Identifying Data Sharing and Data Reuse in Full-text NIMH-funded papers

Most recent version available here: 10.5281/zenodo.3894807

Travis Riddle[1], Francisco Pereira[1], Adam Thomas[1]

[1]Data Science and Sharing Team, National Institute of Mental Health, Bethesda, MD

Key Points

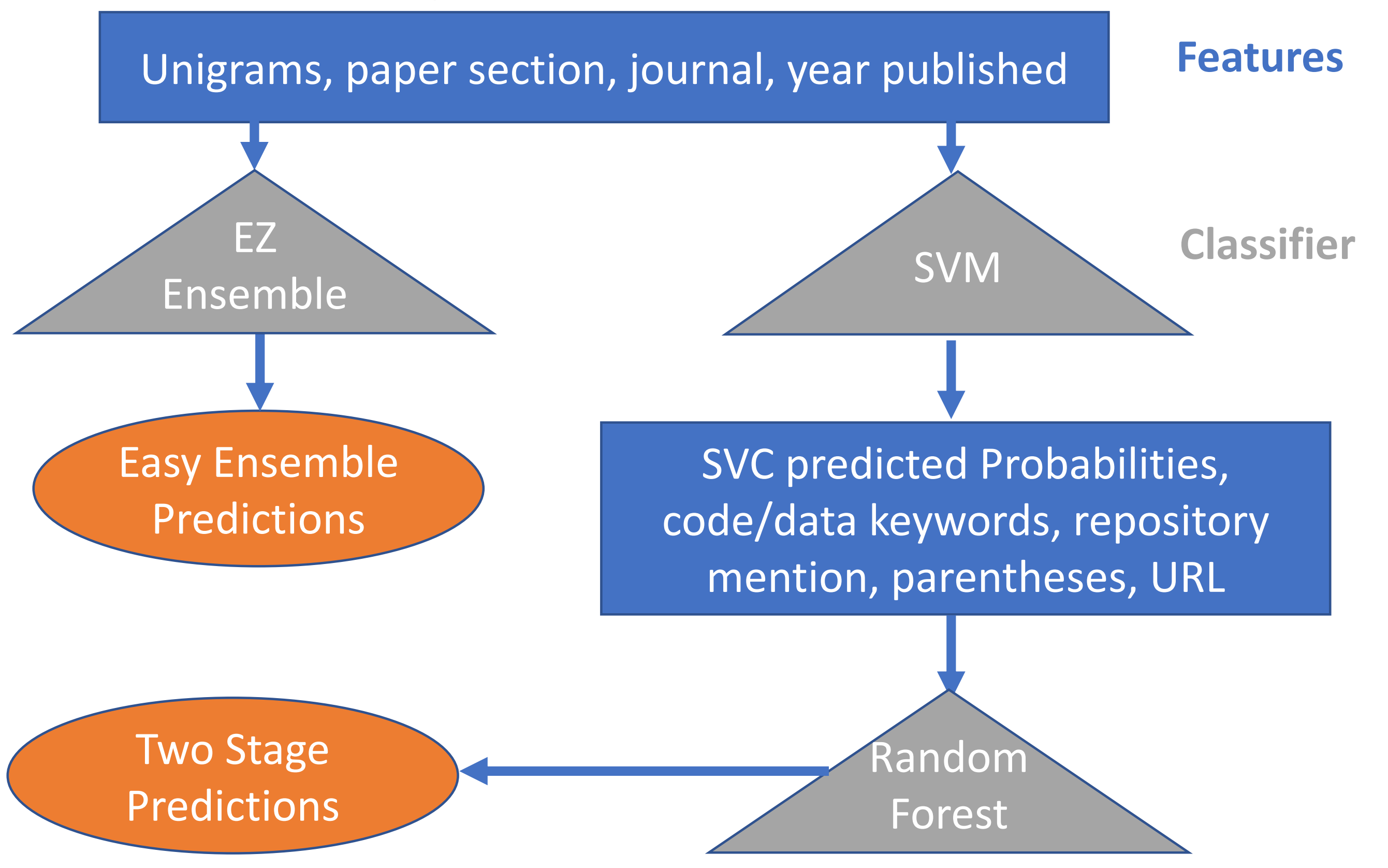
- We developed a series of classification systems to identify data sharing/reuse statements
- Applying the top-performing classifier to a corpus of 57k NIMH funded papers published since 2008, we extrapolate that 4.5% are predicted to contain data sharing/reuse statements
- By our predictions, a large majority of data statements come from a small handful of investigators

Introduction

- The 2017 Cures Act authorizes the NIH Director to require award recipients to share data in a manner consistent with applicable laws and regulations[1]
- Identifying and measuring data sharing and data reuse serves a number of goals that are important for scientists, funding agencies, and the public more generally.
- The unmet objective of an efficient and accurate system for identification and tracking of datasets is a conspicuous shortcoming of the larger open science community.
- We sought to develop a classification system to automate labeling of data statements in full-text NIMH-funded papers from PubMed Central
- We used our top-performing classifier to predict the presence of data statements and examined the distribution of these predictions as a function of paper meta-data

Materials & Methods

- Data
  - The population of papers is any paper on PubMed Central found to have been listed as a paper produced from an NIMH funded grant on FederalReporter (total = 57,692)
  - Labeled data consist of 3,355 sentences from 1,559 papers. Samples were drawn through a mixture of random samples, regex searches, and active learning.
  - Of these, 375 sentences from 314 papers were found to contain data statements.
  - Paper metadata was obtained through FederalReporter, PubMed Central, and iCite.
- Analysis
  - We tested two classification approaches using 3-fold cross validation
    - Easy Ensemble[2]
    - Two stage classifier with SVM and Random Forest



Unigrams, paper section, journal, year published

Features

EZ Ensemble

SVM

Classifier

Easy Ensemble Predictions

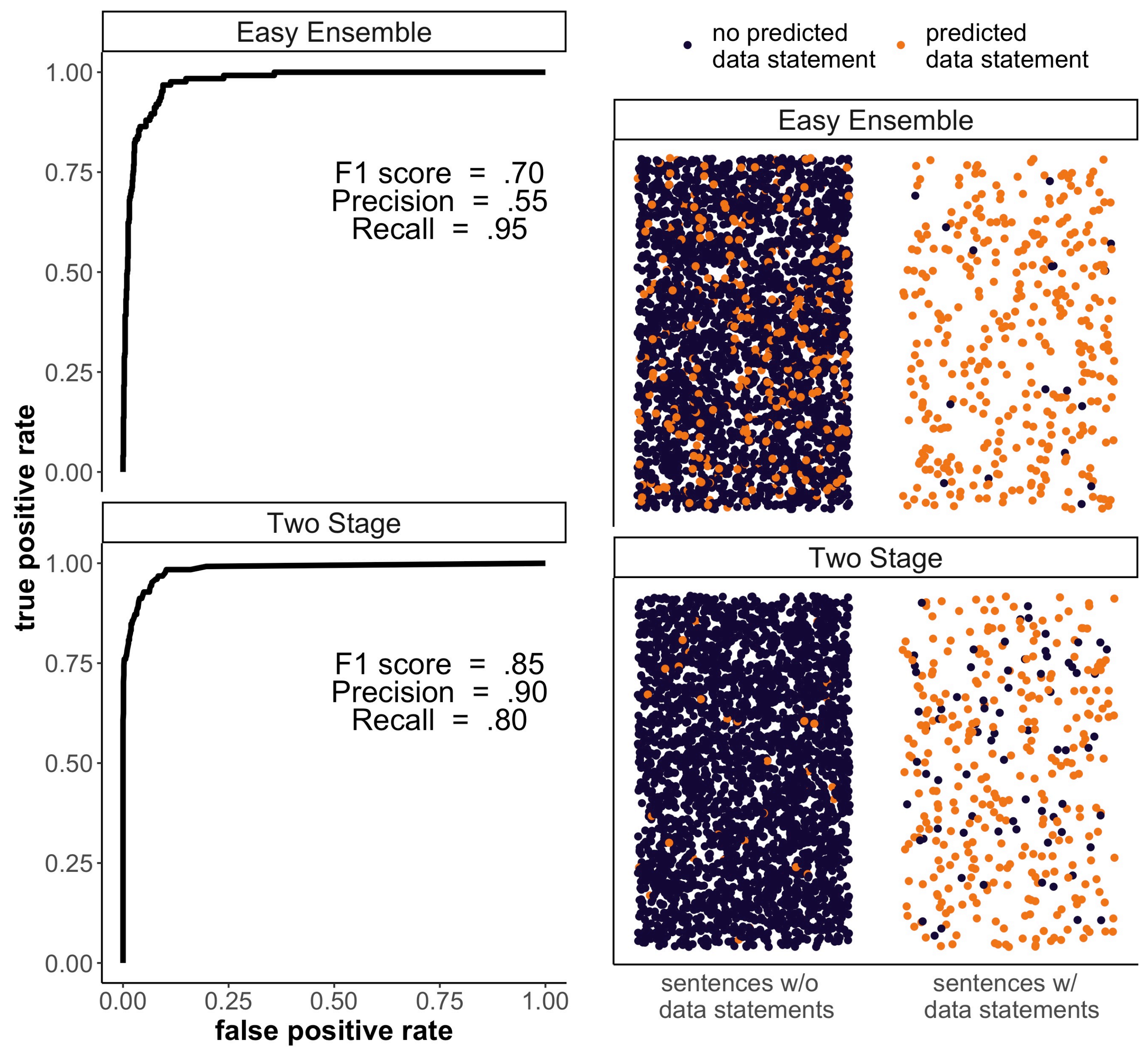
SVC predicted Probabilities, code/data keywords, repository mention, parentheses, URL

Random Forest

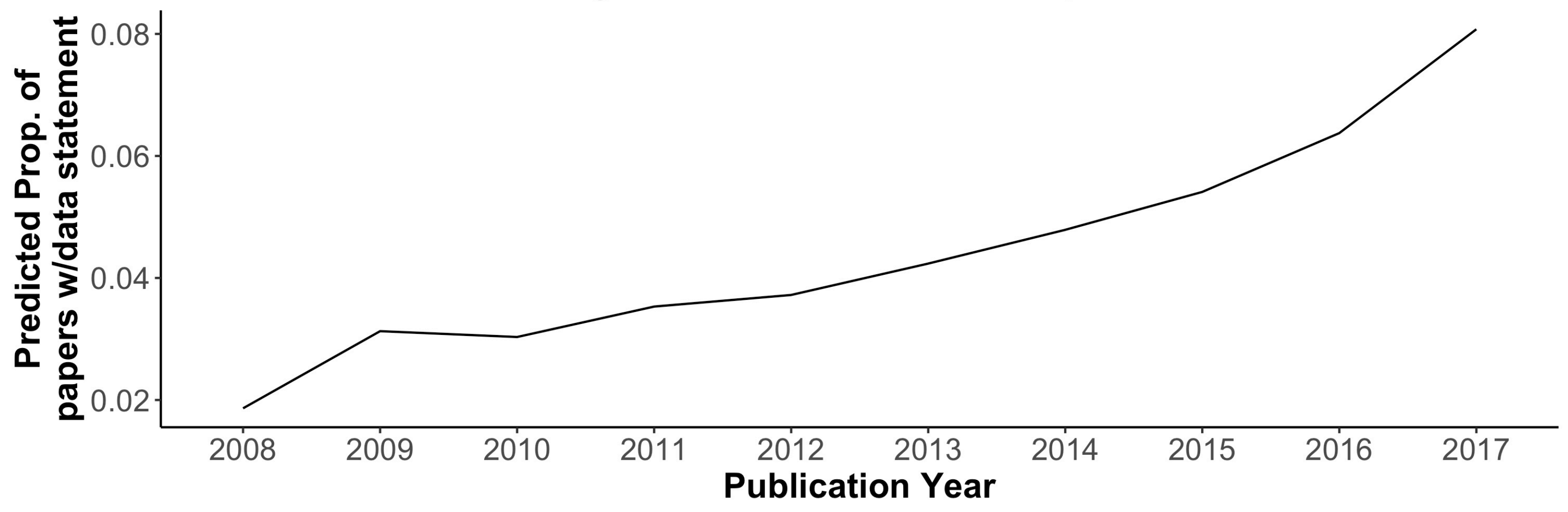
Two Stage Predictions

Results

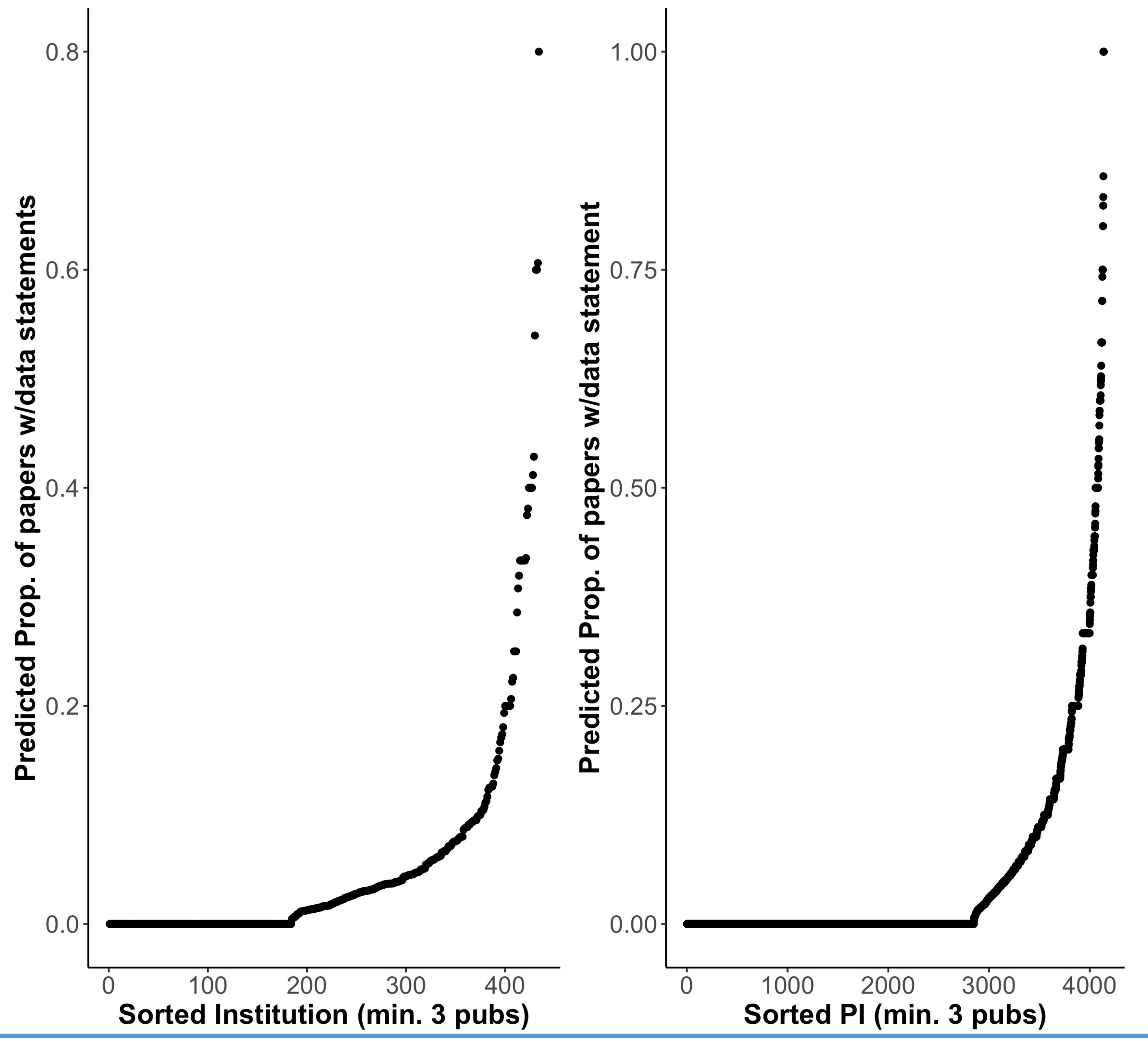
Cross-validation Classifier Performance



We extrapolate that data statements are increasing overtime, but still infrequent



We extrapolate that most data statements come from a small fraction of NIMH-funded investigators and institutions



Error Examples

- **False Alarms**

Detailed NHANES survey operations manuals are available on the NHANES Web site (http://www.cdc.gov/nchs/nhanes.htm).

Sequence data were aligned and variants called by the Picard (http://picard.sourceforge.net) zBWA GATK pipeline.

- **Misses**

European and African-American participants from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project were obtained from the National Institutes of Mental Health repository (http://www.nimhgenetics.org).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Conclusion

- Though data sharing is expected for those with NIMH funding, these results suggest that rates of data sharing are likely very low.
- Of the 4,139 PIs with at least 3 publications represented in our data, we extrapolate that just 210 (5%) have data statements in more than a third of their papers.
- Our classifier could be improved by more consistent handling of software tools and other data resources (e.g. manuals), or by using features derived from parsing sentences
- Future directions could capitalize on list of data DOIs (e.g. datacite), though a rough perusal of our data suggests that DOIs are not consistently cited.

1. Majumder, M.A. et al (2017). Sharing data under the 21st century Cures act. *Genetics in Medicine, 19,* 1289-1294
2. X. Y. Liu, J. Wu and Z. H. Zhou, (2009) Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B,* 39(2), 539-550

NIH National Institute of Mental Health

CMN

DSST

Data & code:
github.com/riddlet/ohbm_2020_poster