A UCSD View on Replication and Reproducibility for CPS & IoT

Alex Yen, Bryse Flowers, Wenshan Luo, Nitish Nagesh,
Peter Tueller, Ryan Kastner, and Pat Pannuto
University of California San Diego

ABSTRACT

Reproducibility and replicability (R&R) are important for research. Many communities are beginning efforts to reward, incentivize, and highlight projects as a motive to adopt R&R practices. This is clearly a good direction – we should all aim to make our research sound, replicable, and reproducible. Yet, this involves a lot of effort to document, debug, and generally make the systems that we build more usable. Interfacing with the *Physical* world and building custom *Things* exacerbates these challenges. Therein lies the dilemma: how does the CPS/IoT community reward and incentivize R&R efforts? This paper looks into the question of R&R in CPS/IoT. We survey efforts in other fields spanning computing to healthcare and highlight similarities and differences to CPS/IoT. We then discuss several exemplar CPS/IoT projects related to UCSD's research and highlight the R&R efforts in these projects, the potential ways that they could be improved, and best practices. We finish with recommendations and insights for R&R tailored to the CPS/IoT community.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**.

KEYWORDS

Reproducibility, replication, open science

ACM Reference Format:
Alex Yen, Bryse Flowers, Wenshan Luo, Nitish Nagesh,, Peter Tueller, Ryan Kastner, and Pat Pannuto. 2021. A UCSD View on Replication and Reproducibility for CPS & IoT. In *Benchmarking Cyber-Physical Systems and Internet of Things (CPS-IoTBench2021), May 18, 2021, Nashville, TN, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3458473.3458821

1 INTRODUCTION

There is a move towards reproducibility and replicability (R&R) in science and engineering. In practice, this manifests as a mixture of top-down directives, such as open data mandates from funding agencies, and bottom-up initiatives, such as the early conference artifact programs [13]. The former began with sub-communities producing open datasets and demonstrating their value, and the latter is now evolving into a standard, with the release of an ACM-wide badging and artifact proposal late last year [11].

CPS-IoTBench2021, May 18, 2021, Nashville, TN, USA

ACM ISBN 978-1-4503-8439-1/21/05.
https://doi.org/10.1145/3458473.3458821

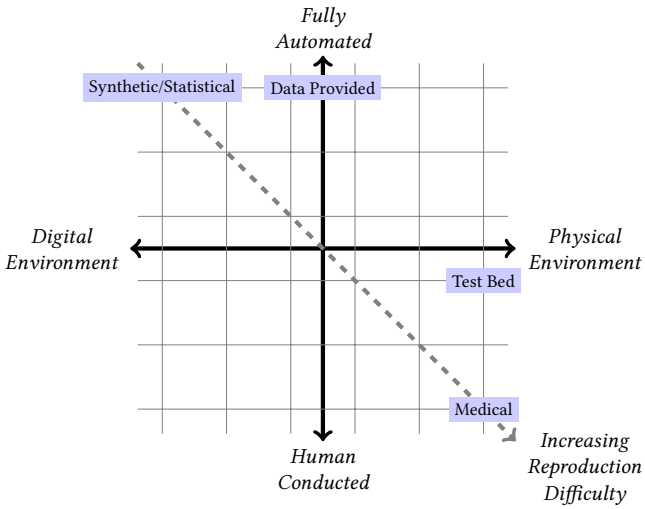

Figure 1: Reproducibility segmentation based on level of system automation and level of environmental automation.

Cyber-physical systems and Internet of Things (CPS/IoT) devices by definition live at the boundary between the digital and physical world. As an interstitial discipline, the nature of CPS/IoT research is that salient work can reflect disparate domains: sometimes it looks like theoretical computer science, other times it can resemble a medical study, and often it lies somewhere in the middle. Figure 1 captures one view of this; many disciplines can site the majority of their work in one region, whereas CPS/IoT covers the whole area. In such a broad field, a one-size-fits-all solution to R&R is not likely. As a discipline, we must strive to develop and follow best practices that make our results more broadly usable and verifiable by the community. However, we must also ensure that our unique constraints are considered in the emerging structural mechanisms – we must balance R&R standards with the broad base of research.

There is a spectrum of what it means to be reproducible. Following good scientific protocol when collecting results is crucial. Making the data available to the broader community is valuable. Having that data documented makes it much more likely that another group will use it. Having the data analysis code easily executable provides confidence that your analysis is sound. Allowing others to easily verify the results using your analysis should be commended. Yet, all of these things take a tremendous amount of time and effort to do well. *Making things replicable, reproducible, and broadly usable is a lot of effort!* Clearly there is value to the community, but how should efforts be balanced with other demands on a research project? In essence, this boils down to the question: how much does the community value reproducibility? And how does the community properly reward and incentivize those efforts which in many cases are tremendous undertakings?

Table 1: R&R efforts in an array of CS and CPS/IoT venues.

Venue	Artifacts in CFP?	ACM Badging?	Artifact Awards?
SPLASH	Y	Y	Y
IPSN	Y	N	Y
CGO	Y	Y	N
CoNEXT	Y	Y	N
SIGCOMM	Y	Y	N
ISFPGA	Y	Y	N
MobiCom	Y	Y	N
PLDI	Y	N	N
ICFP	Y	N	N
BuildSys	N	N	N
SenSys	N	N	N
e-Energy	N	N	N

This paper provides recommendations for enhancing and incentivizing reproducibility in CPS and IoT. We start in Section 2 by surveying efforts in communities including software, hardware, and health and summarizing the measures they have adopted to incentivize and standardize R&R. Next, in Section 3, we present several case studies related to CPS/IoT research projects at UCSD. We discuss efforts to make these projects reproducible along with the unique challenges faced in reproducibility in these contexts. We describe potential ways to make the projects more reproducible and discuss the differences between current efforts in other disciplines. In Section 4, we conclude with a set of recommendations that we believe will help incentivize R&R efforts, particularly for the CPS/IoT community by highlighting the challenges in applying measures adopted by other communities to CPS/IoT.

2 A SURVEY OF R&R EFFORTS

R&R is of interest across all disciplines. For CPS/IoT-related work, the physical world can be a challenge (e.g. due to uncontrollable environmental factors) as well as the use of physical systems (e.g. accessibility for R&R from custom-designed research hardware). While these challenges are not unique to CPS/IoT, spanning the digital and physical world can allow more latitude on the spectrum of feasible R&R. In this section, we discuss R&R influences from structural organizations, such as the National Science Foundation (NSF) and National Institutes of Health (NIH), efforts within the Association Computing for Machinery (ACM) organization, and insights from the machine learning (ML) community towards a synthesis of today's best practices and existing incentives for reproducibility.

2.1 Structural Incentives

Many communities are beginning to call for greater efforts towards reproducibility. However, the time spent towards making one's own work reproducible might be considered misallocated; the personal benefits and incentives from enabling reproducible work are minimal. Because of this contradiction, we look towards the R&R incentives initiated by overarching organizations, such as the NIH and NSF, as examples for initiatives and solutions towards this problem.

Reproducibility issues in biomedical research have garnered significant attention in recent years. One study indicates that irreproducible pre-clinical research exceeds 50%, which results in approximately $28B spent per year on pre-clinical, irreproducible research in the United States alone [18]. The high scientific and financial stake prompted the NIH, the largest funder of biomedical research in the US, to use and develop various strategies improve research reproducibility. The Rigor and Reproducibility policy [10] ensures that researchers use unbiased and well-controlled procedures for (1) experimental design, (2) methodology, (3) analysis, (4) interpretation, and (5) result reports. Grant proposals exist as an opportunity for funding specifically towards reproducibility efforts. In addition, the NIH funds training for researchers on rigor and reproducibility [2, 6, 7], and also funds projects to understand and characterize experimental systems [1, 4].

The NSF also provides funding for R&R concepts. The Computer and Information Science and Engineering (CISE) Community Research Infrastructure (CCRI) [34] provides funding opportunities for three different award categories: (1) Planning Community Infrastructure awards, (2) Medium Community Infrastructure awards, and (3) Grand Community Infrastructure awards. These programs focus on community building, which shares some aspects with R&R but has a much broader goal. Trust-Hub [33] is an example of a CCRI award for the hardware security community that funds the development of benchmarks, tools, and metrics to validate SoC security as well as to create a web-portal to disseminate and promote research and development in securing electronics devices and systems. Similar efforts would strengthen R&R in the CPS/IoT communities.

In summary, there are various external funding opportunities for reproducibility efforts, none or few of which directly address reproducibility issues for CPS/IoT research.

2.2 Communities, Committees, & Conferences

In this section, we discuss reproducibility efforts within software and hardware communities. We also discuss ML reproducibility efforts, which we denote as a community separate from typical software and hardware communities. We show a slice of the existing efforts of various venues in Table 1.

2.2.1 Software-only communities. CS researchers have identified issues with reproducibility standards, elicited the consequences of poor research reproducibility, and suggested collaborative efforts towards reproducibility efforts. Vitek et al. [42] discussed issues that hinder the best practices for reproducibility, including poor statistical analyses and lacking baseline comparisons. They recommended documentation and open-sourced benchmarks that other researchers can access, which are standardized in some conferences today during artifact submission; typically artifacts are compressed, downloadable files or links to GitHub repositories or websites that contain the paper-submitted project. The Artifact Evaluation Committee (AEC) guidelines [13] serve as a baseline for artifact submission guidelines that are complementary to a paper submission or acceptance. These guidelines share recommendations for artifact evaluation on the committee side as well as reproduction guidelines for the submission side. Bajpai et al. [15] mentioned a lack of incentive for reproduction efforts and proposed the ability

to both upload research artifacts and highlight reproducible papers. The ACM Badging system [11] has provided some awareness for reproducibility importance, which some conferences have adopted.

Many software-defined communities, such as conferences in ACM SIGPLAN, adapted some form of artifact evaluation for artifacts submitted alongside papers. Most conferences in ACM SIGPLAN follow the AEC guidelines. 2020 conferences under ACM SIGPLAN, such as PLDI [37], ICFP [24], SPLASH [40], and CGO [16], all follow the base AEC guidelines at the minimum. Other software conferences, such as SIGCOMM [39] and CoNEXT [17], developed their own guidelines in addition to incorporating artifact submissions and the ACM badging system.

Most aformentioned conferences use the standard ACM Badging system to highlight efforts in research reproducibility. SPLASH, CGO, CoNEXT, and SIGCOMM visually labels papers with ACM badges depending on the category of the badge. SPLASH serves as an example of an increasingly successful conference with improved efforts in artifact submissions. SPLASH follows the ACM Badging system and also underscores papers with distinguished artifacts. Through their R&R efforts, SPLASH noted an increase from 44 submitted artifacts in 2019 to 67 submits in 2020, in which 87 papers were conditionally accepted in 2020; out of 67 artifact submissions, 49 have the minimum “functional” artifact badge.

2.2.2 Hardware-aware communities. There has been comparatively less effort in many CPS/IoT venues to promise and reward R&R. Out of the surveyed conferences in this hardware-related research space, only ISFPGA [26] has made efforts similar to the software community. Out of about 26 accepted papers for ISFPGA 2020, six papers had all available ACM artifact badges; two papers only acquired an “artifact available” badge; ISFPGA follows both the AEC guidelines and the ACM badging system. In contrast, MobiCom 2020 [30] yielded four papers with "artifact available" badges out of about 62 accepted papers; only one paper acquired an “artifact evaluated” badge. IPSN 2020 [25] did not adopt AEC guidelines nor the ACM badging system but did award a paper for having the best research artifact.

2.2.3 ML communities. It is difficult to reproduce the different steps involved in obtaining a machine learning model, and the associated test results calls for a systemic approach to solve the ML reproducibility crisis [23]. Top-tier ML conferences, such as NeurIPS [8], have incorporated a code submission policy [36], a ML reproducibility checklist,[1] and a community-wide reproducibility challenge since 2019. A ML code completion checklist[2] is now part of the official NeurIPS 2020 submission process. The ML reproducibility checklist serves to check aspects related to the models, algorithms, code, and experimental results of the paper. The accepted papers are subject to a reproducibility challenge with the aim of replicating experiments in the paper, analyzing evidence of reproducibility in experiments, and verifying the validity of the authors’ findings. The participants in the reproducibility challenge then submit a report explaining aspects that could be reproduced, detailing experimental methodology, implementation details, and result analysis, along with resource utilizations such as time and development effort. Similar reproducibility challenges and workshops were held at ICML 2017 [3], ICLR 2018 [5], and ICLR 2019 [35]. The latest ML Reproducibility Challenge 2020 and Spring 2021 [9] expanded its scope to also include papers published in ACL, EMNLP, CVPR, and ECCV.

To create a common comparison metric, ML/AI benchmarks developed by industry and academia are used. The MLPerf Training benchmark [29] and MLPerf Inference benchmark [38] were developed upon a consensus between more than 200 ML engineers from 30 different organizations to measure the performance of ML frameworks, ML hardware accelerators, and ML cloud platforms. Compared to the MLPerf Training benchmark, the AIBench Training benchmark [41] is more comprehensive with a higher number of task models and reduced benchmarking cost.

2.3 Summary of Best Practices

After surveying multiple venues and suggestions regarding reproducibility efforts, we discuss the observed best practices.

Amongst different communities and venues, we see an emergence of community standards for assays and result reporting; in various Programming Language (PL) and ML venues, conferences advertise a "Call for Artifacts" in which research artifacts are evaluated based on their ability to reproduce research results. Many PL venues set guidelines for reproducibility evaluation. Conferences such as PLDI and SPLASH adopted AEC guidelines to set standards and expectations for artifact submissions and evaluations; these guidelines include artifact ease-of-use in addition to artifact-paper result consistency. Other efforts include external tools to improve an artifact’s reproducibility via versioning and data management.

3 CASE STUDIES

Reproducibility in the CPS/IoT space is different from those defined in the software and ML communities discussed previously. To understand these differences, we consider two case studied from active projects. The first case study considers the difficulties of research replication with machine learning for wireless research. The second case study discusses an approach to methodically reproduce the research and development process as outlined in Figure 2.

3.1 Case 1: Machine Learning for Wireless

The application of ML to wireless communications and networking is an exciting new direction for meeting the throughput, latency, cooperative networking, and spectrum sharing challenges for 5G and beyond. However, ML is data intensive, and the conclusions drawn from experiments about the efficacy of ML applications for wireless are conditioned upon the data used to train and evaluate a model. There are two typical approaches for generating this data: utilize theory to create synthetic data from statistical models of wireless channels or collect the data from over-the-air captures. This section discusses the advantages and drawbacks of each approach, utilizing examples from a small subset of the research in this area to aid discussion of the difficulties and necessity of reproducible research on this topic.

3.1.1 Digital Environment. Using a statistical channel model, such as Additive White Gaussian Noise (AWGN), is a common practice in wireless communications research and has several advantages.

[1] https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf
[2] https://github.com/paperswithcode/releasing-research-code

First, the assumptions are very clearly listed (e.g. if only AWGN is assumed, then the reader would know that multi-path effects are not considered). Second, under the assumptions for the chosen model, known factors can be theoretically proven (e.g. the symbol error rate for various modulations). These two points allow for methodical replication of prior work, either via extension into a different statistical model or demonstration of assumptions where the application would no longer work.

Many Radio Frequency Machine Learning (RFML)[3] papers are based upon these statistical models. The majority of work in this area has been on signal classification, and more specifically modulation classification, where a classifier is provided with a wireless signal and asked to determine which modulation (e.g. BPSK, QAM16, etc.) was used to create that signal. AWGN is a commonly used channel model because synthetic data can be easily and cheaply created for experimentation. However, prior work has shown that these RFML models [12, 20] break down when the channel model's assumptions are incorrect for the operational environment. It was shown in [20] that if the RFML model's training distribution did not account for errors in signal detection[4] then that the model's accuracy would significantly degrade; over-simplistic assumptions about operating conditions can inflate the estimated model performance.

The *implementation* of a statistical model could also lead to results that are not replicable. An AWGN channel is parameterized by the noise power that is typically swept to show varying performance levels across a range of Signal-to-Noise ratios (SNR). A logical way to implement this parameter sweep is to create a grid of SNR values in the specified range (e.g. 0-30 dB). However, the real world is continuous, not discrete, and [12] showed that model performance actually suffered when the SNR was sampled from outside of the discrete set of values that the model was trained on.

Due to the aforementioned issues, along with a host of others, it is common to hear calls for training and evaluating models on "real" data. This reduces the likelihood that oversimplified assumptions are made about the operational environment (i.e. [20]) or that an artifact of the data generation methodology impacts results (i.e. [12]). The easiest way to accomplish this is through the release of an open source dataset that was sampled from the real world. However, the utility of released datasets can be negatively impacted by two factors. First, while datasets can be multi-purpose, they are often quite opinionated, especially those originally intended for supervised learning tasks that require human labeling; a community must first coalesce around a set of common problem definitions before a dataset geared towards those problems can be impactful. Second, the idea that datasets aid in reproduction, and hopefully extension, of prior research in ML for wireless communications is predicated on the assumption that the learning mechanism being researched is predominantly passive; many interesting research directions invalidate this assumption and instead leverage the ability to interact with, and/or learn from, an *environment* in some fashion. Therefore, we reserve discussion of curating datasets until Section 3.2 and instead next discuss replicating wireless test beds.

3.1.2 Physical Environment. Once processes move from the digital to the analog or physical domains, they become nearly impossible to entirely replicate as too many external, uncontrollable factors have influence over the outcomes. Seemingly mundane variables, like the weather in an outdoor case, or a fan in an indoor case, can affect wireless propagation. Although the induced effects from such factors are on a smaller scale, these differences in the wireless channel have been shown to severely impact the performance of Radio Identification[5] [14]. This begs the question of what constitutes a *reasonable* reproduction of research in "ML for Wireless" in which there is likely no one-size-fits-all answer. Developed algorithms must eventually generalize to some super-set of environmental conditions beyond the laboratory to be useful in the real world; algorithms should be generalized such that environmental factors minimally influence the performance of the algorithm or system. However, it is unreasonable to expect performance evaluations conducted in differing radio environments to yield *exactly* reproducible results due to variations between intra- and inter-environmental differences. While the methodology and conclusions can be generalizable and reproducible, the results may differ.

Building a large scale wireless test bed is a herculean effort fraught with both engineering challenges as well as administrative and regulatory ones. Thus, replicating wireless experiments in the physical world requires non-trivial effort. Thankfully, many academic test beds are being opened up for community usage. Ideally, these test beds will lead to the creation of open source datasets collected in a real environment, providing a common set of RF environments where the next generation of wireless communications algorithms can be developed and compared head-to-head against others proposed in the literature. By providing these capabilities, these community test beds will likely aid in R&R of all wireless research, while also lowering the development burden for conducting new research and thus increase the speed of innovation.

3.2 Case 2: Environmental Monitoring

Environmental monitoring is a uniquely difficult problem in CPS within the context of reproducibility. The locations that scientists wish to study the most are often extremely remote and difficult to observe. For example, we have active field work related to the monitoring of mangrove forests [21], coral reefs [32], and Mayan temples in the jungles of Guatemala [19]. When it comes to performing reproducible work on these projects, relying on physical access is impractical. Also, because these locations have a dynamic nature, it is not guaranteed that an experiment could be repeated exactly, even if physical access were possible.

Instead, scientists rely on datasets gathered from these locations to perform any sort of reproducible analysis, such as species classification and segmentation in the case of the mangrove forests and coral reefs. However, datasets cannot represent every detail perfectly; since other feasible options are not available to conduct reproducible work, it is important that the collection methodologies

[3] RFML is a colloquial term used for defining the application of ML to unprocessed wireless signals (i.e. raw IQ data sampled at the baseband).

[4] Wireless communications typically synchronize to a known preamble of a communication in order to estimate the needed sampling rate and center frequency to properly receive a signal; yet, with blind signal classification, this is difficult because little is known *a priori* about the target communication, and errors in this estimation can change the data distribution being used in RFML models.

[5] Hardware imperfections can cause two radios, even of the same model, to perform slightly differently. Radio Identification uses these imperfections to determine which radio transmitted a wireless signal without the need for higher level packet processing.

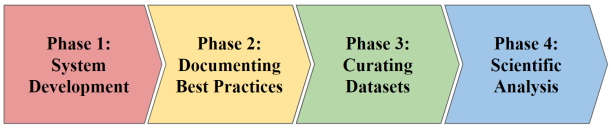

Figure 2: Generic workflow for UCSD field R&D.

and organization methods underpinning these datasets are trustworthy and reliable. More often than not, papers do not elaborate on their methods of data collection for reproducibility and provide a dearth of data that is insignificantly useful.

We have developed a model workflow for our projects that relies on gathering data from the field, as seen in Figure 2. Phase 1 is the development of the instrument that will collect data in the field. For example, this might entail a camera-equipped UAV to image mangrove forests from above. Phase 2 is the organization of the best methods and practices for gathering data with this system in the field. Phase 3 is the curation of the data that is collected in the field. This is more than just a raw data dump; it includes metadata organization like timestamps and sensor statuses, as well as clear methods for accessing and using the data for Phase 4. This phase can consist of image segmentation and species classification as previously mentioned, but it can also include temperature modeling, ocean current modeling, or feature detection in sonar images.

Typically, papers describe Phases 1 and 4 well, but describe Phases 2 and 3 only at a high level, if at all. However, Phase 4 is only reproducible within the context of the phases leading up to it, and if one part of the flow is missing, then it is difficult to say with confidence that the final analysis is valid. We concede that Phases 2 and 3 are easy for an author to minimize, since the artifacts generated in these phases are either not called for, or are not rewarded. Additionally, it is difficult to go back and recreate best practices and quality datasets after the analysis is complete; one must plan for their creation in Phase 1, just as one plans for the type of analyses in Phase 4 during system development.

In this mangrove forest monitoring project, we quickly recognized a need for formalizing best practices for our data collection process. Using a UAV equipped with downward-facing RGB and multispectral cameras, we gather images from above a forest and create photomosaics which can be used for monitoring the health and growth of the mangroves. Due to the number of mangrove forests, their remote locations, and the frequency with which they need to be imaged, it is impractical for a single team to gather all the data. In order to scale the system that we developed (Phase 1), maintained, and standardized in our scientific analysis (Phase 4), we developed and published a field manual on a peer-reviewed online platform [22]. The manual goes into great detail about the specifics of the data collection process, such as altitude, camera configurations, take off and landing methods, and exact methods of data logging, labeling, and saving. This enables other researchers to reproduce our collection process and validate our methods, provided they have access to a similar system or have the resources to create one. Additionally, it allows us to integrate data from a diverse set of collection efforts into a coherent dataset where they can be effectively analyzed and compared.

Because environmental monitoring analysis relies on data from relatively inaccessible locations, any chance of reproducibility relies on the curation of coherent and complete datasets. Datasets, especially when they consist of data from discrete collectors and locations, rely on the standardization, documentation, and publication of best practices. The system development phase must be built around the inclusion of all these other components if it is expected to deliver reproducible results.

4 RECOMMENDATIONS

The creation and dissemination of quality artifacts is a lot of work. Our recommendations for supporting R&R distill down to two major principles: first, people putting in the work deserve a more predictable return on the time and effort invested, and second, that the diversity of systems which fall under the CPS/IoT umbrella be considered to ensure equitable evaluation of R&R contributions.

4.1 Progressive Reward Mechanisms

In practice, there are three primary stakeholders: the users creating artifacts, the reviewers evaluating them, and the community members leveraging artifacts. We need to increase the expected value of investing time and effort in R&R for each of these categories.

4.1.1 Badges. Today, badges certifying degrees of reproducibility are becoming the norm across computer science sub-disciplines. Badges are useful to the community at large, as they reduce the uncertainty for users who would seek to replicate or build on a line of work. We must recognize that badges provide comparatively little benefit to the person receiving them, however. In time, promotion and hiring committees may come to value these certifications, but such systemic change will not come quickly or equally.[6]

4.1.2 Awards for New Work. While badges are becoming commonplace, as Table 1 shows, artifact awards remain rare. Awards, however, can cut through the bureaucracy of evolving merit criteria—community-specific awards are already commonplace, what matters is less often the award title, but than an individual is award-winning in their community.

4.1.3 Awards for Historic Efforts. Much as test-of-time awards for publications today reward impactful ideas, we should celebrate impactful artifacts. Award committees should be generous in considering historical artifacts, many may not have formal publications or other easy handles. Longitudinal rewards inspire critical reflection on prior efforts and help shape future work.

4.1.4 The Case for a Platform Track. The scarcity that ensures the value of awards also introduces uncertainty. For artifacts that require significant labor, we need a more reliable and universally accepted reward: a publication. There is a growing awareness of the need to capture and reward parts of the research ecosystem beyond research papers. Conferences now invite "Industry Papers," "Experience Papers," and "Negative Results." We suggest that "Platform Papers" should be added to this list. There are often rich insights for real-world systems challenges that are rejected in the traditional

[6]Indeed, there yet remain numerous institutions that de-emphasize the conference publications of their computer science faculty in favor of journals.

review process as "just engineering." We should not need to wait 10-20 years for a retrospective research publication [27, 28] to capture and disseminate key design principles.

4.1.5 The Artifact Review Load. Artifact evaluation committees are largely in their infancy. During this bootstrapping phase, volunteers are willing to donate labor to help establish systems. While artifact evaluation committees are often run by faculty, who receive external service credit from their institution for the work, committee membership is largely graduate students, who are beginning to rightly ask what benefit external service work provides them?

4.1.6 Structural Incentives. Things that take time and effort cost money. As example, when NIH added data sharing requirements, funding for data archival and dissemination was also made available to projects. As noted by the National Academies, as R&R expectations grow, so too must financial and structural support [31].

4.2 Mind the Breadth of CPS/IoT

What makes a best artifact best? If we begin to use awards as rewards, it will become important to find a means to fairly evaluate contributions across a wide space of R&R capability—pure software will be more perfectly replicable, but does that always make it a better artifact? The definition of a reproducible artifact becomes context-dependent. Where two papers may both evaluate their work by demonstrating insights synthesized from a physical-world data, one's contribution may be a new machine learning algorithm which generates new insights from extant data—fully and purely reproducible—while another's may be a new sensing system to collect such data, that happens to use machine learning to extract a result—only reproducible if one reproduces the sensing artifact.

REFERENCES

[1] 2016. Tools for Cell Line Identification. https://grants.nih.gov/grants/guide/pa-files/PA-16-186.html. Accessed: 2021-02-18.

[2] 2017. National Institute of General Medical Sciences Ruth L. Kirschstein National Research Service Award (NRSA) Predoctoral Institutional Research Training Grant (T32). https://grants.nih.gov/grants/guide/pa-files/PAR-17-341.html. Accessed: 2021-02-18.

[3] 2017. Reproducibility in Machine Learning Workshop ICML 2017. sites.google.com/view/icml-reproducibility-workshop/icml2017. Accessed: 2021-02-20.

[4] 2018. Better Defining Growth Medium to Improve Reproducibility of Cell Culture. https://grants.nih.gov/grants/guide/pa-files/PA-18-816.html. Accessed: 2021-02-18.

[5] 2018. ICLR 2018 Reproducibility Challenge. https://www.cs.mcgill.ca/~jpineau/ICLR2018-ReproducibilityChallenge.html. Accessed: 2021-02-20.

[6] 2018. Training Modules to Enhance the Rigor and Reproducibility of Biomedical Research. https://grants.nih.gov/grants/guide/rfa-files/RFA-GM-18-002.html. Accessed: 2021-02-18.

[7] 2019. Administrative Supplements to NIGMS Predoctoral Training Grants for Development of Curricular or Training Activities to Enhance Predoctoral Training. https://grants.nih.gov/grants/guide/notice-files/NOT-GM-19-015.html. Accessed: 2021-02-18.

[8] 2019. Thirty-third Conference on Neural Information Processing Systems. https://nips.cc/Conferences/2019. Accessed: 2021-02-20.

[9] 2020. ML Reproducibility Challenge 2020 and Spring 2021. https://paperswithcode.com/rc2020. Accessed: 2021-02-20.

[10] 2020. Rigor and Reproducibility. https://www.nih.gov/research-training/rigor-reproducibility. Accessed: 2021-02-18.

[11] ACM. [n.d.]. Artifact Review and Badging - Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current. Online; accessed 03-February-2021.

[12] D. Adesina, et al. 2019. Practical Radio Frequency Learning for Future Wireless Communication Systems. In *2019 IEEE Military Communications Conference (MILCOM)*. https://doi.org/10.1109/MILCOM47813.2019.9020807

[13] AEC [n.d.]. Guidelines for Packaging AEC Submissions. https://artifact-eval.org/guidelines.html. Online; accessed 03-February-2021.

[14] A. Al-Shawabka, et al. 2020. Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel on Radio Fingerprinting. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 646–655. https://doi.org/10.1109/INFOCOM41043.2020.9155259

[15] Vaibhav Bajpai, et al. 2017. Challenges with Reproducibility. In *Proceedings of the Reproducibility Workshop* (Los Angeles, CA, USA) *(Reproducibility '17)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3097766.3097767

[16] CGO. [n.d.]. Code Generation and Optimization. https://cgo-conference.github.io/cgo2020/. Online; accessed 20-February-2021.

[17] CoNEXT. [n.d.]. International Conference on Emerging Networking EXperiments and Technologies. https://conferences2.sigcomm.org/co-next/2020/. Online; accessed 03-February-2021.

[18] Leonard P. Freedman, et al. 2015. The Economics of Reproducibility in Preclinical Research. *PLOS Biology* 13, 6 (06 2015), 1–9. https://doi.org/10.1371/journal.pbio.1002165

[19] Quentin Kevin Gautier, et al. 2020. Low-cost 3D scanning systems for cultural heritage documentation. *Journal of Cultural Heritage Management and Sustainable Development* (2020).

[20] S. C. Hauser, et al. 2017. Signal detection effects on deep neural networks utilizing raw IQ for modulation classification. In *2017 IEEE Military Communications Conference (MILCOM)*.

[21] Astrid J Hsu, et al. 2020. Driven by Drones: Improving Mangrove Extent Maps Using High-Resolution Remote Sensing. *Remote Sensing* 12, 23 (2020), 3986.

[22] Astrid J Hsu, et al. 2019. Drone Flight Manual: UCSD Mangrove Imaging Procedure. Version 1.2. (2019).

[23] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

[24] ICFP. [n.d.]. International Conference on Functional Programming. https://icfp20.sigplan.org/. Online; accessed 20-February-2021.

[25] IPSN. [n.d.]. Information Processing in Sensor Networks. http://ipsn.acm.org/2020/. Online; accessed 11-February-2021.

[26] ISFPGA. [n.d.]. International Symposium on Field-Programmable Gate Arrays. https://isfpga.org/. Online; accessed 10-February-2021.

[27] Philip Levis. 2012. Experiences from a Decade of TinyOS Development. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation (OSDI'12)*. USENIX Association, USA, 207–220.

[28] Jason Lowe-Power, et al. 2020. The gem5 Simulator: Version 20.0+. arXiv:2007.03152 [cs.AR]

[29] Peter Mattson, et al. 2020. MLPerf Training Benchmark. arXiv:1910.01500 [cs.LG]

[30] MobiCom. [n.d.]. Mobile Computing. https://sigmobile.org/mobicom/2020/. Online; accessed 10-February-2021.

[31] Engineering National Academies of Sciences and Medicine. 2019. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[32] Perry Naughton, et al. 2015. Scaling the annotation of subtidal marine habitats. In *Proceedings of the 10th international conference on underwater networks & systems*.

[33] NSF. [n.d.]. CCRI: ENS: Enhancement of Trust-Hub, a Web-based Portal to support the Cybersecurity Research Community. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2016624&HistoricalAwards=false. Online; accessed 18-February-2021.

[34] NSF. [n.d.]. CISE Community Research Infrastructure (CCRI). https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12810. Accessed: 2021-02-18.

[35] Joelle Pineau, et al. 2019. ICLR Reproducibility Challenge. *ReScience C* (2019).

[36] Joelle Pineau, et al. 2020. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206* (2020).

[37] PLDI. [n.d.]. Programing Language Design and Implementation. https://pldi20.sigplan.org/. Online; accessed 20-February-2021.

[38] V. J. Reddi, et al. 2020. MLPerf Inference Benchmark. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. 446–459. https://doi.org/10.1109/ISCA45697.2020.00045

[39] SIGCOMM. [n.d.]. Special Interest Group on Data Communications. https://conferences.sigcomm.org/sigcomm/2020/. Online; accessed 17-February-2021.

[40] SPLASH. [n.d.]. Systems, Programming, Languages, and Applications: Software for Humanity. https://2020.splashcon.org/. Online; accessed 03-February-2021.

[41] Fei Tang, et al. 2020. AIBench Training: Balanced Industry-Standard AI Training Benchmarking. arXiv:2004.14690 [cs.AI]

[42] J. Vitek and T. Kalibera. 2011. Repeatability, reproducibility and rigor in systems research. In *2011 Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT)*. 33–38. https://doi.org/10.1145/2038642.2038650