What are you studying?

Predicting academic journal usage and programs of study at a university

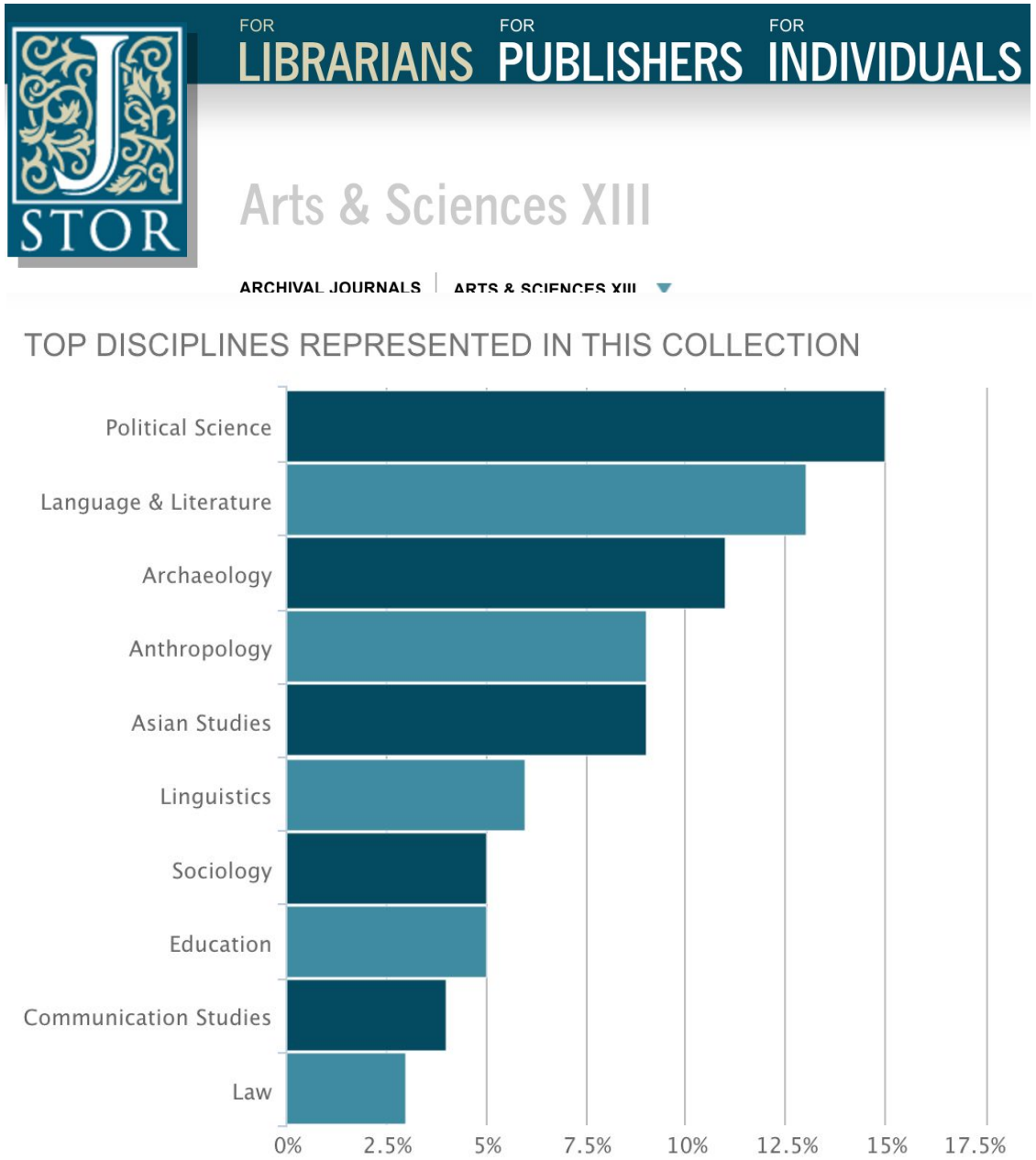
NYC-DAT-34

Hannah Begley

5-31-16

Background

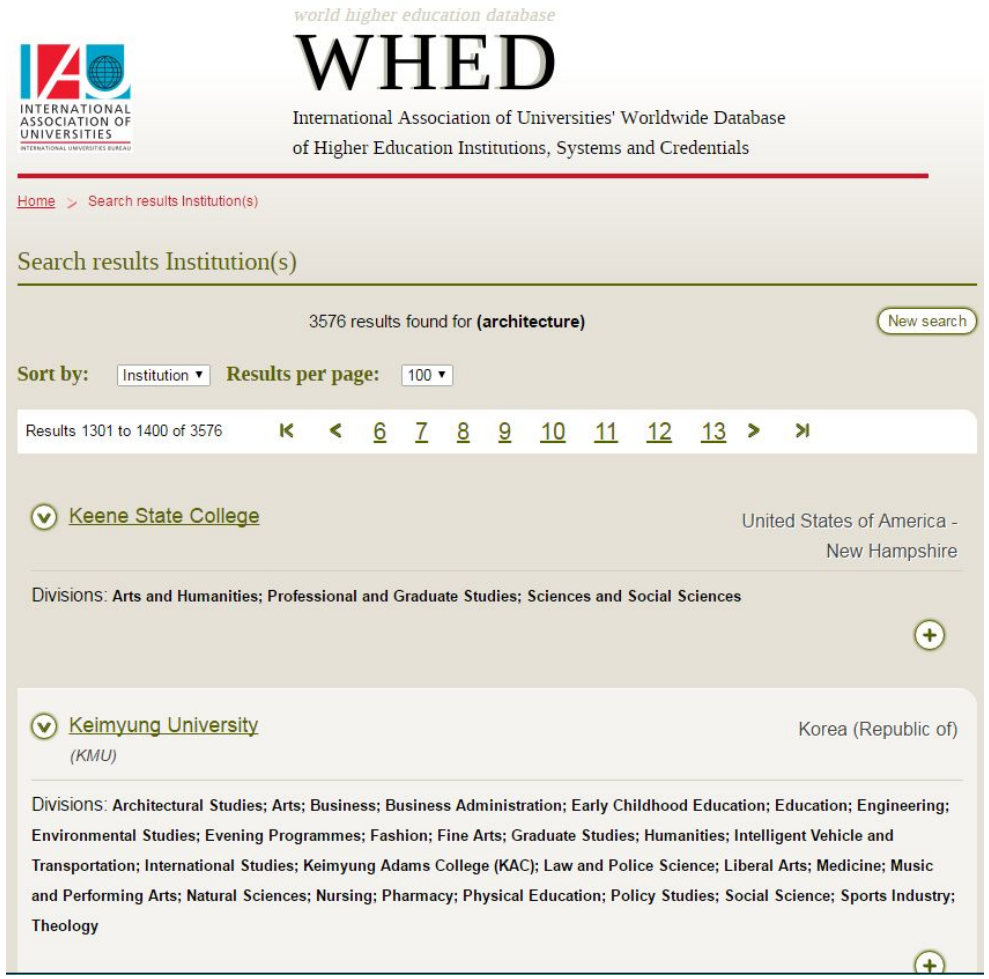
- JSTOR: not-for-profit digital library created in 1995 to provide greater access to scholarly content by digitizing journals back to their original issue
- Access to JSTOR journals is typically through subscriptions to collections of journals from multiple disciplines
- Aggregated content can make it difficult to identify what content drives collection interest at a particular institution



http://about.jstor.org/content/arts-sciences-xiii?cid=dsp_ASXVApril2016

Integrating WHED data on Fields of Study offered at institutions with JSTOR usage data

- WHED collects information about fields of study at universities worldwide
- WHED records information for 660 Fields of Study of varying specificity, from history to sericulture
- WHED Field of Study information for 10,000 institutions has been integrated with JSTOR CRM institution accounts
- WHED Fields of Study have been categorized as belonging to 80 JSTOR disciplines



http://www.whed.net/results_institutions.php accessed 5-16-16

Integrating WHED data on Fields of Study offered at institutions with JSTOR usage data

Goal: To explore, and hopefully verify, that WHED Field of Study data will indicate areas of academic interest and thus will be reflected in an institution's usage of content within those disciplines.

Hypothesis:

- Having academic programs within a particular discipline will lead to higher usage of content within that discipline;

- Content usage in a particular discipline can be used to determine whether an institution has, or does not have, a field of study in that discipline;

- Content usage can be predicted by examining the academic programs offered at that institution.

Desired Outcome:

Confidence that WHED Field of Study data can reasonably be used to:

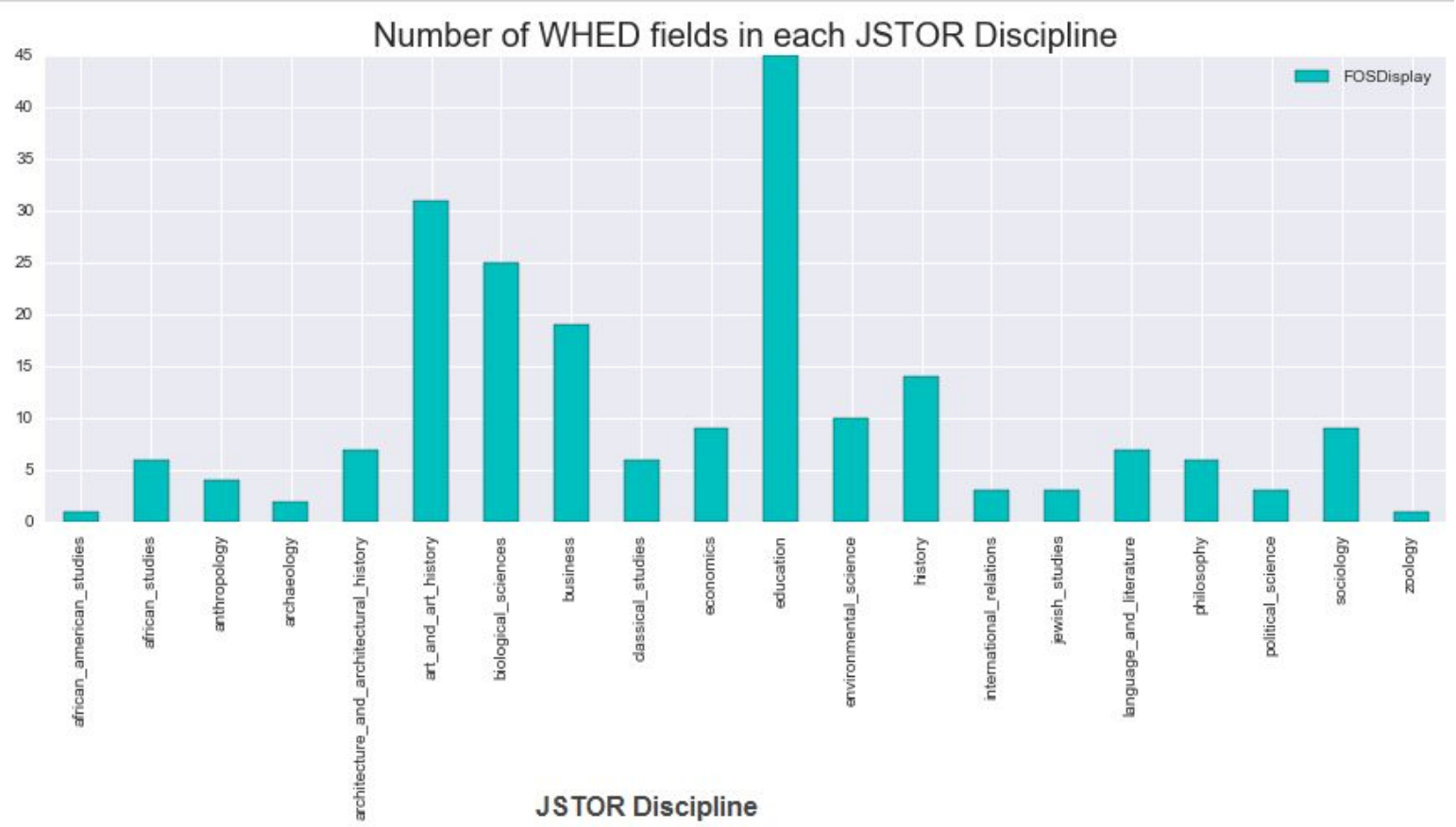
- Predict interest in future projects
- Identify promising institutions for partnerships
- Validate that JSTOR content is known to researchers within an institution and is used accordingly.

Selecting the data

Selected 22 JSTOR disciplines, explore usage data for journals within that discipline

Total data: 6.7M rows

Selected some disciplines that grouped many WHED Fields of Study and some disciplines that represent just one WHED Field of Study



Selecting the data

Main Datasets:
Usage data
data for a 3-year period of usage by journal for each subscribing institution for a selection of 22 of the 80 JSTOR disciplines.
Size: 3 datasets (train, fit, test) that together total 6,765,337 rows

Academic programs data
A dataset from the World Higher Education Database (WHED) containing information on academic programs offered at different universities worldwide.
Size: 456,860 rows

Fields of Study to Disciplines Map
A dataset that “maps” the academic fields of study from the WHED dataset to the relevant JSTOR disciplines.
Size: 850 rows

Supplemental Datasets
Usage of all JSTOR content over the same 3-year priod.

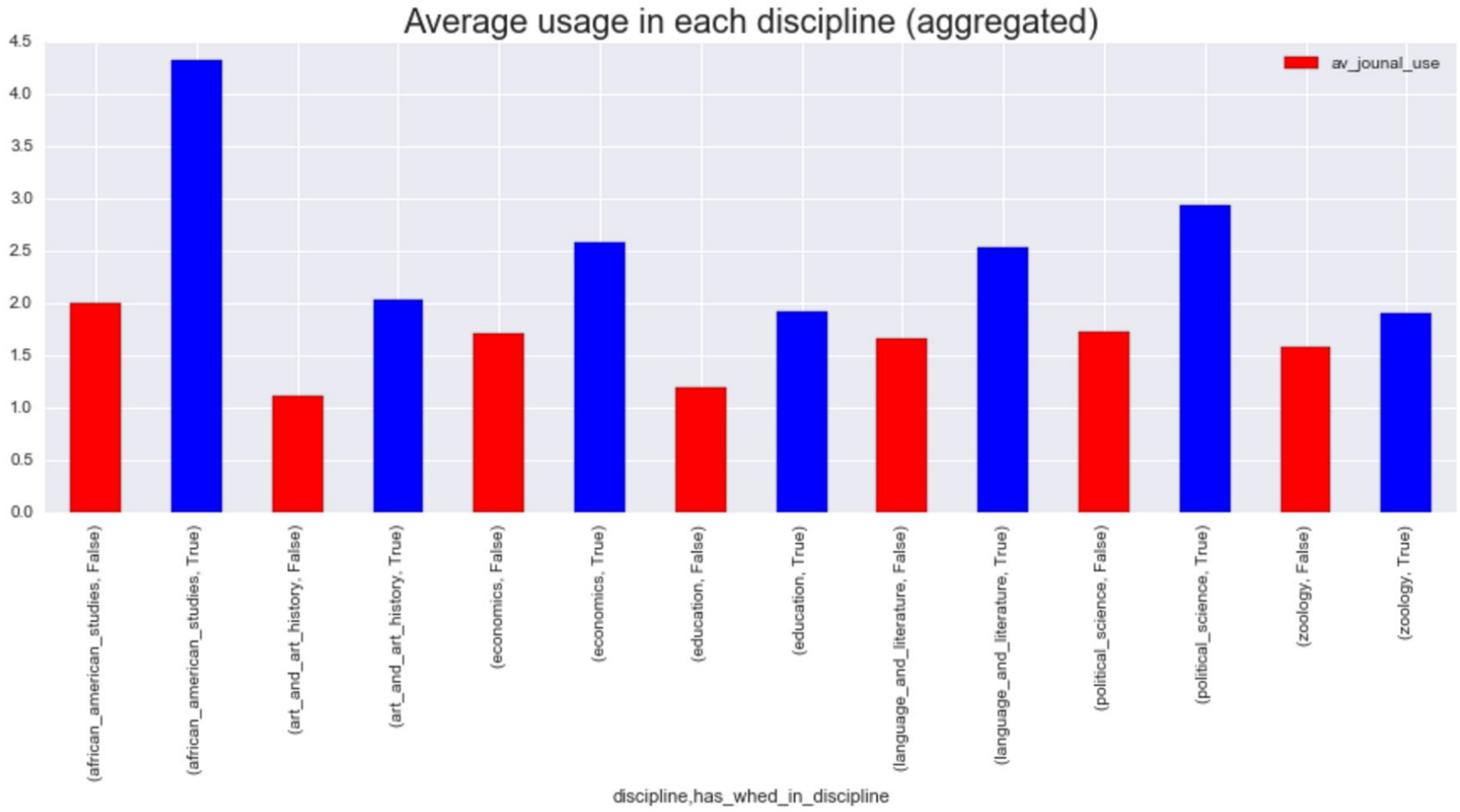
Usage of the collection containing the target journal over the same 3-year priod.

WHED Fields of Study Summary: number of fields of study listed for each institution across all disciplines.

Exploring Data

Institutions with academic programs in a particular discipline have higher average usage of content within that discipline

Average usage for journals in a discipline at a given institution is higher (blue) if they have WHED Fields of Study in that discipline



Modeling
Predicting usage for journals within a particular JSTOR discipline

Numerical Features

	journal_access	collection_access	total_jstor_access	FOS_Total	umber_collections	numberRelevantFOS	journal_in_collection	journal_in_total	FOS_prop
count	26994.000000	26994.000000	26994.000000	21104.000000	26994.000000	20806.000000	26924.000000	26991.000000	20806.000000
mean	1.934795	7.383660	9.473979	51.028478	9.867711	0.905075	0.236550	0.189850	0.022076
std	1.274426	2.132115	2.217857	47.951096	6.127985	1.658628	0.121475	0.106964	0.053555
min	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	0.876145	6.089046	8.090709	23.000000	5.000000	0.000000	0.148280	0.106607	0.000000
50%	1.809715	7.670735	9.583213	39.000000	8.000000	0.000000	0.241204	0.190276	0.000000
75%	2.861069	8.921902	11.042970	63.000000	16.000000	1.000000	0.327928	0.268088	0.025641
max	7.410370	13.036780	14.856137	661.000000	23.000000	22.000000	1.000000	0.735569	1.000000

All features

```
Index([u'collection', u'devnations', u'discipline',
       u'institution_classification', u'institution_community',
       u'institution_country', u'institution_name', u'institution_rank',
       u'institution_state', u'journalid', u'publication', u'region',
       u'sitename', u'system_id', u'UniqueJournal', u'UniqueIns',
       u'journal_access', u'collectionID', u'collection_access',
       u'total_jstor_access', u'FOS_Total', u'umber_collections', u'CRMID',
       u'numberRelevantFOS', u'has_whed_in_discipline', u'av_jounal_use',
       u'journal_in_collection', u'journal_in_total', u'FOS_prop'],
      dtype='object')
```

Modeling
Predicting usage for journals within a particular JSTOR discipline

Tested linear regression models, including LinearRegression, Ridge, and Lasso models with a variety of Alpha scores from .0001 to 10000.

Models explained 70-76% of variance in cross validation

```
features:  Index([u'sitename', u'discipline', u'institution_classification',
       u'institution_country', u'institution_rank', u'institution_state',
       u'devnations', u'collection_access', u'total_jstor_access',
       u'FOS_Total', u'umber_collections', u'numberRelevantFOS', u'FOS_prop'],
      dtype='object')
Mean Absolute Error: 0.46789411291
r2 value: 0.767567354958
```

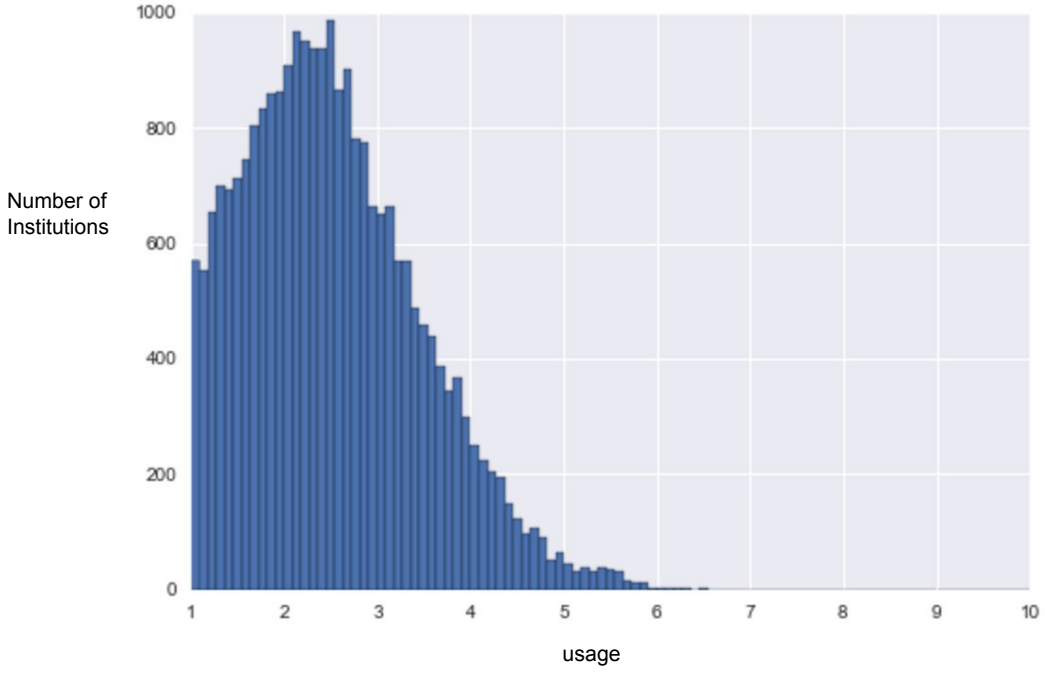
WHED Field of Study data accounts for very small (.004) difference in a model's success

```
Same analysis as above, but without WHED information
features:  Index([u'sitename', u'discipline', u'institution_classification',
       u'institution_country', u'institution_rank', u'institution_state',
       u'devnations', u'collection_access', u'total_jstor_access',
       u'umber_collections', u'has_whed_in_discipline'],
      dtype='object')
Mean Absolute Error: 0.471724636239
r2 value: 0.763775685695
```

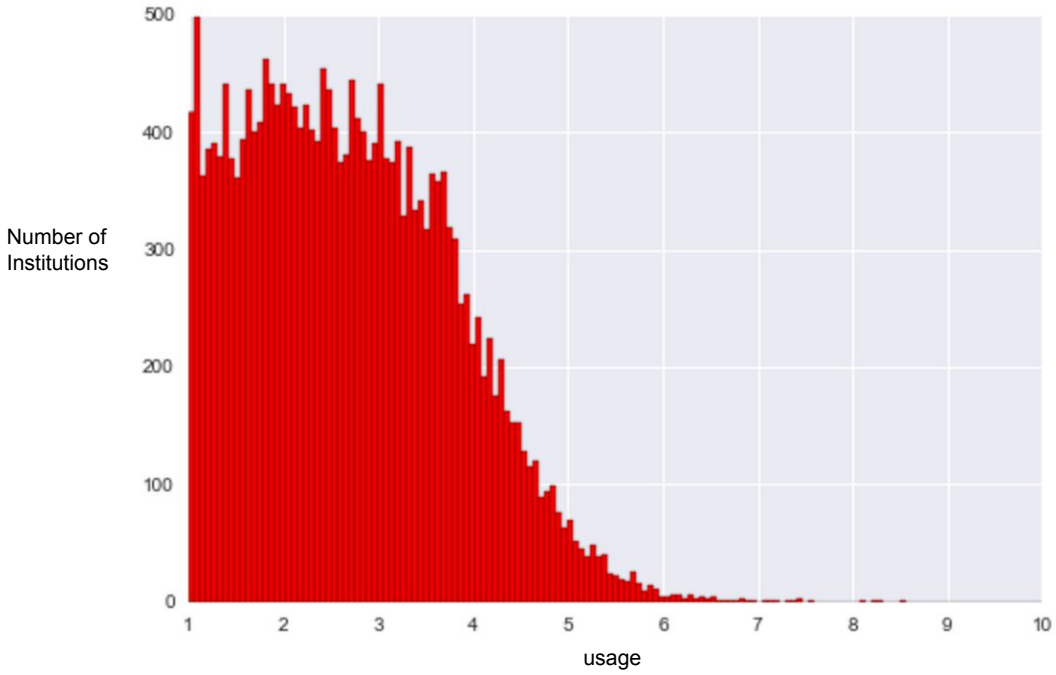
Results
Predicting usage for journals within a particular JSTOR discipline

- Predicting content usage for journals in disciplines not seen in training or fitting of data:
  - Data size: 27,687 predictions
  - Mean Absolute Error: .50 (on log scale shown below)
  - Predicted usage is clustered around mean usage when compared to actual usage

Predicted



Actual:



Modeling
Predicting if institution has WHED Fields of Study in a discipline

First tried to use linear regression. These scores were unsatisfactory (r2 value of .15 - .25).

Logistic Regression: recall of .58 and precision of .66 for all data, and .84 recall and .70 precision when usage in lowest 15% was removed.

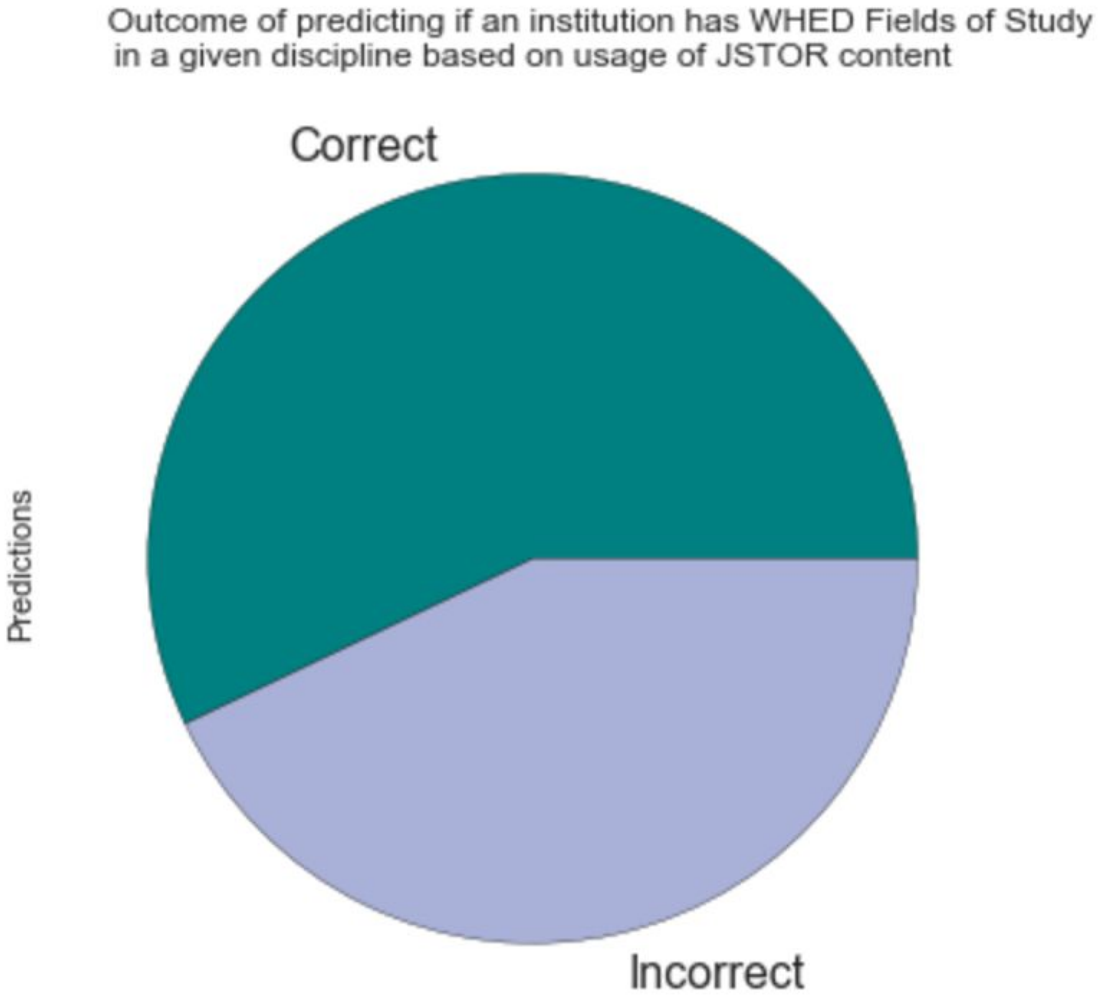
```
predict whether institution has or does not have WHED FOS in discipline
features:  Index([u'sitename', u'discipline', u'institution_classification',
       u'institution_country', u'institution_rank', u'institution_state',
       u'devnations', u'journal_access', u'collection_access',
       u'total_jstor_access', u'umber_collections'],
      dtype='object')
Recall: 0.575067192078
Precision: 0.660212125035

predict whether institution has or does not have WHED FOS in discipline, with low usage removed
features:  Index([u'sitename', u'discipline', u'institution_classification',
       u'institution_country', u'institution_rank', u'institution_state',
       u'devnations', u'journal_access', u'collection_access',
       u'total_jstor_access', u'umber_collections'],
      dtype='object')
Recall: 0.844258112329
Precision: 0.69729265942
```

Results
Predicting if institution has WHED Fields of Study in a discipline

- Predicting if institution has WHED Fields of Study in disciplines not seen in training or fitting of data:
  - Data size: 6,046 predictions
  - Observed success: 58% correct predictions
    - True Positive 3267

      False Positive 2313
      False Negative 283
      True Negative 183
    - recall: .92
    - precision: .58
  - Lots of false positives because many institutions do not have a WHED Field of Study in a given discipline, and model was weighted to return positive (to “catch” all True Positives)
  - Lingering question- more false negatives than true negatives- why might that be?



Conclusions

Goal: To validate whether Fields of Study in a discipline allows us to make predictions about usage of content in that discipline

Outcome:

- Including WHED Field of Study information does not substantially increase ability to accurately predict usage for a particular journal, or for journals within a particular discipline
  - Best usage predictor is total JSTOR usage, providing an increase in r2 of .2
- Graphically, institutions with Fields of Study in a discipline do on average have higher usage. Indicates need for further exploration, as trend does appear to exist.

Goal: To predict if an institution has WHED Fields of Study in a discipline given their usage of content in that discipline

Outcome:

- Same lingering question as above- though there is clear evidence that institutions with Fields of Study in a discipline tend to have higher usage in journals of that discipline, this increase in usage doesn't seem to aid in predicting the presence of a WHED Field of Study
- Another avenue to explore in more depth

Implementation Plans

Ultimate outcome: Dashboard allowing users to select certain criteria based either on JSTOR usage data or WHED Field of Study data to explore trends within that area

- Update quarterly
- All datasets are in data warehouse, making reproduction straightforward

More work needed before unleashing on “broad” audience:

- Are certain disciplines more amenable to prediction than others?
- Is examining all usage at a given institution and comparing that to WHED Field of Study profile more successful than usage across disciplines?
- Can usage data be refined?

Thank you!

Full analysis: https://github.com/hannahbegley/DAT/blob/master/Begley_5-31-16.ipynb