Applying Machine Learning to Compare Research Grant Programs

Khor, K.A.; Ko, G.; Walter, T.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Applying Machine Learning to Compare Research Grant Programs

Khor, K. A.*, Ko, Giovanni**and Theseira, Walter***

*mkakhor@ntu.edu.sg
Research Support Office and Bibliometrics Analysis, Nanyang Technological University, 76 Nanyang Drive, Singapore 637331

**gko@ntu.edu.sg
School of Social Sciences, Nanyang Technological University, 14 Nanyang Dr, Singapore 637332

*** waltertheseira@suss.edu.sg
School of Business, Singapore University of Social Sciences, 463 Clementi Road, Singapore 599494

Introduction

The flagship scientific funding programs in the United States and European Union collectively manage budgets exceeding US$40 billion dollars annually, with much of that awarded competitively to investigator-led research projects. Despite these enormous investments, evaluating performance across different funding programs is difficult. While a large literature compares research productivity cross-nationally (May, 1997; Adams, 1998; King, 2004), less is known about why some programs are more successful than others. In particular, do programs succeed because they select high-impact scientists, or because they select high-impact research areas?

A first step to distinguishing between these effects is to establish a common set of research funding areas to all programs studied. Controlling for such a common set, we can credibly attribute a program's superior performance to factors such as the grant selection process or greater institutional research support, rather than selection of high-output research areas. Alternatively, we may find that a program succeeds only because it made targeted bets on areas of science with high productivity; controlling for common research areas then reveals possible weaknesses in the program's grant selection process.

Ideally, we would determine this common set of research funding areas by asking program grant evaluation panels how they would classify and approve research projects approved by *other* funding agencies: Would they agree to fund a given project, and if so, how would the project fit into the program's scientific funding structure? Since grant selectors are too busy to entertain these questions, the usual method is to manually classify research projects according to scientific evaluation panels or research objectives. But manual classification requires expensive, well-trained research assistants, is subjective and potentially inconsistent, and is not scalable.

This paper applies Machine Learning classification algorithms to map research projects funded by the US National Institutes of Health (NIH) and the Singapore National Research Foundation (NRF) into the research funding structure used by the European Research Council (ERC). Our application of Machine Learning answers the question: "If a research project funded by the NIH or NRF were counterfactually funded instead by the ERC, which ERC

funding panel would have assessed and funded the project?" The Machine Learning approach has several advantages over research manpower: it is scalable, relatively low-cost, and is reproducible.

Literature Review

Most papers that compare research outcomes and productivity across nations examine aggregate measures to understand the research strengths of each nation (May, 1997; Adams, 1998; King, 2004; Cimini, et al., 2014). A common thread is that nations seem to have clear-cut research strengths in certain fields, where their researchers generally publish more papers and their papers are cited more. More developed nations also generally conduct more impactful research and have a broader base of strong research disciplines than less-developed nations. Less clear, however, is where these differential research productivities come from. King (2004) notes differences in "value-for-money" of research spending among the G8 and EU15, but does not identify a causal mechanism for the differential returns on research spending across nations. Auranen and Nieminen (2010) attempt to identify institutional differences in research funding mechanisms in eight countries that could lead to differential university research outputs, but do not directly test whether more competitive funding programs or a higher level of research funding causes better research outcomes. Due to the complexity of teasing apart causal factors at the macro-level, it may be more instructive to seek causal explanations for research productivity differentials at the level of individual grant funding programs. But to proceed, there first needs to be a consistent method for classifying funded projects for comparison.

There is a small but growing literature applying Machine Learning methods to automate classification of research projects (Yau et al 2014; Freyman et al 2016). These papers are motivated by a similar problem to ours: While research projects are usually labelled by administrators when a database is created, such labels are limited in scope and cannot address new research questions readily.[1] These papers demonstrate Machine Learning research project classification can be very accurate because the language used in scientific abstracts is highly discipline-specific; even when classifying interdisciplinary research, Machine Learning models can identify distinct underlying disciplines (Freyman et al 2016), and can cluster related papers accurately based on their common topics (Yau et al 2014). Our paper differs in that instead of identifying underlying labels, our Machine Learning model categorizes research abstracts from one agency using the implied classification scheme from a different agency.

Methodology

We classify a common set of research projects for three high-impact, high-risk early career grant programs aimed at achieving highly innovative research breakthroughs: The US NIH Director's New Innovator Award (NIH-NIA), the ERC Starting Grant (ERC-StG), and the Singapore NRF Fellowship (NRFF). These programs are ideal for cross-national comparison due to three shared features. First, all provide extraordinarily large and stable grant awards, exceeding $1.5 million US dollars over 5 years per project. Second, all are open to researchers of any nationality, provided they conduct their research in an institution in each programs' respective jurisdictions. Third, they all fund Life Sciences. A cross-national evaluation is therefore meaningful because all three programs compete for the same pool of

[1] For example, in Freyman et al (2016), National Science Foundation records were labelled by "the Directorate, Division, and Program(s) that funded the award", but not any "science tag or socioeconomic objective tag".

high-calibre, internationally mobile, early-career life scientists. Table 1 provides a summary overview of the programs.

Table 1: Comparison of the NIH-NIA, ERC-StG and NRFF early-career grant programmes

	NIH-NIA	**ERC-StG**	**NRFF**
Year Initiated	2007	2007	2008
Maximum Award	US$1.5 million for direct costs + additional budget for indirect costs	Up to €1.5 million + up to €0.5 million in start-up costs	Up to S$3 million (excluding PI salary)
Duration of Award	Up to 5 years		
Awardees granted annually	30 – 55	192 – 461 (Life Sciences and Physical Sciences)	6 – 12
Annual Aggregate Funding Quantum	US$73 – 132 million	€116 – 134 million	Not Available

Source: NRF Website (NRF, 2018), NIH Website (NIH, 2018), ERC Website (ERC, 2018)

Because the three programs we study have broad scope, the application and peer review processes are structured to group research projects according to scientific sub-disciplines. The review structure also facilitates decisions by program administrators to prioritize funding to specific areas of science.

The NIH-NIA requires applicants to choose one primary and one secondary area out of nine areas of Life Sciences; peer reviewers are assigned based on these choices. The NIH-NIA bypasses the standard peer review "study sections" used by the NIH for other competitive grants (Li and Agha 2015). After peer review, while applications across all areas compete for a "single source of funds", the NIH-NIA explicitly states that "programmatic considerations", in addition to the peer review scores, determine final funding decisions.

The ERC-StG requires applicants to choose one panel out of twenty-five panels, of which nine are for the Life Sciences and ten are for the Physical Sciences and Engineering. Peer reviewers are assigned based on the panel choice. The ERC-StG adopts a "bottom up" approach to scientific funding, assigning grant funds to panels proportionate to actual demand for grant funding. Grant awards are by design distributed fairly evenly across panels. This means ERC-StG funding decisions are effectively made at the panel level. The ERC does not appear to use the panel structure to prioritize funding in particular research areas.

The NRFF does not require applicants to designate specific sub-areas of science. Instead, the NRFF operates a unified grant call in all areas of Science and Technology. Applications are subsequently directed by the NRFF program office to relevant experts for peer review. The NRFF uses the peer review process to shortlist applicants, who are then invited to Singapore for a final round of interviews, after which grant funding decisions are made.

Data

Our data consists of 5718 research grant abstracts issued by the three agencies from 2008 to 2016. Because the ERC-StG provides the most structured framework for classifying research projects, and has a large database of 5308 research abstracts labelled with the ERC panel awarding funding, we chose to "train" the Machine Learning model on the ERC-StG. We wish to assign the 410 NIH-NIA/NRFF abstracts to ERC Panels using the "trained" Machine Learning model.

Machine Learning Algorithms

Our Machine Learning classification model is designed to build a predictive model from the ERC "training" dataset, to predict the appropriate ERC panel accurately for each NIH-NIA/NRFF abstract. Performance is evaluated by accuracy in classifying out-of-sample data; we discuss additional details on performance measurement shortly in the results section.[2] We compare the performance of three common Machine Learning classification models (Tan, et al. 2018): Multinomial Naïve Bayes (MNB), Multinomial Logistic Regression (MLR) and Support Vector Machines (SVM).[3] These models have low computational requirements and empirically demonstrated good classification accuracy.

We reduce the linguistic content of grant abstracts to a numerical set of features interpretable by Machine Learning models using the bag-of-words model. We first reduce every abstract in our corpus to a set of individual words. We remove punctuation and stop words – articles and pronouns such as "the", "this". We lemmatize remaining words, replacing them with their base dictionary form, e.g. "compared" and "comparing" is lemmatized to "compare". The outcome is a set of individual words that represent the semantic content of the abstract – a unigram set.[4] The corpus becomes a matrix where columns represent semantic features, e.g. "evolution", "engineer", and each abstract is a row. The cell values represent the importance of the feature to each abstract.[5]

Feature importance is commonly represented using either a yes-no indicator or a text frequency (TF) count. We normalise text frequency by multiplying by the inverse document frequency (TF-IDF) (Robertson 2004). Common semantic features, e.g. "model" are given lower weight, while rare features, e.g. "phlebotomy", which help classify abstracts more accurately, are given higher weight. To reduce noise and avoid overfitting, we exclude features that appear too frequently (in over 30% of abstracts) and too rarely (less than 5 abstracts).

Results and Discussion

We evaluate model performance by cross-validation, applying the trained classification model to a 1-in-10 holdback sample from the ERC abstract dataset. This allows us to directly compare model predictions against the actual panel of the abstract in the holdback sample. We score performance with the Precision and Recall metrics (Sokolova and Lapalme, 2009). Precision is the proportion of model-classified abstracts that are correctly predicted, while

[2] For a technical discussion on how Machine Learning fits the data and the intuition behind how these models are scored, refer to Arlot and Celisse (2010)

[3] For a technical discussion, refer to Domingos and Pazzani (1997) and Pang, et al. (2002) for MNB, Jurafsky and Martin (2008) for MLR, and Joachims (1998) for SVMs

[4] We also consider a bigram set, where pairs of words, e.g. "regulate process" form the semantic content. Higher order sets require considerably more computing power and generate little improvement in classification accuracy in our context.

[5] For a more detailed overview, refer to Khan, et al. (2010)

Recall is the proportion of all abstracts that are predicted correctly. Because there is a Precision-Recall tradeoff (Buckland and Gey 1994), and both measures are important, we employ the F1 metric – the harmonic mean of Precision and Recall – as our preferred summary performance measure.

Table 2: Mean classification accuracy scores for each classification model

	Mean F1 Score	Mean Precision	Mean Recall
Multinomial Logistic Regression	0.714	0.732	0.716
Multinomial Naïve Bayes	0.662	0.727	0.674
Support Vector Machine	0.731	0.743	0.732

Table 2 shows that SVM outperforms MNB and MLR in every metric across all panels. On average, 74.3 percent of SVM classifications are correctly assigned (Precision), and 73.2 percent of all abstracts are correctly predicted (Recall); the F1 Score is likewise the highest at 0.731. To benchmark model performance against purely random assignment, consider there are nineteen ERC panels in our training dataset, and as grants are evenly distributed across panels, purely random assignment should only be accurate about five percent of the time. Our model performance is somewhat worse than Yau et al. (2014) and Freyman et al. (2016), who achieve F1 scores over 0.8. However, Freyman et al. (2016) use a much larger data set (over 400,000 abstracts in total), and the corpus used by Yau et al. (2014) was much more diverse, which may contain more distinct features for classification. We use SVM predictions for further analysis of the NIH- and NRF-funded grant projects.

Figure 1: Predicted mapping of NIH and NRF projects onto ERC Life Sciences panels, with actual ERC project distribution

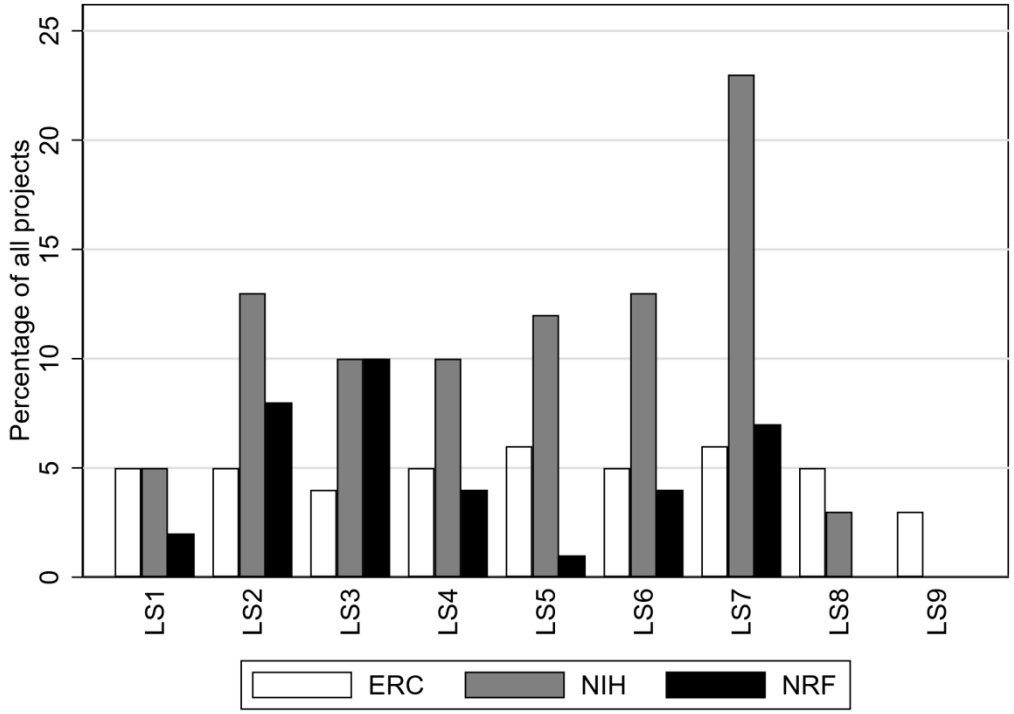

Figure 2: Predicted mapping of NIH and NRF projects onto ERC Physical Sciences panels, with actual ERC project distribution

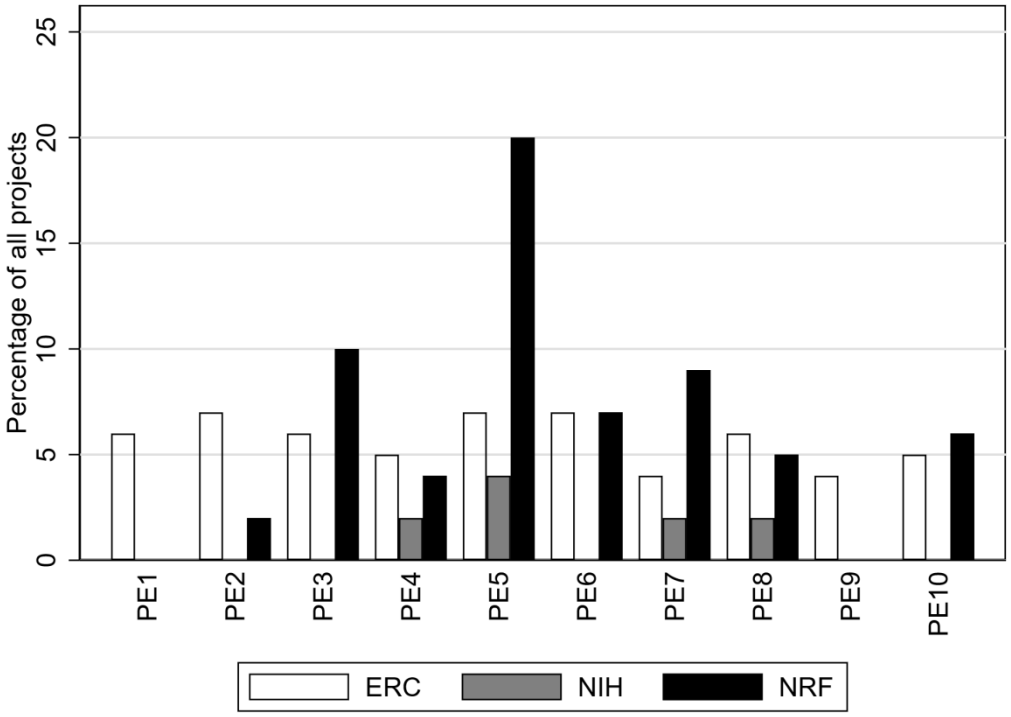

Table 3: Mapping of ERC Panel numbers to discipline

Panel	Description	Panel	Description
LS1	Molecular and Structural Biology and Biochemistry	PE1	Mathematics
LS2	Genetics, Genomics, Bioinformatics and Systems Biology	PE2	Fundamental Constituents of Matter
LS3	Cellular and Developmental Biology	PE3	Condensed Matter Physics
LS4	Physiology, Pathophysiology and Endocrinology	PE4	Physical and Analytical Chemical Sciences
LS5	Neurosciences and Neural Disorders	PE5	Synthetic Chemistry and Materials
LS6	Immunity and Infection	PE6	Computer Science and Informatics
LS7	Diagnostic Tools, Therapies and Public Health	PE7	Systems and Communication Engineering
LS8	Evolutionary, Population and Environmental Biology	PE8	Products and Processes Engineering
LS9	Applied Life Sciences and Non-Medical Biotechnology	PE9	Universe Sciences
		PE10	Earth System Science

Figures 1 and 2 apply the Machine Learning model to map research projects funded by NIH-NIA/NRFF into the ERC panel structure. The results show that the NIH-NIA and NRFF target specific scientific areas for funding, unlike the ERC which distributes grants evenly across panels.

The NIH-NIA concentrates funding in the Life Sciences panels LS2, LS3, LS4, LS5, LS6 and LS7. Projects in LS7 (Diagnostic Tools, Therapies and Public Health) are by far the largest recipient of NIH-NIA grants. These six Life Sciences panels correspond to disciplines associated with medical biology, biomedicine and medical technologies, consistent with the NIH role as the US federal medical research agency. Presumably, Life Sciences research in LS1, LS8 and LS9, which are less readily applied to medical research, are funded by the US National Science Foundation instead.

The NRFF concentrates funding in PE5 (Synthetic Chemistry and Materials), with LS3, LS4, LS7, PE3, PE6, PE7 and PE10 receiving most of the remaining NRFF funds. In contrast to the NIH, which has a clear thematic research goal, and the ERC, which funds research areas evenly by design, the NRFF makes targeted bets in research areas, but without a clear disciplinary pattern. One possibility is that potential commercial applications from the three largest discipline groups funded by the NRFF, PE5, PE3 and LS3, could support the NRF's RIE2020[6] goals in the Technology Domains of Advanced Manufacturing and Engineering, and Health and Biomedical Sciences (NRF, 2016).

Conclusion

Given sufficient training data, Machine Learning models can classify research abstracts according to a common classification scheme with high accuracy (F1 scores exceeding 0.7). While the Support Vector Model was the best performer, all commonly used classification models exhibited high accuracy and are not computationally intensive. This demonstrates Machine Learning text classification is a viable and scalable alternative to using research manpower for conducting matched analysis of grant programs.

There are some caveats. Because Machine Learning text classification models rely on the statistical presence of text features, abstracts which strategically misrepresent the underlying research may generate systematic errors in classification. For example, abstracts which contain strategically inserted keywords that appeal to funding priorities (but which are only minimally related to the research) may generate misclassification, although arguably, a non-expert research assistant (or even the grant funders) may likewise be misled. We are exploring robustness checks to investigate this further.

We have shown the three grant funding agencies studied differ widely in how they utilize funding structure design to prioritize key research areas. Now that the strategic funding areas of the NIH-NIA/NRFF have been identified using the ERC panel framework, as a next step, we can estimate differences in cross-national research productivity, controlling for research funding strategy. This allows us to address one of the central questions in the research funding program literature, namely, whether differences in scientific output across programs and nations are due to more productive topics being funded, or more productive researchers being funded.

[6] Research Innovation Enterprise 2020 Plan, a blueprint created by the Singapore National Research Foundation to guide the agency's research funding priorities from 2016 to 2020

References

Adams, J. (1998). Benchmarking international research. *Nature*, 396, 615–618.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.

Auranen, O., & Nieminen, M. (2010). University research funding and publication performance—An international comparison. *Research Policy*, 39(6), 822-834.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.

Cimini G, Gabrielli A, Sylos Labini F (2014) The Scientific Competitiveness of Nations. *PLOS ONE*, 9(12). Retrieved 19 August 2018 from: https://doi.org/10.1371/journal.pone.0113470

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2), 103–130.

European Research Council (2018). Starting Grants. Retrieved 15 August 2018 from: https://erc.europa.eu/funding/starting-grants.

Freyman, C. A., Byrnes, J. J., and Alexander, J. (2016). Machine-learning-based classification of research grant award records. *Research Evaluation*, 25(4), 442–450.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C. (Ed.), *Machine Learning*: ECML-98, (pp. 137–142), Berlin, Heidelberg: Springer Berlin Heidelberg.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2nd edition.

Khan, A., Baharudin, B., Lee, L. H., and Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, 1(1), 4–20.

King, D. A. (2004). The scientific impact of nations. *Nature*, 430:311–316.

Li, D. and Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, 348(6233), 434–438.

May, R. M. (1997). The Scientific Wealth of Nations. *Science*, 275(5301):793–796.

National Institutes of Health (2018). NIH Director's New Innovator Award. Retrieved 15 August 2018 from: https://commonfund.nih.gov/newinnovator.

National Research Foundation (2018). NRF Fellowship and NRF Investigatorship. Retrieved 15 August 2018 from: https://www.nrf.gov.sg/funding-grants/nrf-fellowship-and-nrf-investigatorship.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503–520.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427 – 437.

Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. (2018). *Introduction to Data Mining*. Pearson, second edition edition.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, pp. 412–420. Nashville, Tennessee, USA: Morgan Kaufmann Publishers.

Yau, C.-K., Porter, A., Newman, N., and Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767–786.