DOI:10.1145/2994031

Ronald F. Boisvert

Incentivizing Reproducibility

A SCIENTIFIC RESULT is not truly established until it is independently confirmed. This is one of the tenets of experimental science. Yet, we have seen a rash of recent headlines about experimental results that could not be reproduced. In the biomedical field, efforts to reproduce results of academic research by drug companies have had less than a 50% success rate,^a resulting in billions of dollars in wasted effort.^b In most cases the cause is not intentional fraud, but rather sloppy research protocols and faulty statistical analysis. Nevertheless, this has led to both a loss in public confidence in the scientific enterprise and some serious soul searching within certain fields. Publishers have begun to take the lead in insisting on more careful reporting and review, as well as facilitating government open science initiatives mandating sharing of research data and code.

But what about experimental computer science? Fortunately, we haven't been in the headlines. But, it is rare for research studies in computing to be reproduced. On the surface this seems odd, since we have an advantage over science done in wet labs. For us, the object of study is often software, so it, along with the associated experimental scaffolding, is a collection of bits that can be easily shared for the purpose of audit and inspection, for an assisted attempt at replication, or for building upon the work to advance science further or to transfer technologies to commercial use. Certainly the situation is a bit more complex in practice, but there is no reason for us not to be leaders in practices that

enable audit and reuse when technically and legally possible.

Some communities within ACM have taken action. SIGMOD has been a true pioneer, establishing a reproducibility review of papers at the SIGMOD conference since 2008. The Artifact Evaluation for Software Conferences initiative has led to formal evaluations of artifacts (such as software and data) associated with papers in 11 major conferences since 2011, including OOPSLA, PLDI, and ISSTA. Here the extra evaluations are optional and are performed only after acceptance. In 2015 the *ACM Transactions on Mathematical Software* announced a Replicated Computational Results initiative,^c also optional, in which the main results of a paper are independently replicated by a third party (who works cooperatively with the author and uses author-supplied artifacts) before acceptance. The *ACM Transactions on Modeling and Computer Simulation* is also now doing this, and the *Journal of Experimental Algorithmics* has just announced a similar initiative. In all cases, successfully reviewed articles receive benefits, such as a brand on the paper and extra recognition at the conference.

To support efforts of this type, the ACM Publications Board recently approved a new policy on Result and Artifact Review and Badging.^d This policy defines two badges ACM will use to highlight papers that have undergone independent verification. *Results Replicated* is applied when the paper's main results have been replicated using artifacts provided by the author, or *Results Reproduced* if done completely independently.

Formal replication/reproduction is sometimes impractical. However,

both confidence in results and downstream reproduction are enhanced if a paper's artifacts (that is, code and datasets) have undergone a rigorous auditing process such as those being undertaken by ACM conferences. The new ACM policy provides two badges that can be applied here: *Artifacts Evaluated—Functional*, when the artifacts are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation, and if, in addition the artifacts facilitate reuse and repurposing at a higher level, then *Artifacts Evaluated—Reusable* can be applied. When artifacts are made publicly available, further enhancing auditing and reuse, we apply an *Artifacts Available* badge. ACM is working to expose these badges in the ACM Digital Library on both the landing pages for articles and in search results.

Replication of results using author-supplied artifacts is no doubt a weak form of reproducibility, but it is an important first step. We believe that auditing that goes beyond traditional refereeing will help raise the bar for experimental research in computing, and that the incentives that we provide will encourage sharing and reuse of experimental artifacts.

This policy is but the first deliverable of the ACM Task Force on Data, Software and Reproducibility. Ongoing efforts are aimed at surfacing software and data as first-class objects in the DL, so it can serve as both a host and a catalog for not just articles, but the full range of research artifacts deserving preservation. ■

Ronald F. Boisvert (boisvert@acm.org) is chair of the ACM Publications Board's Digital Library Committee.

a *Nature Reviews Drug Discovery* 10, 643-644 (September 2011), doi:10.1038/nrd3545

b *Nature*, June 9, 2015, doi:10.1038/nature.2015.17711

c <http://toms.acm.org/replicated-computational-results.cfm>

d <http://www.acm.org/publications/policies/artifact-review-badging>