Improving Reproducibility in Machine Learning Research
(*A Report from the NeurIPS 2019 Reproducibility Program*)

Joelle Pineau JPINEAU@CS.MCGILL.CA
School of Computer Science, McGill University (Mila)
Facebook AI Research
CIFAR

Philippe Vincent-Lamarre PHILVLAM@GMAIL.COM
Ecole de bibliothèconomie et des sciences de l'information,
Université de Montréal

Koustuv Sinha KOUSTUV.SINHA@MAIL.MCGILL.CA
School of Computer Science, McGill University (Mila)
Facebook AI Research

Vincent Larivière VINCENT.LARIVIERE@UMONTREAL.CA
Ecole de bibliothéconomie et des sciences de l'information,
Université de Montréal

Alina Beygelzimer BEYGEL@YAHOO-INC.COM
Yahoo! Research

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARIS.FR
Télécom Paris,
Institut Polytechnique de France

Emily Fox EBFOX@CS.WASHINGTON.EDU
University of Washington
Apple

Hugo Larochelle HUGOLAROCHELLE@GOOGLE.COM
Google
CIFAR

Abstract

One of the challenges in machine learning research is to ensure that presented and published results are sound and reliable. Reproducibility, that is obtaining similar results as presented in a paper or talk, using the same code and data (when available), is a necessary step to verify the reliability of research findings. Reproducibility is also an important step to promote open and accessible research, thereby allowing the scientific community to quickly integrate new findings and convert ideas to practice. Reproducibility also promotes the use of robust experimental workflows, which potentially reduce unintentional errors. In 2019, the Neural Information Processing Systems (NeurIPS) conference, the premier international conference for research in machine learning, introduced a reproducibility program, designed to improve the standards across the community for how we conduct, communicate, and evaluate machine learning research. The program contained three components: a code submission policy, a community-wide reproducibility challenge, and the inclusion of the Machine Learning Reproducibility checklist as part of the paper submission process. In this paper, we describe each of these components, how it was deployed, as well as what we were able to learn from this initiative.

Keywords: Reproducibility, NeurIPS 2019

1. Introduction

At the very foundation of scientific inquiry is the process of specifying a hypothesis, running an experiment, analyzing the results, and drawing conclusions. Time and again, over the last several centuries, scientists have used this process to build our collective understanding of the natural world and the laws that govern it. However, for the findings to be valid and reliable, it is important that the experimental process be repeatable, and yield consistent results and conclusions. This is of course well-known, and to a large extent, the very foundation of the scientific process. Yet a 2016 survey in the journal Nature revealed that more than 70% of researchers failed in their attempt to reproduce another researcher's experiments, and over 50% failed to reproduce one of their own experiments (Baker, 2016).

In the area of computer science, while many of the findings from early years were derived from mathematics and theoretical analysis, in recent years, new knowledge is increasingly derived from practical experiments. Compared to other fields like biology, physics or sociology where experiments are made in the natural or social world, the reliability and reproducibility of experiments in computer science, where the experimental apparatus for the most part consists of a computer designed and built by humans, should be much easier to achieve. Yet in a surprisingly large number of instances, researchers have had difficulty reproducing the work of others (Henderson et al., 2018).

Focusing more narrowly on machine learning research, where most often the experiment consists of training a model to learn to make predictions from observed data, the reasons for this gap are numerous and include:

- Lack of access to the same training data / differences in data distribution;
- Misspecification or under-specification of the model or training procedure;
- Lack of availability of the code necessary to run the experiments, or errors in the code;
- Under-specification of the metrics used to report results;
- Improper use of statistics to analyze results, such as claiming significance without proper statistical testing or using the wrong statistic test;
- Selective reporting of results and ignoring the danger of adaptive overfitting;
- Over-claiming of the results, by drawing conclusions that go beyond the evidence presented (e.g. insufficient number of experiments, mismatch between hypothesis & claim).

We spend significant time and energy (both of machines and humans), trying to overcome this gap. This is made worse by the bias in the field towards publishing positive results (rather than negative ones). Indeed, the evidence threshold for publishing a new positive finding is much lower than that for invalidating a previous finding. In the latter case, it may require several teams showing beyond the shadow of a doubt that a result is false for the research community to revise its opinion. Perhaps the most infamous instance of this is that of the false causal link between vaccines and autism. In short, we would argue that it is always more efficient to properly conduct the experiment and analysis in the first place.

In 2019, the Neural Information Processing Systems (NeurIPS) conference, the premier international conference for research in machine learning, introduced a reproducibility program, designed to improve the standards across the community for how we conduct, communicate, and evaluate machine learning research. The program contained three components: a code submission policy, a community-wide reproducibility challenge, and the inclusion of the Machine Learning Reproducibility checklist as part of the paper submission process.

In this paper, we describe each of these components, how it was deployed, as well as what we were able to learn from this exercise. The goal is to better understand how such an approach is implemented, how it is perceived by the community (including authors and reviewers), and how it impacts the quality of the scientific work and the reliability of the findings presented in the conference's technical program. We hope that this work will inform and inspire renewed commitment towards better scientific methodology, not only in the machine learning research community, but in several other research fields.

2. Background

There are challenges regarding reproducibility that appear to be unique (or at least more pronounced) in the field of ML compared to other disciplines. The first is an insufficient exploration of the variables that might affect the conclusions of a study. In machine learning, a common goal for a model is to beat the top benchmarks scores. However, it is hard to assert if the aspect of a model claimed to have improved its performance is indeed the factor leading to the higher score. This limitation has been highlighted in a few studies reporting that new proposed methods are often not better than previous implementations when a more thorough search of hyper-parameters is performed (Lucic et al., 2018; Melis et al., 2017), or even when using different random parameter initializations ((Bouthillier et al., 2019; Henderson et al., 2018).

The second challenge refers to the proper documentation and reporting of the information necessary to reproduce the reported results (Gundersen and Kjensmo, 2018). A recent report indicated that 63.5% of the results in 255 manuscripts were successfully replicated (Raff, 2019). Strikingly, this study found that when the original authors provided assistance to the reproducers, 85% of results were successfully reproduced, compared to 4% when the authors didn't respond. Although a selection bias could be at play (authors who knew their results would reproduce might have been more likely to provide assistance for the reproduction), this contrasts with large-scale replication studies in other disciplines that failed to observe similar improvement when the original authors of the study were involved (Klein et al., 2019). It therefore remains to be established if the field is having a reproduction problem similar to the other fields, or if it would be better described as a reporting problem.

Thirdly, as opposed to most scientific disciplines where uncertainty of the observed effects are routinely quantified, it appears like statistical analysis is seldom conducted in ML research (Forde and Paganini, 2019; Henderson et al., 2018).

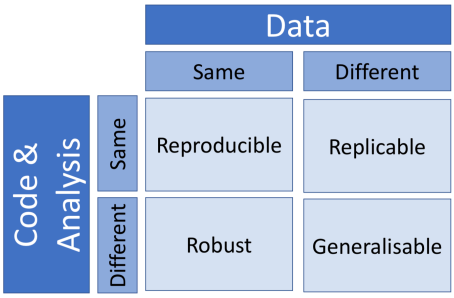


Figure 1: Reproducible Research. Adapted from: https://github.com/WhitakerLab/ReproducibleResearch

2.1 Defining Reproducibility

Before going any further, it is worth defining a few terms that have been used (sometimes interchangeably) to describe reproducibility & related concepts. We adopt the terminology from Figure 1, where Reproducible work consists of re-doing an experiment using the same data and same analytical tools, whereas Replicable work considers different data (presumably sampled from similar distribution or method), Robust work assumes the same data but different analysis (such as reimplementation of the code, perhaps different computer architecture), and Generalisable work leads to the same conclusions despite considering different data and different analytical tools. For the purposes of our work, we focus primarily on the notion of Reproducibility as defined here, and assume that any modification in analytical tools (e.g. re-running experiments on a different computer) was small enough as to be negligible. A recent report by the National Academies of Sciences, Engineering, and Medicine, provides more in-depth discussion of these concepts, as well as several recommendations for improving reproducibility broadly across scientific fields (National Academies of Sciences, Engineering, and Medicine, 2019).

2.2 The Open Science movement

"Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks" (Vicente-Sáez and Martínez-Fuentes, 2018). In other words, Open science is a movement to conduct science in a more transparent way. This includes making code, data and scientific communications publicly available, increasing the transparency of the research process and improving the reporting quality in scientific manuscripts. The implementation of Open science practices has been identified as a core factor that could improve the reproducibility of science (Munafò et al., 2017). As such, the NeurIPS reproducibility program was designed to incorporate elements designed to encourage researchers to share the artefacts of their research (code, data), in addition to their manuscripts.

2.3 Code submission policies

It has become increasingly common in recent years to require the sharing of data and code, along with a paper, when computer experiments were used in the analysis. It is now standard expectation in the Nature research journals for authors to provide access to code and data to readers (Nature Research, 2021). Similarly, the policy at the journal Science specifies that authors are expected to satisfy all reasonable requests for data, code or materials (Science - AAAS, 2018). Within machine learning and AI conferences, the ability to include supplementary material has now been standard for several years, and many authors have used this to provide the data and/or code used to produce the paper. More recently, ICML 2019, the second largest international conference in machine learning has also rolled-out an explicit code submission policy (ICML, 2019).

2.4 Reproducibility challenges

The 2018 ICLR reproducibility challenge paved the way for the NeurIPS 2019 edition. The goal of this first iteration was to investigate reproducibility of empirical results submitted to the 2018 International Conference on Learning Representations (ICLR, 2018). The organizers chose ICLR for this challenge because the timing was right for course-based participants: most participants were drawn from graduate machine learning courses, where the challenge served as the final course project. The choice of ICLR was motivated by the fact that papers submitted to the conference were automatically made available publicly on OpenReview, including during the review period. This means anyone in the world could access the paper prior to selection, and could interact with the authors via the message board on OpenReview. This first challenge was followed a year later by the 2019 ICLR Reproducibility Challenge (Pineau et al., 2019).

Several less formal activities, including hackathons, course projects, online blogs, open-source code packages, have participated in the effort to carry out re-implementation and replication of previous work and should be considered in the same spirit as the effort described here.

2.5 Checklists

The Checklist Manifesto presents a highly compelling case for the use of checklists in safety-critical systems (Gawande, 2010). It documents how pre-flight checklists were introduced at Boeing Corporation as early as 1935 following the unfortunate crash of an airplane prototype. Checklists are similarly used in surgery rooms across the world to prevent oversights. Similarly, the WHO Surgical Safety Checklist, which is employed in surgery rooms across the world, has been shown to significantly reduce morbidity and mortality (Clay-Williams and Colligan, 2015).

In the case of scientific manuscripts, reporting checklists are meant to provide the minimal information that must be included in a manuscript, and are not necessarily exhaustive. The use of checklists in scientific research has been explored in a few instances. Reporting guidelines in the form of checklists have been introduced for a wide range of study design in health research (The EQUATOR Network, 2021), and the Transparency and Openness Promotion (TOP) guidelines have been adopted by multiple journals across disciplines (Nosek

et al., 2015). There are now more than 400 checklists registered in the EQUATOR Network. CONSORT, one of the most popular guidelines used for randomized controlled trials was found to be effective and to improve the completeness of reporting for 22 checklist items (Turner et al., 2012). The ML checklist described below was significantly influenced by Nature's Reporting Checklist for Life Sciences Articles (Checklist, 2021). Other guidelines are under development outside of the ML community, namely for the application of AI tools in clinical trials (Liu et al., 2019) and health-care (Collins and Moons, 2019).

2.6 Other considerations

Beyond reproducibility, there are several other factors that affect how scientific research is conducted, communicated and evaluated. One of the best practices used in many venues, including NeurIPS, is that of double-blind reviewing. It is worth remembering that in 2014, the then program chairs Neil Lawrence and Corinna Cortes ran an interesting experiment, by assigning 10% of submitted papers to be reviewed independently by two groups of reviewers (each lead by a different area chair). The results were surprising: overall the reviewers disagreed on 25.9% of papers, but when tasked with reaching a 22.5% acceptance rate, they disagreed on 57% of the list of accepted papers. We raise this point for two reasons. First, to emphasize that the NeurIPS community has for many years already demonstrated an openness towards trying new approaches, as well as looking introspectively on the effectiveness of its processes. Second, to emphasize that there are several steps that come into play when a paper is written, and selected for publication at a high-profile international venue, and that a reproducibility program is only one aspect to consider when designing community standards to improve the quality of scientific practices.

3. The NeurIPS 2019 code submission policy

The NeurIPS 2019 code submission policy, as defined for all authors (see Appendix, Figure 6), was drafted by the program chairs and officially approved by the NeurIPS board in winter 2019 (before the May 2019 paper submission deadline.)

The most frequent objections we heard to having a code submission policy (at all) include:

- **Dataset confidentiality**: There are cases where the dataset cannot be released for legitimate privacy reasons. This arises often when looking at applications of ML, for example in healthcare or finance. One strategy to mitigate this limitation is to provide complementary empirical results on an open-source benchmark dataset, in addition to the results on the confidential data.

- **Proprietary software**: The software used to derive the result contains intellectual property, or is built on top of proprietary libraries. This is of particular concern to some researchers working in industry. Nonetheless, as shown in Figure 2a, we see that many authors from industry were indeed able to submit code, and furthermore despite the policy, the acceptance rate for papers from authors in industry remained high (higher than authors from academia (Figure 2b)). By the camera-ready deadline, most submissions from the industry reported having submitted code (Figure 2a,b).

- **Computation infrastructure**: Even if data and code are provided, the experiments may require so much computation (time & number of machines) that it is impractical for any reviewer, or in fact most researchers, to attempt reproducing the work. This is the case for work on training very large neural models, for example the AlphaGo game playing agent (Silver et al., 2016) or the BERT language model (Devlin et al., 2018). Nonetheless it is worth noting that both these systems have been reproduced within months (if not weeks) of their release.
- **Replication of mistakes**: Having a copy of the code used to produce the experimental results is not a guarantee that this code is correct, and there is significant value in reimplementing an algorithm directly from its description in a paper. This speaks more to the notion of Robustness defined above. It is indeed common that there are mistakes in code (as there may be in proofs for more theoretical papers). Nonetheless, the availability of the code (or proof) can be tremendously helpful to verify or re-implement the method. It is indeed much easier to verify a result (with the initial code or proof), then it is to produce from nothing (this is perhaps most poignantly illustrated by the longevity of the lack of proof for Fermat's last theorem (Wikipedia, 2020).)

It is worth noting that the NeurIPS 2019 code submission policy leaves significant time & flexibility, in particular it says that it: *"**expects** code **only for accepted papers**, and only **by the camera-ready deadline**"*. So code submission is not mandatory, and the code is not expected to be used during the review process to decide on the soundness of the work. Reviewers were asked as a part of their assessment to report if code was provided along the manuscript at the initial submission stage. About 40% of authors reported that they had provided code at this stage which was confirmed by the reviewers (if at least one reviewer indicated that the code was provided for each submission) for 71.5% of those submissions (Figure 2d). Note that authors are still able to provide code (or a link to code) as part of their initial submission. In Table 1, we provide a summary of code submission frequency for ICML 2019, as well as NeurIPS 2018 and 2019. We observe a growing trend towards more papers adding a link to code, even with only soft encouragement and no coercive measures.

While the value of having code extends long beyond the review period, it is useful, in those cases where code is available during the review process, to know how it is used and perceived by the reviewers. When surveying reviewers at the end of the review period, we found:

Q. Was code provided (e.g. in the supplementary material)? Yes: 5298

If provided, did you look at the code? Yes: 2255

If provided, was the code useful in guiding your review? Yes: 1315

If not provided, did you wish code had been available? Yes: 3881

We were positively surprised by the number of reviewers willing to engage with this type of artefact during the review process. Furthermore, we found that the availability of code at submission (as indicated on the checklist) was positively associated with the reviewer score ($p < 1e-08$).

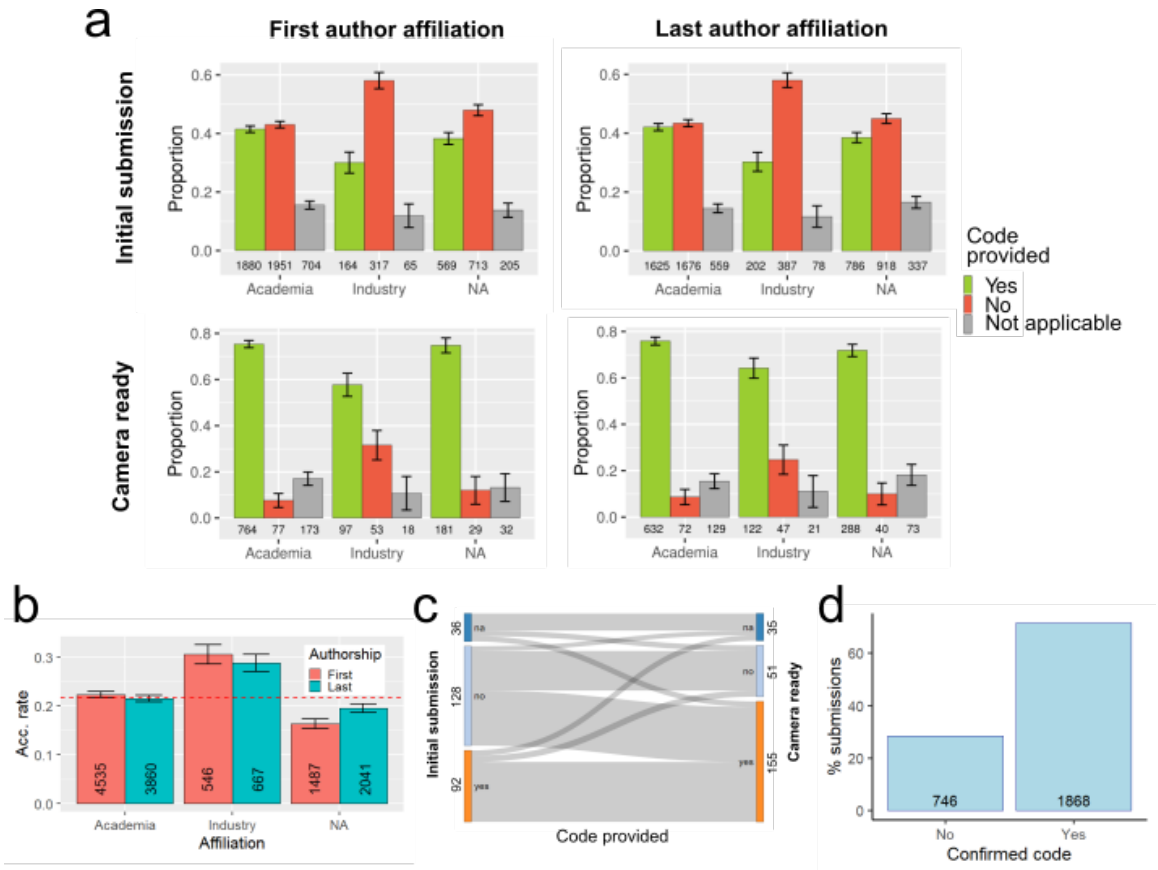


Figure 2: (a) Link to code provided at initial submission and camera-ready, as a function of affiliation of the first and last authors. (b) Acceptance rate of submissions as a function of affiliation of the first and last authors. The red dashed line shows the acceptance rate for all submissions. (c) Diagram representing the transition of the code availability from initial submission to camera-ready only for submissions with an author from the industry (first or last). All results presented here for code availability are based on the author's self-response in the checklist. (d) Percentage of submissions reporting that they provided code on the checklist subsequently confirmed by the reviewers.

Conference	# papers submitted	% papers accepted	% papers w/code at submission	% papers w/code at camera-ready	Code submission policy
NeurIPS 2018	4856	20.8		<50%	"Authors may submit up to 100MB of supplementary material, such as proofs, derivations, data, or source code."
ICML 2019	3424	22.6	36%	67%	"To foster reproducibility, we highly encourage authors to submit code. Reproducibility of results and easy availability of code will be taken into account in the decision-making process."
NeurIPS 2019	6743	21.1	40%	74.4%	"We expect (but not require) accompanying code to be submitted with accepted papers that contribute and present experiments with a new algorithm." See Appendix, Fig. 6

Table 1: Code submission frequency for recent ML conferences. Source for number of papers accepted and acceptance rates: https://github.com/lixin4ever/Conference-Acceptance-Rate. ICML 2019 numbers reproduced from the ICML 2019 Code-at-Submit-Time Experiment.

Conference	# papers submitted	Acceptance rate	# papers claimed	# participating institutions	# reports reviewed
ICLR 2018	981	32.0	123	31	n/a
ICLR 2019	1591	31.4	90	35	26
NeurIPS 2019	6743	21.1	173	73	84

Table 2: Participation in the Reproducibility Challenge. Source for number of papers accepted and acceptance rates: https://github.com/lixin4ever/Conference-Acceptance-Rate

4. The NeurIPS 2019 Reproducibility Challenge

The main goal of this challenge is to provide independent verification of the empirical claims in accepted NeurIPS papers, and to leave a public trace of the findings from this secondary analysis. The reproducibility challenge officially started on Oct.31 2019, right after the final paper submission deadline, so that participants could have the benefit of any code submission by authors. By this time, the authors' identity was also known, allowing collaborative interaction between participants and authors. We used OpenReview (OpenReview.net, 2021) to enable communication between authors and challenge participants.

As shown in Table 2, a total of 173 papers were claimed for reproduction. This is a 92% increase since the last reproducibility challenge at ICLR 2019 (Pineau et al., 2019). We had participants from 73 different institutions distributed around the world (see Appendix, Figure 7), including 63 universities and 10 industrial labs. Institutions with the most participants came from 3 continents and include McGill University (Canada), KTH (Sweden), Brown University (US) and IIT Roorkee (India). In those cases (and several others), high participation rate occurred when a professor at the university used this challenge as a final course project.

All reports submitted to the challenge are available on OpenReview [1] for the community; in many cases with a link to the reimplementation code. The goal of making these available is to two-fold: first to give examples of reproducibility reports so that the practice becomes more widespread in the community, and second so that other researchers can benefit from the knowledge, and avoid the pitfalls that invariably come with reproducing another team's work. While many readers may be looking for a simple answer to the question *Is this paper reproducible?* There is rarely such a concise outcome to a reproducibility study. Most reports produced during the challenge offer a much more detailed & nuanced account of their efforts, and the level of fidelity to which they could reproduce the methods, results & claims of each paper. Similarly, while some readers may be looking for a "reproducibility score", we have not found that the findings of most reproducibility studies lend themselves to such a coarse summary.

Once submitted, all reproducibility reports underwent a review cycle (by reviewers of the NeurIPS conference), to select a small number of high-quality reports, which will be published in an upcoming edition of the journal ReScience (ReScience C, 2021). This provides a lasting archival record for this new type of research artefact.

5. The NeurIPS 2019 ML reproducibility checklist

The third component of the reproducibility program involved use of the Machine Learning reproducibility checklist (see Appendix, Figure 8). This checklist was first proposed in late 2018, at the NeurIPS conference, in response to findings of recurrent gaps in experimental methodology found in recent machine learning papers. An earlier version (v.1.1) was first deployed as a trial with submission of the final camera-ready version for NeurIPS 2018 papers (due in January 2019); this initial test allowed collection of feedback from authors and some minor modifications to the content of the checklist. The edited version 1.2 was then deployed during the NeurIPS 2019 review process, and authors were obliged to fill it both at the initial paper submission phase (May 2019), and at the final camera-ready phase (October 2019). This allowed us to analyze any change in answers, which presumably resulted from the review feedback (or authors' own improvements of the work). The checklist was implemented on the CMT platform; each question included a multiple choice "Yes, No, not applicable", and an (optional) open comment field.

Figure 3 shows the initial answers provided for each submitted paper. It is reassuring to see that 97% of submissions are said to contain Q#. *A clear description of the mathematical setting, algorithm, and/or model.* Since we expect all papers to contain this, the 3% no/na answers might reflect margin of error in how authors interpreted the questions. Next, we notice that 89% of submissions answered to the affirmative when asked Q#. *For all figures and tables that present empirical results, indicate if you include: A description of how experiments were run.* This is reasonably consistent with the fact that 9% of NeurIPS 2019 submissions indicated "Theory" as their primary subject area, and thus may not contain empirical results.

One set of responses that raises interesting questions is the following trio:

Q#. *A clear definition of the specific measure or statistics used to report results.*

1. https://openreview.net/group?id=NeurIPS.cc/2019/Reproducibility_Challenge

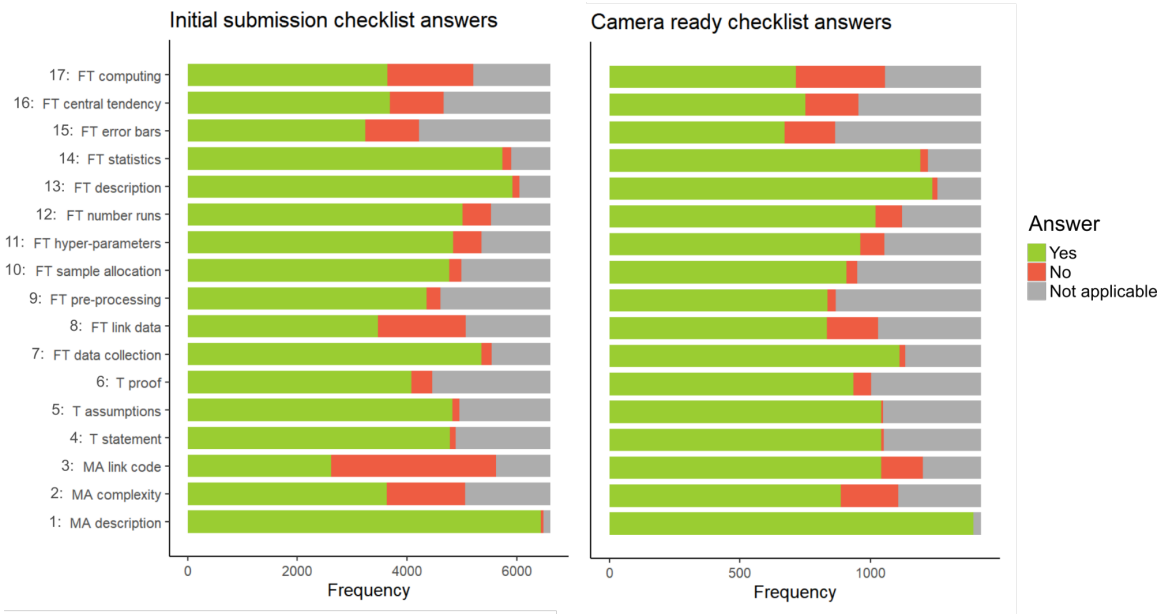


Figure 3: Author responses to all checklist questions for NeurIPS 2019 submitted papers.

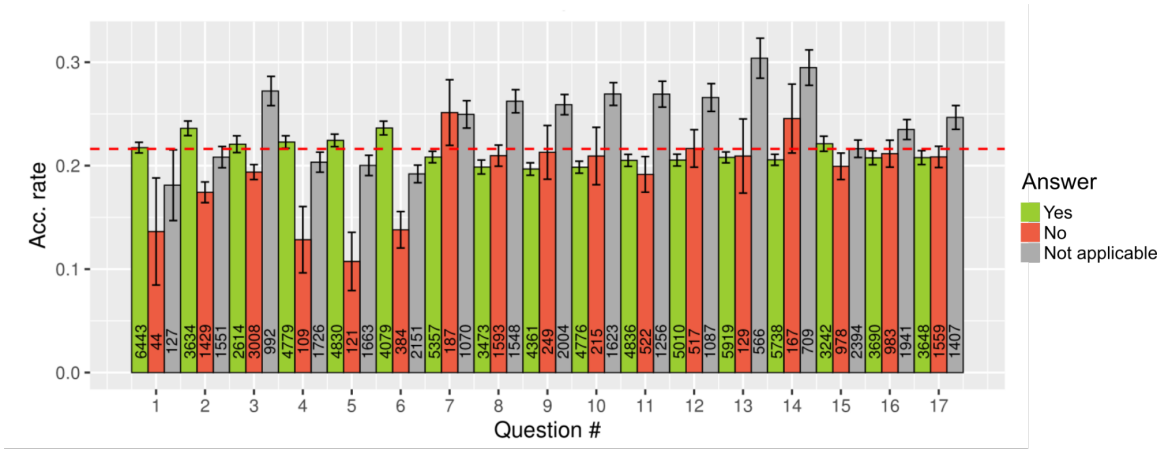


Figure 4: Acceptance rate per question. The numbers within each bar show the number of submissions for each answer. See Fig. 3 (and in Appendix Fig. 8) for text corresponding to each Question # (x-axis). The red dashed line shows the acceptance rate for all submissions.

> Q#. *Clearly defined error bars.*
>
> Q#. *A description of results with central tendency (e.g. mean) & variation (e.g. stddev).*

In particular, it seems surprising to have 87% of papers that see value in clearly defining the metrics and statistics used, yet 36% of papers judge that error bars are not applicable to their results.

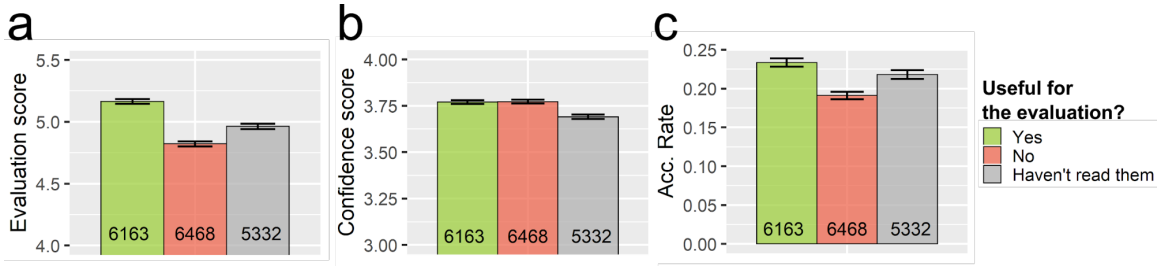


Figure 5: Perceived usefulness of the ML reproducibility checklist vs the review outcomes. (a) Effect on the paper score (scale 1-10). (b) Effect on the reviewer confidence score (scale of 1 to 5, where 1 is lowest). (c) Effect on the final accept/reject decision.

As shown in Figure 4, many checklist answers appear to be associated with a higher acceptance rate when the answer is "yes". However, it is too early to rule out potential covariates (e.g. paper's topic, reviewer expectations, etc.) At this stage, it is encouraging that answering "no" to any of the questions is not associated with a higher acceptance rate. There seems to be a higher acceptance rate associated with "NA" responses on a subset of questions related to "Figures and tables". Although it is still unclear at this stage why this effect is observed, it disappears when we only include manuscripts for which the reviewers indicated that the checklist was useful for the review.

Finally, it is worth considering the reviewers' point of view on the usefulness of the ML checklist to assess the soundness of the papers. When asked *"Were the Reproducibility Checklist answers useful for evaluating the submission?"*, 34% responded Yes.

We also note, as shown in Figure 5, that reviewers who found the checklist useful gave higher scores. And that those who found the checklist useful or not useful were more confident in their assessment than those who had not read the checklist. Finally, papers where the checklist was assessed as useful were more likely to be accepted.

6. Discussion

We presented a summary of the activities & findings from the NeurIPS 2019 reproducibility program. Perhaps the best way to think of this effort is as a case study showing how three different mechanisms (code submission, reproducibility challenge, reproducibility checklist) can be incorporated into a conference program in an attempt to improve the quality of scientific contributions. At this stage, we do not have concluding evidence that these processes indeed have an impact on the quality of the work or of the papers that are submitted and published.

However we note several encouraging indicators:

- The number of submissions to NeurIPS increased by nearly 40% this year, therefore we can assume the changes introduced did not result in a significant drop of interest by authors to submit their work to NeurIPS.

- The number of authors willingly submitting code is quickly increasing, from less than 50% a year ago, to nearly 75%. It seems a code submission policy based on volun-

tary participation is sufficient at this time. We are not necessarily aiming for 100% compliance, as there are some cases where this may not be desirable.

- The number of reviewers indicating that they consulted the code, or wished to consult it is in the 1000's, indicating that this is useful in the review process.
- The number of participants in the reproducibility challenge continues to increase, as does the number of reproducibility reports, and reviewers of reproducibility reports. This suggests that an increasing segment of the community is willing to participate voluntarily in secondary analysis of research results.
- One-third of reviewers found the checklist answers useful, furthermore reviewers who found the checklist useful gave higher scores to the paper, which suggests the checklist's use is useful for both reviewers and authors.

The work leaves several questions open, which would require further investigation, and a careful study design to elucidate:

- What is the long-term value (e.g. reproducibility, robustness, generalization, impact of follow-up work) of the code submitted?
- What is the effect of different incentive mechanisms (e.g. cash payment, conference registration, a point/badge system) on the participation rate & quality of work in the reproducibility challenge?
- What is the benefit of using the checklist for authors?
- What is the accuracy of the ML checklist answers (for each question) when filled by authors?
- What is the measurable effect of the checklist on the quality of the final paper, e.g. in terms of soundness of results, clarity of writing?
- What is the measurable effect of the checklist on the review process, in terms of reliability (e.g. inter-rater agreement) and efficiency (e.g. need for response/rebuttal, discussion time)?

A related direction to explore is the development of tools and platforms that enhance reproducibility. Throughout this work we have focused on processes & guidelines, but stayed away from prescribing any infrastructure or software tooling to support reproducibility. Software containers, such as Docker, can encapsulate operating systems components, code and data into a single package. Standardization of such tools would help sharing of information and improve ease of reproducibility.

In conclusion, one aspect worth emphasizing is the fact that achieving reproducible results across a research community, whether NeurIPS or another, requires a significant cultural and organizational changes, not just a code submission policy or a checklist. The initiative described here is just one step in helping the community adopt better practices, in terms of conducting, communicating, and evaluating scientific research. The NeurIPS

community is far from alone in looking at this problem. Several workshops have been held in recent years to discuss the issue as it pertains to machine learning and computer science (SIGCOMM, 2017; ICML, 2017, 2018; ICLR, 2019). Specific calls for reproducibility papers have been issued (ECIR, 2020). An open-access peer-reviewed journal is dedicated to such papers (ReScience C, 2021), which was used to publish select reports in ICLR 2019 Reproducibility Challenge (Pineau et al., 2019). And in the process, many labs are changing their practices to improve reproducibility of their own results.

Acknowledgments

We thank the NeurIPS board and the NeurIPS 2019 general chair (Hanna Wallach) their unfailing support of this initiative. Without their courage and spirit of experimentation, none of this work would have been possible. We thank the many authors who submitted their work to NeurIPS 2019 and agreed to participate in this large experiment. We thank the program committee (reviewers, area chairs) of NeurIPS 2019 who not only incorporated the reproducibility checklist into their task flow, but also provided feedback about its usefulness. We thank Zhenyu (Sherry) Xue for preparing the data on NeurIPS papers & reviews for the analysis presented here. We thank the OpenReview team (in particular Andrew McCallum, Pam Mandler, Melisa Bok, Michael Spector and Mohit Uniyal) who provided support to host the results of the reproducibility challenge. We thank CodeOcean (in particular Xu Fei) for providing free compute resources to reproducibility challenge participants. Thank you to Robert Stojnic for valuable comments on an early version of the manuscript. Finally, we thank the several participants of the reproducibility challenge who dedicated time and effort to verify results that were not their own, to help strengthen our understanding of machine learning, and the types of problems we can solve today.

References

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533 (7604):452, May 2016. doi: 10.1038/533452a. URL `https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970`.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/bouthillier19a.html`.

Nature Checklist. Nature checklist. 2021. URL `https://media.nature.com/original/nature-assets/ng/journal/v49/n10/extref/ng.3933-S2.pdf`.

Robyn Clay-Williams and Lacey Colligan. Back to basics: Checklists in aviation and healthcare. *BMJ Quality & Safety*, 24(7):428–431, July 2015. ISSN 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2015-003957.

Gary S. Collins and Karel G. M. Moons. Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181):1577–1579, April 2019. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(19)30037-6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, October 2018.

ECIR. Call for Reproducibility papers. April 2020. URL `ecir2020.org`.

Jessica Zosa Forde and Michela Paganini. The Scientific Method in the Science of Machine Learning. *arXiv:1904.10922 [cs, stat]*, April 2019. URL `http://arxiv.org/abs/1904.10922`.

Atul Gawande. *Checklist manifesto, the (HB)*. Penguin Books India, 2010.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

ICLR. ICLR 2018 Reproducibility Challenge. 2018. URL `https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html`.

ICLR. Reproducibility in Machine Learning Workshop. 2019. URL `https://sites.google.com/view/icml-reproducibility-workshop/home`.

ICML. Reproducibility in Machine Learning Workshop. 2017. URL `https://sites.google.com/view/icml-reproducibility-workshop/icml2017/talks-and-abstracts`.

ICML. Reproducibility in Machine Learning Workshop. 2018. URL `https://sites.google.com/view/icml-reproducibility-workshop/icml2018/home`.

ICML. Call for papers. 2019. URL `https://icml.cc/Conferences/2019/CallForPapers`.

Richard A Klein, Corey L Cook, Charles R Ebersole, Christine Vitiello, Brian A Nosek, Christopher R Chartier, Cody D Christopherson, Samuel Clay, Brian Collisson, Jarret Crawford, et al. Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. 2019.

Xiaoxuan Liu, Livia Faes, Melanie J Calvert, and Alastair K Denniston. Extension of the consort and spirit statements. *The Lancet*, 394(10205):1225, 2019.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.

Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and replicability in science*. National Academies Press, 2019.

Nature Research. Reporting standards and availability of data, materials, code and protocols, 2021. URL `https://www.nature.com/nature-research/editorial-policies/reporting-standards`.

Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

OpenReview.net. Neurips 2019 reproducibility challenge, 2021. URL `https://openreview.net/group?id=NeurIPS.cc/2019/Reproducibility_Challenge`.

Joelle Pineau, Koustuv Sinha, Genevieve Fried, Rosemary Nan Ke, and Hugo Larochelle. ICLR Reproducibility Challenge 2019. *ReScience C*, 5(2):5, May 2019. doi: 10.5281/zenodo.3158244. URL `https://zenodo.org/record/3158244/files/article.pdf`.

Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, pages 5486–5496, 2019.

ReScience C. The rescience journal. reproducible science is good. replicated science is better., 2021. URL `http://rescience.github.io/`.

Science - AAAS. Science journals: Editorial policies., 2018. URL `https://www.sciencemag.org/authors/science-journals-editorial-policies`.

SIGCOMM. Reproducibility '17: Proceedings of the reproducibility workshop. 2017. URL `https://dl.acm.org/doi/proceedings/10.1145/3097766`.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

The EQUATOR Network. Enhancing the quality and transparency of health research., 2021. URL `https://www.equator-network.org/`.

Lucy Turner, Larissa Shamseer, Douglas G Altman, Laura Weeks, Jodi Peters, Thilo Kober, Sofia Dias, Kenneth F Schulz, Amy C Plint, and David Moher. Consolidated standards of reporting trials (consort) and the completeness of reporting of randomised controlled trials (rcts) published in medical journals. *Cochrane Database of Systematic Reviews*, (11), 2012.

Rubén Vicente-Sáez and Clara Martínez-Fuentes. Open science now: A systematic literature review for an integrated definition. *Journal of business research*, 88:428–436, 2018.

Wikipedia. Fermat's Last Theorem. March 2020. URL `https://en.wikipedia.org/wiki/Fermat%27s_Last_Theorem`. Page Version ID: 944945734.

Appendix A.

As an experiment, NeurIPS-2019 will use the following Code Submission Policy.

1. The policy only applies to papers that **contribute and present experiments with a new algorithm (or a modification to an existing algorithm)**. That is, a paper is **not** covered by this policy if:
   a. The paper is not claiming the contribution of any novel algorithm.
   b. The paper presents a new algorithm but only analyzes it theoretically (i.e., no experimental results are presented).

2. Code submission for papers covered by this policy is **expected but not enforced.**

3. The policy **accepts a reimplementation** by the authors that isn't the code originally run to produce the results reported in the paper (what is instead requested is the equivalent of an official implementation of the paper's contribution).

4. The policy **accepts code that isn't "executable" as is** as it has dependencies going beyond the algorithm itself and that cannot be released. Such dependencies would include
   a. Dataset that cannot be released (e.g., for privacy reasons).
   b. Specialized hardware that might not be commonly accessible (e.g., specialized accelerators or robotic platforms).
   c. Non-open sourced or non-free libraries, which do not include the algorithm that is claimed as the scientific contribution of the paper (e.g., paid-for mathematical programming solvers, commercial simulators, MATLAB).
   The authors will be asked to explain what dependencies are not released and why.

5. The policy expects code **only for accepted papers,** and only **by the camera-ready deadline (October 27, 2019)**.

After the camera-ready deadline, NeurIPS intends to measure the percentage of accepted papers for which code was not released, despite being covered by the policy.

Figure 6: The NeurIPS 2019 code submission policy. Reproduced (with permission) from: [ADD URL

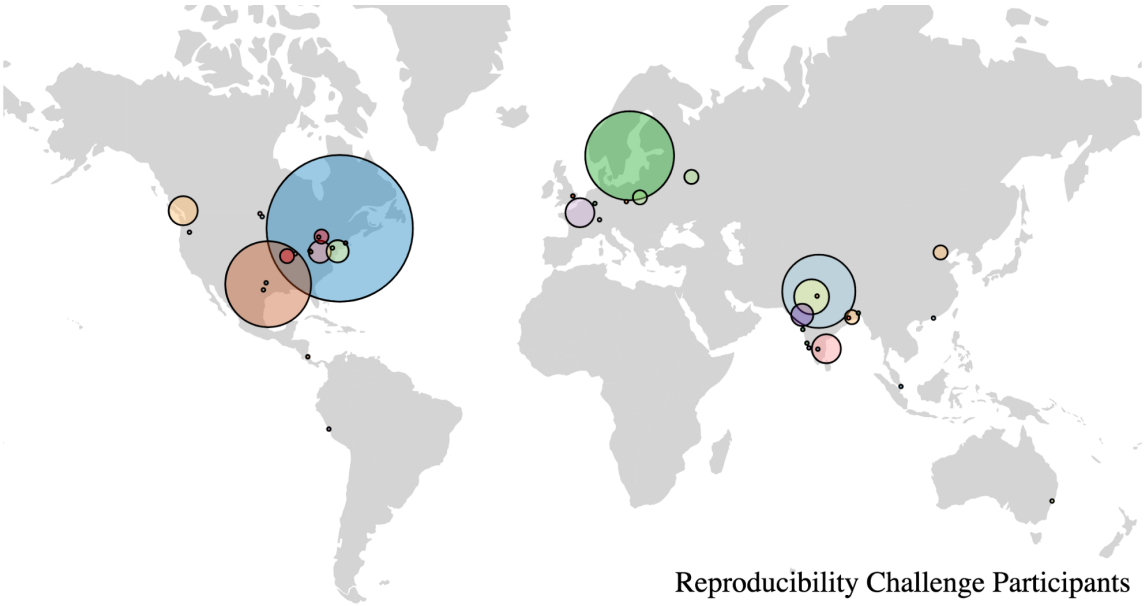
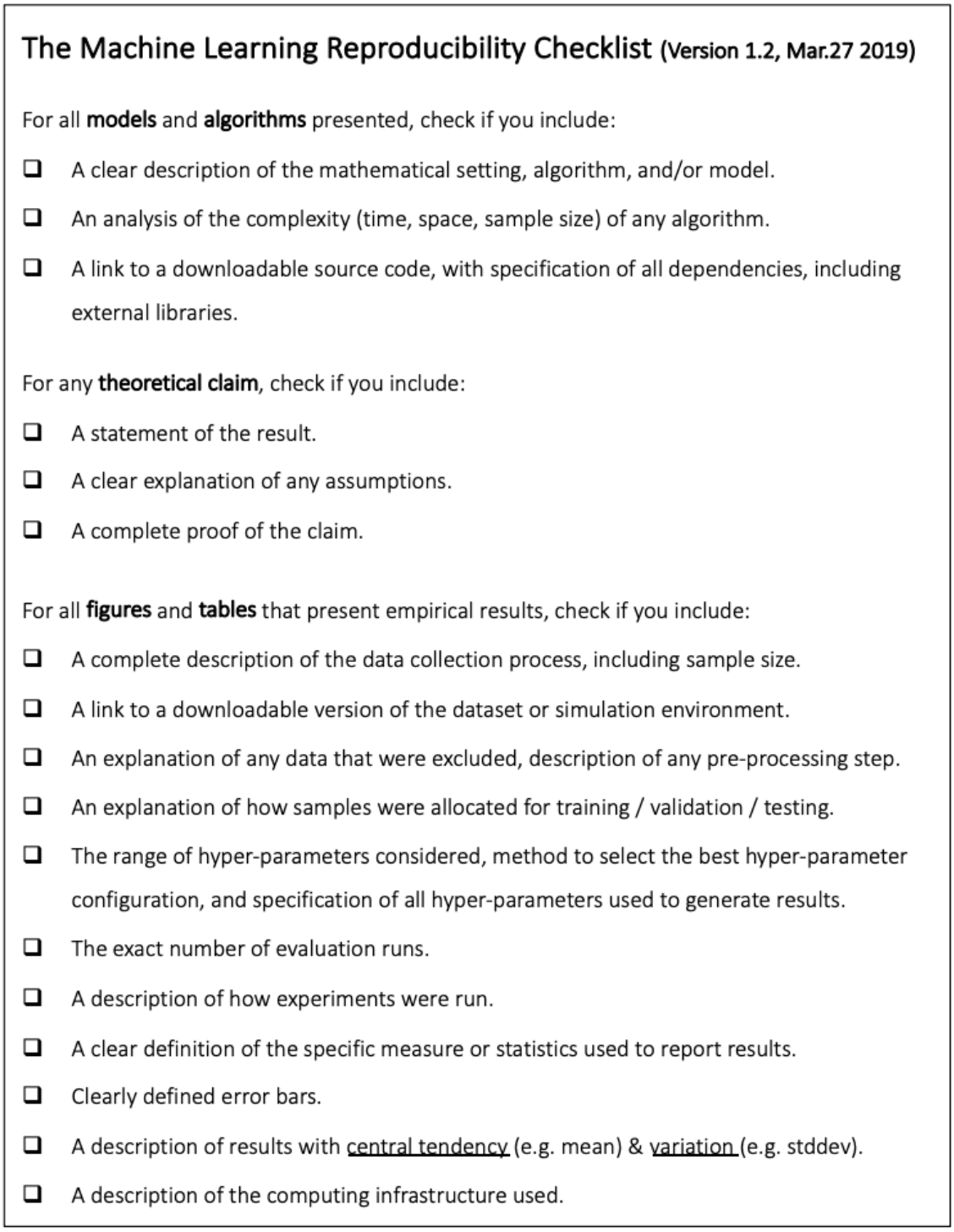


Figure 7: NeurIPS 2019 Reproducibility Challenge Participants by geographical location.



The Machine Learning Reproducibility Checklist (Version 1.2, Mar.27 2019)

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.
- ❑ A link to a downloadable source code, with specification of all dependencies, including external libraries.

For any **theoretical claim**, check if you include:

- ❑ A statement of the result.
- ❑ A clear explanation of any assumptions.
- ❑ A complete proof of the claim.

For all **figures** and **tables** that present empirical results, check if you include:

- ❑ A complete description of the data collection process, including sample size.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ An explanation of any data that were excluded, description of any pre-processing step.
- ❑ An explanation of how samples were allocated for training / validation / testing.
- ❑ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ❑ The exact number of evaluation runs.
- ❑ A description of how experiments were run.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ Clearly defined error bars.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. stddev).
- ❑ A description of the computing infrastructure used.

Reproduced from: *www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf*

Figure 8: The Machine Learning Reproducibility Checklist, version 1.2, used during the NeurIPS 2019 review process.