TAKEAWAYS

- We use ~56k NIMH funded full-text papers on Pubmed Central to estimate prevalence of shared code or data
- By 2018, ~14% of all papers from 2018 mentioning at least one code or data repository, with this growth largely driven by increased use of Github
- We explored basic machine learning approaches to classify shared data and data reuse

Data sharing and reuse in the mental health sciences

Travis Riddle[1] & Adam Thomas[1]

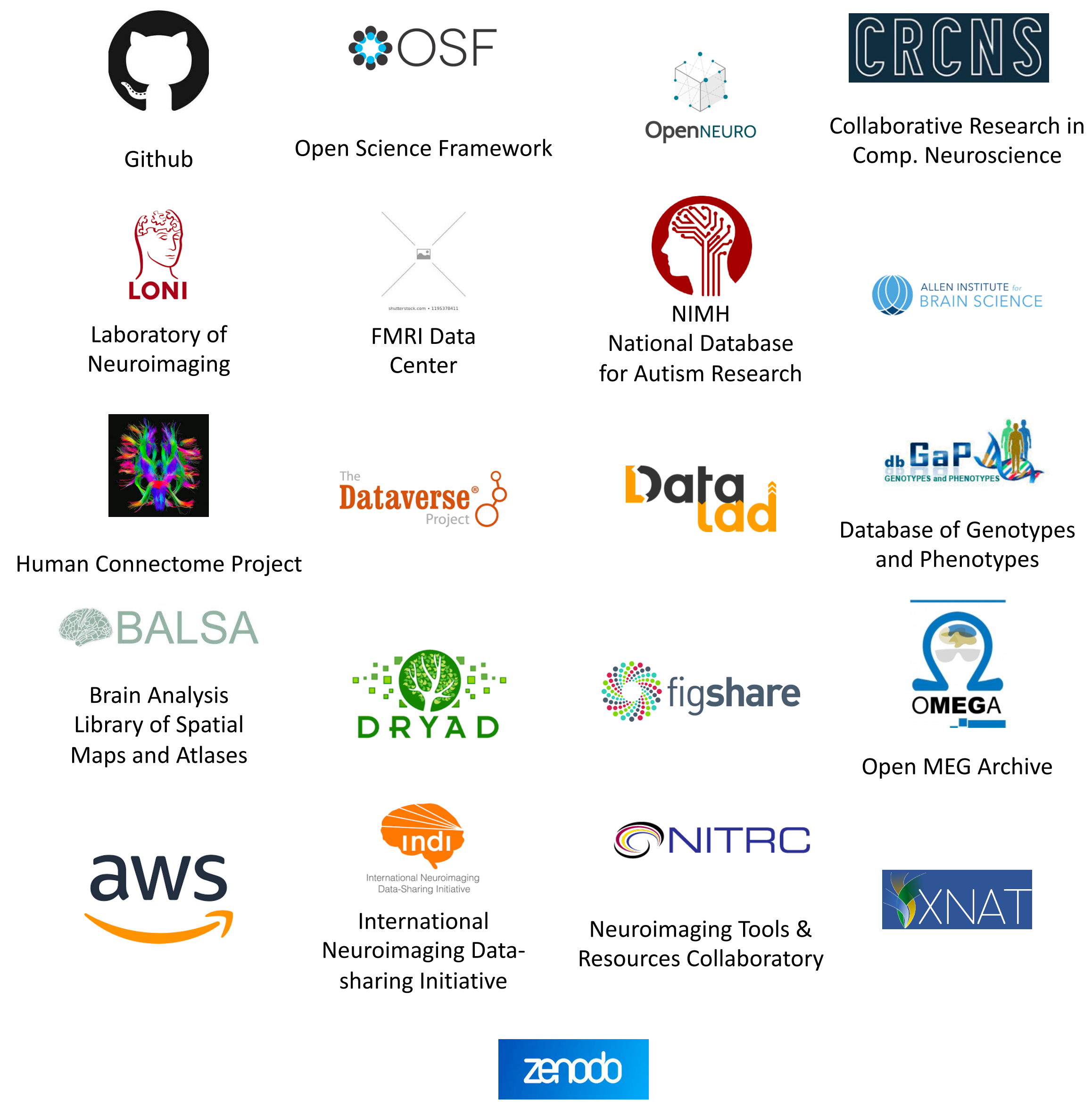
Introduction

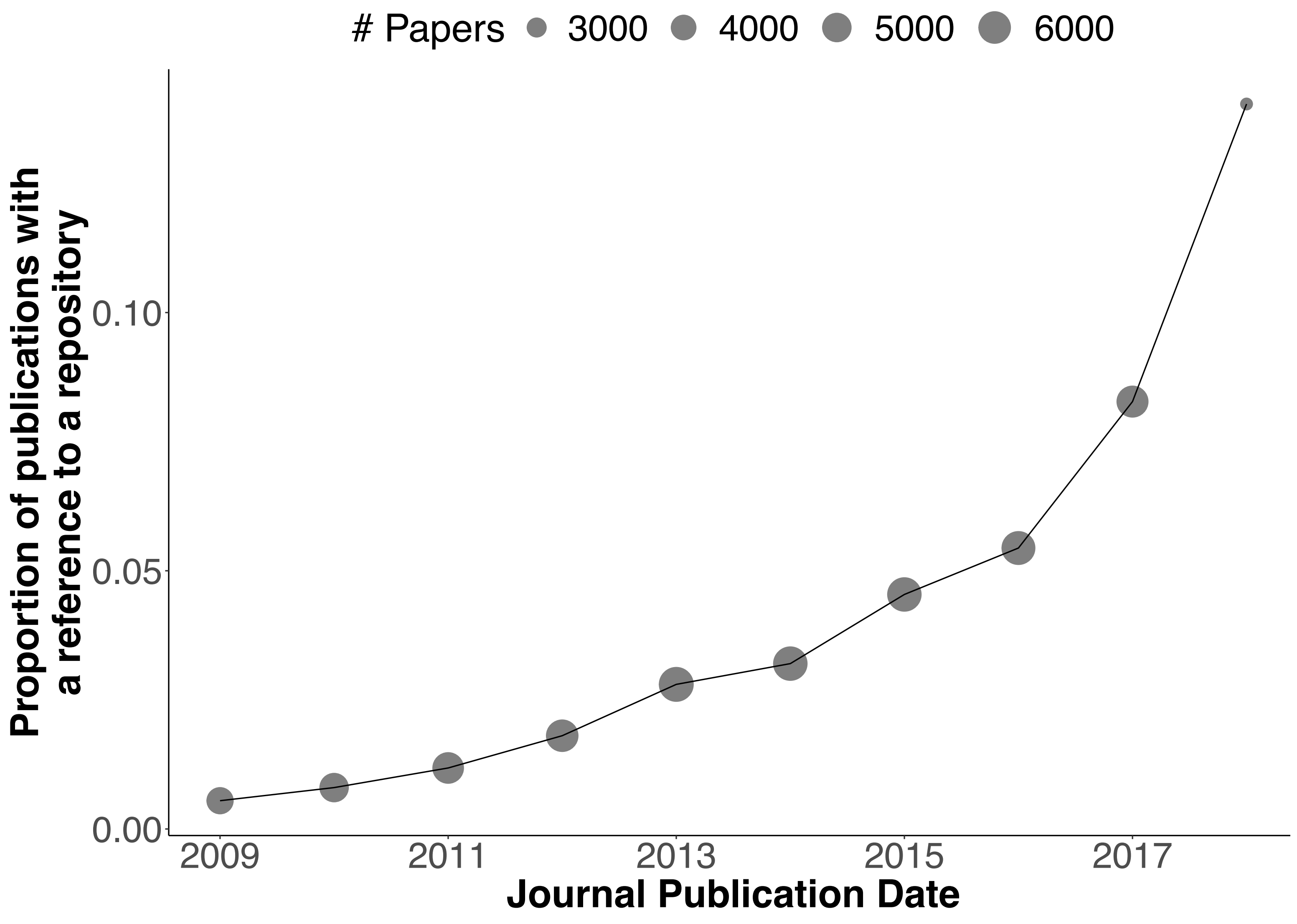
- The 2017 Cures Act authorizes NIH Director to require award recipients to share data in a manner consistent with applicable laws and regulations[2]
- We sought to estimate how often researchers in the mental health sciences share data or make use of open datasets.

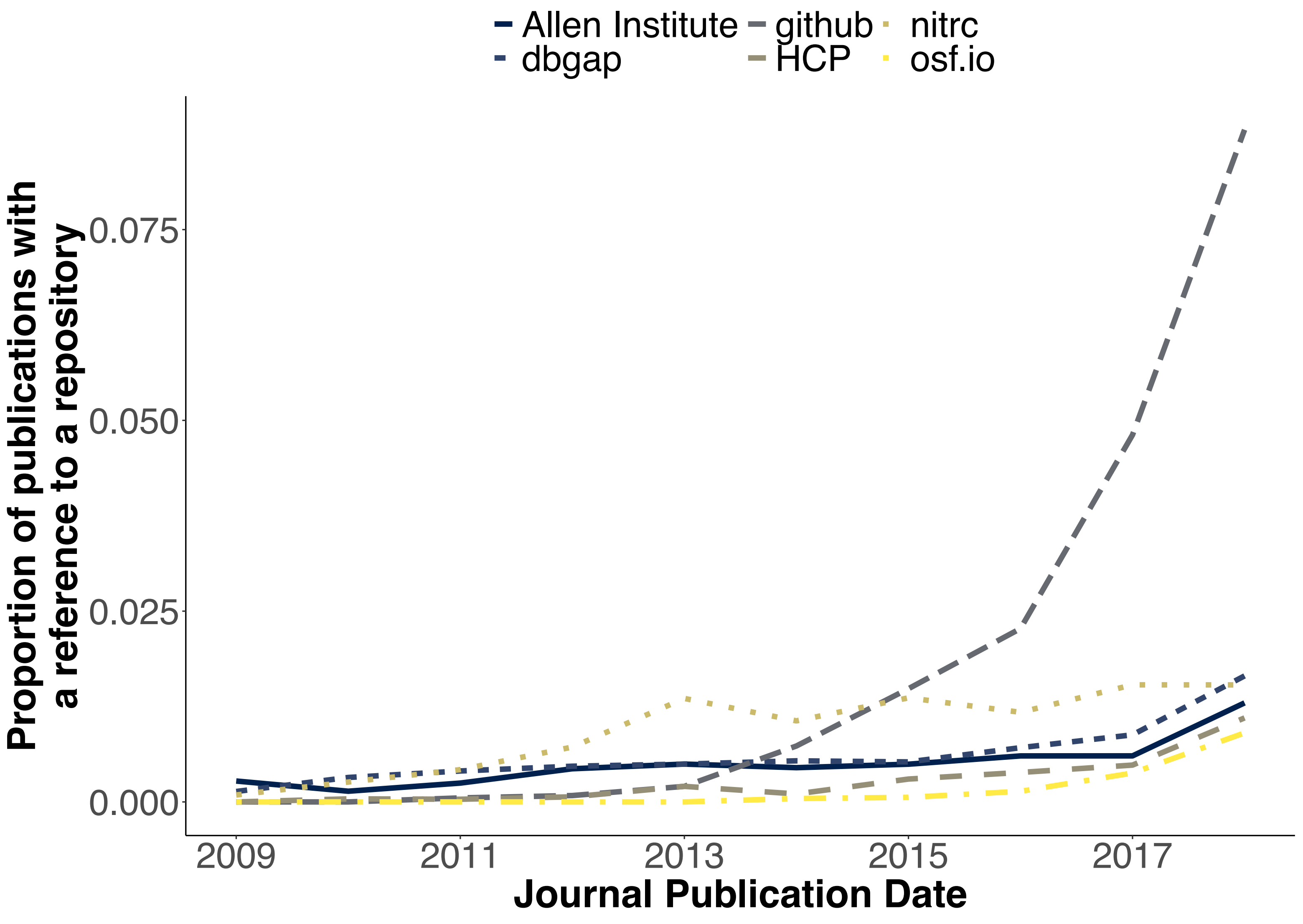
Data & Methods

- Using Federal RePORTER, we identified ~56k NIMH funded full-text papers in PMC
- We used simple string searches to look for mentions of a set of popular data repos
- Because this process does not precisely & exclusive identify data sharing/reuse, we are currently working on methods to detect these behaviors based on the text of papers

Results


Github
OSF
Open Science Framework
OpenNEURO
CRCNS
Collaborative Research in Comp. Neuroscience
LONI
Laboratory of Neuroimaging
FMRI Data Center
NIMH National Database for Autism Research
ALLEN INSTITUTE for BRAIN SCIENCE
Human Connectome Project
The Dataverse Project
DataLad
dbGaP GENOTYPES and PHENOTYPES
Database of Genotypes and Phenotypes
BALSA
Brain Analysis Library of Spatial Maps and Atlases
DRYAD
figshare
OMEGA
Open MEG Archive
aws
indi International Neuroimaging Data-Sharing Initiative
International Neuroimaging Data-sharing Initiative
NITRC
Neuroimaging Tools & Resources Collaboratory
XNAT
zenodo

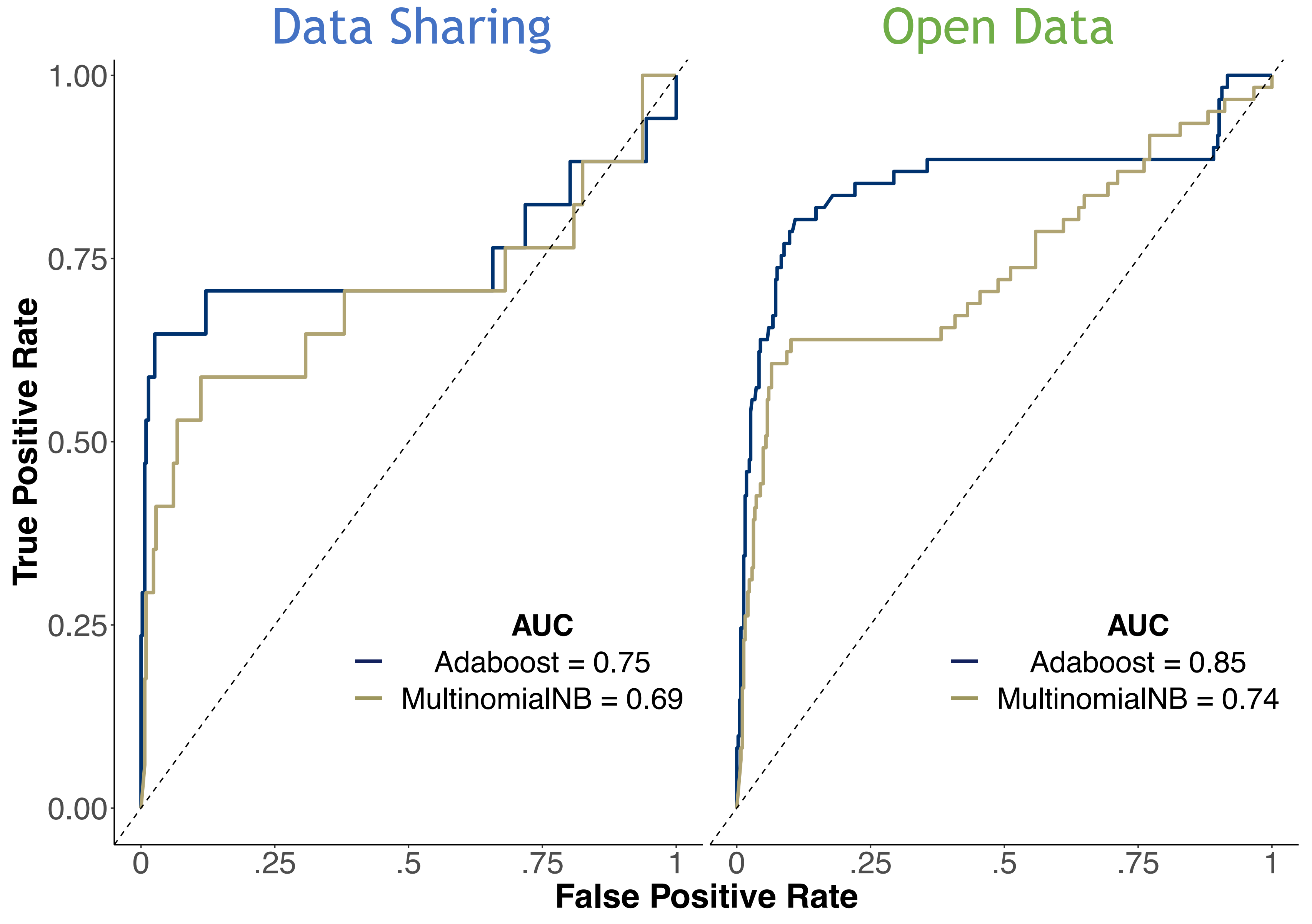




We labeled of 1,351 paragraphs as:

data sharing (n=66): *cases where the authors indicate that the data that they generated for the paper are deposited and available in a public repository*

open data (n=184): *cases where the authors make reference to a specific shared dataset or repository. It need not be actually used in any analysis, but it needs to be specific*



Misclassified Examples

False Positives

Data Sharing	Open Data
Determine the prevalence of muscle weakness using the two 2014 Foundation for the National Institutes of Health (FNIH) Sarcopenia Project criteria and its relationship to physical limitations, basic and instrumental activities of daily living (ADL). 🤷‍♀️	*A.J.N and J.C.M developed the study concept and experimental design. A.J.N and J.W programmed the experiment. A.J.N conducted data analyses and collected data from participants. All authors drafted the manuscript.* 🤷‍♀️
----	----
Integrity of raw sequencing reads was confirmed using FastQC. Reads were aligned to the mm10 reference genome using bwa mem v0.7.5a–r406. Coverage and quality summaries were computed using the Picard suite (http://broadinstitute.github.org/picard).	*a) Graph of the FC values for each paired language region for each group. b) Visual depiction of the ROIS and the significantly different connections between Patients and TD Controls. The brain networks were visualized with BrainNet Viewer (http://nitrc.org/projects/bnv/). * indicates language connections with a significant difference between groups.*

False Negatives

Data Sharing	Open Data
To test the proposed framework, two different sets of data were used. First, 120 T1-weighted MRI scans acquired from 3 subjects in a test-retest experiment were applied to evaluate the reproducibility of our proposed method. The data was downloaded from http://dx.doi.org/10.6084/m9.figshare.929651. 40 scans were acquired for each subject in 20 separate days in a month by using a GE MR750 3T scanner. ADNI-recommended T1-weighted imaging protocol were applied and one can refer to for more details.	*All analyses were stratified by genetic ancestry. An outline of the primary analyses can be found in Supplemental Table S2 and more detailed descriptions of the planned analyses are provided in our published protocol. All code and documents relevant for running the analyses and meta-analysis are available in the public repository at https://github.com/achorton/SD_5HTTLPR. All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/m64ue/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617699167.This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges.*

Notes

[1] Data Science & Sharing Team, National Institute of Mental Health, National Institutes of Health http://cmn.niimh.nih.gov/dsst

[2] Majumder, M.A. et al (2017). Sharing data under the 21st century Cures act. *Genetics in Medicine, 19,* 1289-1294