The State of Reproducible Research in Computer Science

Jorge Ramón Fonseca Cacho
Department of Computer Science
University of Nevada, Las Vegas
Email: Jorge.FonsecaCacho@unlv.edu

Kazem Taghva
Department of Computer Science
University of Nevada, Las Vegas
Email: kazem.taghva@unlv.edu

***Abstract*—Reproducible research is the cornerstone of cumulative science and yet is one of the most serious crisis that we face today in all fields. This paper aims to describe the ongoing reproducible research crisis along with counter-arguments of whether it really is a crisis, suggest solutions to problems limiting reproducible research along with the tools to implement such solutions by covering the latest publications involving reproducible research.**

***Index Terms*—Docker, Improving Transparency, OCR, Open Science, Replicability, Reproducibility**

I. Introduction

Reproducible Research in all sciences is critical to the advancement of knowledge. It is what enables a researcher to build upon, or refute, previous research allowing the field to act as a collective of knowledge rather than as tiny uncommunicated clusters. In Computer Science, the deterministic nature of some of the work along with the lack of a laboratory setting that other Sciences may involve should not only make reproducible research easier, but necessary to ensure fidelity when replicating research.

> **"Non-reproducible single occurrences are of no significance to science."**
>
> **— Karl Popper** [1]

It appears that everyone loves to read papers that are easily reproducible when trying to understand a complicated subject, but simultaneously hate making their own research easily reproducible. The reasons for this vary from fear of poor coding critiques to outright laziness of the work involved in making code portable and easier to understand, to eagerness to move on to the next project. While it is true that hand-holding should not be necessary as one expects other scientist to have a similar level of knowledge in a field, there is a difference between avoiding explaining basic knowledge and not explaining new material at all.

II. Understanding Reproducible Research

Reproducible Research starts at the review process when a paper is being considered for publication. This traditional peer review process does not necessarily mean the research is easily reproducible, but is at minimum credible and shows coherency. Unfortunately not all publications maintain the same standards when it comes to the peer review process. Roger D. Peng, one of the most known advocates for reproducible research, explains that requiring a reproducibility test at the peer review stage has helped the computational sciences field publish more quality research. He further states that reproducible data is far more cited and of use to the scientific community [2].

As define by Peng and other well-known authors in the field, reproducible research is research where, "Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results" [3]. On the other hand, Replication is "A study that arrives at the same scientific findings as another study, collecting new data (possibly with different methods) and completing new analyses" [3]. Barba compiled these definitions after looking at the history of the term used throughout the years and different fields in science. It is important to differentiate the meanings of reproducible research and Replication because both involve different challenges and both have proponents and opponents in believing that there is a reproducible crisis.

III. Collection of Challenges

Reproducible research is not an individual problem with an individual solution. It is a collection of problems that must be tackled individually and collectively to increase the amount of research that is reproducible. Each challenge varies in difficulty depending on the research. Take Hardware for example, sometimes a simple a budgetary concern with hardware used for a resource intensive experiment such as genome sequencing can be the limiting factor in reproducing someone else's research. On the other side, it could be a hardware compatibility issue where the experiment was ran on ancient hardware that no longer exists and cannot be run on modern hardware without major modifications.

As mentioned in our past research some of the main difficulties when trying to reproduce research in computational sciences include "missing raw or original data, a lack of tidied up version of the data, no source code available, or lacking

the software to run the experiment. Furthermore, even when we have all these tools available, we found it was not a trivial task to replicate the research due to lack of documentation and deprecated dependencies" [4].

Another challenge in reproducible research is the lack of proper data analysis. This problem is two-folded. Data Analysis is critical when trying to publish data that will be useful in reproducing research by organizing it correctly and publishing all steps of the data processing. Data Analysis is also critical to avoid unintentional bias in research. This is mostly due to a lack of proper training in data analysis or lack of using correct statistical software that has been shown to improve reproducibility [5].

IV. Statistics: Reproducible Crisis

Many have gone to say that reproducible research is the greatest crisis in science today. Nature published a survey where 1,576 researchers where asked if there is a reproducible crisis across different fields and 90% said there is either a slight crisis(38%) or a significant crisis(52%) [6]. Baker and Penny then asked follow up questions regarding what contributed to the problem and found Selective Reporting, Pressure to Publish on a deadline, poor analysis of results, insufficient oversight, and unavailable code or experimental methods were the top problems; however, the surveyed people do also mention that they are taking action to improve reproducibility in their research [6].

We ran a similar survey at the University of Nevada, Las Vegas, but only targeted the Graduate Students since we wanted to know how the researchers and professors of tomorrow are taking reproducible research into consideration. The survey involved three main questions, and two additional questions based on the response to the third question, the tables in this paper represent the results.

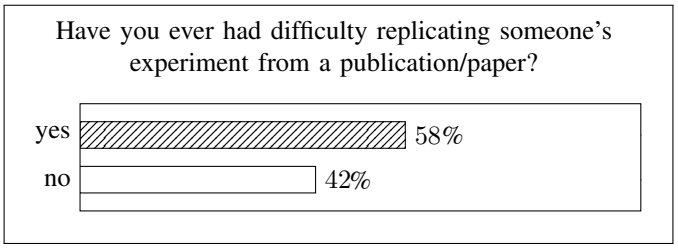

Fig. 1: First Survey Question.

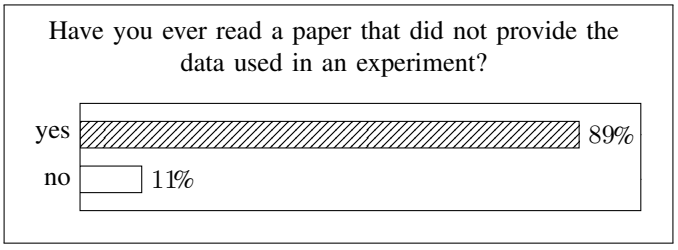

Fig. 2: Second Survey Question.

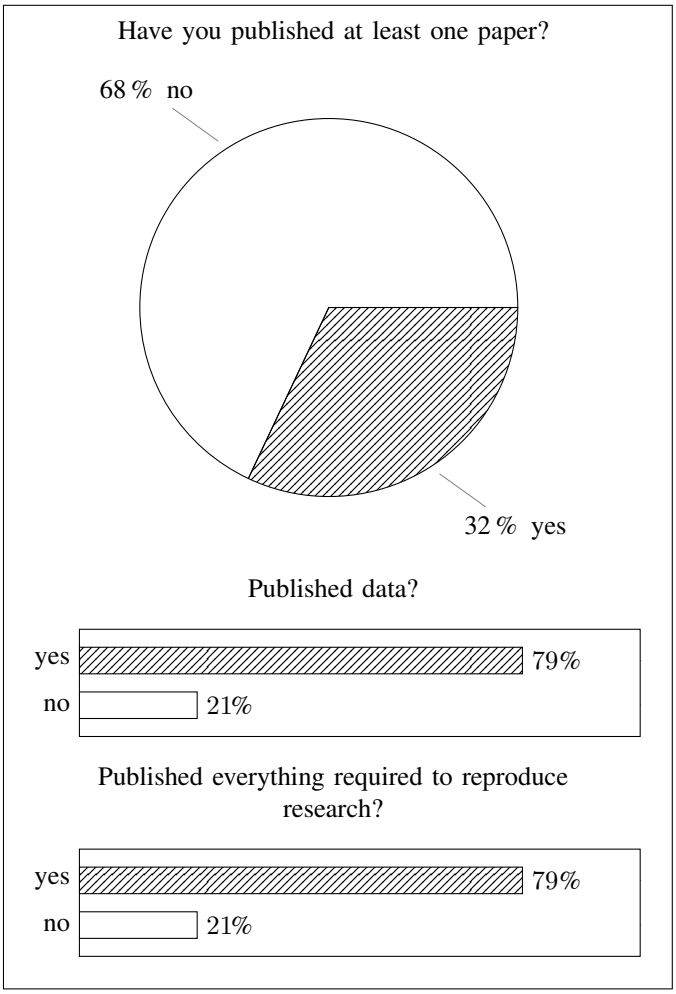

Fig. 3: Third Survey Question and follow up questions if graduate student answered yes.

The survey is in line with what other surveys on similar subject have concluded, such as Baker's survey [6]. Baker has published what he calls the "raw data" spreadsheet of his survey for everyone to scrutinize, analyze, and potentially use for future research. This is not always the case, as Rampin et al. mention, the majority of researchers are forced to "rely on tables, figures, and plots included in papers to get an idea of the research results" [7] due to the data not being publicly available. Even when the data is available, sometimes what researchers provide is either just the raw data or the tidy data [8]. Tidy data is the cleaned up version of the raw data that has been processed to make it more readable by being organized and potentially other anomalies or extra information has been removed. Tidy data is what one can then use to run their experiments or machine learning algorithms on. One could say that the data Baker published is the tidy data rather than the actual raw data. When looking at the given spreadsheet we can see that each recorded response has a `responseid`. Without any given explanation for this raw data, one could conclude that invalid, incomplete, or

otherwise unacceptable responses were removed from the data set as the first four `responseid`'s are 24, 27, 36, and 107. What happened to the rest of the responseids between 36 and 107? As one can see, sometimes providing the data without documentation, or providing just the tidy data, can complicate reproducing the experiment. Furthermore, if only raw data is given by a publication then one could be completely be lost on how to process such data into something usable. Take our survey for example, it has nice looking bar and a pie charts, but did the reader stop to question who was considered to be a valid 'researcher' among the graduate student surveyed? As one can see two-thirds of the graduate students questioned have not published a paper. This could be because they are newer students or because they are not researchers and are doing graduate degrees that do not necessarily require reading or writing publications. So is our survey flawed? Only by looking at the data would one be able to deduce this as the charts could be cherry-picked to show a bias from the original data such. Similarly, the questions could be loaded to encourage specific answers reinforcing the author's hypothesis. The same questions can be asked about Baker's survey on who was considered a researcher. The data and questionnaire indicate that many questions regarding the person taking the survey were asked, most likely to solve this problem, but the threshold they used to remove responses they did not consider came from a researcher is not provided with the "raw data". Ultimately, there is more to a reproducible research standard than just asking the researchers to provide their data. An explanation is necessary along all intermediate steps of processing the data, but how many researchers would really take the time to explain this when they could be working on the next breakthrough? After all, as we mentioned, a line has to be drawn between hand-holding and giving necessary explanations.

Well-known scientist are not exempt from reproducible research. When the University of Montreal tried to compare their new speech recognition algorithm with the benchmark algorithm in their field from a well-known scientist, they failed to do so due to lack of source code. Then when they tried to recreate the source code from the published description, they could not get the claimed performance that made it leading edge in the first place. [9]. Because machine learning algorithms rely on training data, not having the ideal data can greatly influence the performance of said algorithm. This makes both the data and the source code as important to have when reproducing an experiment. Unfortunately, people tend to report on the edge cases when they get "really lucky" in one run [9]. After their experience reproducing the benchmark algorithm, Hutson went on to run a survey where they found that, "of 400 artificial intelligence papers presented at major conferences, just 6% included code for the papers' algorithms. Some 30% included test data, whereas 54% included pseudocode, a limited summary of an algorithm" [9].

There is no simple solution to distributing data in experiments due to potential copyright issues or sometimes sheer size of the data used. For example, our Google 1T experiment uses Copyrighted Data that is 20+ Gigabytes, something not so easily hosted even if the data was not copyrighted [10]. Some datasets only allow distribution through the official channels which only allows researchers to link to it, such as the TREC-5 File used in several of our experiments [11], [12]. This forces us to link to it and hope it remains available for as long as our publication is relevant. This can be both good and bad. Having the data public is a move in the right direction, and having it hosted in only one location can allow for any modifications, or new inclusions, to the dataset, but also increases the risk of losing the dataset if that single host is ever abandoned or shut down.

V. Standardizing Reproducible Research

As Nosek et al. discuss in the article, *Promoting an open research culture*, "In the present reward system, emphasis on innovation may undermine practices that support verification" [13]. He argues that current culture encourages novel results over null results that tend to rarely be published. Publishing good results help move science forward, publishing bad results, such as an algorithm that did not perform as good as hoped, is still important to avoid other researchers attempting the same mistakes. It could be argued that both are just as important.

Nosek et al. also discuss a possible standard with guidelines for reproducible research with different levels where Transparency of the process (data, design and analysis, source code), and even citations standards are maintained. The idea of "Preregistration of studies" is introduced where this could help combat the lack of publishing experiments that produced "statistically insignificant results" [13]. This works by informing the Journal where the study intends to publish its results about the work that is being started, which then will force the researchers to report, regardless of outcome, after some time what happened with that research. This does not necessarily force a researcher to 'pick' a journal before even starting the research since these preregistrations could be transferred between journals as long as the record is kept somewhere. We propose that maybe a repository of ongoing research could be maintained by a non-profit, neutral, organization in order to encourage transparency of all research happening, regardless if the research ends in a novel result, or discover, or a null one where nothing of significance was gained. Reporting failure is just as important as reporting success in order to avoid multiple researches doing the same failed experiment. Similar guidelines can be seen implemented by Academic libraries trying to "lead institutional support for reproducible research" [14]. For example, New York University has appointed a special position in order to bring reproducible research practices into the research stage in order to ensure that practices and techniques are implemented early on to foster publications that are more reproducible [15]. The concept of requiring mandatory data archiving policies is not new and has been shown to greatly improve reproducibility [16], but such a task has also been shown to create disagreement with authors.

VI. Tools to help Reproducible Research

Historically reproducible research began to be a quantifiable concern in the early 1990s at Stanford with the use of Makefiles [17]. CMake's Makefiles is one of the original methods created to help with reproducible research since it made it easier to compile a program that may have otherwise required a compilation guide. Since then other tools have been developed among with other ways to manage source code. There are many popular solutions to source code management and distribution for reproducible research. Among these, one of the most popular ones is the Git Repository system like the popular Github.com [18]. Using Github, one can publish versioned code that anyone can fork and contribute to. Furthermore, it adds transparency to the workflow of the code development.

Even when both, the code and the data, are available, having the same environment [15] is critical to replicate an experiment as both "reproducibility and replicability are fundamental requirements of scientific studies" [19] What this means in applicable terms is the hardware or dependencies surrounding the source code as these can become outdated or deprecated making it very complicated to run old code. Solutions to these exists such as virtual environments, or containers, that can be frozen and easily ran again. Practical applications of these include Vagrant, Docker [20], [21] and the Apache Foundation's Maven and Gradle.

Maven and Gradle are aimed at Java developers where an Ant script (the equivalent of a CMake file, but for Java with a few more features) is not enough. Maven projects contain a POM file that includes documentation to build code, run tests, and explain dependencies required to run the program among other documentation. [22]. What makes Maven special is that it will download and compile automatically all required dependencies from online repositories that are ideally maintained. Docker, on the other hand, is Container technology which is a barebones virtual machine template that can be used to create the necessary environment for a program to run including all dependencies and data and then stored in a publicly available repository that not only includes the instructions to create the Docker container, but also has a frozen image that can be downloaded to run the experiment without requiring any form of installation. For Further information, see our paper describing Docker and its application to reproducible research [4]. The only downside is the efficiency cost of running a virtual machine, that while bare-bone, still has a performance cost. However, the ability to download an image work on it, then push the new image to a global repository in a building block method is a great solution.

However, the above solutions require that the users either start working on them from the beginning or to take the time to modify their work in order to get it working with one of the solutions. For example, both CMake and Ant require the researchers to either start coding and add lines to their makefiles as they go or to go back and take the time to make them when their code is complete. For Docker Containers or other VM like software, it requires starting development inside such VMs, which may mean sacrificing some efficiency, or to go back and create, test, and run their implementations on such Virtual Machines. Among many reasons, researchers not having the time or wanting to go back and do this contributes to source code never leaving the computer where it was original made and ran. A solution to this problem, where the amount of code or data is small, was proposed in ReproZip and ReproServer. The idea is to automate the creation of a distributable bundle that "works automatically given an existing application, independently of the programming language" [7]. The author of this tool mentions it works by packing all the contents, be it code or databases, even hardware OS information. Then when someone wishes to reproduce another researcher's experiment, they can unpack into an isolated environment such as Docker or Vagrant. ReproServer furthers this idea by allowing a web interface where for simple experiments they can host the backend environments and all the user must do is upload the package created by ReproZip. The downfall to this is that because it is being run on their servers they must implement limitations based on their hardware. For non-intensive tasks, this is a friendly environment and a simple solution.

VII. Reproducible Research: Not a crisis?

One counter-argument to the reproducible research crisis given in a letter by Voelkl and Würbel states, "a more precise study conducted at a random point along the reaction norm is less likely to be reproducible than a less precise one" [23]. The argument being that reproducible research is important, but should not be done at the cost of a precise study. This is a valid point for non-deterministic research, such as Machine Learning and Training Data, where it is important to provide the learning algorithm detailed data to try and achieve the best results; however, this should not be a problem for deterministic research or where the exact training data can be given in a controlled environment. In short, reproducible research is important, but should not limit an experiment.

Others such as Fanneli, argue that "the crisis narrative is at least partially misguided" and that issues with research integrity and reproducibility are being exaggerated and is "not growing, as the crisis narrative would presuppose" [24]. The author references his previous works and studies showing that only 1-2% of researches falsify data [25] and that reproducible research, at least in terms of ensuring that the research is valid and reliable, is not a crisis,

> "To summarize, an expanding metaresearch literature suggests that science–while undoubtedly facing old and new challenges–cannot be said to be undergoing a reproducibility crisis, at least not in the sense that it is no longer reliable due to a pervasive and growing problem with findings that are fabricated, falsified, biased, underpowered, selected, and irreproducible. While these problems certainly exist and need to be tackled, evidence does not

suggest that they undermine the scientific enterprise as a whole. [24]"

A natural solution, that is currently happening and we would like to present is the idea of Reproducible Balance. Research that is currently not reproducible, if interesting and relevant enough, will be made reproducible by other scientist who in turn will ensure proper reproducibility in a new publication to stand out. An example of this is Topalidou's undertaking of a computational model created by Guthrie et al [26]. Here a highly cited paper with no available code or data was attempted to be reproduced by contacting the authors for the original source code, only to be met by "6000 lines of Delphi (Pascal language)" code that did not even compile due to missing packages [27]. After they recoded it in Python, which included a superior reproduction after the original was found to have factual errors in the manuscript and ambiguity in the description they ensured that the new model was reproducible by creating a dedicated library for it, using a versioning system for the source code (git) posted publicly (github) [27]. This is a prime example of the collective field attempting to correct important research by making it reproducible.

VIII. Conclusion

As Daniele Fanelli comments, "Science always was and always will be a struggle to produce knowledge for the benefit of all of humanity against the cognitive and moral limitations of individual human beings, including the limitations of scientists themselves" [24]. Some argue that reproducible research can hinder science, and others that it is key to cumulative science. This paper reported on the current state, importance, and challenges of reproducible research along with suggesting solutions to these challenges and commenting on available tools that implement these suggestions. At the end of the day, reproducible research has one goal: To better the scientific community by connecting all the small steps by man, into advancements as whole for mankind.

IX. Acknowledgements

Ben Cisneros for his contributions in helping run the survey and generating the graphics in this Publication.

References

[1] K. Popper, *The logic of scientific discovery*. Routledge, 2005.
[2] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.
[3] L. A. Barba, "Terminologies for reproducible research," *arXiv preprint arXiv:1802.03311*, 2018.
[4] J. R. Fonseca Cacho and K. Taghva, "Reproducible research in document analysis and recognition," in *Information Technology-New Generations*. Springer, 2018, pp. 389–395.
[5] J. T. Leek and R. D. Peng, "Opinion: Reproducible research can still be wrong: Adopting a prevention approach," *Proceedings of the National Academy of Sciences*, vol. 112, no. 6, pp. 1645–1646, 2015.
[6] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature News*, vol. 533, no. 7604, p. 452, 2016.
[7] R. Rampin, F. Chirigati, V. Steeves, and J. Freire, "Reproserver: Making reproducibility easier and less intensive," *arXiv preprint arXiv:1808.01406*, 2018.
[8] H. Wickham *et al.*, "Tidy data," *Journal of Statistical Software*, vol. 59, no. 10, pp. 1–23, 2014.
[9] M. Hutson, "Artificial intelligence faces reproducibility crisis," 2018.
[10] J. R. Fonseca Cacho, K. Taghva, and D. Alvarez, "Using the google web 1t 5-gram corpus for ocr error correction," in *16th International Conference on Information Technology-New Generations (ITNG 2019)*. Springer, 2019, pp. 505–511.
[11] J. R. Fonseca Cacho, "Improving ocr post processing with machine learning tools," 2019, phd diss., University of Nevada, Las Vegas.
[12] J. R. Fonseca Cacho and K. Taghva, "Aligning ground truth text with ocr degraded text," in *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 2019, pp. 815–833.
[13] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen *et al.*, "Promoting an open research culture," *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015.
[14] F. Sayre and A. Riegelman, "The reproducibility crisis and academic libraries," *College & Research Libraries*, vol. 79, no. 1, p. 2, 2018.
[15] V. Steeves, "Reproducibility librarianship," *Collaborative Librarianship*, vol. 9, no. 2, p. 4, 2017.
[16] T. H. Vines, R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J.-S. Moore, B. T. Moyers, S. Renaut, D. J. Rennison *et al.*, "Mandated data archiving greatly improves access to research data," *The FASEB journal*, vol. 27, no. 4, pp. 1304–1308, 2013.
[17] J. F. Claerbout and M. Karrenbach, "Electronic documents give reproducible research a new meaning," in *SEG Technical Program Expanded Abstracts 1992*. Society of Exploration Geophysicists, 1992, pp. 601–604.
[18] K. Ram, "Git can facilitate greater reproducibility and increased transparency in science," *Source code for biology and medicine*, vol. 8, no. 1, p. 7, 2013.
[19] P. Patil, R. D. Peng, and J. T. Leek, "A visual tool for defining reproducibility and replicability," *Nature human behaviour*, p. 1, 2019.
[20] L.-H. Hung, D. Kristiyanto, S. B. Lee, and K. Y. Yeung, "Guidock: Using docker containers with a common graphics user interface to address the reproducibility of research," *PloS one*, vol. 11, no. 4, p. e0152686, 2016.
[21] A. Hosny, P. Vera-Licona, R. Laubenbacher, and T. Favre, "Algorun, a docker-based packaging system for platform-agnostic implemented algorithms," *Bioinformatics*, p. btw120, 2016.
[22] O. Dalle, "Olivier dalle. should simulation products use software engineering techniques or should they reuse products of software engineering? part 1," *SCS Modeling and Simulation Magazine*, vol. 2, no. 3, pp. 122–132, 2011.
[23] B. Voelkl and H. Würbel, "Reproducibility crisis: are we ignoring reaction norms?" *Trends in pharmacological sciences*, vol. 37, no. 7, pp. 509–510, 2016.
[24] D. Fanelli, "Opinion: Is science really facing a reproducibility crisis, and do we need it to?" *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2628–2631, 2018.
[25] ——, "How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data," *PloS one*, vol. 4, no. 5, p. e5738, 2009.
[26] M. Guthrie, A. Leblois, A. Garenne, and T. Boraud, "Interaction between cognitive and motor cortico-basal ganglia loops during decision making: a computational study," *Journal of neurophysiology*, vol. 109, no. 12, pp. 3025–3040, 2013.
[27] M. Topalidou, A. Leblois, T. Boraud, and N. P. Rougier, "A long journey into reproducible computational neuroscience," *Frontiers in computational neuroscience*, vol. 9, p. 30, 2015.