COMP 551 Applied Machine Learning Project Report: Reproducibility and Replicability of Empirical Results in Machine Learning

Gavin McCracken, Nikesh Muthukrishnan, Yuanhang Yang
Email: {firstname}.{lastname}@mail.mcgill.ca

Abstract—**In this report, we demonstrate our work in surveying the issue of replicability and reproducibility, both in general science and machine learning. We provide various proposed definitions of the terms replicability and reproducibility, and present a comparison across them. We survey the history of replicability and reproducibility issues in scientific works, and researchers' approaches to them. Finally, we suggest a few Do's and Don'ts for the machine learning community to improve replicability and reproducibility in research works.**

I. Introduction

Our approach to studying replicability and reproducibility is threefold: we begin this report by reminding the readers of the definition of scientific methods, and introduce the various definitions of the terms *reproducibility* and *replicability*. We proceed to linking the two terms to the scientific method, and show their importance. In the second part, we review the history of replicability and reproducibility, including an introduction to the replicability crisis event and scientist's work on this issue. Finally, we provide a few practices that various researchers have suggested in order to improve the replicability and reproducibility of scientific works.

Throughout the scope of this project, our goal is to answer the following questions: **What are replicability and reproducibility, why are they key to scientific progression in all fields, and how does one ensure publications in the field of machine learning are replicable and reproducible?**

II. The Importance of Replicability and Reproducibility in Science

To define replicability and reproducibility, one must look to the scientific method, in which reproducibility and replicability are key, as well as some historical examples.

The editors of the *Biostatistics* journal define replicability and reproducibility as follows [29]. They claim that one must first understand the difference between a reproducible result and replicable result. They go on to say that the reproduction of a scientific finding means that starting with the dataset the publishing scientist gathered, and using the published parameters, you can reproduce the same results, p-values, confidence intervals, tables and figures as those reported by the scientist [29]. In other words, the results are reproducible. However, the replication of a scientific finding means that you can replicate the exact study, using the same apparatus, same source code, same parameters, same dataset and etc to replicate the results [29]. They also address the issue that in some scientific fields, the investigations may be extremely difficult to replicate due to lack of time or resources and state that there must be a minimum standard between replication and nothing. They suggest "reproducible research", as an acceptable candidate, which they claim "requires that datasets and computer code be made available to others for verifying published results and conducting alternative analyses" [29]. As one could imagine, in a field like machine learning where datasets can require millions upon millions of computations to train a model, there are indeed difficulties with "time and resources". This stems from the fact that massive datasets require thousands of computational units to train a complex model on; something that most laboratories do not have access to.

The editors of the *Nature* journal, are making a similar push. Stating that "papers in Nature journals should make computer code accessible where possible." and that a key part of the replicability and reproducibility of their research papers is that components of publications should be available to peers who wish to validate the techniques and results" [3]. They go on to say that a key element of many papers is the source code used by authors in models, simulations and data analysis and that in an ideal world, it would always be made publicly available. They do however acknowledge that in some cases this is not possible, as proper publication of the code would require many hours, as well as extra funding to perform proper segregation, encapsulation and rendering of the source code into a shareable output [3]. Naturally, this concept of making components of publications publicly available applies to both datasets and source code used in machine learning publications.

While the importance of reproducibility may or may not be obvious at first glance, we can find examples even in modern times in which entire bodies of knowledge would have been proven wrong if not for a lack of reproducibility. Somewhat recent high-profile cases include the 1998 vaccination study by Andrew Wakefield, the OPERA experimental results from 2011, and a study on cold fusion in 1989 by Fleischmann and Stanley Pons.

The anti-vaccination movement commonly cites a fraudulent study from 1998 by Andrew Wakefield that was published in The Lancet which has since been retracted [42]. This study found that there was a link between autism and the measles, mumps and rubella (MMR) vaccination. Many large epidemiological studies were undertaken to test its replicability and reproducibility, and it was deemed fraudulent because it was both irreproducible, and irreplicable. One review on the MMR vaccinations safety by the Cochrane Library claimed, and correctly, that Wakefields study had damaged public health [14]. As a result, Andrew Wakefield was found guilty of scientific misconduct and struck off the United Kingdom's medical register. However, unfortunately the retraction of this publication, the large number of papers disagreeing with it's finding, and the staunch rejection of it in the entire medical community, have not been enough to persuade the entire public of vaccine safety. The damage of this publication is still being realized today.

The OPERA experiment made headlines by claiming they measured neutrinos travelling faster than the speed of light. Many physicists tried to replicate and reproduce the experiment, and all of them found that neutrinos did not travel faster than the speed of light. Eventually, the OPERA team discovered faults in their scientific methodology. They had a clock ticking too fast, and an improperly attached fibre optic cable. Their publication has since been revised to state that after accounting for these methodology errors, their findings are in agreement with the fact that neutrinos do not travel faster than the speed of light [10].

Cold fusion is a hypothesized type of nuclear reaction that would occur near room temperature. In 1989 Martin Fleischmann and Stanley Pons reported that they were measuring small amounts of nuclear reaction byproducts, including neutrons and tritium in an experiment designed to electrochemically induce the fusion of deuterium [16]. The experiment was attempted to be replicated around the world, and while some findings initially showed favour, they were later retracted, and the physical chemistry community eventually formed the theories that currently exist today, under which there is currently no accepted theoretical model that would allow cold fusion to occur. Martin Fleischmann and Stanley Pons never retracted their claims however.

The progression of science as a whole can clearly be seen to depend on replicability. According to Goldhaber and Nieto in 2010 [17], the scientific method is a group of techniques to be applied when investigating phenomenona, acquiring new knowledge or correcting or integrating previous knowledge. The Oxford Online Dictionaries state that the scientific method is: ”A method of procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses” [1]. It follows from this definition that reproducibility and replicability are key to the correct function of the scientific method. Theodore Garland Jr. from University of California Riverside claims that the strongest tests of hypothesis are those from ”carefully controlled and replicated experiments”. He goes on to claim that ”whatever methods are used to test a hypothesis, in an ideal world they should be replicated at least once” [2].

While the scientific community holds high-profile studies, or those with far-reaching implications, to a higher standard with regards to reproducibility, this is not necessarily the case for smaller studies. This has led to a replicability crisis.

III. A replicability crisis

A. The replicability crisis

In his 1963 book *Little science, big science... and beyond* (which was reproduced in [13]), historian of science Derek de Solla Price presented the idea that science may soon encounter an inevitable saturation stage, where growth of the field halts. Price first noted that science is a fast-growing ”modern and contemporaneous” field, as most of the scientific discoveries in the world at the time were made in the most recent decades. According to Price's own work [31], where he used numbers of scientific journals as a benchmark index, the data indicated a growth rate of about 5.6% per year and a doubling time of 13 years. Almost half a century later, a research [23] published on *Scientometrics* stepped up and said that there has been no indicators that that rate had decreased in the past 50 years at all. In fact, some research deducted that by the year of 2009, the scientific world had already welcomed its 50-millionth published journal article, quite a number since journals first appeared in 1665 [19]. Unfortunately, the rapid growth of science may have contributed to a broader acceptance of publications that do not have such great reproducibility or replicability. A ”publish or die” culture has been created, with a constant pressure on scientists to publish [6] [4]. Relating to this, according to Begley and Ioannidis [6], who believed that the majority of the discoveries made while new data and scientific publications are produced at this unprecedented rate will ”not stand the test of time”, when researchers don't follow scientifically rigorous practices.

The series of events that sparked the so-called replicability crisis started in 2011, in the field of psychology, when social psychologist Diederik Stapel admitted to large-scale research fraud after being accused of scientific misconduct. Researchers reacted to this ”terrible shock” by studying similar cases in psychological researches and suggesting ways to reduce risk of fraud [40]. But this case, although high-profile, was only one of the many events that would bring psychology under the spotlight repeatedly. Some research that was conducted to investigate psychologists' unwillingness to share research data led to the conclusion that researchers are less likely to share data when reanalysis of data would lead to contrasting conclusions, and that mandatory data archiving policies are important [43]. Another study more unambiguously claimed that psychologists can practically present any hypothesis with statistically significant result, through various corner-cutting practices [37]. Events like these triggered a high-profile study on the Questionable Research Practices (QRPs) among psychological researchers; this study used anonymous questionnaires to survey over 2000 psychologists, and found that the percentage of the survey participants who have engaged in QRPs were surprisingly high... This finding suggests that some questionable practices may constitute the prevailing research norm” [20].

This series of events finally led to the ”Reproducibility Project: Psychology” by Open Science Collaboration. In this project, 100 experimental and correlational studies published in three psychology journals were selected and their experiments redone. Five metrics were used to measure reproduction results: significance, P values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The report from the Open Science Collaboration [9], although written professionally with neutral choice of words, arrived at "a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes." We find the wording of this conclusion quite restrained, given that on almost every measurement discussed, the strength of published conclusions backed by replicated experiments significantly declined, compared to original experiment results. Sometimes the results from this study is referred to as the "replicability crisis" itself.

Not really to psychologists' or other scientists' pleasure, similar events were being found and dissected by similar studies in the general field of science. Science misconducts were being observed from time to time, in various fields. Some examples range from an investigation on the science misconduct of a former Bell Lab physicist [34], to the so-called "Korean cloning fraud" where a large portion of data in cloning experiments were found fabricated [33]. The science community has been trying to detect how prevalent an issue misconduct is, but there had been disagreements upon this question: while some scientists and historians of science believe there is a disturbingly high percentage of scientists involved in fraudulent researches [25], another point of view is held by others, that sometimes disagreement on the level of scientific rigor can be explained by a difference in scientific misconduct [39].

After the shocking findings of [9] for psychology, it seemed like general science needed a large-scale study to self-diagnose on replicability issues. This time, it was *Nature* who led the survey on replicability of scientific works with 1,576 researchers participating [4]. The result showed that 90% of researchers surveyed agreed that there is currently a replicability crisis to some degree. "More than 70% have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments." Interestingly, despite these seemingly high figures, less than 31% said they would link failure in replicating results to wrong conclusions in original works and would still trust the published literature, a point of view we will discuss shortly in section III-B. This survey also showed, among other things, that respondents all presented a certain level of doubt in the replicability in their fields, but that level of doubt (or confidence) varied by field, with chemistry and physics scoring higher and medicine among the lowers. When asked about potential causes of lack of replicability, both "selective reporting" and "pressure to publish" were chosen by more than 60% of researchers.

B. Researchers' works, in response to the crisis

The 2015 *Nature* survey, together with various other scholars' researches, showed that the replicability issue in the field of science was spreading to a scale and depth that not everyone is prepared for. This called for the science community's collective attention on the issue. As Begley and Ioannidis said in [6], "No single party is solely responsible, and no single solution will suffice".

First of all, there are scientists who aren't fully convinced that science is in the soup. While praising the initiative in [4], psychologist Schwarz advocates *conceptual replications* - where original studies' hypothesis are tested in experiments with tweaked parameters rather than identical to the original ones - should be highly valued [8]. He believes this is actually a better measure of the scientific robustness of the original study as it examines the stability of a result across different content domains. Machine learning scholar Drummond advocates the same concept under a different brand name, *reproducibility*, even more, saying reproducibility is actually the only valuable practice in comparison to fully identical replications [15]. He believes that identical replications of original experiments actually cannot establish the underlying mechanism, as well as tweaked reproduced experiments.

But there are scientists who do not share this optimistic point of view. Nosek, who led the study in [4], believes that although common and valuable, conceptual replications cannot be a substitute to full replication in psychology, because conceptual replications makes the assumption that the original and modified experiments address the same underlying phenomenon, but this is not always the case [27]. In the field of computational science, there are scholars who agree with Nosek. Peng [30] actually maintains that what he calls *reproducibility* is the potential minimum threshold for judging scientific claims. What he means by reproducibility, "calls for both the data and the computer code used to analyze the data to be made available to others", so that other scientists can re-do exactly the original experiment; this aligns with what Nosek would believe is necessary practice in psychology. Pashler and Harris [28] went further, surveying and disagreeing with three arguments that the replicability crisis is overblown. They argued that conceptual replication would introduce undesirably biased results, and that low alpha level in hypothesis testing is not an excuse for replication errors, and that erroneous literature cannot like some say be pruned out automatically over time, but should be addressed with systematic reform of scientific practices.

It may be worthwhile to present another kind of optimistic point of view: that failure to replicate studies may be a good thing sometimes. Very recently, Leyser [24] pointed out in a short article published on University of Cambridge's website that, sometimes when two supposedly identical experiments produce different results, it can lead to the discovery of previously unknown differences, and potentially exciting new findings. She raised the example of the paper in [26], where the exciting phenomenon of *reversible co-suppression of homologous genes in transgenotes* were discovered in an unexpected way. We the authors of this report can also add to this point of view with an example from the Artificial Intelligence domain: researchers noticed that although Upper Confidence bounds for Trees (UCT) has notable success in the game of Go, it appeared difficult to reproduce the level of performance in minimax-search-dominated games. This led to Ramanujan and Selman's finding [32] that, in partial game settings, decisions made by Monte Carlo Tree Search family algorithms can be superior to those of minimax search in regions of the search space with no or few terminal nodes, such as Mancala games where there is only one terminal state: a board with all empty pits.

In this time where the awareness of replicability is being refreshed, no matter which definition of replicability they used, scientists are definitely working their ways to make science stronger with this opportunity. Begley and Ioannidis [6] proposed several recommended practices to improve the quality control of preclinical research works, including promoting more standardized research practices and data disclosure. More such recommendations, especially those pertinent to the field of machine learning, are discussed in section IV.

We conclude this section with the observation that replicability issues are indeed getting more and more attention in the scientific community. Taking the field of medicine as an example, when querying the keywords *replicability* and *reproducibility* on PubMed database using [11], we plotted the total number and percentage of PubMed papers that contain the keywords *replicability* and *reproducibility*. The plots 1 and 2 serve as "exhibit A" that the awareness have clearly been raised on these issues, and we believe that the machine learning community could also take this opportunity to make our field stronger.

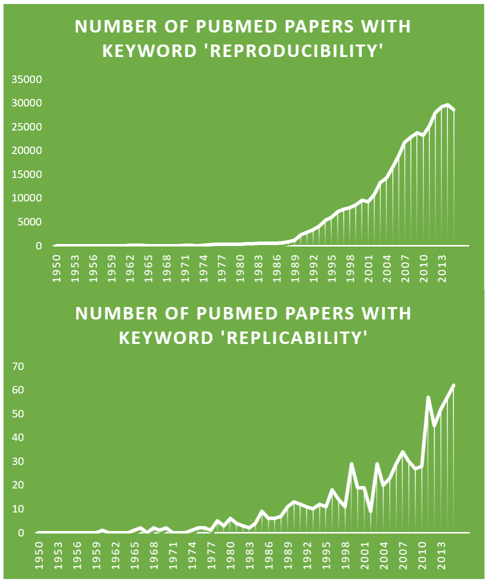


Fig. 1. Number of PubMed papers with keywords 'replicability' and 'reproducibility', from 1950 to 2015.

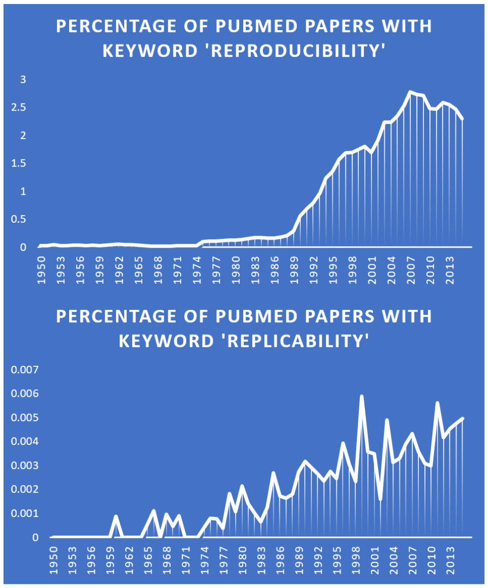


Fig. 2. Percentage of PubMed papers with keywords 'replicability' and 'reproducibility', from 1950 to 2015.

IV. Recommended practices

There are several factors that can harm or aid replicability of an experiment and these are explored in this section.

A. Size of Dataset

The study, "Low Replicability of Machine Learning Experiments is not a Small Data Set Phenomenon," conducted by Remco Bouckaert [7], demonstrates the effect of data set size on statistical replication of an experiment. In general, testing an experiment on a small data set can lead to biased results, and thus low replicability is expected. However, from the Bouckaert study, it was found that a large dataset in certain circumstances can also decrease replicability. Bouckaert's study consists of comparing two learning algorithms, A and B, and training them on a dataset D, and observing the results from a statistical point of view. The size of the dataset D was varied for different experiments along with various allocations of D for the training set and test set. Then hypothesis testing was performed on the experiment, where the null hypothesis indicated that A and B have the same performance.

The major definitions Bouckaert used for his study are Type 1 Replicability and Power of Test. Type 1 Replicability is defined as the "probability that when performing the same experiment twice ... the same outcome" is achieved and the null hypothesis holds. Power of Test is defined as the probability of correctly concluding there is a difference in both algorithms.

To study the effect of dataset sizes as well as various train/test splits, the algorithms A and B were selected as simple learning algorithms. The input for the experiment is $x \in 0, 1$ and the output is $y \in 0, 1$. A always predicts 1 and B always assign the value of x to y. Based on the nature of the problem and the algorithms A and B, the difference between the observed results achieved can differ by a maximum of 50%.

Figure 3 illustrate the results of Bouckaert's study. The x-axis indicates the difference between the observed results from A and B (denoted as Δ) and the y-axis indicates the percentage of both the Type 1 Replicability (plot with local minimum) and the Power of Test (monotonically increasing plot). The size of the dataset is specified by the plot colour (red for 1000, green for 2000, and blue for 10000). Figure 3 illustrates the results for a 90%/10% train/test split.

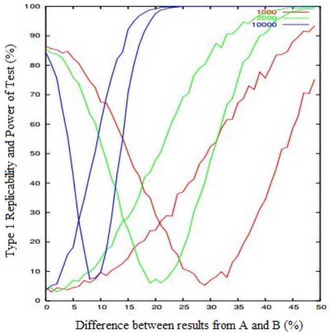


Fig. 3. Bouckaert's Results for 90%/10% train/test split. Image from [7].

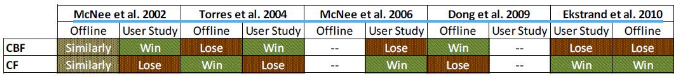
The major takeaways from his experiment are the following:

- Independent of the train/test split, the Type 1 Replicability of an experiment when Δ is zero (identical results are obtained from A and B) decreases with larger data sizes. In Figure 3, the Type 1 Replicability decreases from 86.5% to 85.6% to 84.4% with increasing data size.
- The worst case replicability, indicated by the local minimum of the Type 1 Replicability plot, increases with larger test sizes. This is due to both algorithm containing overlapping test data when more data is partitioned for the test set.
- Although resampling and k-fold cross-validation are computationally expensive, they increase Type 1 Replicability as well as the worst case replicability. Performing resampling and k-fold cross-validation multiple times increases Type 1 replicability.

B. Butterfly Effect

A robust algorithm is generally expected to be replicable, despite minor changes (slight variations in computer hardware, or slight difference variable initialization). On the contrary, for significant changes, the results achieved by algorithm can be excepted to be non-replicable. In some situations, although only minor changes are present, the results are not similar. This phenomenon is referred to as the "butterfly effect".

The "butterfly effect" has been documented in many fields, but in the machine learning world, the study by TechLens on recommender systems is one of the more prominent ones [5]. TechLens analyzed two major research paper recommendation approaches: Content-based filtering (CBF) and collaborative filtering (CF). Over several years and a series of online and user population studies, various researchers conducted studies to determine which approach is the better approach. The results found from these researches fluctuated over the years. In some studies, the CBF approach proved to outperform the CF approach and in other studies, the CF approach proved to perform better than the CF approach. Figure 4 illustrates the results of TechLens study.



	McNee et al. 2002		Torres et al. 2004		McNee et al. 2006		Dong et al. 2009		Ekstrand et al. 2010	
	Offline	User Study	Offline	User Study	Offline	User Study	Offline	User Study	User Study	Offline
CBF	Similarly	Win	Lose	Win	--	Lose	Win	--	Lose	Lose
CF	Similarly	Lose	Win	Lose	--	Win	Lose	--	Win	Win

Fig. 4. Results of TechLens CBF and CF evaluations. Image from [5].

In the case of the TechLens study, the authors acknowledge the variations in the reproducibility of the algorithms by providing several potential reasons. The variations can be due to different dataset, different user populations, or slight changes in the implementation of each algorithm. However, Beel, Langer and Lommatzsh state that these reasons only account for some of the variations, and overall the differences in reproducibility results are unexpected, thus, due to the "butterfly effect" [5].

Butterfly effect could be due to small difference in implementation that unexpectedly caused a wide difference in results. In the research paper recommender systems studies, Beel, Langer and Lommatzsh discovered several small variations that led to dramatic difference in results. These variations includes:

- Layout difference. Recommender systems are generally evaluated based on a user's click-through rate (CTR) on the suggested recommendations. When a more convenient and aesthetically pleasing model is presented with an algorithm, the CTR is generally higher.
- Hour of study. If the algorithm is implemented on a website, the website may have varying web-traffic throughout the day. The study should be performed at an appropriate time, and reproducibility studies should be consistent.
- Test Population. For the research paper recommendation study, student may prefer to read more novice papers and professor may prefer more novel papers. Therefore to ensure reproducibility, the population on which the study is conducted must be considered.

C. Summary of Recommended Practices

As mentioned earlier in II, the replicability of the results of an experiment is crucial to its contribution to the scientific community. In the machine learning community, replicability has not yet been prioritized and there is a lack of criteria for researchers to follow. Currently, methods similar to what Bouckaert mentions: re-sampling and k-fold cross-validation should be performed on the algorithm and that the associated parameters should be well-documented, are the most common guidelines. However, several authors have suggested practices that should be included to ensure better replicability.

Beel, Langer and Lommatzsh suggested several other methods to improve replicability in experiments. A few of the studies in Figure 4, were conducted on private datasets as well as involving or lacking multiple significant features that may have skewed the results. Detailed analysis on the source of the dataset should be included as well as any processing implemented on the dataset for a successful replication of results. Many dataset repositories are available to the public, such as the UCI repository (http://archive.ics.uci.edu/ml/) or the Delve repository (http://www.cs.toronto.edu/ delve/data/datasets.html), and should be used whenever possible for better replication of results. Furthermore, algorithms must be reported in detail by the original research study. There may be variable that are present in the algorithm that are not properly represented in the replication study. Incorporating pseudo-code or open-source software would be beneficial in replicability studies. Also, the values of all hyper-parameters in the study should be included. Finally, the minor details including the layout, hour and test population study should be detailed in the report to avoid any butterfly effects.

Many other researchers have noticed other factors that affected the replicability of their experiments. In a study by Shepperd, Bowes and Hall, they found that researcher bias is a largely influential factor on the variance of replicability and reproducibility of results [35]. They suggested several improvements for replication studies. First, the study should be conducted in group. This slightly alleviates the issue that a researcher's expertise in a field or perspective may influence the results of an experiment. Collaborative work also allows researchers to share their resources and knowledge. In addition, Shepperd, Bowes and Hall emphasize on the importance of documentation and better communication. They describe the lack of documentation in several experiments as the "unwritten setup", and mention that the "unwritten setup" is just as important as the documented setup. Finally, the shared that researchers are highly likely to be biased when the results are known to them. Knowing the results beforehand can often lead to a placebo effect in certain areas of science and is applicable to machine learning as well. They discuss the importance of a blind analysis when attempting to replicate the results of a study.

In a study by Kononenko in "Machine learning for medical diagnosis: history, state of the art and perspective", he describes the replicability problems with his experiment [22]. His algorithm contains a large number of parameters and as a consequence replicating his study has become a very difficult process. In his algorithm, controlling such a large set of variables is described as nearly impossible. Future studies should properly document all required parameters to aid in replication studies.

In Cruz and Wishart paper, "Application of Machine Learning in Cancer Prediction and Prognosis" the importance of feature selection is explored [12]. They describe that many features are favourable and helpful in cancer prediction, however some features such as "site codes", are hospital specific and can cause poor replication if the algorithm is used with a different hospital. Without proper documentation on these such features, other hospitals that attempt to replicate the results and use for their own purpose will not be successful. Furthermore, other values pertinent to some features may be assessed from a pathologist or health care professional. These features are highly inconsistent as different pathologists may assess differently. They proclaim that for better replicability and reproducibility of results, features that are consistent and universal such as age, gender, and biomarker measurement should be used whenever possible. Furthermore, they describe the importance of size of dataset and validation set in replicability, by highlighting issues faced by another study. Models trained on smaller datasets are prone to overtraining. In addition, in a study by Hamamoto et al., their algorithm was able to predict the survival of hepatectomized patients on a test set with 100% accuracy, however, the size of their test set was only 11 patients [18]. A more robust verification should be performed on a larger test set. Cruz and Wishart suggest that there should be at least 5-10 samples available for each feature used in the algorithm.

Sonnenburg et al.discuss the importance of open source in their study "The Need for Open Source Software in Machine Learning" [38]. They highlight that sharing software not only aids in the replicability of experimental results but also allows for quicker validation by granting other researchers the opportunity to detect errors in the proposed algorithm. Furthermore, open source sharing also aids in the advancement of the particular field associated to the proposed algorithm, by allowing other researchers to build upon the proposed algorithm as well as combine it into their own research.

Some machine learning replication studies may be inhibited by hardware limitations. AlphaGo is a machine learning program by Google, designed to play the board game Go [36]. In order to optimize running the complex algorithm on a massive quantity of data, Google designed and built a custom chip called the Tensor Processing Unit (TPU) [21]. Although other researchers may be interested in replicating results achieved by AlphaGo, it would not be possible without the hundreds of TPU which run in parallel. Therefore at the moment, only Google can replicate the results of AlphaGo. In cases such as these, the scientific community must rely on the scientific integrity of Google's publications. Google, or any scientist or laboratory that finds themselves in this position, should strive to ensure replicability themselves, and discuss it in their publication.

Finally, the last recommendation for better replicability in machine learning experiments is suggested by Sun et al, in their study "Application of Machine Learning to Predict Coronary Artery Calcification With a Sibship-Based Design" [41]. Sun et al. suggested that for better replicability of an algorithm, it is important to evaluate the performance of an algorithm on multiple datasets. They implement their own algorithm on multiple datasets and reported the results such that future replication study can validate their replication over multiple datasets as well, instead of on a single dataset.

V. Conclusion

It seems to be the mutual conclusion of many researchers and scientific journals that for replicability to be maintained in computer science, all components of publications should be available to peers. For the case of machine learning and computer science, this would mean making both source code, and the exact dataset used, as well as any other materials required by a model or simulation, publicly available. Doing this also promotes reproducibility because it's easy for scientists to make minor tweaks to the published algorithm, dataset, hyper-parameters or methodology. Furthermore, it is absolutely essential for this methodology to explain itself in detail. It must not be vague. Additionally, as demonstrated, variations that may not present themselves as influential may result in large reproducibility errors in the form of a butterfly effect. These must be addressed, and the hypothesis revised accordingly. Likewise, if the exact dataset is not available for disclosure, the experiment should be performed on an adequate size dataset and documentation should be provided on the relevant features in the dataset. Each algorithm specific feature e.g. hospital codes, should be universal or easily translatable for purposes of replication and reproduction. Finally, in the event that training a model requires thousands of computational units, the publishing researchers should take it upon themselves to replicate their own experiment, and talk about it's own replicability.

VI. Statement of Contributions

1) McCracken: Surveying literature for a definition of replicability in machine learning and the current state of replicability w.r.t. machine learning. Report writing.
2) Muthukrishnan: Surveying literature on the recommended practices for replicability. Report writing.
3) Yang: Surveying literature on the history of replicability and reproducibility issues. Report writing.

We hereby state that all the work presented in this report is that of the authors.

References

[1] Definition of scientific method oxford online. https://goo.gl/dBX3Ad. Accessed: 2017-04-10.

[2] The scientific method as an ongoing process. https://goo.gl/Mo5cPn. Accessed: 2017-04-10.

[3] Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 514(7524):536, 2014.

[4] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[5] Joeran Beel, Stefan Langer, and Andreas Lommatzsch. Exploring the butterfly effect and (non-) reproducibility in recommender systems research. 2014.

[6] C Glenn Begley and John PA Ioannidis. Reproducibility in science. *Circulation research*, 116(1):116–126, 2015.

[7] Remco Bouckaert. Low replicability of machine learning experiments is not a small data set phenomenon. 2005.

[8] Siri Carpenter. Psychology's bold initiative. *Science*, 335(6076):1558–1561, 2012.

[9] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[10] The OPERA Collaboration, T. Adam, N. Agafonova, A. Aleksandrov, O. Altinok, P. Alvarez Sanchez, A. Anokhina, S. Aoki, and A. Ariga et al. Measurement of the neutrino velocity with the opera detector in the cngs beam. 2011.

[11] Alexandru Dan Corlan. Medline trend: automated yearly statistics of pubmed results for any query, 2004.

[12] Joseph Cruz and David Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*, (2):59–77, 2007.

[13] Derek John de Solla Price. *Little science, big science... and beyond*. Columbia University Press New York, 1986.

[14] Rivetti Alessandro Debalini Maria Grazia Demicheli, Vittorio and Carlo Di Pietrantonj. Vaccines for measles, mumps and rubella in children. *The Cochrane Library*, (2), 2012.

[15] Chris Drummond. Replicability is not reproducibility: nor is it good science. 2009.

[16] Martin Fleischmann. Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 261(2), 1989.

[17] Alfred Scharff Goldhaber and Michael Martin Nieto. Photon and graviton mass limits. *Rev. Mod. Phys.*, 82:939–979, Mar 2010.

[18] Okada S Hamamoto I and T Hashimoto. Prediction of the early prognosis of heptectomized patien with hepatocellular carcinoma with a neural network. *Comput Biol Med*, (25):49–59, 1995.

[19] Arif E Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

[20] Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, page 0956797611430953, 2012.

[21] Norm Jouppi. Google supercharges machine learning tasks with tpu custom chip, 2016.

[22] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[23] Peder Olesen Larsen and Markus Von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603, 2010.

[24] Ottoline Leyser. Opinion: The science 'reproducibility crisis' – and what can be done about it, Mar 2017.

[25] Eliot Marshall. How prevalent is fraud? that's a million-dollar question. *Science*, 290(5497):1662–1663, 2000.

[26] Carolyn Napoli, Christine Lemieux, and Richard Jorgensen. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *The plant cell*, 2(4):279–289, 1990.

[27] Brian A Nosek, Jeffrey R Spies, and Matt Motyl. Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012.

[28] Harold Pashler and Christine R Harris. Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, 7(6):531–536, 2012.

[29] Roger D. Peng. Reproducible research and biostatistics. *Biostatistics*, 10(3):405, 2009.

[30] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[31] Derek J Price. Science since babylon. 1961.

[32] Raghuram Ramanujan and Bart Selman. Trade-offs in sampling-based adversarial planning. In *ICAPS*, pages 202–209, 2011.

[33] R Saunders and Julian Savulescu. Research ethics and lessons from hwanggate: what can we learn from

the korean cloning fraud? *Journal of Medical Ethics*, 34(3):214–221, 2008.

[34] Robert F Service. Scientific misconduct. more of bell labs physicist's papers retracted. *Science (New York, NY)*, 299(5603):31, 2003.

[35] Martin Shepperd, David Bowes, and Tracy Hall. Researcher bias: The use of machine learning in software defect prediction. *IEEE Transactions on Software Engineering*, 40(6):603–616, 2014.

[36] David Silver and Demis Hassabis. Alphago: Mastering the ancient game of go with machine learning, 2016.

[37] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[38] Sören Sonnenburg, Mikio Braun, Cheng Soon Ong, Samy Bengio, Leon Bottau, Geoffrey Holmes, Yann LeCun, Klaus-Robert Müller, Fernando Pereira, Carl Edward Rasmussen, Gunnar Rätsh, Bernhard Schölkopf, Alexander Smola, Pascal Vincent, Jason Weston, and Robert Williamson. The need for open source software in machine learning. *Journal of Machine Learning Research*, (8):2443–2466, 2007.

[39] Benjamin K Sovacool. Exploring scientific misconduct: Isolated individuals, impure institutions, or an inevitable idiom of modern science? *Journal of Bioethical Inquiry*, 5(4):271, 2008.

[40] Wolfgang Stroebe, Tom Postmes, and Russell Spears. Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7(6):670–688, 2012.

[41] Yan V. Sun, Lawrence F. Bielak, Patricia A. Peyser, Stephen T. Turner, Patrick F. Sheedy, Eric Boerwinkle, and Sharon L.R. Kardia. Application of machine learning algorithms to predict coronary artery calcification with a sibship-based design. *Genet Epidemiol*, 32(4):350–360, 2008.

[42] Andrew Wakefield. Retraction–ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 375(9713):445, 1998.

[43] Jelte M Wicherts, Marjan Bakker, and Dylan Molenaar. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one*, 6(11):e26828, 2011.