# Artifact Evaluation and Reproducibility (at CHES)

Markku-Juhani O. Saarinen
<markku-juhani.saarinen@tuni.fi>

OPTIMIST Workshop
September 4, 2024
Halifax, Canada

Tampereen yliopisto
Tampere University

# What are (Security) Research Artifacts?

Authors of accepted papers are invited to submit associated artifacts for permanent archiving alongside their papers.

*Examples:*

- **Source Code** (Hardware or Software: Implementations, PoCs, tools)
- **Datasets** (network or side-channel measurement traces, raw study data)
- **Scripts** for data processing, analysis, or simulations used
- **Formal specifications** and **Machine-generated proofs**
- **Build environments** (e.g., VMs, Docker containers, configuration scripts)
- Any other digital data related to the research paper and its results

# Why are we doing this?

**Trustworthiness:**

Third party verification of results to gain confidence in their validity. Document research at a technical level unattainable with a traditional publication format.

**Help the research community:**

Using provided data and tools for further study and education. Allow the community to build, improve, expand, and to correct errors.

**Broader Open Science Goals:**

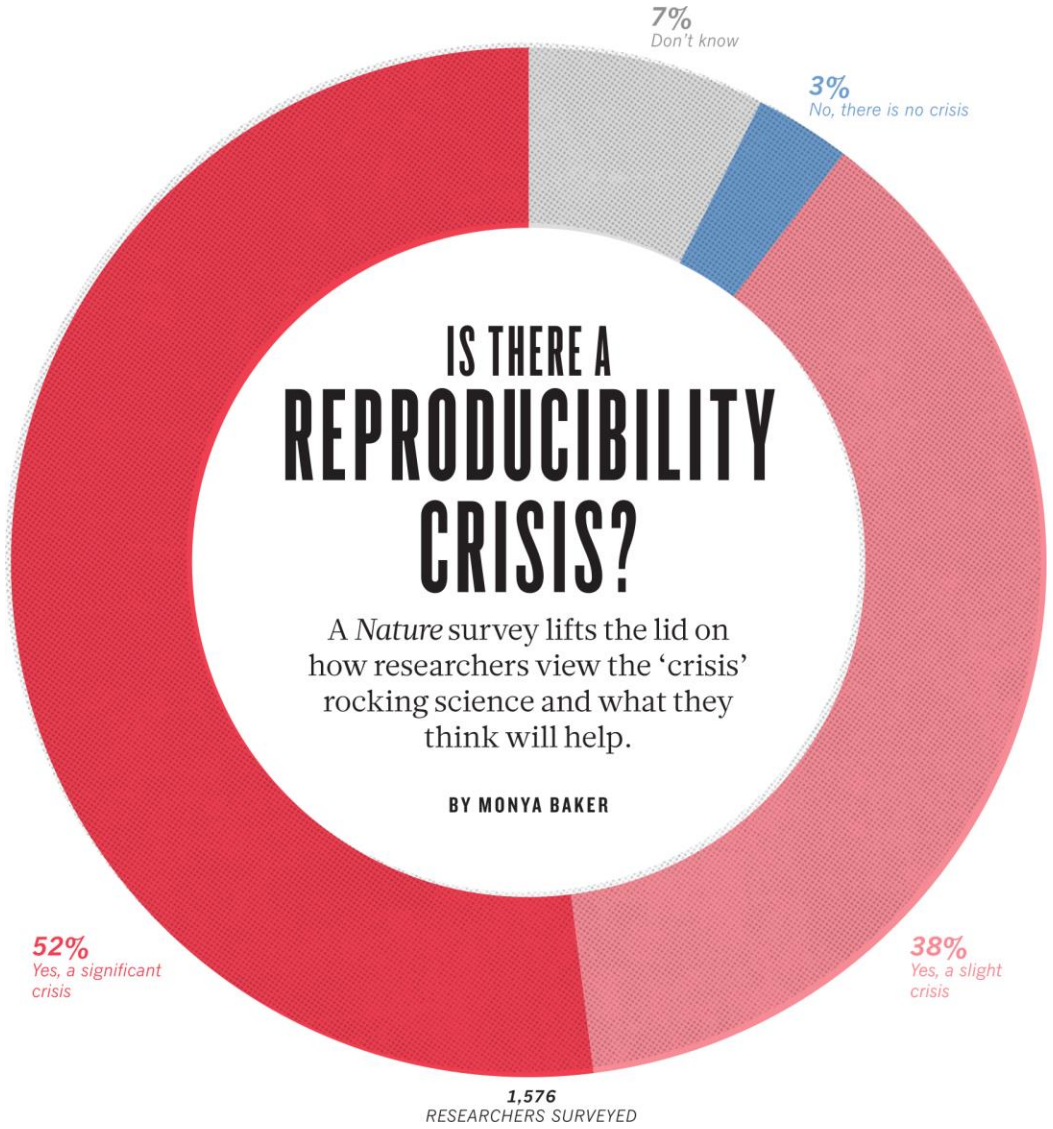
https://open.science.gov/ and https://www.unesco.org/en/open-science

# "Reproducibility Crisis"

A large chunk of peer-reviewed research is not reproducible, and researchers know this.

**Metascience**: (Research on Research)

- Early finding: Everybody is bad at interpreting statistics and p-values.
- More Recently: Leakage and other errors in the use of Machine Learning. *(You need to keep training and testing data separate..)*




Baker, M. "1,500 scientists lift the lid on reproducibility." *Nature* **533**, 452–454 (2016). https://doi.org/10.1038/533452a

# Artifact Evaluation (CS and InfoSec)

- CS Artifact evaluation was "Invented" at ESEC/FSE 2011 [Andreas Zeller]. See: https://cs.brown.edu/~sk/Memos/Conference-Artifact-Evaluation/

- Repeated at ECOOP 2013, and began to spread to other computer science areas, including information security.

- The first "badges" were at OOPSLA 2013 and PDLI 2014. Now adopted by ACM & IEEE.

- ACSAC: 2017-
- WOOT: 2019 -
- USENIX Sec: 2020 -
- **CHES/TCHES: 2021 -**
- NDSS: 2024 -
- SysTEX: 2024 -
- PETS / PoPETs: 2024-
- **IACR Eurocrypt, IACR Crypto, IACR Asiacrypt, IACR FSE/ToSC: 2024-**

# ESEC/FSE ’22:


## A Retrospective Study of One Decade of Artifact Evaluations

Stefan Winter
LMU Munich
Munich, Germany
sw@stefan-winter.net

Christopher S. Timperley
Carnegie Mellon University
Pittsburgh, USA
ctimperley@cmu.edu

Ben Hermann
Technische Universität Dortmund
Dortmund, NRW, Germany
ben.hermann@cs.tu-dortmund.de

Jürgen Cito
TU Wien
Vienna, Austria
juergen.cito@tuwien.ac.at

Jonathan Bell
Northeastern University
Boston, MA, USA
j.bell@northeastern.edu

Michael Hilton
Carnegie Mellon University
Pittsburgh, PA, USA
mhilton@cmu.edu

Dirk Beyer
LMU Munich
Munich, Germany
dirk.beyer@sosy-lab.org

https://doi.org/10.1145/3540250.3549172

# Institutionalizing Artifact Evaluation at IACR

- CHES pioneered Artifact Evaluation within IACR (2021); this year, all other IACR conferences started doing it. This is mainly up to individual PC Chairs.
- A group of Artifact Chairs & other interested people met at Crypto 2024 (August 20) to discuss the future of artifact evaluation.
- We agreed to ask IACR to establish an **Artifact Task Group** to work on guidelines documents, shared policies, and other harmonization.

| | Papers | Artifacts | Rate |
|---|---|---|---|
| **CHES 2024** | 101 | 30 | 29.7% |
| **Crypto 2024** | 143 | 12 | 8.4% |
| **Eurocrypt 2024** | 105 | 13 | 12.4% |

# CHES 2024 Artifact Process

# CHES 2024 Artifact Review Process

The goal is not just to evaluate artifacts, but also help to improve them.

The review is an interactive / collaborative process between authors and the artifact review committee. (We used a custom-configured HotCRP for this.)

For TCHES 2024, the Artifact submission deadlines were 6 weeks from notification, 2 weeks from camera ready. (Issue 1 deadline was extended.)

Artifacts for TCHES 2024 issues 1,2,3 have already been published, Issue 4 will follow. All IACR Artifacts are archived at: https://artifacts.iacr.org/.

# New in CHES 2024: Reproducibility Badges

CHES 2024 AEC adopted a "Badge System" modeled after Usenix Security. Authors were asked to select the scope of evaluation:


IACR CHES 2024 Artifacts Available


IACR CHES 2024 Artifacts Functional


IACR CHES 2024 Results Reproduced

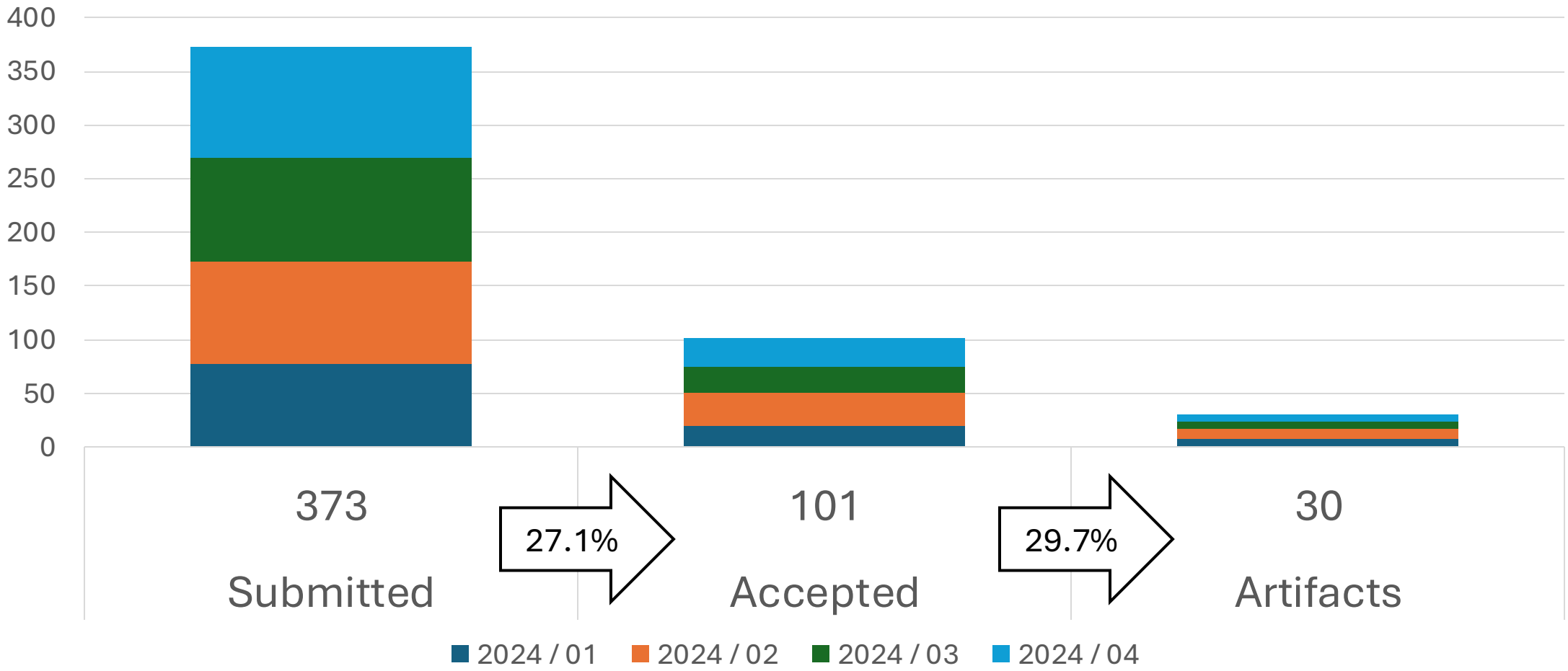
*(IACR still needs nice graphics for badges!)*

This is done to reward authors who put extra effort into polishing their artifacts to make them reproducible. *More on this later..*

# Artifact Copyright and Consent Form

- IACR does **not** require authors to *give up or transfer* the copyright of the artifacts, but IACR does require a permission to distribute them.
- Signed forms are kept by IACR. The form used is here: https://www.iacr.org/docs/copyright_form-artifact-2024-01-22-v1.1.pdf
- All kinds of licenses can be used: CC, BSD, MIT, Apache, etc.
- More restrictive would be potentially ok, as long as we can distribute.
- Quite often 3rd party components were included in artifacts, so different parts of the artifact have different licenses rules.

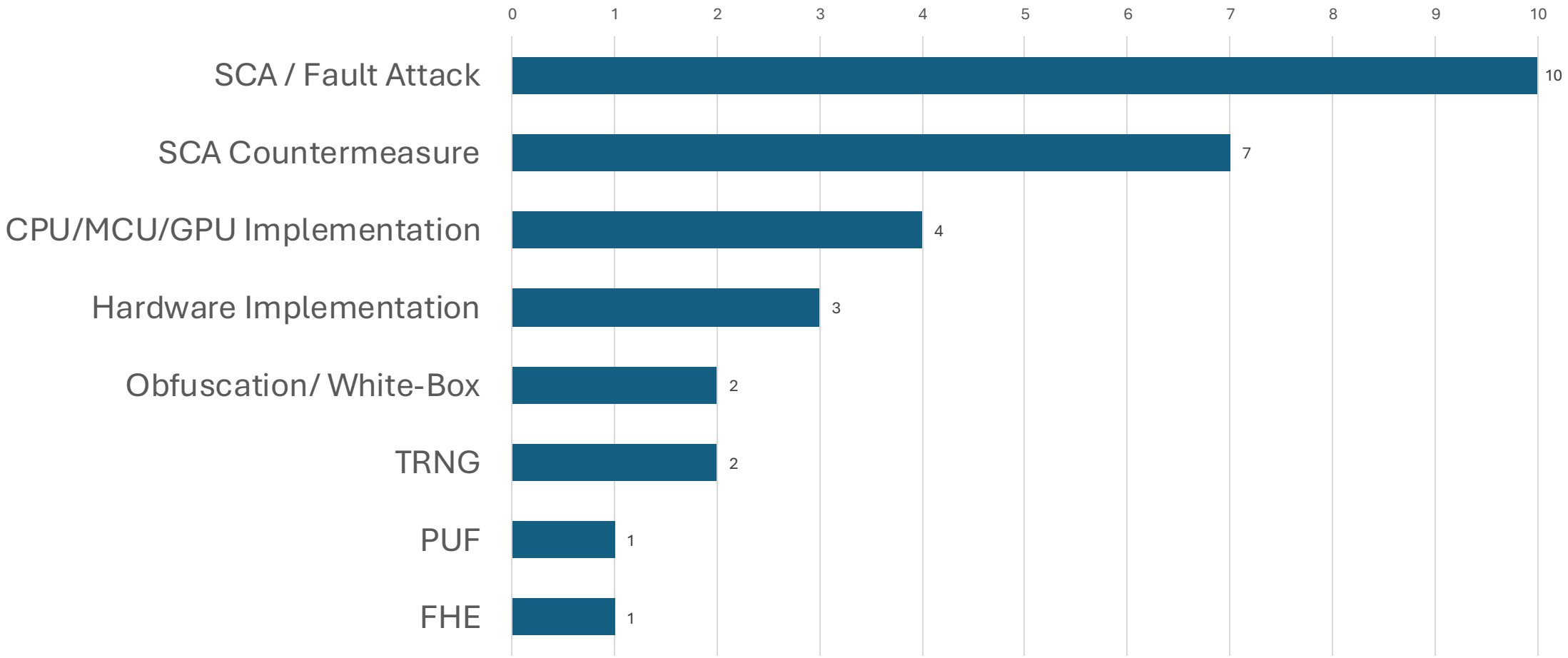
# 30% of CHES 2024 Papers have Artifacts



# All Evaluation Categories Were in Use

| TCHES | Submitted | Accepted | Rate | Artifacts | Rate | Available | Functional | Reproduced |
|---|---|---|---|---|---|---|---|---|
| 2024 / 01 | 77 | 20 | 26.0% | 7 | 35.0% | 1 | 1 | 5 |
| 2024 / 02 | 96 | 31 | 32.3% | 10 | 32.3% | 1 | 4 | 5 |
| 2024 / 03 | 96 | 23 | 24.0% | 6 | 26.1% | 2 | 2 | 2 |
| 2024 / 04 | 104 | 27 | 26.0% | 7 | 25.9% | n/a | n/a | n/a |
| **2024 / All** | **373** | **101** | **27.1%** | **30** | **29.7%** | **4+** | **7+** | **12+** |

# My Rough Categorization of the 30 Artifacts



# CHES 2024 Artifact Evaluation Committee

A mix of PhD students, experienced researchers, and industry practitioners:

- Andrea Basso / University of Bristol
- Gaëtan Cassiers / UCLouvain
- Hao Cheng / University of Luxembourg
- Junhao Huang / BNU-HKBU
- Pantea Kiaei / Apple
- Kris Kwiatkowski / PQShield
- Robin Leander Schröder / Fraunhofer SIT
- Nicolai Müller / Ruhr University Bochum
- Alexander Nilsson / Advenica AB
- Jordi Ribes-González / U. Rovira i Virgili
- Shubhi Shukla / IIT Kharagpur
- Kavya Sreedhar / Stanford
- Thomas Szymkowiak / Tampere University
- Nicola Tuveri / Tampere University
- Rei Ueno / Tohoku University
- Lennert Wouters / KU Leuven

Chair: Markku-Juhani O. Saarinen (Chair) / Tampere University.

# Some Technical Issues in CHES Artifact Evaluation ..

# Hardware & Standard Cell Libraries

- **FPGA**s and **Microcontrollers** are relatively inexpensive and available in most labs, apart from some high-end models.
- **EDA:** Commercial tools for silicon (Synopsys, Cadence) are expensive, and not universally available to researchers.
- **PDK**: Physical Design Kits  may have extremely restrictive NDAs, especially for higher-end technology nodes: Can't reproduce 😔

**Open Source EDA & PDK results are more reproducible:**

- Some de-facto conventions are emerging for open-source hardware area and timing estimates via **Yosys**, **OpenSTA**, **OpenRoad,** etc. Researchers can include these estimates along with others.

# Dependencies: Is it usable in 2044? 2064?

- Many/most code artifacts required installation of external dependencies such as libraries. Sometimes these, too, have to be compiled from source as they don't have "standard packages."
- While it may be possible to reproduce results now, the APIs and behavior of these dependencies will change over time.
- Often there were dependencies to experimental tools from the same research group. Not sure how long these are maintained.

**How to address this?**

- One partial solution is to package a snapshot of a development environment as a **Docker** container or some other kind of VM.
- At bare minimum, **document all versions** of the environment used.

# SCA Procedures and Hypothesis Testing

- TVLA (Test Vector Leakage Assessment) is a well-known physical/statistical experiment to detect side-channel leakage in crypto implementations. [Goodwill et al, 2011 – Also ISO/IEC 17825.]

- TVLA is one of the few accepted methods to obtain "positive assurance" i.e., an argument **for** the security of an implementation. Hence it is often used in implementation papers as evidence to demonstrate security.

- The most common experiment design uses a large number of Welch's t-tests (one for each time point) to compare synchronized "fixed" and "random" key traces. If time points have the same means, it is a PASS.

# SCA and Reproducibility

- Originally critical value C=±4.5 was used for each t-test. When traces get longer, satisfying all t-tests gets harder. Do you even have a p-value anyway?

- Such errors in TVLA experiment design were pointed out e.g. in [Whitnall, Oswald. ASIACRYPT 2019]. Authors now use various methods to adjust critical values.

- Laboratory procedures: Sampling frequency? Target frequency? Connectors? Low/high pass filters? Amplitude normalization? Trigger jitter? ( ISO 20085 *"Test tool requirements and test tool calibration methods"* not entirely satisfactory. )

- We can't be sure if some people don't just repeat the TVLA until PASS 😭 How to distinguish between "debug" and "for real" experiments anyway?

# The Most Cited Paper of All Time (300,000+)!

PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT*

By OLIVER H. LOWRY, NIRA J. ROSEBROUGH, A. LEWIS FARR, AND ROSE J. RANDALL

(*From the Department of Pharmacology, Washington University School of Medicine, St. Louis, Missouri*)

(Received for publication, May 28, 1951)

- Vast majority of all-time top 100 cited papers describe experimental methods or software that have become essential in their fields.
- In cryptography it is probably the RSA paper. We don't have many "standard laboratory procedures." Especially for SCA we need more.

# Traces, VMs, and Very Large Artifacts

- IACR runs on volunteer effort; most things are done with no budget at all. However, we are currently hosting everything ourselves.

- We had to make some very large supplementary data sets available "only by request" due to technical/financial constraints

- USENIX Sec. and some others recommend: https://zenodo.org/ Operated by CERN in Switzerland with public money. Offers DOIs for all data sets. It is being considered.

# On Badges and Definitions ..

Badge Model:
**Usenix Security 2024**


Text used in artifact call:

**IACR CHES 2024 Artifacts Available:**

*To earn this badge, the AEC must judge that artifacts associated with the paper have been made available for retrieval. Other than making the artifacts available, this badge does not mandate any further requirements on functionality, correctness, or documentation. This is intended for authors who simply wish to make some supplementary material available that supports their paper. Examples include data sets, large appendices, and other documentation.*

Badge Model:
**Usenix Security 2024**


**IACR CHES 2024 Artifacts Functional:**

*To earn this badge, the AEC must judge that the artifacts conform to the expectations set by the paper in terms of functionality, usability, and relevance. The AEC will consider four aspects of the artifacts in particular.*

***Documentation:*** *are the artifacts sufficiently documented to enable them to be exercised by readers of the paper?*

***Completeness:*** *do the submitted artifacts include all of the key components described in the paper?*

***Exercisability:*** *do the submitted artifacts include the scripts and data needed to run the experiments described in the paper, and can the software be successfully executed?*

***Reusability:*** *means that the artifacts are not just functional but of sufficient quality that they could be extended and reused by others.*

Badge Model:
**Usenix Security 2024**


Text used in artifact call:

**IACR CHES 2024 Results Reproduced:**

*To earn this badge, the AEC must judge that they can use the submitted artifacts to obtain the main results presented in the paper. In short, is it possible for the AEC to independently repeat the experiments and obtain results that support the main claims made by the paper? The goal of this effort is not to reproduce the results exactly but instead to generate results independently within an allowed tolerance such that the main claims of the paper are validated.*

[**Note:** It has been suggested to change the word "reproduced" for "reproducible".]

# ACM is especially formal about artifacts..

https://www.acm.org/publications/policies/artifact-review-and-badging-current

*"ACM Task Force on Data, Software, and Reproducibility in Publication"* running since 2017.

ACM has **five** different badges (for conferences to use):

# ACM's Definitions (easily confused words)

**Repeatability** (Same team, same experimental setup)

The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.

**Reproducibility** (Different team, same experimental setup)

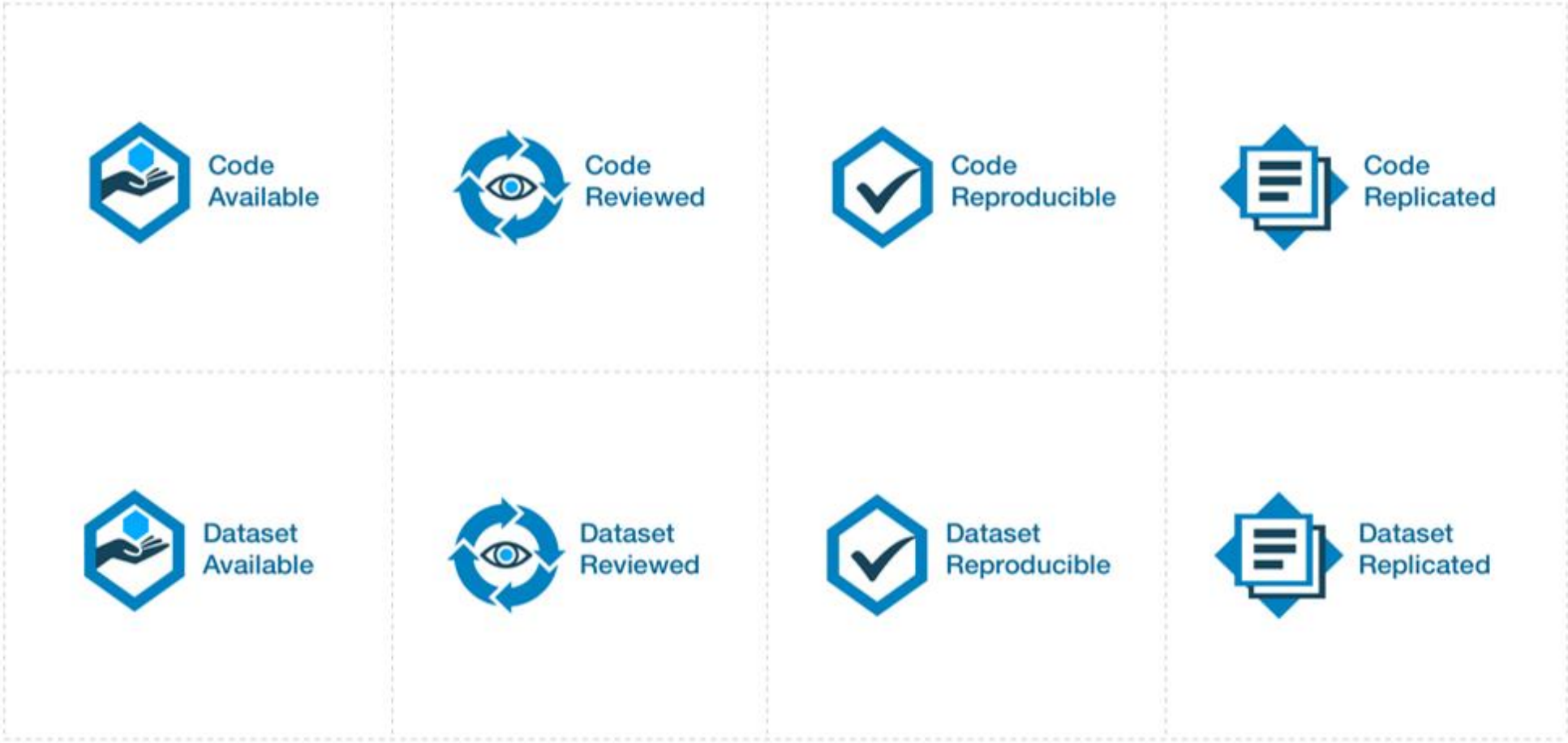
The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

**Replicability** (Different team, different experimental setup)

The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

# IEEE (Xplore) Reproducibility Badges

https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-content#reproducibility-badges



# IEEE Badge Definitions

1. **Available:** The code and/or datasets, including any associated data and documentation, provided by the authors is reasonable and complete and can potentially be used to support reproducibility of the published results.

2. **Reviewed:** The code and/or datasets, including any associated data and documentation, provided by the authors is reasonable and complete, runs to produce the outputs described, and can support reproducibility of the published results.

3. **Reproducible:** This badge signals that an additional step was taken or facilitated to certify that an independent party has regenerated computational results using the author-created research objects, methods, code, and conditions of analysis. Reproducible assumes that the research objects were also reviewed.

4. **Replicated:** This badge signals that an independent study, aimed at answering the same scientific question, has obtained consistent results leading to the same findings (potentially using new artifacts or methods). This badge is awarded by the publisher of the original work that is being badged.

# PETS is also doing badges

Proceedings on
Privacy Enhancing Technologies
Symposium

PoPETs | Go to PETS

Home | CFPs | Authors | Reviews | Proceedings | Awards | Sponsorship | Code of conduct

## PoPETs Artifact Review

PoPETs reviews and publishes digital artifacts related to its accepted papers. This process aids in the reproducibility of results and allows others to build on the work described in the papers. Artifact submissions are requested from authors of all accepted papers, and although they are optional, we strongly encourage you to submit your artifacts for review.

Possible artifacts include (but are not limited to):

- Source code (e.g., system implementations, proof of concepts)
- Datasets (e.g., network traces, raw study data)
- Scripts for data processing or simulations
- Machine-generated proofs
- Formal specifications
- Build environments (e.g., VMs, Docker containers, configuration scripts)

Artifacts are evaluated by the artifact review committee. The committee evaluates the artifacts to ensure that they provide an acceptable level of utility. Issues considered include software bugs, readability of the documentation, appropriate licensing, and the reproducibility of the results presented in the paper. After your artifact has been approved by the committee, we will accompany the paper link on petsymposium.org with a link to the artifact along with the obtained artifact badges so that interested readers can find and use your hard work.

# Authors seem proud of their badges..


**Terrapin Attack: Breaking SSH Channel Integrity By Sequence Number Manipulation**

Fabian Bäumer
*Ruhr University Bochum*

Marcus Brinkmann
*Ruhr University Bochum*

Jörg Schwenk
*Ruhr University Bochum*

**Abstract**

The SSH protocol provides secure access to network services, particularly remote terminal login and file transfer

**1 Introduction**

**Secure Shell (SSH).** While TLS is commonly used to secure user-facing protocols such as web, email, or FTP, SSH is

https://terrapin-attack.com/TerrapinAttack.pdf

# Trend: Need to justify *not* publishing artifacts

Announcement and Call for Papers www.usenix.org/sec25/cfp

## 34th USENIX Security Symposium

### August 13–15, 2025, Seattle, WA, USA

*Sponsored by USENIX, the Advanced Computing Systems Association*

usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

The USENIX Security Symposium brings together researchers, practitioners, system programmers, and others interested in the latest advances in the security and privacy of computer systems and networks. The 34th USENIX Security Symposium will be held on August 13–15, 2025, in Seattle, WA, USA.

Summary of main changes from previous editions

1. Two submission cycles instead of three.
2. New open science policy: Research results should be available to the public or explain why this is not possible. The artifact evaluation process is adjusted to accommodate this.
3. New guidelines for ethics considerations.
4. Extra page to discuss ethics considerations and compliance with open science policy.
5. Revisions are reviewed within the same submission cycle instead of the next.
6. New approach to presenting accepted papers (see the public RFC at https://github.com/USENIX-Security-2025/conference-format about the plans for this new model).

## Important Dates

**Cycle 2**

- Paper submissions due: **Wednesday, January 22, 2025**
- Early reject notification: **Tuesday, March 4, 2025**
- Rebuttal period: **April 7–14, 2025**
- Notification to authors: **Wednesday, April 30, 2025**
- Shepherding/revision period: **Thursday, May 1, 2025–Thursday, May 29, 2025**
- Artifacts due for availability verification: **Thursday, May 29, 2025**
- Shepherding/revision author notification: **Thursday, June 5, 2025**
- Final papers due: **Thursday, June 12, 2025**

## Symposium Topics

Refereed paper submissions are solicited in all areas relating to systems research in security and privacy. This topic list is not meant to be exhaustive; USENIX Security is interested in all aspects of computing systems security and privacy. Papers without a clear application to security or privacy of computing systems, however, will be considered out of scope and may be rejected without full review.

# Thank You !

Time for some questions and discussion..

These slides: https://mjos.fi/doc/20240904-optimist-artifact.pdf