What Do Machine Learning Researchers Mean by "Reproducible"?

Edward Raff[1,2], **Michel Benaroch**[3], **Sagar Samtani**[4], **Andrew L. Farris**[1]

[1]Booz Allen Hamilton
[2]University of Maryland, Baltimore County
[3]Syracuse University
[4]Indiana University

Abstract

The concern that Artificial Intelligence (AI) and Machine Learning (ML) are entering a "reproducibility crisis" has spurred significant research in the past few years. Yet with each paper, it is often unclear what someone means by "reproducibility". Our work attempts to clarify the scope of "reproducibility" as displayed by the community at large. In doing so, we propose to refine the research to eight general topic areas. In this light, we see that each of these areas contains many works that do not advertise themselves as being about "reproducibility", in part because they go back decades before the matter came to broader attention.

1 Introduction

The Artificial Intelligence (AI) and Machine Learning (ML) communities are increasingly concerned with the "reproducibility" of their fields. This has come on the heels of a reproducibility crisis noted in many others. We will refer to this overarching concern, that the science of research is being done with some error rate, as a generic *scientific rigor* concern. This concern is justified, and it is increasingly challenging to evaluate the state of research around scientific rigor due to confused and incompatible usage of the same few terms like "reproducibility" (Plesser 2018).

Due to confusing and often inconsistently used terminology in the literature, it is challenging to understand precisely what issues of scientific rigor the community is tackling. In light of these issues, we propose a new formulation of current scientific rigor research by surveying the current articles by the topics they cover. In doing so, we observe that many historical works tackled these very issues – with different motivations and no particular thematic name like "reproducibility" as it was not an urgent concern at the time.

In this article, we will expand the ACM's proposed terminology of Repeatability, Reproducibility, and Replicability which we find useful, although still insufficient to capture the breadth of work done to date. Our contribution classified current AI / ML research in scientific rigor into eight aspects we label as *repeatability*, *reproducibility, replicability*, *adaptability*, *model selection*, *label/data quality*, *meta & incentive*, and *maintainability*. These eight aspects are defined in Table 1. We propose these aspects based on our review of 101 papers published since 2017 and reflect the focus of the community at large. Table 1 also shows for each aspect the proportion of papers focused primarily on that aspect (though many papers touch on multiple aspects).

The rest of this article is organized as follows. We will summarize the eight main topic areas of scientific rigor in Section 2, with sub-areas included based on our literature review. Based on this survey of the literature, we propose relationships for how these rigors interact in Section 3, which we find informative as a macro-level picture of the scope of scientific rigor. Finally, we conclude in Section 4.

2 The Current Scope of Work

Our literature survey identifies at least eight primary aspects of scientific rigor studied in the AI/ML literature. Each major sub-section will repeat one of the eight rigors defined in Table 1, and include further delineation for nuanced subcategories that are present or noteworthy in the literature. A key criterion for being included in Table 1 is that the paper must self-identify itself as being about "repeatability, reproducibility, or replicability" since those are the three preexisting terminologies used (interchangeably) in the prior literature. These delimitations reflect the current scope of what researchers actively consider "reproducibility" consider worthy of study and effort. As our bibliography will show though, many more papers exist in these topical areas that were published before 2017, and thus before the AI/ML communities started to put renewed effort into the issue of scientific rigor. We include such articles in the discussion of each section to establish the full scope of available work and to connect the current reproducibility-themed motivation to its historical precedents, as the historical literature is often unknown to existing researchers on this topic.

Before we detail these aspects, it is worth noting that many existing articles are best summarized as opinion pieces with varying degrees of formalization of their arguments. Most of these articles propose strategies or arguments on how to obtain "reproducibility", without evidence of effect (Gundersen and Kjensmo 2018; Matsui and Goya 2022; Publio, Esteves, and Zafar 2018; Tatman, Vanderplas, and Dane 2018; Sculley et al. 2018; Vollmer et al. 2020; Drummond 2009, 2018; Raff and Farris 2022; Lin 2022). The contents of our article are focused on works that study issues,

Topics	Main Concern	% of papers
Repeatability	Can the results be obtained by the original authors using their original code and data.	12.9
Reproducibility	Can a different team obtain the results using the code and data provided by the original authors	16.8
Replicability	Can a different team, using different code and/or data, obtain the same results or results congruent with the original publication.	15.8
Adaptability	Can the original authors using the original code obtain qualitatively similar results on new/different data.	4.0
Model Selection	Given a set of two or more models, what process can be used to meaningfully and reliably determine which model to select for use.	19.8
Label/Data Quality	Given a process for labeling data, how can we ensure that the process results in meaningfully same labels over time and that the process of labeling has minimal errors.	4.0
Meta & Incentive	What are the motivators, or lack thereof, for scientific rigor.	13.9
Maintainability	What are the issues and remediations in running the same AI/ML solution as the people, code, and data are all altered in their nature over time.	12.9

Table 1: Eight primary topics that have been collectively described as "reproducibility" in the literature, determined by our manual review. The first three are based on the ACM's guidelines, and the rest are informed by surveying and categorizing the themes of existing literature.

incentives, or interventions to rigor issues in AI/ML — and thus go beyond opinion or thought pieces on the topic, of which there are many. These thought pieces are valuable in spring motivation and growth in the field, but most are disconnected from the long literature on the topic, so we prefer not to focus on their opinions which may well change with new literature[1].

Since there is no canonically accepted "home" for AI/ML papers on reproducibility, we find that such published literature is scattered across various subfields and specialized conferences. In many cases, we find common themes in the nature of issues that occur across fields and domains, and in some aspects, the literature on issues impacting scientific rigor directly goes back to the 1990s. Our categorization is based on a review of all literature we are aware of that tackles scientific rigor issues, even if they did not use terms like "reproducible" as they often pre-date the larger academic concern itself. One recommendation that we would put forth for *all major AI/ML conferences is to create a track for scientific rigor studying all eight proposed rigor topics to further incentivize and organize this important work.*

2.1 Repeatability

Repeatability concerns the authors who obtain the same results using the original source code and data. Interesting questions in repeatability include how to develop code and systems that make it easy for the developer to keep track of how they came to their experimental results from an experimental design perspective (Gardner, Brooks, and Baker 2018; Paganini and Forde 2020). In Human-Computer Interaction (HCI) research, there has been significant study on the iterative development nature of computational notebooks (e.g., Jupyter) that are widely used in AI/ML development processes. These notebooks can be prone to many subtle code errors/issues due to their fluidity and out-of-order execution. Enhanced tools can ensure the exact execution sequence to generate a result (Head et al. 2019; Kery and Myers 2018; Courtès et al. 2024). Many simple factors, like using a random-number seed (i.e., for a pseudo-random-number generator (PRNG)), are important for obtaining instantaneous repeatability. Furthermore, many mathematical operations are not guaranteed to produce identical numerical results due to floating point errors and differences in numerical stability of different implementations and hardware (Arteaga, Fuhrer, and Hoefler 2014; Schlögl, Hofer, and Böhme 2023).

Other factors, such as software version conflicts, are often thought to be factors of repeatability but often lead to conflation. For example, does capturing software versions via a container system lead to repeatability or reproducibility? We argue that it would be reproducibility as a higher-level concept in our categorization, which we will detail further in Section 3. A second distinction we make is that of instantaneous repeatability vs. repeatability over time. In this immediate section, we consider instantaneous repeatability, where the question is how to ensure repeatable results as the software/algorithm is being developed, and we find that there is surprisingly little beyond the work noted in the prior paragraph. When time is added as a factor, we consider this to be distinguishable as the maintainability rigor that we will detail in Section 2.8.

2.2 Reproducibility

Reproducibility alters repeatability by requiring that a different individual/team be able to produce the same results using the original source code and data. This is a high focus of the AI/ML community and incentivization of Open Source Software (OSS) by major conferences and paper submission questionnaires/guidelines. Current work can be divided into those that explore surface-level issues such as unquantified proposals or exact procedure reproductions, vs those that at-

[1]Indeed, our own understanding have evolved in the discovery of the wide and deep literature on scientific rigor.

tempt to quantify or better understand why a reproduction does(not) work.

Surface Reproducibility Surface-level studies of reproducibility report on the scale of the reproducibility challenge without examining whether their attempts at improving reproducibility work. The only large-scale study we are aware of found that 74% of the code released by the broader scientific community (beyond AI/ML) ran without issue (Trisovic et al. 2022). Toward remediating this in machine learning, many have proposed techniques like Docker to try and capture the exact conditions to re-run the experiments (Forde et al. 2018a,b). (Gardner et al. 2018) looked at enhancing reproducibility by standardizing data access and execution environments for MOOCs. However, the project appears to be abandoned and stresses the importance of repeatability/reproducibility over time, which we note forms the aspect of maintainability we discuss later in Section 2.8.

Reproducibility In Depth A major factor in Reproducibility, and the discovery of non-reproducible work, is errors in the original comparisons being made. There are cases where reproducibility may be strictly achievable but meaningless due to an error in the fundamental approach being taken or experimental setup. In a seminal example of metric learning, it was found that papers had multiple changes occurring simultaneously in comparison to prior baselines (new layers like Batch-Norm, optimizers, etc.) beyond just the proposed metric learning changes, which produced misleadingly large effect sizes (Musgrave, Belongie, and Lim 2020). In general, many other works have identified similar issues nuanced to the subdomain being studied (Lu, Raff, and Holt 2023; Liu et al. 2020; Chen, Belouadi, and Eger 2022; Ito et al. 2023). More serious instances have determined a subfield of research being constructed around unsound methodologies (Lin et al. 2022; Kapoor and Narayanan 2023; Hullman et al. 2022; Raff and Holt 2023).

Thematically similar to (Musgrave, Belongie, and Lim 2020) are the multiple realizations of insufficient baseline evaluation that have occurred in many works since. Such work includes studies that use similar baseline errors/lack of adjustment (Rao et al. 2022), studies that expand the set of baselines against an overly broad prior conclusion (Huang et al. 2022; Wang et al. 2022), and studies that demonstrate that decades-old methods are still competitive when given the chance to run on larger modern datasets (Liu, Hu, and Lin 2022). Another example is the effectiveness of linear models in natural language processing tasks, which are orders of magnitude faster and capable of comparable results (Lin et al. 2023). A unique aspect shown by (Chen et al. 2018) is that many improvements prescribed to one family of algorithms are actually applicable to prior approaches and would perform just as well using an "out-of-date" method. They showed this by applying improvements from seq2seq modeling to Recurrent Neural Networks and found that the improvements were still effective, allowing a Pareto improvement in combined approaches.

2.3 Replicability

Replicability concerns the ability of a different person/team to produce qualitatively similar results from the original article by writing their own code and potentially different data. The aspect of replicability is highly understudied, likely due to the challenges this aspect presents. Replicability can be subdivided into empirical replicability and theoretical replicability.

Empirical Replicability Empirical Replicability requires re-implementing a target method's code from scratch, which is a labor-intensive process. Notable work in this direction was done by (Raff 2019), who attempted to reimplement 255 papers, and computed features to quantify what properties correlated with a replicable paper. Smaller scale replications have also been performed (Belz et al. 2022), including a volunteer effort by ReproducedPapers.org collecting some (most are reproduction attempts) Replicability attempts (Yildiz et al. 2021) based on which a thorough study in IR has been performed (Wang et al. 2022). (Chen et al. 2022; Ganesan et al. 2021) identified issues with a specific common baseline method XML-CNN in multi-label learning. Famously, (Henderson et al. 2018) replicated recent reinforcement learning results and discovered various aspects, such as the seed and scale of rewards, that significantly altered the perception of improvement.

(Johnson, Pollard, and Mark 2017) Replicate studies in mortality prediction in a healthcare context, highlighting the difficulty of producing comparable results when replication also requires collecting new data of the same intrinsic nature (that is, patient data in this context). Textual descriptions presented in the original studies were found to be insufficient for collecting new data that would replicate. (Hegselmann et al. 2018) extended this observation by showing how to produce replicable data collection schemes for survival analysis against medical repositories such as SEER.

We are not aware of other work within AI/ML on empirical replicability. This state of affairs is common to other (relatively) code-free disciplines such as medicine (Ioannidis 2005) economics (Camerer et al. 2016) social sciences (Camerer et al. 2018). In these disciplines, replication studies are necessary and representative because they are the least costly way to evaluate a result. Other aspects of scientific rigor have a significantly lower barrier to entry, largely because AI/ML has a large open-source culture.

Theoretical Replicability More recent work has advanced a theoretical definition of replicability in terms of constraints on the output distribution as a function of the input distribution. (Impagliazzo et al. 2022; Kalavasis et al. 2023; Bun et al. 2023) have developed much of this foundation by showing various desirable statistical properties such as that Total Variation (TV) between outputs drawn from the same distribution using the same algorithm (i.e., a congruent definition of replicability) is equivalent to results in approximate differential privacy and robust statistics. This idea has since been expanded to bandits (Esfandiari et al. 2023), optimization (Zhang et al. 2023), clustering (Esfandiari et al. 2023), and reinforcement learning (at an exponential increase in runtime) (Karbasi et al. 2023).

Lastly, a unique approach to the question of replicability was studied by (Ahn et al. 2022), which focuses on the difference between the computational precision of floating points and the underlying symbolic math. From this perspective, they are able to suggest conditions about what statements can be rigorously tested and concluded about the math based on floating-point errors that would accumulate and cause issues otherwise.

2.4 Model Selection

Model selection deals with the common task of AI/ML papers: given two competing methods (one of which may be the paper's own proposal), how do we conclude which method is better? As the AI/ML literature has advanced significantly through the presentation of empirically "better" algorithms, it is not surprising that most historical and current work has focused on the question of model selection. This includes how to pick and evaluate criteria to decide "better", how to build benchmarks for a problem, and the process to determine "better" given criteria in a statistically sound way. There are also multiple resurgences of this issue as ML is incorporated into other fields, and comparisons that may be invalid in a new field occur as both communities begin to merge and discover what is/is not acceptable (Hoefler 2022).

Evaluation Criteria & Methodology Before selecting the "better" of one or more methods, it is necessary first to determine how the quality of a method is determined. The scope of evaluation metrics and scores is larger than that of scientific rigor, and this article is concerned with cases where an invalid or errant procedure was identified and remediated. The literature in this direction is old, starting in the late 1990s on the various pros and cons of metrics like Area Under the Curve (AUC) for evaluation (Bradley 1997; Hand 2009; Lobo, Jiménez-Valverde, and Real 2008). Likewise, work has addressed issues in scoring from leaderboards (Blum and Hardt 2015), and subtle issues in using cross-validation to produce test scores (Varma and Simon 2006; Bergmeir, Hyndman, and Koo 2018; Varoquaux 2018; Bates, Hastie, and Tibshirani 2021; Mathieu and Preux 2024). Niche examples of the evaluation concern also exist. For example, three decades of malware detection performed subtle train/test leakage by adjusting for a target false-positive rate incorrectly (Nguyen et al. 2021) and time series anomaly detection scores being overly generous to "near hits" (Kim et al. 2022).

Building Problem-Specific Benchmark Suites It is becoming increasingly popular to build benchmarks of multiple datasets, pre-prepared evaluation code, and methodology for specific problem domains (Blalock et al. 2020; Eggensperger et al. 2021; Sun et al. 2020; Saul et al. 2024; Liu et al. 2024; Ordun et al. 2021; Kebe et al. 2021). Such a benchmark construction is popular, although it has yet to evolve into a science of how to build benchmarks, with limited study at a macro level (Koch et al. 2021). Some domains may require additional thought to how methods are compared, especially when they are measuring a non-stationary objective like human preferences in Information Retreival (Breuer and Maistro 2024).

Selection Determination Much of the ML literature presents raw results and makes a nonscientifically rigorous statement of being "better" by some metric (i.e., **bold** numbers in a table are better, and our method has more bold numbers in the table). There are two approaches to developing improved comparisons.

One is to devise better statistical tests to compare two methods when a single test set is available, first seriously studied by Dietterich (1998) with many follow-up works shortly after (Alpaydin 1999; Bouckaert 2003; Bouckaert and Frank 2004). Different perspectives on this include using one test run to make a conclusion (Dror, Shlomov, and Reichart 2019), or including sources of variation in model performance (e.g., hyperparameter values) and comparing the distribution of model results (Bouthillier, Laurent, and Vincent 2019; Bouthillier et al. 2021; Cooper et al. 2021). Others have introduced computational budget for training and parameter tuning as a conditional factor that impacts the conclusion of "best" (Dodge et al. 2019).

The second option is to use multiple datasets to perform a single test of whether one algorithm is better than another (Guerrero Vázquez et al. 2001; Hull 1994; Pizarro, Guerrero, and Galindo 2002). The use of a nonparametric Wilcoxon test has been found to be effective in multiple studies (Demšar 2006; Benavoli, Corani, and Mangili 2016). (Dror et al. 2017) extended this to make a conclusion about how many datasets and which one method performs better. Other recent work has proposed using meta-analysis methods to draw conclusions about a single method tested under multiple conditions (Soboroff 2018). Notably, work using multiple datasets to make decisions based on a single evaluation metric implicitly contributes to the Adaptability question, which we explore next. Interestingly, we note the field of programming languages has also proposed quantile regression as a better method of analyzing results (de Oliveira et al. 2013).

2.5 Adaptability and its Second-Class Status

Adaptability is the study of a different person/team, using the original code but applying it to their own and different data. Very little work on scientific rigor in AI/ML focuses on Adaptability. To be clear, many prior works have studied the question of generalization in machine learning, of which there is recent evolution due to the advance of deep learning (Zhang et al. 2017). However, generalization assumes some form of intrinsic relationship (usually I.I.D.) between the training and testing distribution. Under Adaptability, there is no direct train/test split to compare. Instead, it is a question of the methodology's effectiveness on an entirely different statistical distribution at training and test time. Thus, our concern is more focused on the practical, real-world issues that enable or inhibit a *method* to generalize. Our contention is the lack of study on adaptability is one of the most glaring shortfalls in the current scientific rigor literature, with significant room for researchers to define and develop new

ways of studying the problem[2].

The work we have found can broadly be described as including adaptability to new datasets or specialized subsets to better understand the overall behavior and utility of a set of algorithms (Marchesin, Purpura, and Silvello 2020; Rahmani et al. 2022). The other work that tackles adaptability is from an HCI perspective in validating a method's utility as population preferences evolve (Roy, Maxwell, and Hauff 2022).

Though it has not been presented as a part of the literature on scientific rigor, considerable effort in the Adaptability question has been advanced by Decision Tree-based literature. In particular, the long-standing effectiveness of tree ensembles has led to numerous studies investigating the persistent efficacy of tree ensembles (Grinsztajn, Oyallon, and Varoquaux 2022; Wainberg, Alipanahi, and Frey 2016; Bagnall et al. 2020). Despite little work on the adaptability question, we note that many works in Model Selection make use of the adaptability argument as a component of their study or an otherwise latent concern.

One way that others could seek to understand adaptability is to see if they adapt to crossing a small "chasm" of change in the problem. This can be done by taking methods developed in one context and applying them to a highly related problem (ideally with a minor to no modification necessary). Two prior works that we are aware of have demonstrated surprising failures of methods that fail to cross these small chasms. (Riquelme, Tucker, and Snoek 2018) studied extensions of Thompson Sampling for reinforcement learning that work well in supervised settings, but a modest adaption to sequential decision making causes simple Thompson Sampling to outperform the various previous improvements. (Liu et al. 2021) found that the original hyperparameters for a multi-label prediction algorithm were kept when the method was adapted to a new task. Subsequent works compared to this original parameterized version rather than re-tuning to the new task. When properly accounted for, all subsequent methods failed to improve on the original method.

2.6 Label & Data Quality

Label & Data Quality is focused on the reliability of data and label acquisition, error rates, and working to understand how they occur, detect them, or work around them. The distinction we make from research in inferring a single label from labelers is that scientific rigor is concerned with the process of how labels are collected, defined, and have impacted research conclusions (e.g., inferring a 99% accurate model when labels have a 5% noise level would imply a failure in process). Many works today identify these issues long after dataset construction, in part due to the high accuracies now being achieved, making the errors more pronounced. For example, the process for deriving labels of ImageNet had rules incongruent with the nature of the data (e.g. assuming that only one class is present) and error-prone steps in the labeling pipeline (Beyer et al. 2020). Label quality issues also include leakage from the train / test set (Barz and Denzler 2020).

Some of the most insightful research results have come from replicating dataset construction and labeling processes for prior datasets and then characterizing and discovering why differences in results occur. This includes detection cases where recreation is implicitly made more challenging than the original dataset (Engstrom et al. 2020). Although there is a long history of research inferring a single correct label from multiple labelers (Whitehill et al. 2009; Lin, Mausam, and Weld 2014; Ratner et al. 2016; Yoshimura, Baba, and Kashima 2017; Ratner et al. 2020), this literature is generally not framed as a scientific rigor issue. While these methods have been utilized in work from a rigor perspective (Beyer et al. 2020), we are not aware of work that bridges a longitudinal study of the replicability of these various label inference procedures.

2.7 Meta and Incentives

Very few papers have studied incentives for scientific rigor. The study of the scientific process itself is often termed metascience and, when applied to AI/ML research, would fall into this category. Such research could include basic studies of incentives, drivers of scientific rigor, and surveys across various AI/ML research domains. Sample studies focused on drivers such as the rate of data and code sharing in computational linguistics (Wieling, Rawee, and van Noord 2018) and the use of statistical testing (Dror et al. 2018). Related work has found that code sharing and replica research are correlated with higher citations (Raff 2022; Obadage, Rajtmajer, and Wu 2024), although most meta-studies have looked at the rate of code sharing in their sub-disciplines (McDermott et al. 2021; Olszewski et al. 2023; Arvan, Pina, and Parde 2022; Arvan, Doğruöz, and Parde 2023; Cavenaghi et al. 2023). A unique aspect of code availability is studied by (Storks et al. 2023), who perform a user study with students on the time and difficulty factors for students to reproduce the results of three NLP papers. Another study focuses on how evaluation and comparison practices evolve throughout the Machine Translation community (Marie, Fujita, and Rubino 2021). The last work we are aware of challenged the treatment of replicability as a binary "yes/no" question and instead suggested a survival model, where replicability is a function of time/effort (Raff 2021) and quantifying a reproducibility score (Belz, Popovic, and Mille 2022).

2.8 Maintainability

Maintainability is similar to Repeatability, in that we are concerned with producing the same results with the original authors (though new users could also occur) using the original code and data. The key difference that distinguishes maintainability is that time is a factor, as the ability to repeat results degrades over time as nuances of labels (Inel, Draws, and Aroyo 2023) or dependency versions change (Connolly et al. 2023)[3]. Maintainability can also deal with the code

[2] Anecdotally, we have had significant trouble publishing work attempting to tackle adaptability problems in other sub-domains when we were trying to use it for real-world needs (Raff, McLean, and Holt 2023).

[3] In software development, this notion is often termed "bit rot".

itself changing over time. The focus on the aspect of maintainability within AI/ML was started by the seminal work of (Sculley et al. 2015). A key area of maintainability deals with adapting known "code smells" while considering ML-specific concerns and factors that practitioner surveys consider most important (Gesi et al. 2022). Another key area of maintainability is the quality of the results as the code itself changes. It is well known that scientific algorithms may produce different results by different (but supposedly equivalent) implementations (Hatton 1993). Multiple studies have found that AI/ML is no exception to this history, with large and statistically significant changes in accuracy when using allegedly equivalent algorithms and changing just the implementation or the runtime platform (e.g. GPU hardware) (Coakley, Kirkpatrick, and Gundersen 2022; Gundersen, Shamsaliei, and Isdahl 2022; Pham et al. 2020; Zhuang et al. 2021). (Zhou, Chen, and Lipton 2023) found that in many medical time series tasks, it may be beneficial to train on all historical data in some cases vs. training a sliding window of recent data. They also looked at models that experienced "shocks" of sudden degradation in time.

The study of maintainability is surprisingly minimal in our community despite the rapid adoption, abandonment, and evolution of frameworks used within the field. Torch, Tensorflow, JAX, Theano, and many more frameworks have come and gone through major revisions over time. These changes and re-implementation of algorithms are fertile ground for maintenance issues and, thus, their study, which directly impacts researchers and the developers of these frameworks. Studying how to build maintinable code in AI/ML is still nascent (Gilbertson et al. 2024; Papi et al. 2024)

3 Connections between Rigor Types

Having defined a set of eight rigor types that are being worked on, we further elaborate on our perception of connections between these rigors. In particular, there are direct and indirect relationships, which are summarized in Figure 1 with solid and dashed lines, respectively.

3.1 Direct Relationships

The most obvious, and intuitive connections are from repeatability to reproducibility to repeatability, as each requires a progressive step of difficulty from the prior. If a single person/team cannot repeat their own experiments, there is no reason to believe that a different person with the same code would be able to reproduce those results. Extended further, if they cannot reproduce the results with the original code, there is no special reason to believe that by writing their own code or using different data, they would be able to replicate the results.

Less obvious are the interactions between maintainability, repeatability, and replicability. The first is the two-way relationship between repeatability and maintainability. If an AI/ML system is not repeatable, it cannot be maintainable, as repeatability is the property that we want to maintain. Similarly, if it cannot be maintained, it may not be repeatable *over time*. A simple case is the use of Docker to gain repeatability, which is predicated on the repeatability of Docker

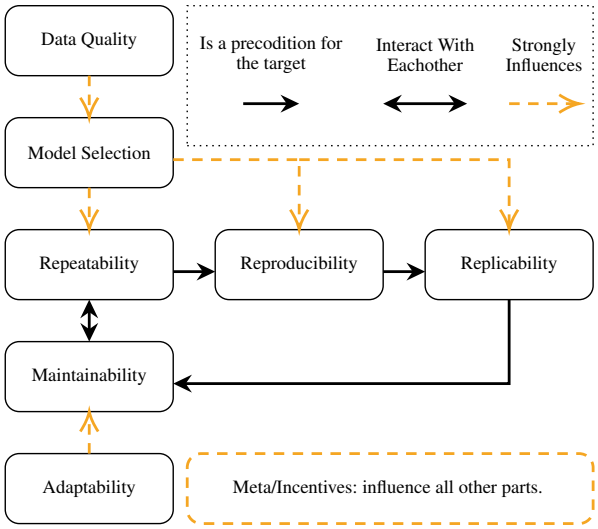

Figure 1: Connections on how rigor types influence each other. Solid lines indicate hard dependencies, while dashed lines show influencing effects.

containers. This assumption is true on short time horizons, but changes in software, hardware, and eventually deprecation of tools like Docker itself do not make it true in perpetuity. The time-based evolution that maintainability requires then directly implies the replicability of a method. If a system is replicable, meaning that the code or data can change as well as the people, it satisfies the requirement of maintainability over a single point in time. Thus, maintainability involves iterated replicability over time and instantaneous repeatability at any point in time.

3.2 Indirect Relationships

Beyond the general influence of meta- and incentives-based rigor having a relationship to all parts of scientific rigor, we can further draw other connections that are of particular note. The most straightforward of these is that of model selection on repeatability, reproducibility, and replicability, each of which will often incorporate the model selection task as part of the motivation for why the proposed work should be used (i.e., it was demonstrated to "be better" than something prior). Thus, by its nature, different approaches to model selection will influence each. For example, the use of random search as a hyperparameter tuning method (Bergstra and Bengio 2012) is potentially a hindrance to replicability due to higher variance, even if it is easily repeatable and reproducible given the original code with initial seed values for the pseudorandom number generator.

Upstream from this concern is then label and data quality, which will influence what features are selected. This is particularly notable as many datasets reach high accuracies where "errors" in the model's predictions are discovered to be either 1) correct and that the test data were mislabeled or 2) that the test instance was inherently ambiguous (Barz and Denzler 2020). This creates a new kind of noise in the selection process, and can thus alter conclusions on the merits

of what is considered. This is particularly true for the eventual selection of the downstream model under replicability, where the data in use may be different.

Finally, we note that a method that is adaptable is more likely to be maintainable. The nature of one method being effective in many others is the observation that many small details on the implementation can vary, while still producing quantitatively similar results, an often observed phenomenon in decision tree literature (Quinlan 1993; Breiman et al. 1984; Quinlan 2006; Raff 2017). This provides some inherent "robustness" to issues that often cause maintainability problems, such as changes in low-level libraries like BLAS/LAPACK or new hardware.

4 Conclusions

We have synthesized eight current directions in the literature of scientific rigor for machine learning, disentangling them from the commonly repeated moniker of "reproducibility" and thus quantified the proportion of each type as studied today. These rigor types have been further characterized by their interactions/dependencies with each other.

References

Ahn, K.; Jain, P.; Ji, Z.; Kale, S.; Netrapalli, P.; and Shamir, G. I. 2022. Reproducibility in Optimization: Theoretical Framework and Limits. 1–51. ArXiv: 2202.04598.

Alpaydin, E. 1999. Combined 5 × 2 cv F Test for Comparing Supervised Classification Learning Algorithms. *Neural Comput.*, 11(9): 1885–1892. Publisher: MIT Press Place: Cambridge, MA, USA.

Arteaga, A.; Fuhrer, O.; and Hoefler, T. 2014. Designing Bit-Reproducible Portable High-Performance Applications. In *2014 IEEE 28th International Parallel and Distributed Processing Symposium*, 1235–1244.

Arvan, M.; Doğruöz, A. S.; and Parde, N. 2023. Investigating Reproducibility at Interspeech Conferences: A Longitudinal and Comparative Perspective. In *Proc. INTERSPEECH 2023*, 3929–3933.

Arvan, M.; Pina, L.; and Parde, N. 2022. Reproducibility in Computational Linguistics: Is Source Code Enough? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2350–2361. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Bagnall, A.; Flynn, M.; Large, J.; Line, J.; Bostrom, A.; and Cawley, G. 2020. Is rotation forest the best classifier for problems with continuous features? arXiv:1809.06705.

Barz, B.; and Denzler, J. 2020. Do We Train on Test Data? Purging CIFAR of Near-Duplicates. *Journal of Imaging*, 6(6): 41. ArXiv: 1902.00423.

Bates, S.; Hastie, T.; and Tibshirani, R. 2021. Cross-validation: what does it estimate and how well does it do it? *arXiv*, 1–38. ArXiv: 2104.00673.

Belz, A.; Popovic, M.; and Mille, S. 2022. Quantified Reproducibility Assessment of NLP Results. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16–28. Dublin, Ireland: Association for Computational Linguistics.

Belz, A.; Shimorina, A.; Popović, M.; and Reiter, E. 2022. The 2022 ReproGen Shared Task on Reproducibility of Evaluations in NLG: Overview and Results. In Shaikh, S.; Ferreira, T.; and Stent, A., eds., *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, 43–51. Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics.

Benavoli, A.; Corani, G.; and Mangili, F. 2016. Should We Really Use Post-Hoc Tests Based on Mean-Ranks? *Journal of Machine Learning Research*, 17(5): 1–10.

Bergmeir, C.; Hyndman, R. J.; and Koo, B. 2018. A Note on the Validity of Cross-validation for Evaluating Autoregressive Time Series Prediction. *Computational Statistics & Data Analysis*, 120(C): 70–83. Publisher: Elsevier Science Publishers B. V. Place: Amsterdam, The Netherlands, The Netherlands.

Bergstra, J.; and Bengio, Y. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13: 281–305. ISBN: 1532-4435.

Beyer, L.; Hénaff, O. J.; Kolesnikov, A.; Zhai, X.; and Oord, A. v. d. 2020. Are we done with ImageNet? *arXiv*. ArXiv: 2006.07159.

Blalock, D.; Gonzalez Ortiz, J. J.; Frankle, J.; and Guttag, J. 2020. What is the State of Neural Network Pruning? In *Proceedings of Machine Learning and Systems 2020*, 129–146.

Blum, A.; and Hardt, M. 2015. The Ladder: A Reliable Leaderboard for Machine Learning Competitions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 1006–1014. JMLR Workshop and Conference Proceedings.

Bouckaert, R. R. 2003. Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, 51–58. Washington, DC, USA: AAAI Press. ISBN 978-1-57735-189-4.

Bouckaert, R. R.; and Frank, E. 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In Dai, H.; Srikant, R.; and Zhang, C., eds., *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, 3–12. Berlin, Heidelberg: Springer. ISBN 978-3-540-24775-3.

Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Sepah, N.; Raff, E.; Madan, K.; Voleti, V.; Kahou, S. E.; Michalski, V.; Serdyuk, D.; Arbel, T.; Pal, C.; Varoquaux, G.; and Vincent, P. 2021. Accounting for Variance in Machine Learning Benchmarks. In *Machine Learning and Systems (MLSys)*. ArXiv: 2103.03098.

Bouthillier, X.; Laurent, C.; and Vincent, P. 2019. Unreproducible Research is Reproducible. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 725–734. PMLR. Series Title: Proceedings of Machine Learning Research.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7): 1145–1159. Publisher: Pergamon.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. 1984. *Classification and Regression Trees*. CRC press.

Breuer, T.; and Maistro, M. 2024. Toward Evaluating the Reproducibility of Information Retrieval Systems with Simulated Users. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*, ACM REP '24, 25–29. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705304.

Bun, M.; Gaboardi, M.; Hopkins, M.; Impagliazzo, R.; Lei, R.; Pitassi, T.; Sivakumar, S.; and Sorrell, J. 2023. Stability Is Stable: Connections between Replicability, Privacy, and Adaptive Generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, 520–527. New York, NY, USA: Association for Computing Machinery. ISBN 9781450399135.

Camerer, C. F.; Dreber, A.; Forsell, E.; Ho, T.-H.; Huber, J.; Johannesson, M.; Kirchler, M.; Almenberg, J.; Altmejd, A.; Chan, T.; Heikensten, E.; Holzmeister, F.; Imai, T.; Isaksson, S.; Nave, G.; Pfeiffer, T.; Razen, M.; and Wu, H. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280): 1433–1436. Publisher: University Library of Munich, Germany.

Camerer, C. F.; Dreber, A.; Holzmeister, F.; Ho, T. H.; Huber, J.; Johannesson, M.; Kirchler, M.; Nave, G.; Nosek, B. A.; Pfeiffer, T.; Altmejd, A.; Buttrick, N.; Chan, T.; Chen, Y.; Forsell, E.; Gampa, A.; Heikensten, E.; Hummer, L.; Imai, T.; Isaksson, S.; Manfredi, D.; Rose, J.; Wagenmakers, E. J.; and Wu, H. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9): 637–644.

Cavenaghi, E.; Sottocornola, G.; Stella, F.; and Zanker, M. 2023. A Systematic Study on Reproducibility of Reinforcement Learning in Recommendation Systems. *ACM Trans. Recomm. Syst.*, 1(3).

Chen, M. X.; Firat, O.; Bapna, A.; Johnson, M.; Macherey, W.; Foster, G.; Jones, L.; Schuster, M.; Shazeer, N.; Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Chen, Z.; Wu, Y.; and Hughes, M. 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 76–86. Melbourne, Australia: Association for Computational Linguistics.

Chen, S.-A.; Liu, J.-j.; Yang, T.-H.; Lin, H.-T.; and Lin, C.-J. 2022. Even the Simplest Baseline Needs Careful Reinvestigation: A Case Study on XML-CNN. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1987–2000. Seattle, United States: Association for Computational Linguistics.

Chen, Y.; Belouadi, J.; and Eger, S. 2022. Reproducibility Issues for BERT-based Evaluation Metrics. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2965–2989. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Coakley, K.; Kirkpatrick, C. R.; and Gundersen, O. E. 2022. Examining the Effect of Implementation Factors on Deep Learning Reproducibility. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, 397–398.

Connolly, A.; Hellerstein, J.; Alterman, N.; Beck, D.; Fatland, R.; Lazowska, E.; Mandava, V.; and Stone, S. 2023. Software Engineering Practices in Academia: Promoting the 3Rs—Readability, Resilience, and Reuse. *Harvard Data Science Review*, 5(2). Https://hdsr.mitpress.mit.edu/pub/f0f7h5cu.

Cooper, A. F.; Lu, Y.; Forde, J. Z.; and De Sa, C. 2021. Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In *NeurIPS*. ArXiv: 2102.03034.

Courtès, L.; Sample, T.; Zacchiroli, S.; and Tournier, S. 2024. Source Code Archiving to the Rescue of Reproducible Deployment. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*, ACM REP '24, 36–45. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705304.

de Oliveira, A. B.; Fischmeister, S.; Diwan, A.; Hauswirth, M.; and Sweeney, P. F. 2013. Why You Should Care about Quantile Regression. *SIGPLAN Not.*, 48(4): 207–218.

Demšar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7: 1–30. Publisher: JMLR.org.

Dietterich, T. G. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.*, 10(7): 1895–1923. Publisher: MIT Press Place: Cambridge, MA, USA.

Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; and Smith, N. A. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of EMNLP*, 2185–2194. ArXiv: 1909.03004 Issue: 2.

Dror, R.; Baumer, G.; Bogomolov, M.; and Reichart, R. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5: 471–486.

Dror, R.; Baumer, G.; Shlomov, S.; and Reichart, R. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1383–1392.

Dror, R.; Shlomov, S.; and Reichart, R. 2019. Deep Dominance - How to Properly Compare Deep Neural Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785. Association for Computational Linguistics.

Drummond, C. 2009. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada,2009*. Series Title: Evaluation Methods for Machine Learning Workshop, the 26th ICML, June 14-18, 2009, Montreal, Canada.

Drummond, C. 2018. Reproducible research: a minority opinion. *Journal of Experimental and Theoretical Artificial Intelligence*, 30(1): 1–11. Publisher: Taylor & Francis.

Eggensperger, K.; Müller, P.; Mallik, N.; Feurer, M.; Sass, R.; Klein, A.; Awad, N.; Lindauer, M.; and Hutter, F. 2021. HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 1–36. ArXiv: 2109.06716 Issue: NeurIPS.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Steinhardt, J.; and Madry, A. 2020. Identifying Statistical Bias in Dataset Replication. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 2922–2932. Series Title: Proceedings of Machine Learning Research.

Esfandiari, H.; Karbasi, A.; Mirrokni, V.; Velegkas, G.; and Zhou, F. 2023. Replicable Clustering. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Forde, J.; Head, T.; Holdgraf, C.; Panda, Y.; Perez, F.; Nalvarte, G.; Ragan-kelley, B.; and Sundell, E. 2018a. Reproducible Research Environments with repo2docker. In *Reproducibility in ML Workshop, ICML'18*.

Forde, J. Z.; Bussonnier, M.; Fortin, F.-A.; Granger, B. E.; Head, T. D.; Holdgraf, C.; Ivanov, P.; Kelley, K.; Pacer, M. D.; Panda, Y.; Pérez, F.; Nalvarte, G.; Ragan-Kelley, B.; Sailer, Z. R.; Silvester, S.; Sundell, E.; and Willing, C. 2018b. Reproducing Machine Learning Research on Binder. In *Machine Learning Open Source Software 2018: Sustainable communities*.

Ganesan, A.; Gao, H.; Gandhi, S.; Raff, E.; Oates, T.; Holt, J.; and McLean, M. 2021. Learning with Holographic Reduced Representations. In *Advances in Neural Information Processing Systems*. ArXiv: 2109.02157.

Gardner, J.; Brooks, C.; Andres, J. M.; and Baker, R. S. 2018. MORF: A Framework for Predictive Modeling and Replication At Scale With Privacy-Restricted MOOC Data. In *2018 IEEE International Conference on Big Data (Big Data)*, 3235–3244.

Gardner, J.; Brooks, C.; and Baker, R. S. 2018. Enabling End-To-End Machine Learning Replicability : A Case Study in Educational Data Mining. In *Reproducibility in ML Workshop, ICML'18*.

Gesi, J.; Liu, S.; Li, J.; Ahmed, I.; Nagappan, N.; Lo, D.; de Almeida, E. S.; Kochhar, P. S.; and Bao, L. 2022. Code Smells in Machine Learning Systems. ArXiv:2203.00803 [cs].

Gilbertson, C.; Mundt, M.; Teves, J.; Toribio, S.; and Milewicz, R. 2024. Towards Evidence-Based Software Quality Practices for Reproducibility: Practices and Aligned Software Qualities. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*, ACM REP '24, 52–63. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705304.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? In *NeurIPS*. ArXiv:2207.08815 [cs, stat].

Guerrero Vázquez, E.; Yañez Escolano, A.; Galindo Riaño, P.; and Pizarro Junquera, J. 2001. Repeated Measures Multiple Comparison Procedures Applied to Model Selection in Neural Networks. In *Bio-Inspired Applications of Connectionism*, Lecture Notes in Computer Science, 88–95. Springer. ISBN 978-3-540-45723-7.

Gundersen, O. E.; and Kjensmo, S. 2018. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, 1644–1651.

Gundersen, O. E.; Shamsaliei, S.; and Isdahl, R. J. 2022. Do machine learning platforms provide out-of-the-box reproducibility? *Future Generation Computer Systems*, 126: 34–47.

Hand, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1): 103–123.

Hatton, L. 1993. The quality and reliability of scientific software. *Transactions on Information and Communications Technologies*, 4.

Head, A.; Hohman, F.; Barik, T.; Drucker, S. M.; and DeLine, R. 2019. Managing Messes in Computational Notebooks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5970-2.

Hegselmann, S.; Gruelich, L.; Varghese, J.; and Dugas, M. 2018. Reproducible Survival Prediction with SEER Cancer Data. In Doshi-Velez, F.; Fackler, J.; Jung, K.; Kale, D.; Ranganath, R.; Wallace, B.; and Wiens, J., eds., *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85, 49–66. Palo Alto, California: PMLR. Series Title: Proceedings of Machine Learning Research.

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning That Matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Hoefler, T. 2022. Benchmarking Data Science: 12 Ways to Lie With Statistics and Performance on Parallel Computers. *Computer*, 55(8): 49–56.

Huang, J.; Oosterhuis, H.; Cetinkaya, B.; Rood, T.; and de Rijke, M. 2022. State Encoders in Reinforcement Learning for Recommendation: A Reproducibility Study. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2738–2748. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Hull, D. A. 1994. *INFORMATION RETRIEVAL USING STATISTICAL CLASSIFICATION*. Ph.D. thesis, stanford university.

Hullman, J.; Kapoor, S.; Nanayakkara, P.; Gelman, A.; and Narayanan, A. 2022. The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psy-

chology and Machine Learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 335–348. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Impagliazzo, R.; Lei, R.; Pitassi, T.; and Sorrell, J. 2022. Reproducibility in Learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, 818–831. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392648.

Inel, O.; Draws, T.; and Aroyo, L. 2023. Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1): 51–64.

Ioannidis, J. P. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2): 218–228.

Ito, T.; Fang, Q.; Mosteiro, P.; Gatt, A.; and van Deemter, K. 2023. Challenges in Reproducing Human Evaluation Results for Role-Oriented Dialogue Summarization. In Belz, A.; Popović, M.; Reiter, E.; Thomson, C.; and Sedoc, J., eds., *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, 97–123. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Johnson, A. E. W.; Pollard, T. J.; and Mark, R. G. 2017. Reproducibility in critical care: a mortality prediction case study. In Doshi-Velez, F.; Fackler, J.; Kale, D.; Ranganath, R.; Wallace, B.; and Wiens, J., eds., *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, 361–376. PMLR.

Kalavasis, A.; Karbasi, A.; Moran, S.; and Velegkas, G. 2023. Statistical Indistinguishability of Learning Algorithms. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 15586–15622. PMLR.

Kapoor, S.; and Narayanan, A. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9): 100804.

Karbasi, A.; Velegkas, G.; Yang, L.; and Zhou, F. 2023. Replicability in Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kebe, G. Y.; Higgins, P.; Jenkins, P.; Darvish, K.; Barron, R.; Winder, J.; Engel, D.; Raff, E.; Ferraro, F.; Matuszek, C.; and Hamilton, B. A. 2021. A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning. In *NeurIPS*.

Kery, M. B.; and Myers, B. A. 2018. Interactions for Untangling Messy History in a Computational Notebook. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 147–155. ISSN: 1943-6106.

Kim, S.; Choi, K.; Choi, H.-s.; and Lee, B. 2022. Towards a Rigorous Evaluation of Time-series Anomaly Detection. In *AAAI*. ArXiv: 2109.05257v2.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *NeurIPS Dataset & Benchmark track*. arXiv. ArXiv:2112.01716 [cs, stat].

Lin, C.; Mausam; and Weld, D. 2014. To Re(label), or Not To Re(label). *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2: 151–158.

Lin, J. 2022. Building a Culture of Reproducibility in Academic Research.

Lin, L.-C.; Liu, C.-H.; Chen, C.-M.; Hsu, K.-C.; Wu, I.-F.; Tsai, M.-F.; and Lin, C.-J. 2022. On the Use of Unrealistic Predictions in Hundreds of Papers Evaluating Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(77): 7479–7487.

Lin, Y.-C.; Chen, S.-A.; Liu, J.-J.; and Lin, C.-J. 2023. Linear Classifier: An Often-Forgotten Baseline for Text Classification. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1876–1888. Toronto, Canada: Association for Computational Linguistics.

Liu, C.; Gao, C.; Xia, X.; Lo, D.; Grundy, J.; and Yang, X. 2020. On the Replicability and Reproducibility of Deep Learning in Software Engineering. 1(1): 1–34. ArXiv: 2006.14244.

Liu, C.; Saul, R.; Sun, Y.; Raff, E.; Fuchs, M.; Pantano, T. S.; Holt, J.; and Micinski, K. 2024. Assemblage: Automatic Binary Dataset Construction for Machine Learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Liu, J.-J.; Yang, T.-H.; Chen, S.-A.; and Lin, C.-J. 2021. Parameter Selection: Why We Should Pay More Attention to It. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 825–830. Online: Association for Computational Linguistics.

Liu, Y.; Hu, C.; and Lin, J. 2022. Another Look at Information Retrieval as Statistical Translation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2749–2754. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Lobo, J. M.; Jiménez-Valverde, A.; and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2): 145–151. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-8238.2007.00358.x.

Lu, F.; Raff, E.; and Holt, J. 2023. A Coreset Learning Reality Check. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*. arXiv. ArXiv:2301.06163 [cs, stat].

Marchesin, S.; Purpura, A.; and Silvello, G. 2020. Focal elements of neural information retrieval models. An outlook through a reproducibility study. *Information Processing & Management*, 57(6): 102109.

Marie, B.; Fujita, A.; and Rubino, R. 2021. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7297–7306. Online: Association for Computational Linguistics.

Mathieu, T.; and Preux, P. 2024. Statistical comparison in empirical computer science with minimal computation usage. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*, ACM REP '24, 20–24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705304.

Matsui, B. M. A.; and Goya, D. H. 2022. MLOps: A Guide to its Adoption in the Context of Responsible AI. In *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, 45–49.

McDermott, M. B. A.; Wang, S.; Marinsek, N.; Ranganath, R.; Foschini, L.; and Ghassemi, M. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586): eabb1655.

Musgrave, K.; Belongie, S.; and Lim, S.-N. 2020. A Metric Learning Reality Check. In *ECCV*. ArXiv: 2003.08505.

Nguyen, A. T.; Raff, E.; Nicholas, C.; and Holt, J. 2021. Leveraging Uncertainty for Improved Static Malware Detection Under Extreme False Positive Constraints. In *IJCAI-21 1st International Workshop on Adaptive Cyber Defense*. ArXiv: 2108.04081.

Obadage, R. R.; Rajtmajer, S. M.; and Wu, J. 2024. SHORT: Can citations tell us about a paper's reproducibility? A case study of machine learning papers. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*, ACM REP '24, 96–100. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705304.

Olszewski, D.; Lu, A.; Stillman, C.; Warren, K.; Kitroser, C.; Pascual, A.; Ukirde, D.; Butler, K.; and Traynor, P. 2023. "Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, 3433–3459. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700507.

Ordun, C.; Cha, A. N.; Raff, E.; Gaskin, B.; Hanson, A.; Rule, M.; Purushotham, S.; and Gulley, J. L. 2021. Intelligent Sight and Sound : A Chronic Cancer Pain Dataset. In *NeurIPS*.

Paganini, M.; and Forde, J. Z. 2020. dagger: A Python Framework for Reproducible Machine Learning Experiment Orchestration. *arXiv*. ArXiv: 2006.07484.

Papi, S.; Gaido, M.; Pilzer, A.; and Negri, M. 2024. When Good and Reproducible Results are a Giant with Feet of Clay: The Importance of Software Quality in NLP. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3657–3672. Bangkok, Thailand: Association for Computational Linguistics.

Pham, H. V.; Qian, S.; Wang, J.; Lutellier, T.; Rosenthal, J.; Tan, L.; Yu, Y.; and Nagappan, N. 2020. Problems and Opportunities in Training Deep Learning Software Systems: An Analysis of Variance. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, 771–783. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6768-4. Series Title: ASE '20.

Pizarro, J.; Guerrero, E.; and Galindo, P. L. 2002. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48(1): 155–173.

Plesser, H. E. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in neuroinformatics*, 11: 76. Publisher: Frontiers Media S.A.

Publio, G. C.; Esteves, D.; and Zafar, H. 2018. ML-Schema : Exposing the Semantics of Machine Learning with Schemas and Ontologies. In *Reproducibility in ML Workshop, ICML'18*.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*, volume 1. Morgan Kaufmann. ISBN 1-55860-238-0. Series Title: Morgan Kaufmann series in {M}achine {L}earning Publication Title: Morgan Kaufmann San Mateo California Issue: 3 ISSN: 08856125.

Quinlan, J. R. 2006. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4(1996): 77–90.

Raff, E. 2017. JSAT: Java Statistical Analysis Tool, a Library for Machine Learning. *Journal of Machine Learning Research*, 18(23): 1–5.

Raff, E. 2019. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *NeurIPS*. ArXiv: 1909.06674.

Raff, E. 2021. Research Reproducibility as a Survival Analysis. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*. ArXiv: 2012.09932.

Raff, E. 2022. Does the Market of Citations Reward Reproducible Work? In *ML Evaluation Standards Workshop at ICLR 2022*.

Raff, E.; and Farris, A. L. 2022. A Siren Song of Open Source Reproducibility. In *ML Evaluation Standards Workshop at ICLR 2022*.

Raff, E.; and Holt, J. 2023. Reproducibility in Multiple Instance Learning: A Case For Algorithmic Unit Tests. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Raff, E.; McLean, M.; and Holt, J. 2023. *An Easy Rejection Sampling Baseline via Gradient Refined Proposals*. IOS Press. ISBN 9781643684376.

Rahmani, H. A.; Naghiaei, M.; Dehghan, M.; and Aliannejadi, M. 2022. Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2755–2764. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Rao, J.; Wang, F.; Ding, L.; Qi, S.; Zhan, Y.; Liu, W.; and Tao, D. 2022. Where Does the Performance Improvement Come From? - A Reproducibility Concern about Image-Text Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2727–2737. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Ratner, A.; Bach, S. H.; Ehrenberg, H.; Fries, J.; Wu, S.; and Ré, C. 2020. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2): 709–730.

Ratner, A. J.; De Sa, C. M.; Wu, S.; Selsam, D.; and Ré, C. 2016. Data Programming: Creating Large Training Sets, Quickly. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*, 3567–3575. Curran Associates, Inc.

Riquelme, C.; Tucker, G.; and Snoek, J. 2018. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In *International Conference on Learning Representations*. arXiv.

Roy, N.; Maxwell, D.; and Hauff, C. 2022. Users and Contemporary SERPs: A (Re-)Investigation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2765–2775. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Saul, R.; Liu, C.; Fleischmann, N.; Zak, R. J.; Micinski, K.; Raff, E.; and Holt, J. 2024. Is Function Similarity Over-Engineered? Building a Benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Schlögl, A.; Hofer, N.; and Böhme, R. 2023. Causes and Effects of Unanticipated Numerical Deviations in Neural Network Inference Frameworks. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 56095–56107. Curran Associates, Inc.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, 2503–2511. Cambridge, MA, USA: MIT Press. Series Title: NIPS'15.

Sculley, D.; Snoek, J.; Wiltschko, A.; and Rahimi, A. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor.

Soboroff, I. 2018. Meta-Analysis for Retrieval Experiments Involving Multiple Test Collections. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 713–722. New York, NY, USA: ACM. ISBN 978-1-4503-6014-2. Series Title: CIKM '18.

Storks, S.; Yu, K.; Ma, Z.; and Chai, J. 2023. NLP Reproducibility For All: Understanding Experiences of Beginners. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10199–10219. Toronto, Canada: Association for Computational Linguistics.

Sun, Z.; Yu, D.; Fang, H.; Yang, J.; Qu, X.; Zhang, J.; and Geng, C. 2020. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In *Fourteenth ACM Conference on Recommender Systems*, 23–32. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7583-2. Series Title: RecSys '20.

Tatman, R.; Vanderplas, J.; and Dane, S. 2018. A Practical Taxonomy of Reproducibility for Machine Learning Research. In *Reproducibility in ML Workshop, ICML'18*.

Trisovic, A.; Lau, M. K.; Pasquier, T.; and Crosas, M. 2022. A large-scale study on research code quality and execution. *Scientific Data*, 9(1): 60. Number: 1 Publisher: Nature Publishing Group.

Varma, S.; and Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7: 91. Place: London.

Varoquaux, G. 2018. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180: 68–77.

Vollmer, S.; Mateen, B. A.; Bohner, G.; Király, F. J.; Ghani, R.; Jonsson, P.; Cumbers, S.; Jonas, A.; McAllister, K. S. L.; Myles, P.; Grainger, D.; Birse, M.; Branson, R.; Moons, K. G. M.; Collins, G. S.; Ioannidis, J. P. A.; Holmes, C.; and Hemingway, H. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, 368.

Wainberg, M.; Alipanahi, B.; and Frey, B. J. 2016. Are Random Forests Truly the Best Classifiers? *Journal of Machine Learning Research*, 17(110): 1–5.

Wang, X.; MacAvaney, S.; Macdonald, C.; and Ounis, I. 2022. An Inspection of the Reproducibility and Replicability of TCT-ColBERT. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 2790–2800. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8732-3.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J.; and Ruvolo, P. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Wieling, M.; Rawee, J.; and van Noord, G. 2018. Squib: Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics*, 44(4): 641–649. Place: Cambridge, MA Publisher: MIT Press.

Yildiz, B.; Hung, H.; Krijthe, J. H.; Liem, C. C. S.; Loog, M.; Migut, G.; Oliehoek, F. A.; Panichella, A.; Pawełczak, P.; Picek, S.; de Weerdt, M.; and van Gemert, J. 2021. ReproducedPapers.org: Openly Teaching and Structuring Machine Learning Reproducibility. In *Reproducible Research in Pattern Recognition*, Lecture Notes in Computer Science, 3–11. ISBN 978-3-030-76423-4.

Yoshimura, K.; Baba, Y.; and Kashima, H. 2017. Quality Control for Crowdsourced Multi-label Classification Using

RAkEL. In Liu, D.; Xie, S.; Li, Y.; Zhao, D.; and El-Alfy, E.-S. M., eds., *Neural Information Processing*, Lecture Notes in Computer Science, 64–73. Cham: Springer International Publishing. ISBN 978-3-319-70087-8.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*. ArXiv: 1611.03530v2.

Zhang, L.; YANG, J.; Karbasi, A.; and He, N. 2023. Optimal Guarantees for Algorithmic Reproducibility and Gradient Complexity in Convex Optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhou, H.; Chen, Y.; and Lipton, Z. 2023. Evaluating Model Performance in Medical Datasets Over Time. In Mortazavi, B. J.; Sarker, T.; Beam, A.; and Ho, J. C., eds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, 498–508. PMLR.

Zhuang, D.; Zhang, X.; Song, S. L.; and Hooker, S. 2021. Randomness In Neural Network Training: Characterizing The Impact of Tooling. *arXiv*. ArXiv: 2106.11872.