5-4-2018

Can We Predict Reproducible Scholarly Research?

Joseph C. McDade

NORTHERN ILLINOIS UNIVERSITY

Can We Predict Reproducible Scholarly Research?

A Thesis Submitted to the

University Honors Program

In Partial Fulfillment of the

Requirements of the Baccalaureate Degree

With Upper Division Honors

Department Of

Computer Science

By

Joseph McDade

DeKalb, Illinois

December 2018

University Honors Program

Capstone Approval Page

Capstone Title (print or type)

Can We Predict Reproducible Scholarly Research?

Student Name (print or type) Joseph McDade

Faculty Supervisor (print or type) HAMED ALHOORI

Faculty Approval Signature 

Department of (print or type) Computer Science

Date of Approval (print or type) 5/04/2018

HONORS THESIS ABSTRACT

Empirical research should always be backed by substantial and verifiable data so that anyone who wishes to reproduce the study or replicate the study with different data can verify the claims made by the research are accurate. We attempt to use a novel method of discovering reproducible research papers. Using this technique future research can be done to provide an even better understanding of the reproducibility crisis. We collected scholarly data from three different sources and combined them in order to obtain a dataset of 657 papers. The dataset comprises of papers that are verified as reproducible and ones that have been shown to not be reproducible. When the dataset was cleaned it resulted in 237 papers marked reproducible and 36 irreproducible. We then used three different models; Gaussian Naive Bayes, Multinomial Naive Bayes, and Adaboost to classify texts based on structural characteristics of papers and linguistic. Then we used a Long Short-Term Memory Recurrent Neural Network to compare results.

Can We Predict Reproducible Scholarly Research?

Introduction:

An important step in the scientific method is that of peer review. It's allows for an idea to be given credence [1]. Peer review is typically done through an attempt of another researcher to replicate the original researcher's results [2]. When a peer can replicate the results of a study using similar concepts, but their own data the study is considered to be replicable [3], [4]. For a long time this was considered the gold standard for the scientific method. In the modern era this isn't always possible. The time and effort to get a novel dataset of an appropriate size may be too great or too expensive. The emerging standard in computational sciences is reproducibility [5]. Reproducibility is often defined as the ability of researcher to reproduce the results of another using the original methods and dataset provided by the original researcher [3], [4], [6], [7]. With computational sciences becoming the central focus, it is required that reproducibility be the minimum standard [8], [9]. In order for a study to be reproducible the original code used for computations and the original dataset must be made available. There are several guides outlining the standard available [8], [10]–[17]. Some journals have begun adopting these standards and requiring them for publication [18], [19]. However, even with these standards in place Savage and Vickers [20] found that datasets were rarely shared and only 1 of 10 from the Public Library of Science were available. Not all think journals increasing standards is a good thing, it could create a chilling effect and an increase of the "file drawer" problem [21], [22]. Many have gone a step further to try to solve the problem by creating workflow tools that, if used, create a sort of digital paper trail that other researchers can use to better reproduce results [23]–[28] one even on the trending technology of the blockchain [29]. Others have gone even further yet to not only offer tools for research, but the capability to encapsulate the entire research environment including the operating system it was run on through containerization or virtual machines [30], [31].

The phrase reproducibility crisis was born in a paper by Pashler and Wagenmakers in 2012 [32]. However, the problem of reproducibility has been known for longer, author Ioannidis argued that most published research finding are false in 2005 [33], [34]. Ioannidis continues to write about the problem of reproducibility through to even the current year 2018 [35]–[37]. One of the earlier works looking at the the problem was in 2006 when the field of psychology was faced with the problem of irreproducible research. It was shocking for the research community after it was discovered that most of the studies published by psychology researchers in that year have been tagged as irreproducible. A total of 103 out of 141 authors for different academic studies didn't respond with their data over a six month period [38]. In a follow up study [39] published in 2015, it was found that 246 out of 394 contacted authors of papers in APA journals did not share their data upon request. In 2015, Psychology became the first discipline to conduct and publish an open, registered empirical study of reproducibility called the Reproducibility Project. 270

researchers from around the world collaborated to replicate 100 empirical studies from three top Psychology journals. Fewer than half of the attempted replications were successful [40]. It is important to note that while the fields of psychology and biology have been a major focus of the reproducibility crisis, there is no reason to believe other fields aren't also experiencing the same issues. A study by Chang and Li [41] found that in the field of economics they were able to reproduce 22 of 67 papers (33%).

In 2016, Monya Baker surveyed 1,576 scientists. Her findings showed that 52% of those surveyed thought there was a "significant crisis" of reproducibility. Only 24% said they successfully published a replication while 13% had failed to publish. Perhaps the most striking number is that 70% of the researchers had tried and failed to reproduce another scientist's experiment [42]. There are various reasons why science struggles with reproducibility. It takes more time and can be more costly for the original author to take steps to make sure their research is reproducible [43]. Novel research is generally seen as more valuable than reproducing another's results [44]. Kovacevic [45] writes that the ultimate goal of researchers should not just to be published, but to also have the data, software, and anything else needed to reproduce the research available. In some cases it may be that even though you are working with the same code, you would need access to the same hardware types [46] or have differences in training models sensitivity [47].

It is important that studies be reproducible to make sure we are spending money on effective science [48], [49], in 2015 a paper by Freedman, et al. [48] found the cost of irreproducible preclinical research to be $28 billion in the US alone. Making the work done by researchers openly available is seen as an effective way to settle scientific disputes [50]. Antelman [51] conducted a study into the effectiveness of these open access studies and found that open access studies tend to have a greater impact research impact than those that were not open. This also held true across a variety of disciplines. Follow up research agrees and has found that open access articles were cited earlier and more often than non-open access articles [16], [52]–[55].

Previous or Related Work:

A common approach to find the reproducibility of a study and how prevalent the problem is in the field is to attempt to reproduce random studies from multiple journals. The previously mentioned Open Science Collaboration's [40] study did just that by attempting to reproduce 100 different studies across 3 journals. Others have also taken up this approach [56]–[58]. This has the obvious downside of requiring a lot of resources, researchers, and time. Some meta research has been involved with trying to find features of irreproducible work. The Open Science Collaboration [40] attempted to match P values to the reproducibility and found while no single indicator was sufficient in predicting a study's reproducibility, that there was a correlation between P values and reproducibility. An underlying issue with this

technique could be a tendency for P values to be misunderstood [59] that has caused a discussion of changing minimum P values [21], [36], [60]. Building on what the Open Science Collaboration accomplished authors Van Bavel, et al. [61] took the results of the Open Science Collaboration's work and found that contextual sensitivity of the topic also correlated with reproducibility. While this study did not attempt to make predictions on reproducibility, it did use code to make predictions on contextual sensitivity. A study by Dreber, et al. [62] did attempt to predict reproducibility. The prediction was still conducted in a manual fashion allowing a pool of researchers to place "bets" on which studies would succeed. They then attempted to reproduce 44 studies and finished 41 of them. The prediction market had a 71% success rate, which is higher than the random binomial control rate of 50%.

We plan to build on the works done by the Open Science Collaboration and others. Instead of conducting our own manual reproductions of previous work, we will use the findings of these other groups to build a classifier that attempts to predict if a study is reproducible based on several features of the research paper itself. Instead of considering just P values or contextual sensitivity, we plan to look at the structure of the papers along with linguistic characteristics.

Data:

For our experiment we used a dataset found from a similar style of study as the Open Science Collaboration conducted. The Computer Science department at University of Arizona conducted a study resulting in a data set of 601 computer science papers that can be found here http://repeatability.cs.arizona.edu/. The website also contains an unpublished research paper [63] that explains the dataset in a separate section. We felt that we can scrape this data and then build features on top of the existing data. We wrote a python script for scraping the dataset from that page and then stored the results to a csv file. Cleaning the dataset took place in two phases :

- In the first phase downloading all the rows in the webpage and storing them onto pandas dataframe was done. The column labels were maintained from the original site.
- In the second phase we scrape nested hyperlinks that store meta information pertaining to the paper (author, DOI, PDF) and store the hyperlink as a new column.

The dataset has other essential fields that define the reproducibility of a paper. These include columns such as "Code location", if observed carefully these closely relate to the structural features we proposed for building the model. We want to extend the dataset by downloading the full texts for every paper mentioned in the paper. We also observed that there are certain limitations to the current dataset as the downloaded features are filled with N.A's. The next step is to eliminate all the N.A's from the dataset and have only those columns that explain the dataset in its entirety. Also, while cleaning the dataset we had to normalize certain fields, the first

one we did is the location of the code column that had four classes briefly describing the location of the code for the respective paper. We made that into a binary feature so that it simplifies the process of classification. Below are the statistics for the dataset :

1. paper_doi : DOI for the scholarly paper [String]
2. paper_title: Title of the scholarly paper [String]
3. paper_hyp: Hyperlink for the PDF of the paper [String]
4. code_location: Is code present in the scholarly paper [Binary]
5. reproducibility: Is the scholarly paper reproducible or not [Binary]

Upon cleaning the data we had 217 remaining papers marked as reproducible and only 9 marked irreproducible. We then went to https://retractionwatch.com/ where we manually gathered 16 more papers that are marked irreproducible. After downloading the Retraction Watch papers, we had to write a script that extracted the information from the MongoDB style databases and add them to our CSV created from the University of Arizona data. To train our model for two classes we have undersampled the 217 reproducible papers to a randomly chosen 100 papers for our positive class, we then over sampled our 25 irreproducible papers to 100 for our negative class.

For cross validation we have collected 40, new papers. The 40 papers are composed of 20 marked reproducible and 20 marked irreproducible. We retrieved the new papers from The XPhi Replicability Project [64]. The webpage has descriptive analysis of the replication study for 38 papers. Out of the 38, there were only 7 papers that were marked irreproducible and 31 papers as reproducible. We took 20 papers from reproducible set and included them for the final cross validation set. For the 13 other papers we needed for the irreproducible class we again manually retrieved papers from Retraction Watch.

Feature Engineering:

Selecting the right features to fit the curve of the statistical learning model is a challenging task, because there are numerous features available to experiment with when it comes to linguistics aspect of any text and choosing the proper ones requires certain experimentation. To understand the description of the below features let us assume there exists a paper *P*, where:

Feature	Description
Length of the paper	The length of the full text in paper *P*
Number of references	The number of references cited in paper *P*

Availability of sections	Are there multiple subsections in paper *P* (Introduction, Methodology, Conclusion etc)
Availability of dataset	Are there any hyperlinks or sources provided for the dataset used in paper *P*
Presence of images	Are there images present in paper *P*
Presence of links	Are there hyperlinks present in paper *P*
Presence of tables	Are there tables present in paper *P*
Presence of algorithms	Are there algorithms present in paper *P*

Table 1. List of Structural Features and their descriptions

Feature	Description
Word count	The number of words present in paper *P*
Paragraph count	The number of paragraphs present in paper *P*
Average Paragraph length	Average length of paragraphs present in paper *P*
Average word length	Average word length for paper *P*
Average sentence length	Average sentence length for paper *P*
Frequency of words greater than average word length	The number of words in paper *P* whose word length is greater than the average word length
Syntactic complexity	Sentence syntax similarity, all combinations, across paragraphs, mean for paper *P*
Word Concreteness	The text easability principal component score for paper *P*
Yule's K measure of lexical diversity	The lexical richness of the text present in paper *P*
Polysemy for content words	The number of content words (non stop words) for whom there exists a case of polysemy in paper *P*

Table 2. List of Linguistic Features and their descriptions

The features which we are going to employ are therefore being categorized into two categories namely:

1. Structural features
2. Linguistic features

Table 1 shows the Structural features which are representative of the structure of the paper. These might include aspects like presence of certain sections etc that we generally observe when we want to evaluate the claims made by author of paper *P*. Table 2 shows the Linguistic features that are representative of the semantics of the text present in paper *P*. These might help us understand to check if linguistic features offer any contribution towards the final classification model.

Methods:

In order to predict whether a given research paper is reproducible or not there is an underlying task of analyzing the nature of texts. We initially perform structural and linguistic analysis on the aforementioned dataset using Gaussian Naive Bayes, Multinomial Naive Bayes, and AdaBoost classification models. Once completed we then used a RNN/LSTM (Recurrent Neural Network Long Short-Term Memory) [65], [66], using the full text of the papers as input and labelling. In searching through related works, it appears that this idea is a novel way of predicting reproducibility of research.

For validating the performance of our models, we used k-fold cross validation of 10 folds. K-fold is a cross validation that works by dividing the dataset into k parts, using k-1 parts to train the model and the remaining 1 part to validate [67]. Then the dataset is iterated over with the validation part changing such that all k parts are validated against.

The nature of the dataset which we are trying to analyze played a crucial part in determining what machine learning model would be helpful in fitting a line between the data points. The primary goal of the experiment is to try different models and classification techniques to find which ones best predict reproducibility. After finding which techniques work best, we concluded that Gaussian Naive Bayes classifier best classifies new data into two coarse-grained classes "***Reproducible***" and "***Not reproducible***" (also could be 1 and 0). The nature of the classification highly depends on the structural features of the research papers. With the number of times the paper is cited, presence of sections, and code availability being the most important features.

Results:

Experiment 1 - Part A: The first part of Experiment 1 uses the structural features and the results are presented in Table 3.

Model	Train Accuracy	Validation Accuracy	Test Accuracy	Precision	Recall	F-1
AdaBoost	0.86 (+/- 0.10)	0.80 (+/- 0.37)	0.900	0.840	1.00	0.913
Gaussian NB	0.831 (+/- 0.12)	0.85 (+/- 0.40)	0.775	0.667	1.00	0.840
Multinomial NB	0.60 (+/- 0.14)	0.90 (+/- 0.24)	0.700	0.737	0.667	0.700

Table 3. Structural Models

We split the dataset 80/20. The 80% was used for training and validation. The 20% was used for testing the model. After running the experiment using the structural features we found that AdaBoost classification model performs the best at 86% percent accuracy with 95% confidence interval. The model didn't overfit for the training data. The accuracy of the model was consistent on the validation set with 80% accuracy. The accuracy for the test data was 90%. The Precision for the model was 84% while the Recall was 100%. The overall f-statistic for the model was 0.91. The other models that we used for this experiment performed worse. If we observe Table 3. we can see that for both the AdaBoost and Gaussian Naive Bayes models there is low precision and high recall; this is likely the result of the oversampling that was done for the negative class.

Experiment 1 - Part B: The second part of Experiment 1 uses the linguistic features and the results are represented in Table 4.

Model	Train Accuracy	Validation Accuracy	Test Accuracy	Precision	Recall	F-1
Gaussian NB	0.769	0.758	0.667	0.857	0.750	0.800
AdaBoost	0.994	0.708	0.667	0.857	0.750	0.800
Multinomial NB	0.711	0.692	0.444	1.000	0.375	0.545

Table 4. Linguistic Models

The second part of the first experiment included linguistic features that were computed using the full texts for all of the papers. We used the same 80/20 split for the dataset. All the linguistic features were numeric in nature. After running the experiment on linguistic features we found that Gaussian Naive Bayes performed best when compared to the remaining models. As is shown in Table 4 it had a training accuracy of 76%, validation accuracy of 75%, and a test accuracy of 67%. The Precision for the model was 85% while the Recall was 75%. The f-statistic was 0.80. If we observe Table 4 we can see that AdaBoost and the Multinomial classification models both have a much higher accuracy on the training set then the test set indicating that they were overfitting. There was one more key observation that we made; Multinomial Naive Bayes has a higher precision and a lower recall which implies that the model is picky on either one of the classes.

Experiment 2: The next experiment we conducted included using an neural network. This technique is similar to the classification models, but instead of building numerous features from the text we embed the words in the text as integers and feed those word embeddings as input to the neural network. The neural network that we used was LSTM (Long Short Term Memory). The first activation function that was used for the LSTM cell is ReLU (Rectified Linear Unit) function. An activation is a node in the network that can be used for the output of any neural network. The second activation function that was used was sigmoid. Using the LSTM we attempt to train a network model to predict if a paper is reproducible or not. The experiment yielded a train accuracy of 97% and a validation accuracy of 87%. The test accuracy for the model was surprisingly 100%. Although the statistics imply that the model performed well, the loss functions told a

different story altogether. Even though for every epoch the validation accuracy and the train accuracy was increasing we did not observe any significant change in the loss. The oddities might be better explained using a bigger dataset.

Epochs	5
Train Accuracy	0.97
Validation Accuracy	0.87
Test Accuracy	1
Loss Train	0.6467
Loss Val	0.6328
Loss Test	0.613

Table 5. RNN/LSTM

Conclusion:

We performed two classification experiments with different features and using different models and observed that we can predict if or not a newly given research paper could be reproducible or not with an accuracy of 86% accuracy. There are some observations that we made from those experiments. First, the structural features that were extracted from a research paper proved to be more useful for the classification experiment when compared to the linguistic features. Second, we might need different linguistic features that explain or talk about reproducibility in a paper. Although, we didn't display the statistics for the Random Forest Classification model it did reveal that the number of citations, availability of sections, and presence of code were the most important features for the model. Even though we cannot say that is true for the remaining models we feel it is likely they have played an important role for classifying the papers. The code used for the models can be found on Github at https://github.com/akhilpandey95/reproducibility. (Private repo)

Future Work:

The biggest benefit for a future experiment would be a larger dataset. These datasets are costly and time consuming to create. As projects such as the Open Science Collaboration, the University of Arizona's, and other fields continue to research the problem of reproducibility the available dataset should grow. With a large enough dataset we would expect better results from the RNN/LSTM. Since our experiment shows a greater success with structural features than linguistic, a future experiment would likely add more structural features. Due to the surprisingly poor results of the linguistic features, it would be suggested to try other linguistic measures or add them to the structural features to see if there is an improvement.

References:

[1] R. Smith, "Peer review: a flawed process at the heart of science and journals," J. R. Soc. Med., vol. 99, no. 4, pp. 178–182, Apr. 2006.

[2] M. C. Frank and R. Saxe, "Teaching Replication," Perspect. Psychol. Sci., vol. 7, no. 6, pp. 600–604, 2012.

[3] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis, "What does research reproducibility mean?," Sci. Transl. Med., vol. 8, no. 341, pp. 341ps12–341ps12, 2016.

[4] L. A. Barba, "Terminologies for Reproducible Research." Jan-2018.

[5] R. D. Peng, "Reproducible research in computational science," Science, vol. 334, no. 6060, pp. 1226–1227, Dec. 2011.

[6] J. B. Buckheit and D. L. Donoho, "WaveLab and Reproducible Research," in Lecture Notes in Statistics, 1995, pp. 55–81.

[7] C. J. Playford, V. Gayle, R. Connelly, and A. J. G. Gray, "Administrative social science data: The challenge of reproducible research," Big Data & Society, vol. 3, no. 2, p. 205395171668414, 2016.

[8] "Reproducible Research - IEEE Journals & Magazine." [Online]. Available: http://ieeexplore.ieee.org/document/5562471/authors?part=1. [Accessed: 20-Feb-2018].

[9] S. Fomel and J. F. Claerbout, "Guest Editors' Introduction: Reproducible Research," Comput. Sci. Eng., vol. 11, no. 1, pp. 5–7, 2009.

[10] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," PLoS Comput. Biol., vol. 9, no. 10, p. e1003285, Oct. 2013.

[11] V. Bajpai, M. Kühlewind, J. Ott, J. Schönwälder, A. Sperotto, and B. Trammell, "Challenges with Reproducibility," in Proceedings of the Reproducibility Workshop on ZZZ - Reproducibility '17, 2017.

[12] J. B. Asendorpf et al., "Recommendations for increasing replicability in psychology," in Methodological issues and strategies in clinical research (4th ed.)., pp. 607–622.

[13] K. J. Gorgolewski and R. A. Poldrack, "A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research," PLoS Biol., vol. 14, no. 7, p. e1002506, Jul. 2016.
[14] J. P. Mesirov, "Computer science. Accessible reproducible research," Science, vol. 327, no. 5964, pp. 415–416, Jan. 2010.
[15] V. Stodden and S. Miguez, "Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research," Journal of Open Research Software, vol. 2, no. 1, p. 8, Jul. 2014.
[16] J. Vitek and T. Kalibera, "Repeatability, reproducibility, and rigor in systems research," in Proceedings of the ninth ACM international conference on Embedded software - EMSOFT '11, 2011.
[17] F. Chirigati, R. Capone, R. Rampin, J. Freire, and D. Shasha, "A collaborative approach to computational reproducibility," Inf. Syst., vol. 59, pp. 95–97, 2016.
[18] M. McNutt, "Reproducibility," Science, vol. 343, no. 6168, p. 229, Jan. 2014.
[19] V. Stodden, P. Guo, and Z. Ma, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals," PLoS One, vol. 8, no. 6, p. e67111, Jun. 2013.
[20] C. J. Savage and A. J. Vickers, "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals," PLoS One, vol. 4, no. 9, p. e7078, 2009.
[21] H. Campbell and P. Gustafson, "The world of research has gone berserk: modeling the consequences of requiring 'greater statistical stringency' for scientific publication," arXiv [stat.ME], 16-Mar-2018.
[22] N. Keiding, "Reproducible research and the substantive context," Biostatistics, vol. 11, no. 3, pp. 376–378, Jul. 2010.
[23] P. Thomas et al., "Sharing and Preserving Computational Analyses for Posterity with encapsulator," arXiv [cs.DL], 15-Mar-2018.
[24] N. Matloff, R. Davis, L. Beckett, and P. Thompson, "revisit: a Workflow Tool for Data Science," arXiv [stat.AP], 16-Aug-2017.
[25] M. Schwab, N. Karrenbach, and J. Claerbout, "Making scientific computations reproducible," Comput. Sci. Eng., vol. 2, no. 6, pp. 61–67, 2000.
[26] C. Hurlin, C. Pérignon, V. Stodden, F. Leisch, and R. D. Peng, "RunMyCode. org: A research-reproducibility tool for computational sciences," Implementing reproducible research. CRC Press, Boca Raton, FL, pp. 367–381, 2014.
[27] B. A. Wandell, A. Rokem, L. Perry, G. Schäfer, and R. F. Dougherty, "Data management to support reproducible research," 2015.
[28] J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," Genome Biol., vol. 11, no. 8, p. R86, Aug. 2010.
[29] C. Furlanello, M. De Domenico, G. Jurman, and N. Bussola, "Towards a scientific blockchain framework for reproducible data analysis," arXiv [cs.CY], 20-Jul-2017.
[30] C. Boettiger, "An introduction to Docker for reproducible research," Oper. Syst. Rev., vol. 49, no. 1, pp. 71–79, 2015.
[31] V. Stodden and S. Miguez, "Provisioning Reproducible Computational Science," 2014.

[32] H. Pashler and E.-J. Wagenmakers, "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?," Perspect. Psychol. Sci., vol. 7, no. 6, pp. 528–530, Nov. 2012.
[33] J. P. A. Ioannidis, "Why Most Published Research Findings Are False," Chance , vol. 18, no. 4, pp. 40–47, 2005.
[34] J. P. A. Ioannidis, "Contradicted and initially stronger effects in highly cited clinical research," JAMA, vol. 294, no. 2, pp. 218–228, Jul. 2005.
[35] J. P. A. Ioannidis, "Why Science Is Not Necessarily Self-Correcting," Perspect. Psychol. Sci., vol. 7, no. 6, pp. 645–654, 2012.
[36] J. P. A. Ioannidis, "The Proposal to Lower P Value Thresholds to .005," JAMA, Mar. 2018.
[37] J. P. A. Ioannidis, "Meta-research: Why research on research matters," PLoS Biol., vol. 16, no. 3, p. e2005468, Mar. 2018.
[38] J. M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar, "The poor availability of psychological research data for reanalysis," Am. Psychol., vol. 61, no. 7, pp. 726–728, Oct. 2006.
[39] W. Vanpaemel, M. Vermorgen, L. Deriemaecker, and G. Storms, "Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm," Collabra, vol. 1, no. 1, 2015.
[40] O. S. Collaboration, "Estimating the reproducibility of psychological science," Science, vol. 349, no. 6251, p. aac4716, Aug. 2015.
[41] A. C. Chang and P. Li, "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not,'" SSRN Electronic Journal, 2015.
[42] M. Baker, "1,500 scientists lift the lid on reproducibility," Nature, vol. 533, no. 7604, pp. 452–454, May 2016.
[43] Q. Scheitle, M. Wählisch, O. Gasser, T. C. Schmidt, and G. Carle, "Towards an Ecosystem for Reproducible Research in Computer Networking," in Proceedings of the Reproducibility Workshop on ZZZ - Reproducibility '17, 2017.
[44] B. A. Nosek and D. Lakens, "Registered Reports," Soc. Psychol., vol. 45, no. 3, pp. 137–141, 2014.
[45] J. Kovacevic, "How to Encourage and Publish Reproducible Research," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007.
[46] M. N. Mahfoudi, T. Turletti, T. Parmentelat, and W. Dabbous, "Lessons Learned while Trying to Reproduce the OpenRF Experiment," in Proceedings of the Reproducibility Workshop on ZZZ - Reproducibility '17, 2017.
[47] M. Hutson, "Artificial intelligence faces reproducibility crisis," Science, vol. 359, no. 6377, pp. 725–726, Feb. 2018.
[48] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, "The Economics of Reproducibility in Preclinical Research," PLoS Biol., vol. 13, no. 6, p. e1002165, Jun. 2015.
[49] A.-W. Chan et al., "Increasing value and reducing waste: addressing inaccessible research," Lancet, vol. 383, no. 9913, pp. 257–266, Jan. 2014.

[50] V. Stodden, "The Scientific Method in Practice: Reproducibility in the Computational Sciences," SSRN Electronic Journal, 2010.
[51] K. Antelman, "Do Open-Access Articles Have a Greater Research Impact?," Coll. Res. Libr., vol. 65, no. 5, pp. 372–382, 2004.
[52] H. A. Piwowar, R. S. Day, and D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate," PLoS One, vol. 2, no. 3, p. e308, Mar. 2007.
[53] G. Eysenbach, "Citation advantage of open access articles," PLoS Biol., vol. 4, no. 5, p. e157, May 2006.
[54] J. M. Wicherts, M. Bakker, and D. Molenaar, "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results," PLoS One, vol. 6, no. 11, p. e26828, Nov. 2011.
[55] E. C. McKiernan et al., "How open science helps researchers succeed," Elife, vol. 5, Jul. 2016.
[56] T. M. Errington, E. Iorns, W. Gunn, F. E. Tan, J. Lomax, and B. A. Nosek, "An open investigation of the reproducibility of cancer biology research," Elife, vol. 3, Dec. 2014.
[57] R. A. Klein et al., "Investigating variation in replicability: A 'many labs' replication project," Soc. Psychol., vol. 45, no. 3, p. 142, 2014.
[58] M. S. Hagger et al., "A Multilab Preregistered Replication of the Ego-Depletion Effect," Perspect. Psychol. Sci., vol. 11, no. 4, pp. 546–573, Jul. 2016.
[59] D. Colquhoun, "The reproducibility of research and the misinterpretation of p - values," Royal Society Open Science, vol. 4, no. 12, p. 171085, 2017.
[60] D. J. Benjamin et al., "Redefine statistical significance," Nature Human Behaviour, vol. 2, no. 1, pp. 6–10, Sep. 2017.
[61] J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, and D. A. Reinero, "Contextual sensitivity in scientific reproducibility," Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 23, pp. 6454–6459, Jun. 2016.
[62] A. Dreber et al., "Using prediction markets to estimate the reproducibility of scientific research," Proc. Natl. Acad. Sci. U. S. A., vol. 112, no. 50, pp. 15343–15347, Dec. 2015.
[63] C. Collberg, T. Proebsting, and A. M. Warren, "Repeatability and Benefaction in Computer Systems Research - A Study and a Modest Proposal." 27-Feb-2015.
[64] "The XPhi Replicability Project." [Online]. Available: https://sites.google.com/site/thexphireplicabilityproject/home. [Accessed: 01-May-2018].
[65] S. Hochreiter, J. urgen Schmidhuber, and C. Elvezia, "LONG SHORT-TERM MEMORY," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[66] T. Mikolov, M. Karafiát, L. Burget, J. Černock\`y, and S. Khudanpur, "Recurrent neural network based language model," in Eleventh Annual Conference of the International Speech Communication Association, 2010.
[67] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in Encyclopedia of Database Systems, Springer, Boston, MA, 2009, pp. 532–538.

HONORS THESIS ABSTRACT
THESIS SUBMISSION FORM

AUTHOR: Joseph McDade

THESIS TITLE: Can We Predict Reproducible Scholarly Research?

ADVISOR: Daniel Rogness

ADVISOR'S DEPARTMENT: Computer Science

DISCIPLINE: Computer Science YEAR: 2018

PAGE LENGTH: 10

BIBLIOGRAPHY: Yes

ILLUSTRATED:

PUBLISHED (YES OR NO): No

LIST PUBLICATION:

COPIES AVAILABLE (HARD COPY, MICROFILM, DISKETTE):

ABSTRACT (100-200 WORDS): Yes