# Enhancing and Automating University Reporting Of R&D Expenditure Data Using Machine Learning Techniques[1]

**Joshua L. Rosenbloom, Iowa State University, National Bureau of Economic Research**

**Rodolfo Torres, University of Kansas**

**Joseph St. Amand, University of Kansas**

**Adrienne Sadovsky, University of Kansas**

## Higher Education Spending on Research and Development

In 2014, U.S. Colleges and Universities reported spending $67.3 billion on Research & Development (R&D). While this figure constitutes only about 15 percent of the nation's total R&D effort, colleges and universities performed more than half of U.S. basic research.[2] Most of what we know about R&D performed at the nation's colleges and universities—not only aggregate totals, but also expenditures by field of study and by individual institutions – is derived from data collected by the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES) as part of its Higher Education R&D (HERD) survey. The HERD survey continued and expanded a data collection effort that was started in 1972 as the Academic R&D Expenditures Survey.

The data collected by the HERD survey are widely used by both university administrators and academic researchers interested in understanding the nation's scientific enterprise. University leadership is interested in tracking total R&D expenditures and rankings of R&D expenditures as an indicator of research prowess. Most universities work to move up in the rankings by increasing their expenditures. Scholars interested in the political economy of federal science funding have used HERD expenditure data to track the expansion of the nation's cadre of research universities and to assess the tendency of the political system to promote more equal distribution of funds across states and regions (Geiger and Feller 1995; Graham and Diamond 1997; Feller 2001). Others have used more disaggregated data on expenditures at the

[1] The material in this article is based in part upon work supported by the National Science Foundation under Grant Numbers SMA-1547513 and SMA-1547464. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2] National Science Board 2016, chapters 4, 5

discipline level to explore either how funding is related to scientific productivity (Adams and Griliches, Rosenbloom et al 2015) or to uncover factors that influence the allocation of federal R&D funding across institutions (Lanahan et al 2016; Rosenbloom and Ginther forthcoming).

The data collected in the HERD and the earlier Academic R&D Survey are derived from institutional responses to an annual survey distributed by NCSES. Colleges and Universities undoubtedly take different approaches to compiling the necessary data, but at research universities with specialized research administration staff, responsibility for responding to the survey is likely delegated to one or more specialists within the office of sponsored research or institutional research.

The aforementioned method of collecting data on college and university R&D expenditures results in three distinct problems. First, responding to the HERD is costly in terms of the time required to accurately report the requested data. Second, because of the nature of the annual survey and the lead time involved in tabulating responses, the data are available only with a long lag. While the data are useful for retrospective analysis, the lags make them far less valuable for setting institutional strategy or performing real-time analysis of R&D activity. Finally, the effort to classify projects by their purpose and field of study is inevitably subjective, making it problematic to make comparisons across institutions and introducing spurious variation within institutions when responsibility for data collection shifts from one person to another.[3]

To address these problems, we have been engaged in an experiment to apply techniques of machine learning to automate project classification. Successful development of classification algorithms would reduce the cost of responding to the HERD survey, allow for essentially real-time tracking of expenditures, and offer the potential to increase the consistency of classification over time and across institutions. As we describe here, our proof of concept investigation suggests that such approaches are potentially feasible, but require further efforts.

**Application of Machine Learning to Classify Sponsored Research Projects**

With the growth of large data sets and the declining cost of computation, application of machine learning techniques to identify data patterns and make predictions based on these patterns has become increasingly common.[4] The goal of our project is to develop a classification algorithm that can be used to either supplement or replace human judgement in classifying sponsored research projects. To do so, we begin with a set of sponsored projects awards that have already been classified by Research Administration staff at the University of Kansas. In

[3] The survey categorizes projects into 4 different purposes (applied research, basic research, development, and other), and 40 different scientific fields of study (e.g., Bioengineering and Biomedical Engineering, Astronomy and Astrophysics, Political Science and Government, etc.).

[4] For an overview of machine learning and associated terminology see: https://en.wikipedia.org/wiki/Machine_learning

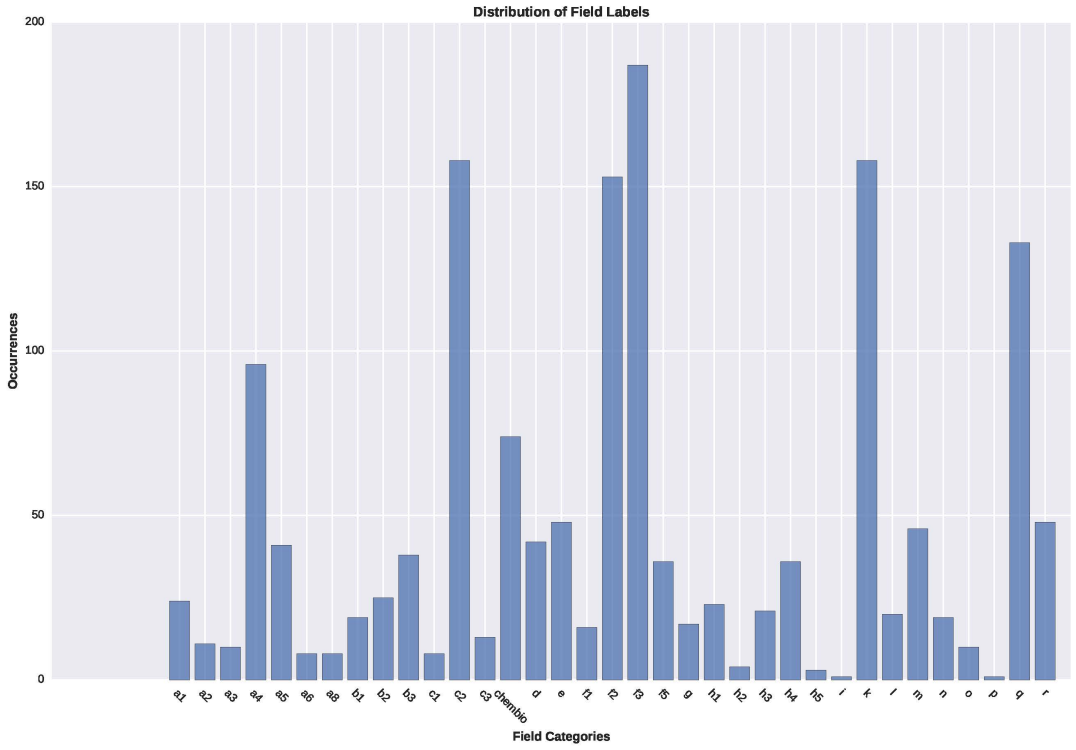
the language of machine learning, this is an example of "supervised learning."

Working with staff in the University of Kansas Office of Research, we obtained a data set of historical sponsored project awards. After dropping awards for which we did not have complete data, we were left with approximately 1,500 projects. For each of these projects, the data included information on the:

- Project sponsor
- Principle Investigator (PI) home unit
- Project abstract describing the project
- Human-assigned classification of the project's purpose and field of study.

Figures 1 and 2 show the distribution of projects across fields of study and purpose based on the human-assigned classification. In addition to the full list of NSF-defined fields of study, our data include KU-specific fields of "Chem-Bio" and "CEBC" (that combines projects across chemistry, chemical engineering and biomedical sciences) that are used for internal institutional purposes.[5] As Figure 1 illustrates, there are some fields for which we do not have a large number of projects. The distribution of projects by purpose is also somewhat uneven, as illustrated in Figure 2.

Figure 1: Distribution of Projects by Field of Study

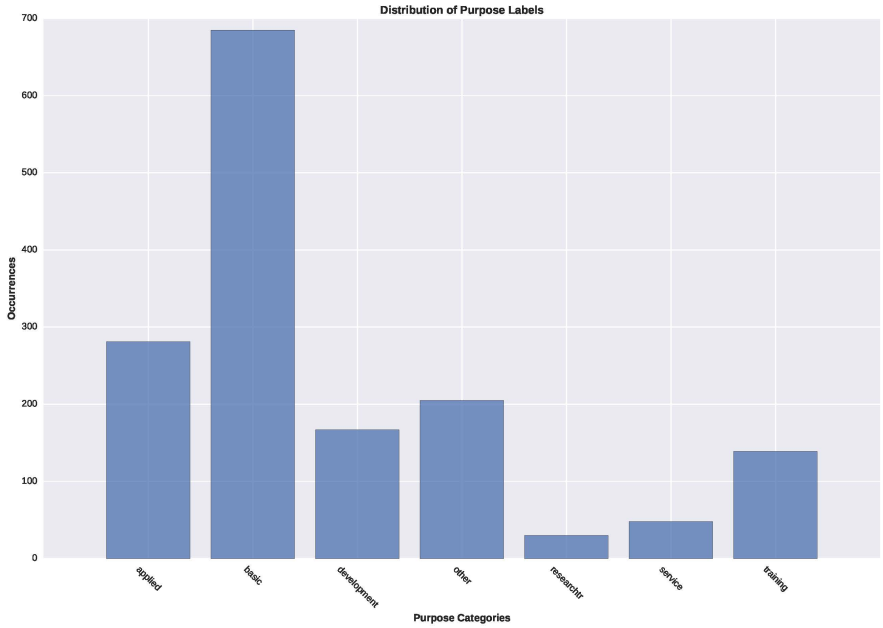



[5] For reporting purposes, expenditures in the Chem-Bio and CEBC categories are split evenly between the Chemistry and Biological and Biomedical Sciences and Chemistry and Chemical Engineering, respectively.

Notes to Figure 1: The University of Kansas Fields of Study are denoted with the following codes.

| Field | Code |
|---|---|
| Computer and Information Sciences | A |
| Aerospace / Aeronautical / Astronautical Engineering | B1 |
| Bioengineering and Biomedical Engineering | B2 |
| Chemical Engineering | B3 |
| Civil Engineering | B4 |
| Electrical, Electronic, and Communications Engineering | B5 |
| Industrial and Manufacturing Engineering | B6 |
| Mechanical Engineering | B7 |
| Metallurgical & Materials Engineering | B8 |
| Other Engineering | B9 |
| Atmospheric Sciences and Meteorology | C1 |
| Geological and Earth Sciences | C2 |
| Ocean Sciences and Marine Sciences | C3 |
| Other Geosciences, Atmospheric, and Ocean Sciences | C4 |
| Agricultural Sciences | D1 |
| Biological and Biomedical Sciences | D2 |
| Health Sciences | D3 |
| Natural Resources and Conservation | D4 |
| Other Life Sciences | D5 |
| Mathematics and Statistics | E |
| Astronomy and Astrophysics | F1 |
| Chemistry | F2 |
| Materials Science | F3 |
| Physics | F4 |
| Other Physical Sciences | F5 |
| Psychology | G |
| Anthropology | H1 |
| Economics | H2 |
| Political Science and Government | H3 |
| Sociology, Demography, and Population Studies | H4 |
| Other Social Sciences | H5 |
| Other Sciences | I |
| Business Management and Business Administration | K |
| Communication and Communications Technologies | L |
| Education | M |
| Humanities | N |
| Law | O |
| Social Work | P |
| Visual and Performing Arts | Q |
| Other Non-S&E Fields | R |

Figure 2: Distribution of Projects by Purpose




The major source of information about each project comes from the proposed statement of work, which is treated as a "bag of words." As a first step, we pre-process the data by standardizing word forms and eliminating "stop-words" (e.g. the, is, to). Individual (or collections of) words are converted to a numerical form based on their frequency of occurrence within and between project abstracts. In machine learning, these numerical representations are referred to as "features." The goal of the machine learning algorithm is to assess which specific combination of features are useful to discriminate between purpose/field categories.

Once the data are processed, we experimented with a selection of commonly used classifiers to identify features that provide predictive power. Following standard practice, we split the data into training and testing samples. The training sample contains approximately 70% of the observations, while the testing sample contains the remaining 30%. The models are trained on the training sample (which is further split into ÛÙÈÐÕÐÕÎ and validation samples) via a cross-validation procedure, which is necessary to prevent the models from over-fitting (i.e. "memı orizing") the data. We estimate the prediction error of the models using the validation samples, and use the testing sample as an assessment of generalization error.

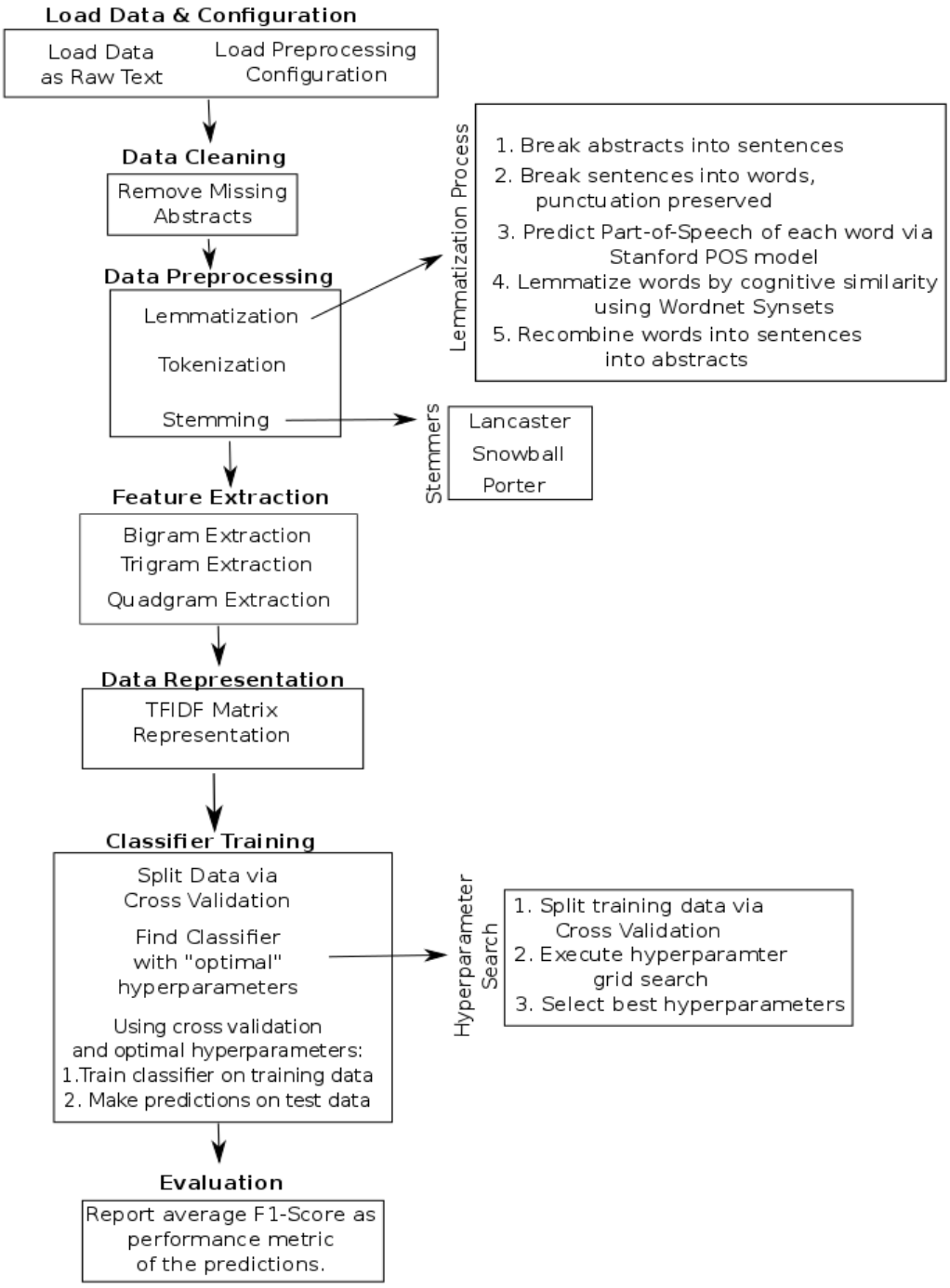



Figure 3: Schematic Representation of Data Analysis Steps

We treat the prediction of purpose and field categories as two separate classification tasks. For each classification problem we tried the following classifiers:

- Decision Tree
- Support Vector Machine
- Logistic Regression
- Random Forest
- Naïve Bayes
- Neural Network

All of these classification schemes are binary: reporting a probability that the project belongs to a particular purpose/field. For each purpose/field, the classifier yields a predicted probability (between 0 and 1) or a categorical determination that the project belongs to that purpose/field.

We first find classifiers for each purpose/field by assessing their performance (as described below) and then assign projects to a single purpose/field using the purpose/field with the highest predicted probability across all the classifiers.

The result of our analysis is a prediction of the purpose/field to which each project should be assigned. Comparing the human- and machine-assigned results produces a two-way contingency table depicted in Figure 4. Projects for which the two classifications agree (T1 and T2) are successful predictions, whereas cases where the assignment is different (F1 and F2) are unsuccessful.

Figure 4: Project Classification Outcomes

| **Actual Outcome** | **Predicted Outcome** | |
|---|---|---|
| | **In Field/Purpose** | **Not in Field/Purpose** |
| **In Field** | T1 | F1 |
| **Not in Field** | F2 | T2 |

A number of measures of the performance of machine learning algorithm are possible. The "accuracy" of the predictions is simply the number of correct predictions as a share of all predictions:

(T1+T2)/(T1+T2+F1+F2)

Where the distribution of outcomes is uneven, however, this measure may not be very illuminating. For example, in a binary classification problem where 90% of observations are not in a field, simply guessing that no projects belong to that field would yield an accuracy of 0.9, but would be a thoroughly uninformative classifier.

To correct for this, two other measures of prediction success have been proposed and are routinely used in evaluating the effectiveness of machine learning algorithms. They are:

- Precision = T1/(T1+F2), and
- Recall = T1/(T1+F1)

Intuitively, Precision measures the share of all projects belonging to a classification that are correctly identified; Recall measures the share of projects that are predicted to be in the field that are correctly predicted. The F-1 score, which is the harmonic mean of Precision and Recall, is generally viewed as the best single summary of classifier performance.

**Results**

Among the different machine learning models, we found that the Logistic Regression classifier provides the best overall performance. Figure 5 summarizes the performance of the classifiers for each field of study, and Figure 6 reports performance for the classifiers for project purpose. In each case, we compare F-1 scores from the cross-validation results to those obtained using the testing sample. In cases where the number of projects was too small, we do not have any projects in the testing sample, so we cannot compute an F-1 score.

Figure 5: Comparison of F-1 Scores for Field of Study in Cross-Validation and Testing Samples

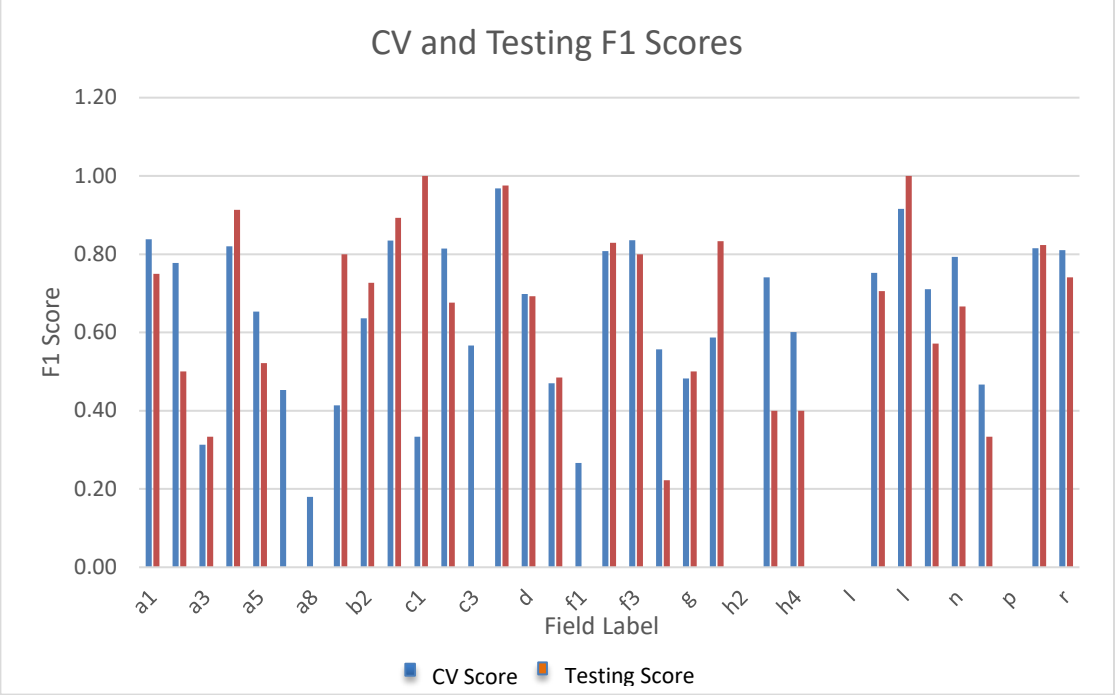
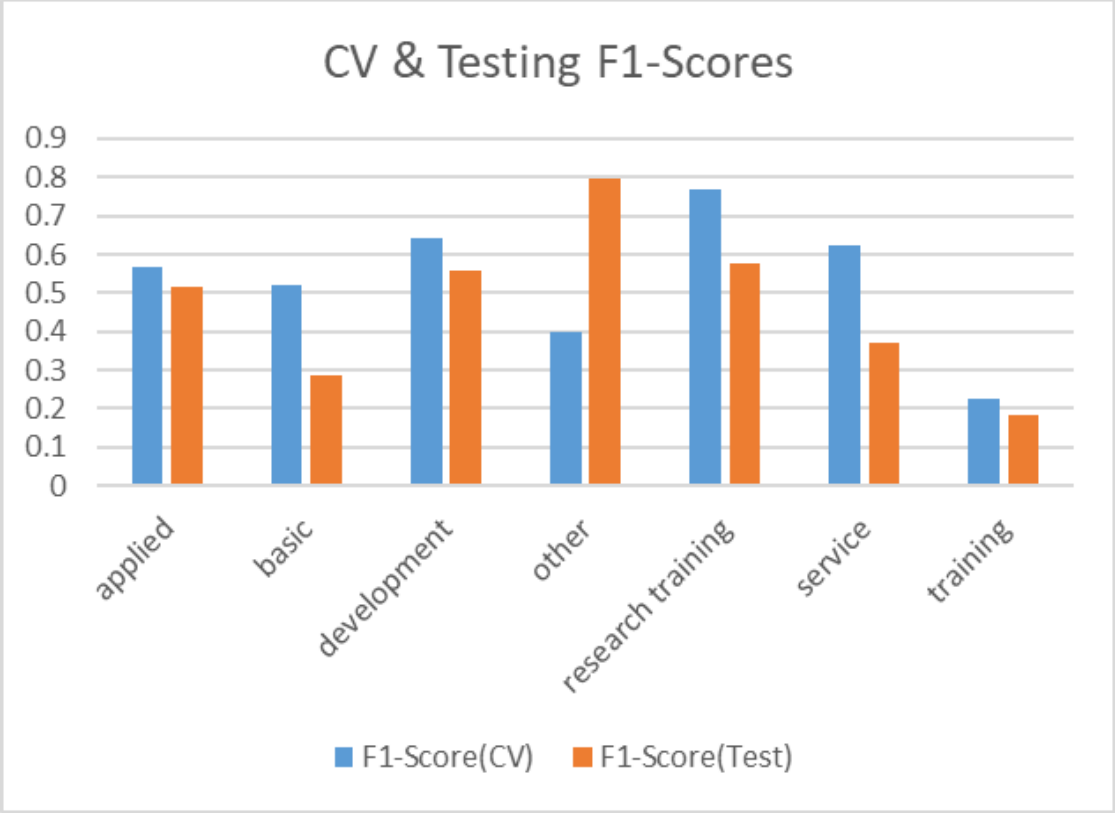


Figure 6: Comparison of F-1 Scores for Purpose in Cross-Validation and Testing Samples




As shown earlier (Figures 1 and 2), the distribution of projects by purpose and field is highly uneven; this difference accounts for much of the variation in classifier performance across the purpose/field categories. Figure 7 plots the relationship between the F-1 score and the number of projects assigned to each field in the training sample. F-1 scores rise sharply as the number of projects increases from 0 to about 40, reaching a range of 0.6-0.9 at this point. For cases with more than 60 projects the F-1 scores are clustered around 0.8.

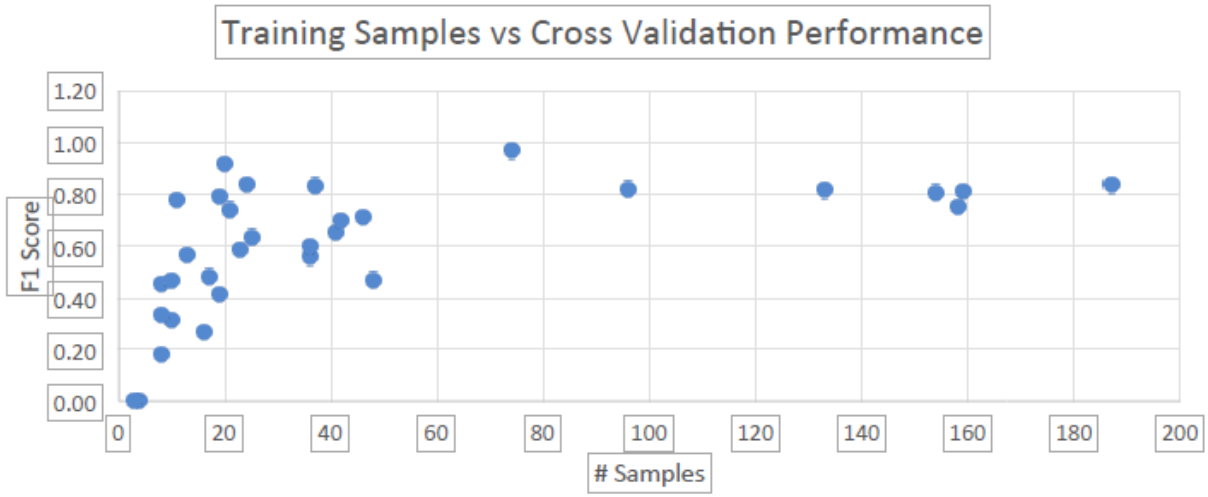



Figure 7: F-1 Scores for Field of Study vs. Number of Projects in Training Sample

Figure 8 shows comparable relationships for project purpose.

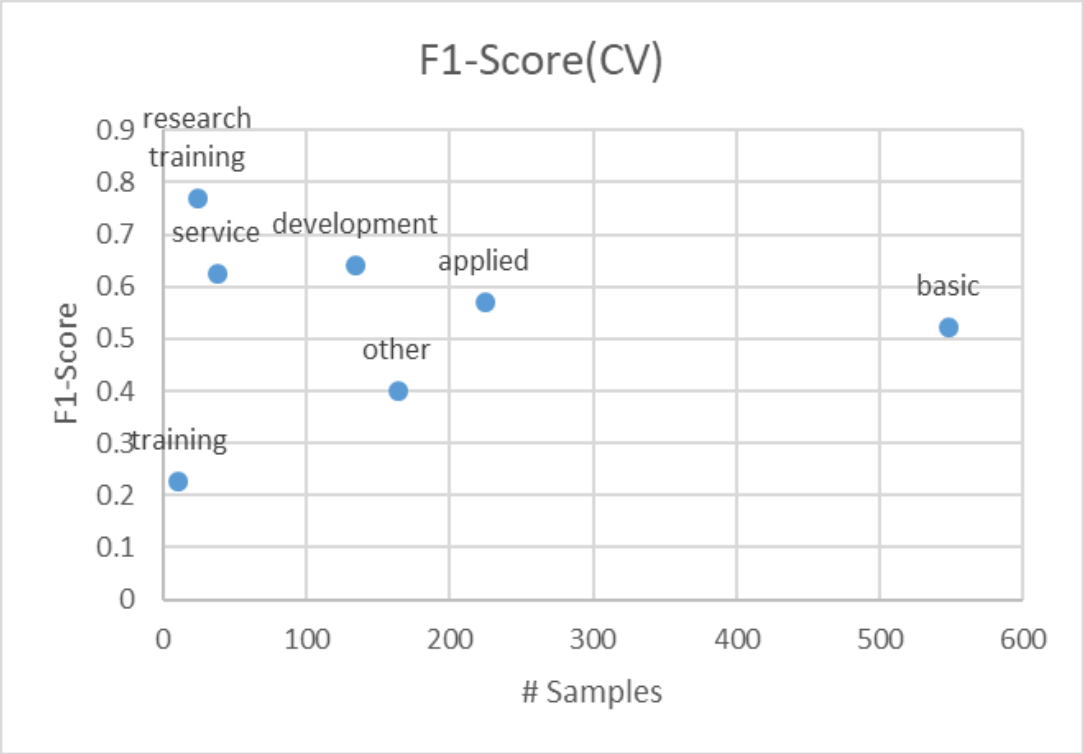



Figure 8: F-Scores for Purpose vs. Number of Projects in Training Sample

**Discussion**

We have not yet succeeded in developing a set of classifiers that will precisely reproduce the human judgements underlying the University of Kansas's response to the HERD survey. But it is not clear that this is the appropriate measure of the project's success.

First, while we have relied on human judgements to train the classifiers, it is not entirely obvious that we should regard the human assessments as constituting the ground truth in this case. Furthermore, different individuals at KU have classified projects for the HERD survey over the years; this may have added additional subjectivity or inconsistency in the classification of projects. It may be that the classifiers are more consistent in their judgment than humans. Evaluating this possibility requires more careful examination of the cases in which the two approaches produce different results. Careful analysis of these cases may help clarify the root of the disagreement and yield additional insights.

Second, the end product of the classification process is an aggregated report on expenditures broken down by field of study and purpose. Rather than focusing on the accuracy of the individual project classifications, it may prove more valuable to look at the extent to which aggregated results from the machine classifiers approximate the aggregated results from the human classifiers.

## Conclusions

As a proof of concept, we believe that the current project has been successful in demonstrating that it is possible to develop reasonably accurate machine-learning classifiers. We believe that many of the problems encountered so far will be reduced by expanding the training data set to include additional examples.

One initial objective of our project was to bring greater uniformity to HERD reporting across institutions. Future goals for this project include assessing the ability of our classifiers to successfully classify projects at other institutions. Adding additional projects from other institutions to the training data set may also offer opportunities to further refine the classifiers we have developed.

## References


Adams, James D. and Zvi Griliches. 1998. "Research Productivity in a System of Universities." Annals of Economics and Statistics 49/50: The Economics and Econometrics of Innovation, 127-62.

Feller, Irwin. 2001. "Elite and/or Distributed Science." In Maryann P. Feldman and Albert N. Link, eds, Innovation Policy in the Knowledge Based Economy. Norwell, MA and Dordrecht, Nether-lands.: Kluwer Academic.

Geiger, Roger and Irwin Feller. 1995. "The Dispersion of Academic Research in the 1980s." Journal of Higher Education 66, 336-60.

Graham, H. Davis and Nancy Diamond. 1997. The Rise of American Research Universities: Elites and Challengers in the Postwar Era. Baltimore, MD: Johns Hopkins University Press.

Lanahan, Loren, Alexandra Graddy-Reed and Maryann P. Feldman. 2016. The Domino Effects of Federal Research Funding. PLOS ONE (June 21) https://doi.org/10.1371/journal.pone.0157325

Rosenbloom, Joshua L., Donna K. Ginther, Ted Juhl and Joseph A. Hep-pert. 2015. "The Effects of Research & Development Funding on Scientific Productivity: Academic Chemistry, 1990-2009." PLOS ONE (September 15) https://doi.org/10.1371/journal.pone.0138176

Rosenbloom, Joshua L. and Donna K. Ginther. Forthcoming. "Show me the Money: Federal R&D Support for Academic Chemistry, 1990-2009." Research Policy.

National Science Board (2016) Science and Engineering Indicators 2016. Arlington, VA: National Center for Science and Engineering Statistics (NCSES) https://www.nsf.gov/statistics/2016/nsb20161/#/report