Transfer Learning: Using Big Data for Classification in Clinical (Stuttering) Data

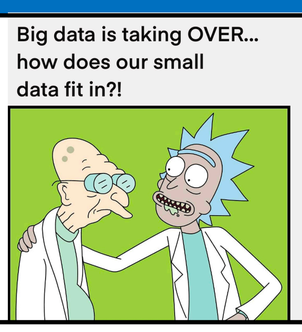
DEPARTMENT OF PSYCHIATRY
MICHIGAN MEDICINE

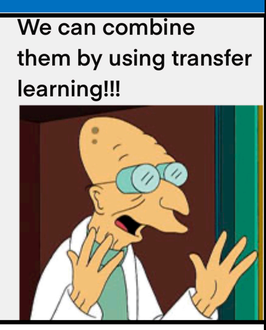
Saige Rutherford, Mike Angstadt, Chandra Sripada, Soo-Eun Chang

@being_saige

MOTIVATION

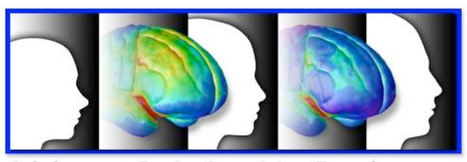
Big datasets have become the new norm in neuroimaging. Most of these big datasets have focused on healthy adult populations. The heterogeneity of clinical populations makes creating datasets of equal size and quality more challenging. There is a strong need for methods that can connect the power of big data with the carefully curated clinical datasets collected over the past ~20 years.





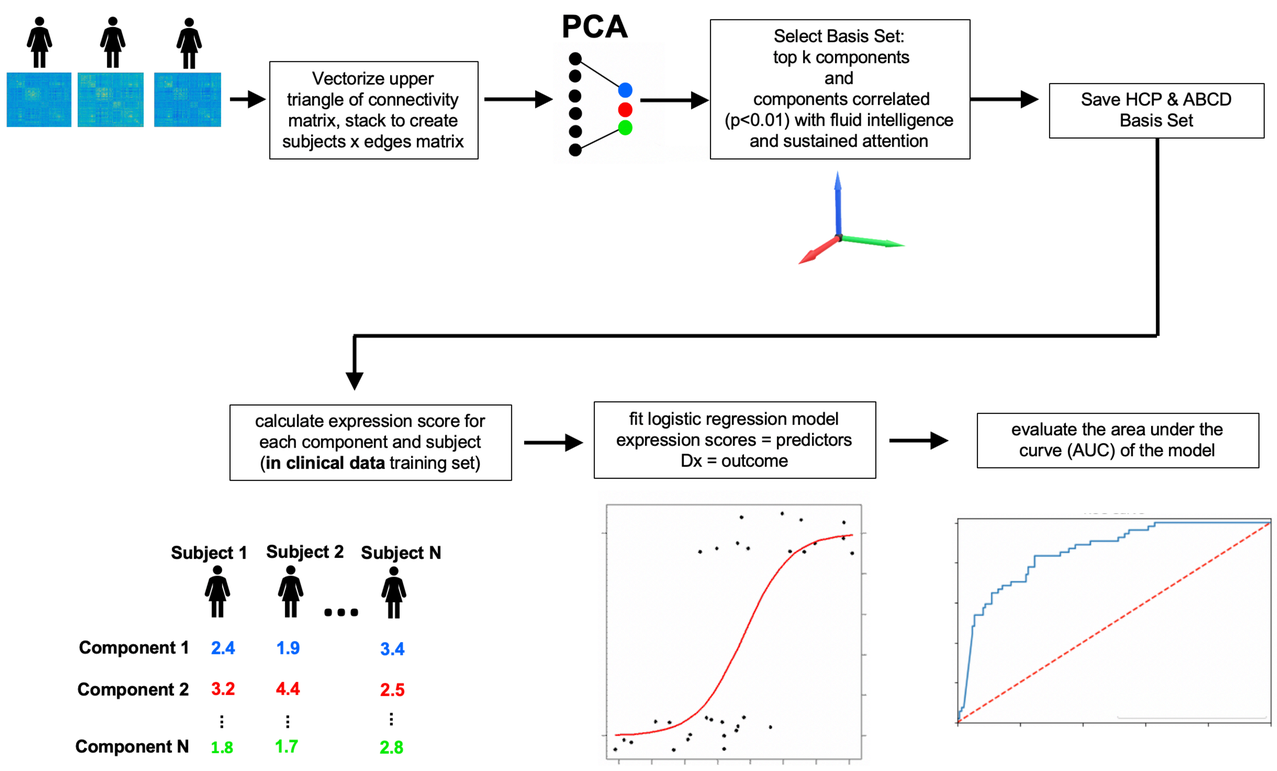
METHODS

Feature Discovery Datasets
ABCD (N=1509) & HCP (N=910)



Out of sample UMich test dataset (N=121) from children who stutter and controls (CWS=61, HC=59)

Inclusion criteria: high quality T1w & ≥5 min of rs-fMRI data after motion censoring.



1. Create a basis set (PCA decomposition) in ABCD, HCP, & In Sample

   2 ways of selecting basis set components: top k (highest variance explaining) or chose components that are sig. correlated with (p<0.01) attention/cognition phenotypes in ABCD & HCP.

2. Train Logistic Regression model using 10-fold cross validation to classify CWS vs. Controls, using the basis set as the predictors and the clinical label as the outcome (control for age/sex/motion)

   ^^You can think of steps 1-2 as Principle Components Regression

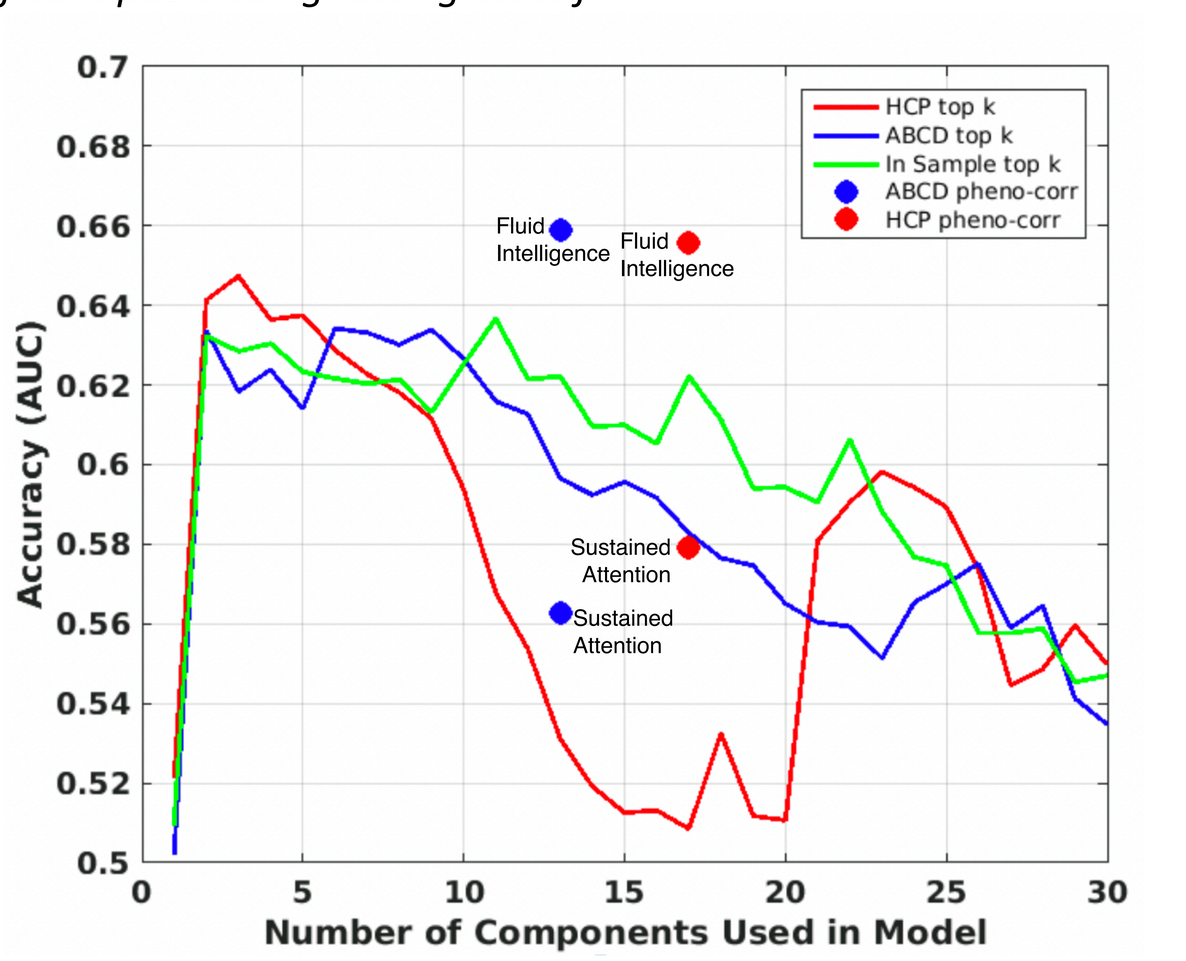
3. Probe models to *attempt* to better understand model accuracy

   "Keep 2 Network" Analysis: What brain networks are contributing most to the prediction?

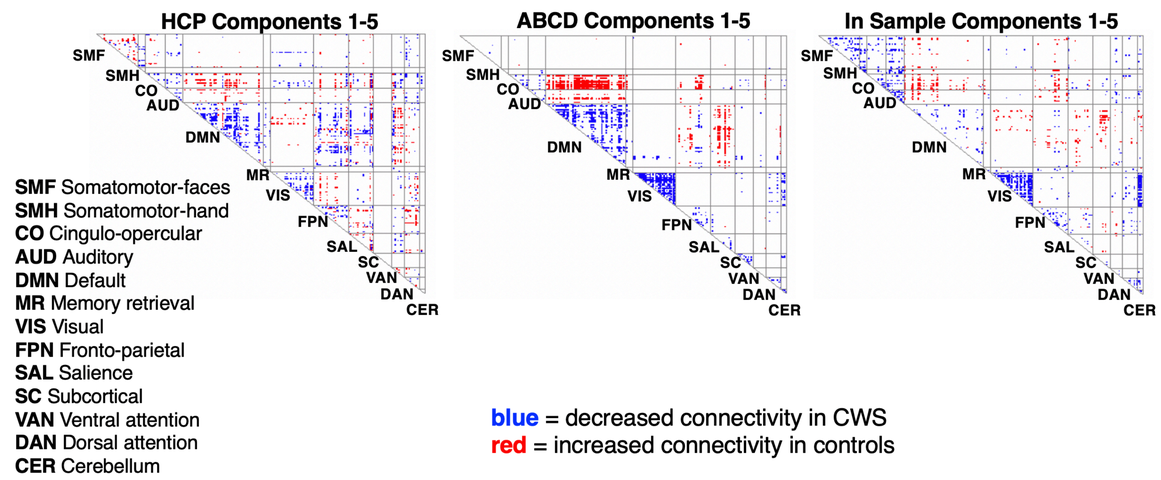
For more detailed methods, figures & additional results, (along with all the code, basis sets, & pretrained models) check out this study's GitHub!
https://github.com/saigerutherford/OHBM2020-stuttering

RESULTS

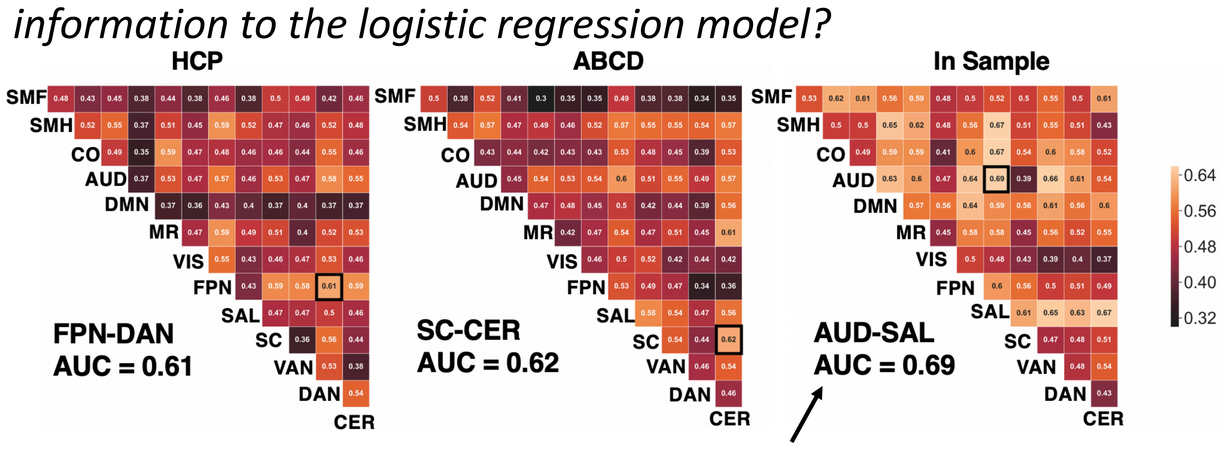
How accurately can we classify children who stutter from fluent peers using resting-state fMRI?



What do these top k basis sets look like?



Keep 2 Networks: What brain networks contribute the most information to the logistic regression model?



TAKE HOME MESSAGES

Unsurprisingly, there is no "best" model. When using a few (~1-10) top variance explaining components (spanning the whole brain), ABCD & HCP basis sets improve classification accuracy compared to the in-sample basis set. However, when using a reduced basis set (from just 2 brain networks, e.g. auditory & salience), the in-sample basis set out performs ABCD & HCP. This suggests that clinical variation has a fine-scale resolution that may be overlooked when searching the full connectome.