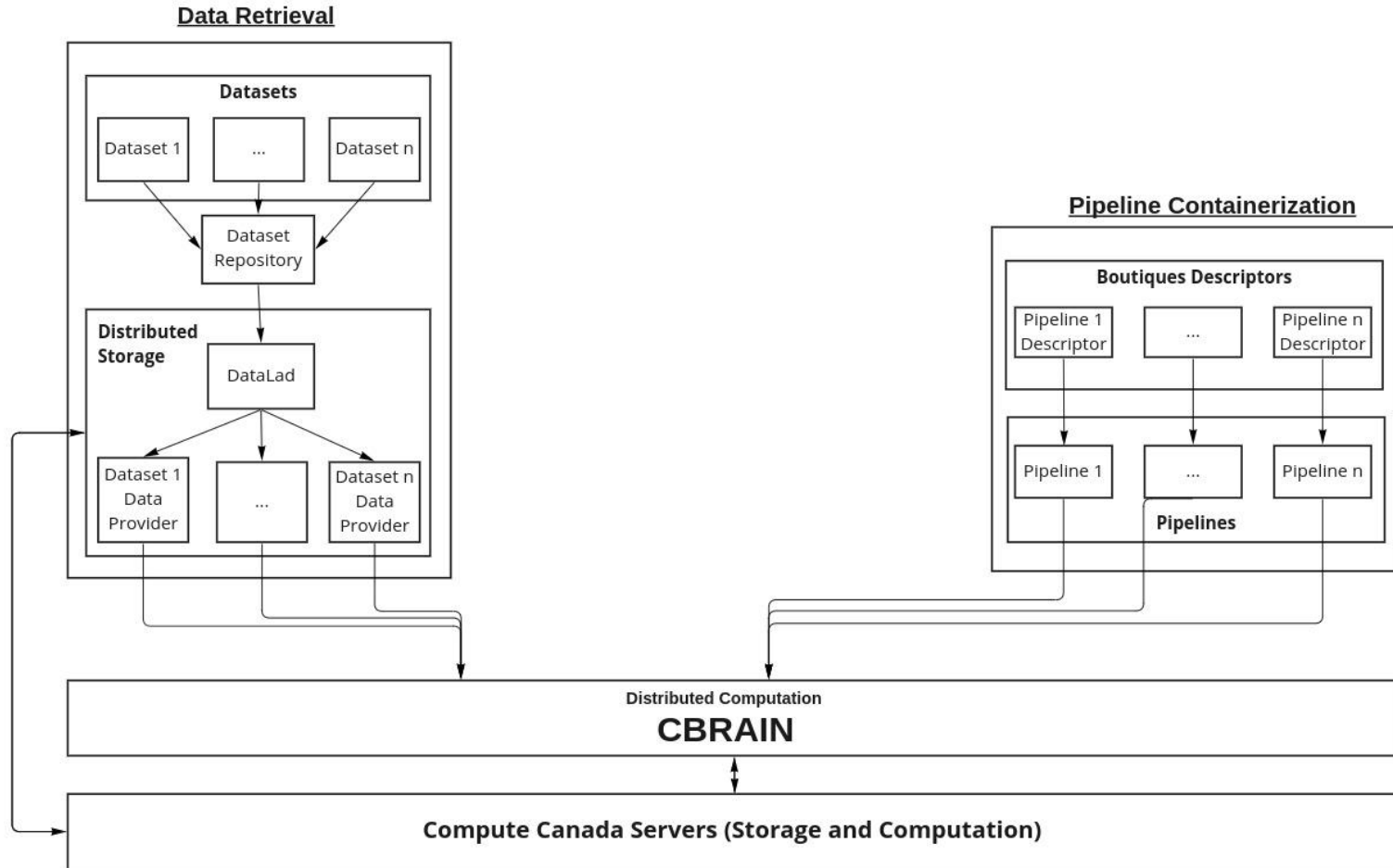
Method



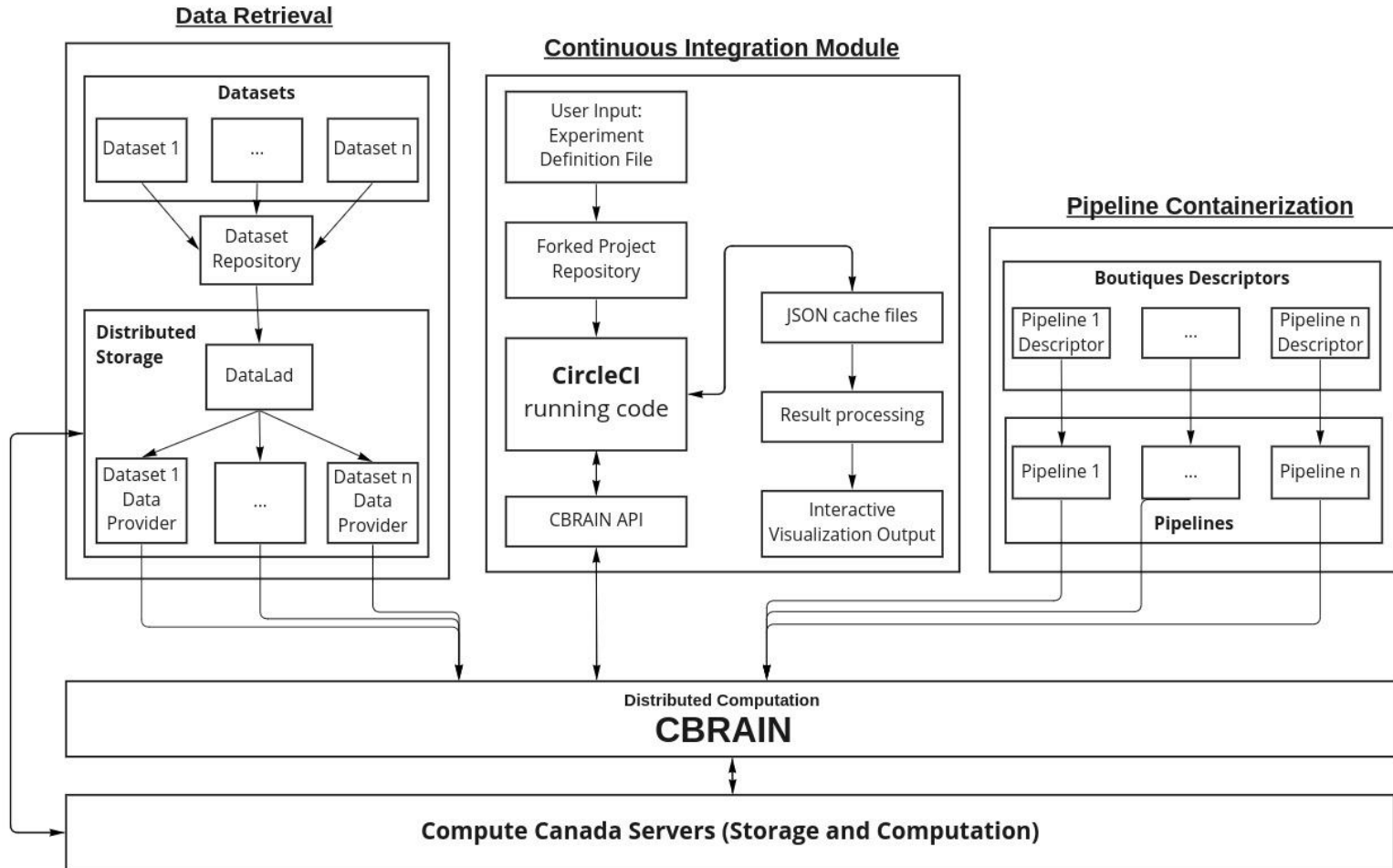
Method



Method

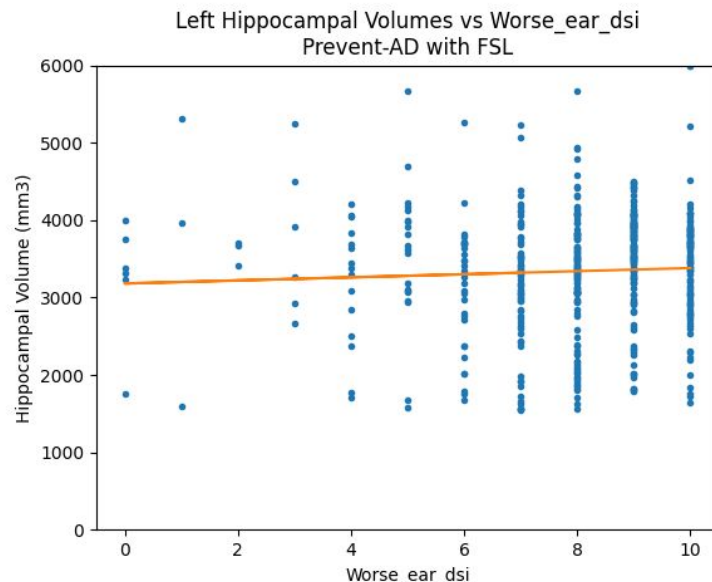
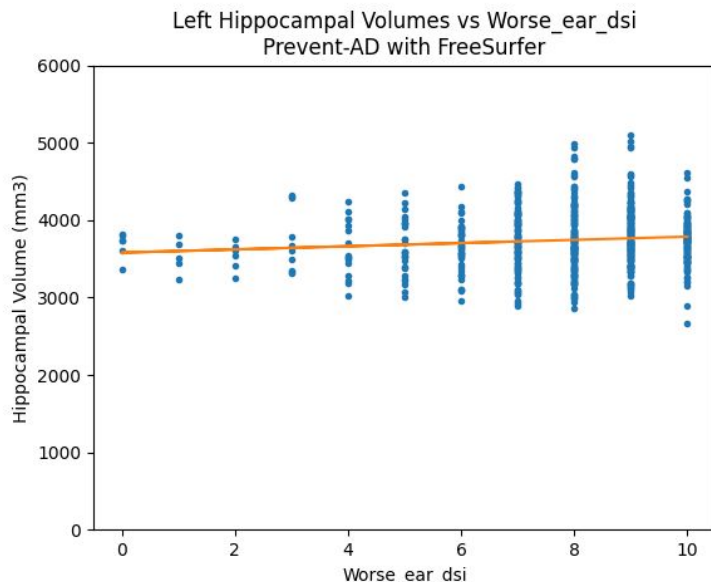


Method



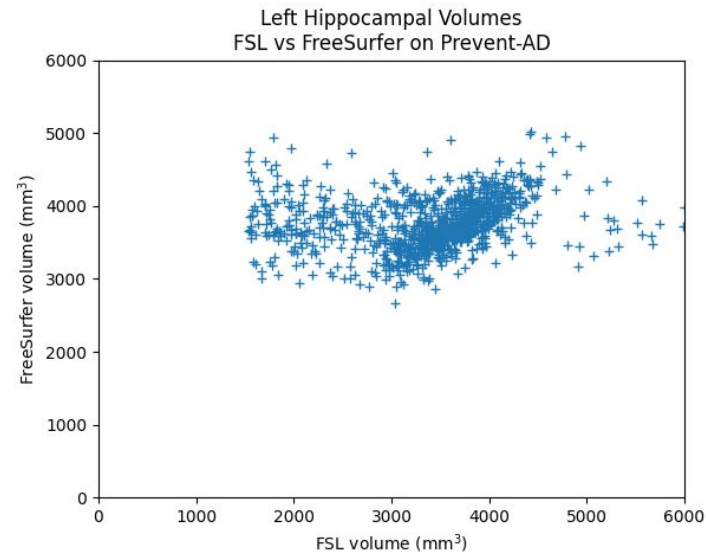
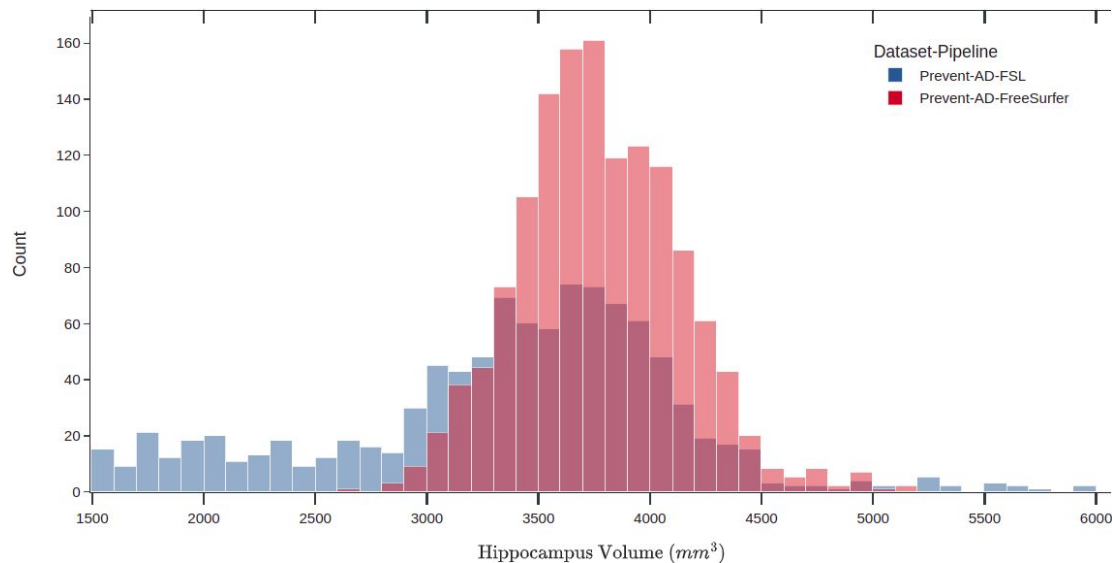
Results I

- Functional prototype of the NeuroCI framework and CI runs: <https://github.com/neurodatascience/NeuroCI>
- Results outside the 1,500-6,000mm³ range are considered outliers and removed [[Honeycutt et al., 1995](#)].



Correlations between left hippocampal volumes measured with FreeSurfer 6.0.0 (left: $r=0.11$, 95% c.i. [0.033, 0.185]) and FSL 5.0.9 (right: $r=0.05$, 95% c.i. [-0.027, 0.127]) and hearing loss on Prevent-AD data. Hearing loss is measured with the worse ear's Dichotic Stimulus Identification (DSI) score. A higher DSI score indicates better central auditory system function.

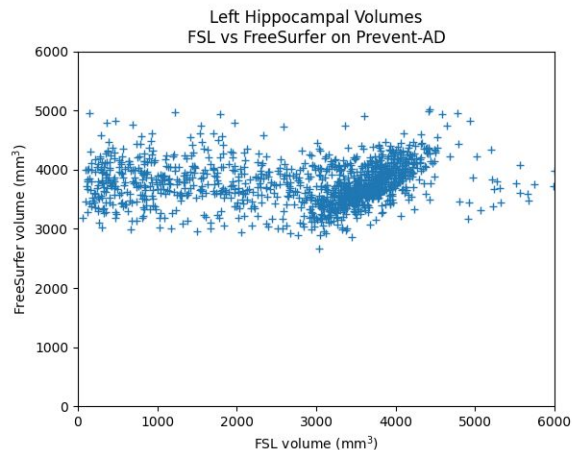
Results II



- Pearson correlation = 0.2003 after outlier removal.
- FSL: larger spread of values, asymmetrical distribution, skewed towards lower volumes
- FreeSurfer: distribution is roughly symmetrical and consists of higher values.
- [\[Gomez-Ramirez et al., 2021\]](#) agreed with.

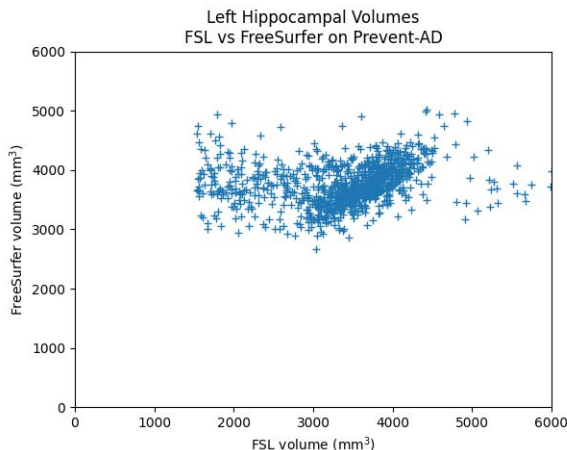
Outlier Removal Via Pipeline Discrepancies

No Outlier Removal



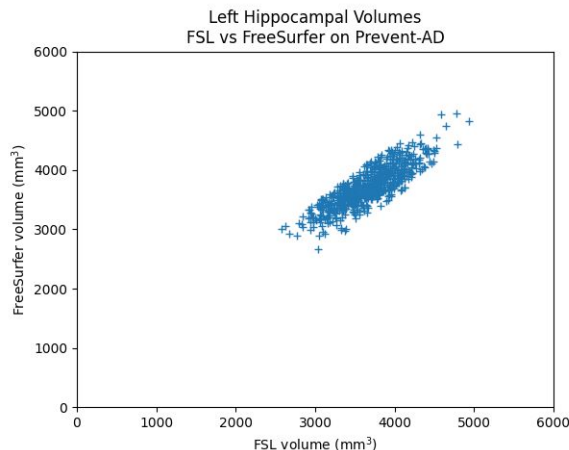
- All successful segmentations.
- 1281 data points plotted.
- Pearson correlation = 0.0429.

Bounded Outlier Removal



- Points under 1500mm³ or above 6000mm³ removed.
- 987 data points plotted.
- Pearson correlation = 0.2003.

Discrepant Outlier Removal



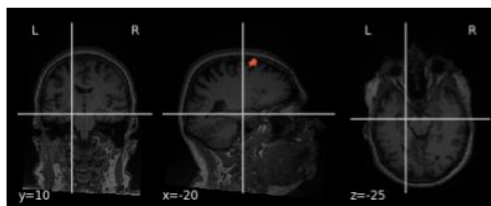
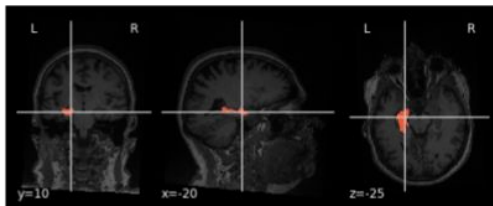
- Points with inter-pipeline discrepancies above 450mm³ removed.
- 636 data points plotted.
- Pearson correlation = 0.84.

Outlier Removal Method experiments

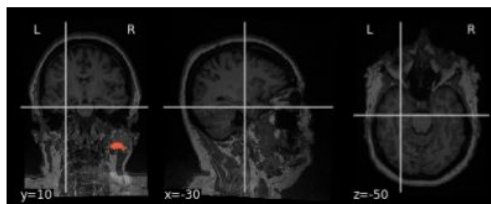
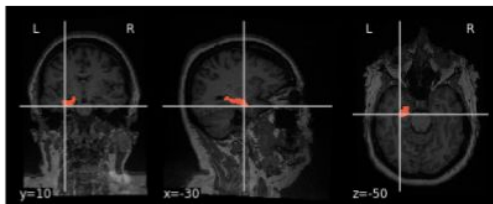
FreeSurfer

FSL

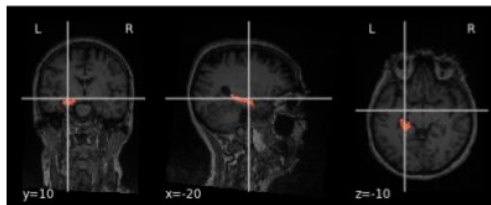
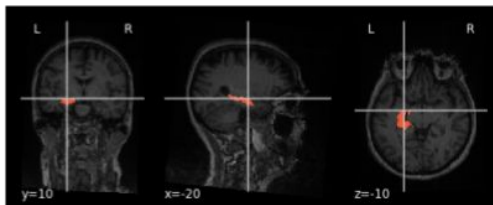
Example #1
discrepancy of
 $13,006.3\text{mm}^3$



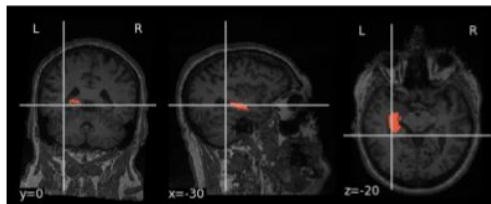
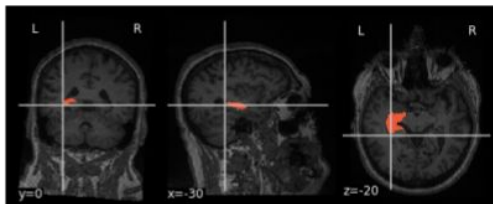
Example #2
discrepancy of
 $5,227.7\text{mm}^3$



Example #3
discrepancy of
 751.6mm^3



Example #4
discrepancy of
 531.4mm^3



#1, #2: Very large discrepancies flagged - FSL segmentation failures.

#3, #4: Significant discrepancies flagged - successful segmentations

We used the median pipeline discrepancy (450mm^3) as our flagging threshold.

Conclusions

- Framework generalizes to other neuro domains:
 - Systematically evaluate result variability in experiments.
 - Pinpoint biases/discrepancies caused by tools & datasets, quantify impact on results.
 - Consolidate knowledge & explain uncertainty in fields with reproducibility concerns.
- Health-related use-cases (eg hearing loss associations with brain structure):
 - Reliable results → Most likely robust biomarkers for researchers and practitioners.
- Future work: additional pipelines (CIVET, ANTs, MINC tools) and datasets (COMPASS-ND, UK-Biobank), Slurm compatibility.

The End

