Prep Work

- Clone https://github.com/EthanDayley/asd-siloing-exploratory
- Ensure that the following are installed on your machine:
    - MySQL
    - Python 3
    - Jupyter
    - The following Python packages:
        - pandas
        - matplotlib
        - numpy
        - scikit-learn
        - gensim
        - phate
        - Wordcloud
- Download GoogleNews-vectors-negative300.bin from https://github.com/mmihaltz/word2vec-GoogleNews-vectors to the word_embeddings folder in the asd-siloing-exploratory repository.
- Create a MySQL database called asd_siloing.
- In your repository, run mysqlsh -u <username> -h localhost -f .\scripts\generate_tables.sql.

Materials Gathering

NOTE: Unless otherwise stated, all command line operations listed below should be conducted from the top level of the local asd-siloing-exploratory repository.

- In a web-browser, navigate to https://www.ncbi.nlm.nih.gov/pmc/ and input a search term.
- From the results page, select the menu labeled “Send to:” and select “File”, with a “PMCID List” format.
- Move the downloaded list to the “search” folder of the local asd-siloing-exploratory repository and rename it to “pmc_result.txt”.
- Run: py -n 90000 .\scripts\select-article-subset.py -i .\search\pmc_result.txt -o .\search\article-test-subset.txt
- Run: py .\scripts\download_abstracts.py
- Run: py .\scripts\fetch_article_info.py
- Run: mysqlsh -u <username> -h localhost -f .\scripts\distinct_journals_query.sql --sql 1> journal_list.txt
- Run: py .\scripts\generate_journal_query.py
- Run: py .\scripts\chunk_journal_query.py
- In a web browser, navigate to https://www.ncbi.nlm.nih.gov/nlmcatalog
- Take the output of the chunk_journal_query script and run each query individually in the NLM Catalog. Make sure to download the results in XML format and move each of them to journal_classification\nlm_catalog_results.
- Run: py .\scripts\extract_source_info.py

Validation

- See the data_analysis\validation.ipynb Juypter notebook

Final Clustering

- See the data_analysis\cluster_analysis.ipynb Jupyter notebook