

Well, It Worked on My Computer: Reproducibility in Security Research

Daniel Olszewski (“Ozzy”)

CSET

August 13th, 2024

Daniel Olszewski (“Ozzy”)

- Graduated from Carroll College in Helena, MT
 - B. A. in Mathematics and Computer Science
- Ph.D. student at the University of Florida
 - Advisor: Dr. Patrick Traynor
- Research Areas:
 - Deepfake analysis, network measurement, cellular security, and reproducibility



The Shoulders of Giants!

- Henry Cavendish is one of the most important scientists in all of history.
 - Credited with discovering hydrogen, conducting the most accurate estimation of Earth's mass, and much more!
- Famously shy...
- Eventually credited with discovering Ohm's Law, Charles' Law of gasses, Dalton's Laws of Partial Pressures, and Thermodynamics...
- *How much was the progress of science slowed by Cavendish not sharing all of his work?*



What If We're All Cavendish?

- The people in this room are among the most brilliant in the world, and are leaders in systems security.
 - Wow, look at those h-indices!
- But what of *impact*?
 - Are our methods and findings being hidden away in our private books?
 - What should we be doing as a community?



Let's Flip This...

- Have you ever...
 - ... tried to write a paper but been unable to compare against prior work?
 - ... tried to recreate a result from another paper but been unable to do so?
 - ... attacked by a reviewer for failing to do so?
 - ... wanted to take your research beyond a paper but couldn't?
- Why is that?



<https://commons.wikimedia.org/wiki/File:Stop-sign.jpg>

*We don't incentivize reproducibility, artifacts, or deployment,
so there is little reason for anyone to do it.*

Is reproducibility a problem or an opportunity?

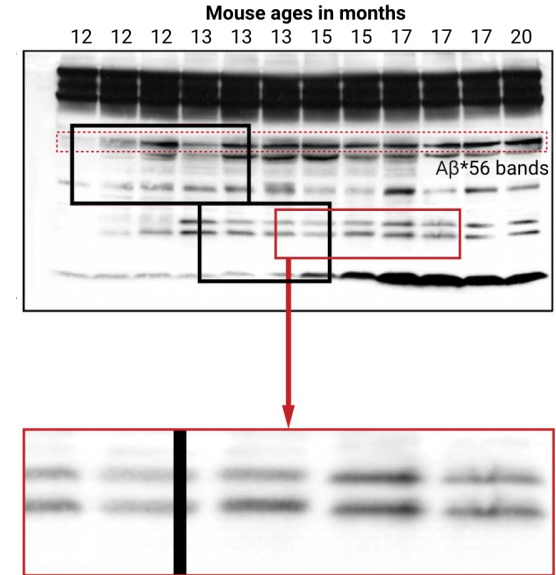
Do Your Homework!

- If you don't do your homework, someone else will have to do it for you.
- Inspired by an artifact that the authors left “reproducing the experiments as an exercise for the reader.”



My Goal Today

- This is not a witch-hunt.
 - Psychology and Medicine
- I want to learn about the things that make real systems robust.
- I want you to get actual credit for your efforts.
- Keep the public trust.



<https://www.science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease>

- Definitions
- Measuring Reproducibility/Case Studies
 - Olszewski, et al. “**Get in Researchers; We’re Measuring Reproducibility**”: **A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences**. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2023.
 - Olszewski, et al., “**Raise Your Hand If You’ve Been Personally Victimized By A Lack Of Reproducibility**”: **On Reproducibility in Tier 2 Security Conferences**, In Submission, 2024.
- Future Work and Open Challenges
- How can we make progress as a community?

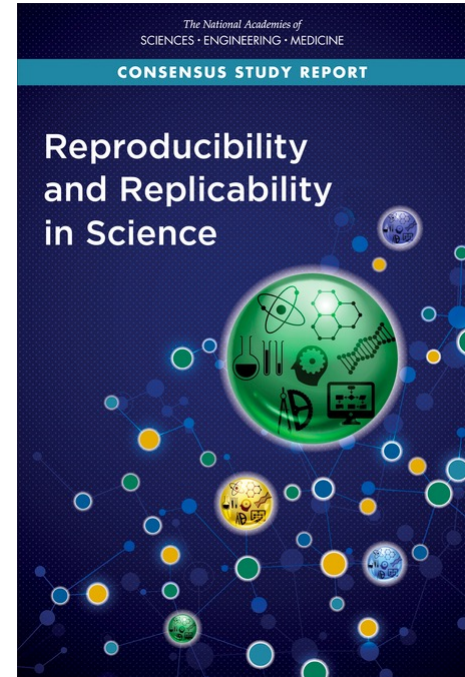


Some Definitions Disagreement

- Even the terms *reproducibility* and *replicability* are not fully agreed upon.
 - Economics and Political Science use the terms with no distinction.
 - Most of science (e.g., Signal Processing, Econometry, Epidemiology, Clinical Studies, Internal Medicine, etc)
 - Computer science and microbiology historically used the opposite definitions of these other fields.
- ACM harmonized its definitions with the National Academies in 2021, so let's use those for our discussion.

Definitions Time

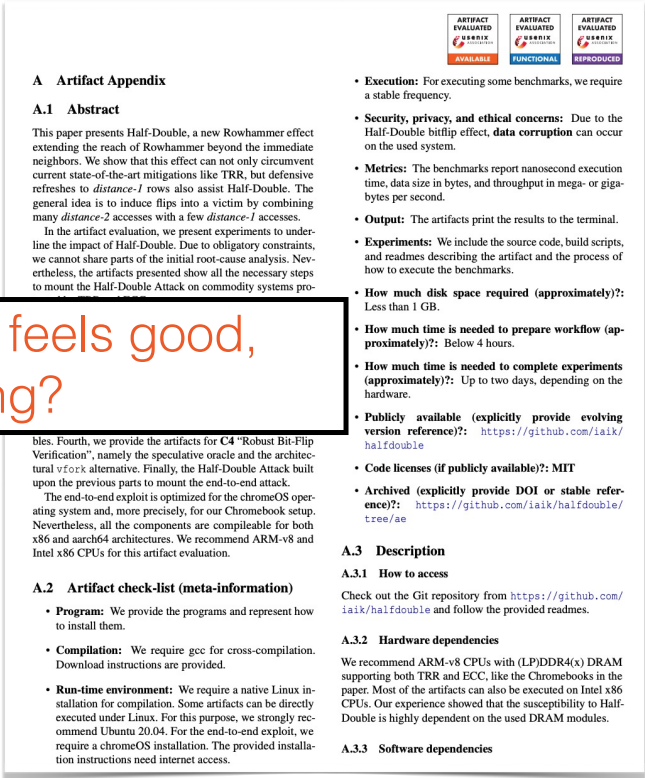
- **Reproducibility:** Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.
- **Replicability:** Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.
- **Generalizability:** The extent that results of a study apply in other contexts or populations that differ from the original one.



Changes and Context

- ACM WiSec was the first conference (to my knowledge) to have a reproducibility initiative in 2017 (<https://wisecdata.ccs.neu.edu/>).
- A few others followed:
 - ACSAC (2017)
 - WOOT (2018)
 - USENIX Security (2020)
 - CCS (2023*), NDSS (2025)
 - IEEE S&P (?)
- USENIX has a tiered process with “Available”, “Functional”, and “Reproducible” distinctions.

Meta Question: Artifact Evaluation feels good, but has it changed anything?



A Artifact Appendix

A.1 Abstract

This paper presents Half-Double, a new Rowhammer effect extending the reach of Rowhammer beyond the immediate neighbors. We show that this effect can not only circumvent current state-of-the-art mitigations like TRR, but defensive refreshes to *distance-1* rows also assist Half-Double. The general idea is to induce flips into a victim by combining many *distance-2* accesses with a few *distance-1* accesses.

In the artifact evaluation, we present experiments to underline the impact of Half-Double. Due to obligatory constraints, we cannot share parts of the initial root-cause analysis. Nevertheless, the artifacts presented show all the necessary steps to mount the Half-Double Attack on commodity systems pro-

bles. Fourth, we provide the artifacts for C4 “Robust Bit-Flip Verification”, namely the speculative oracle and the architectural *vFork* alternative. Finally, the Half-Double Attack built upon the previous parts to mount the end-to-end attack.

The end-to-end exploit is optimized for the chromeOS operating system and, more precisely, for our Chromebook setup. Nevertheless, all the components are compilable for both x86 and aarch64 architectures. We recommend ARM-v8 and Intel x86 CPUs for this artifact evaluation.

A.2 Artifact check-list (meta-information)

- **Program:** We provide the programs and represent how to install them.
- **Compilation:** We require gcc for cross-compilation. Download instructions are provided.
- **Run-time environment:** We require a native Linux installation for compilation. Some artifacts can be directly executed under Linux. For this purpose, we strongly recommend Ubuntu 20.04. For the end-to-end exploit, we require a chromeOS installation. The provided installation instructions need internet access.

A.3 Description

A.3.1 How to access

Check out the Git repository from <https://github.com/iaik/halldouble> and follow the provided readmes.

A.3.2 Hardware dependencies

We recommend ARM-v8 CPUs with (LP)DDR4(x) DRAM supporting both TRR and ECC, like the Chromebooks in the paper. Most of the artifacts can also be executed on Intel x86 CPUs. Our experience showed that the susceptibility to Half-Double is highly dependent on the used DRAM modules.

A.3.3 Software dependencies

ARTIFACT EVALUATION

- AVAILABLE
- FUNCTIONAL
- REPRODUCIBLE

- **Execution:** For executing some benchmarks, we require a stable frequency.
- **Security, privacy, and ethical concerns:** Due to the Half-Double bitflip effect, **data corruption** can occur on the used system.
- **Metrics:** The benchmarks report nanosecond execution time, data size in bytes, and throughput in mega- or giga-bytes per second.
- **Output:** The artifacts print the results to the terminal.
- **Experiments:** We include the source code, build scripts, and readmes describing the artifact and the process of how to execute the benchmarks.
- **How much disk space required (approximately)?**: Less than 1 GB.
- **How much time is needed to prepare workflow (approximately)?**: Below 4 hours.
- **How much time is needed to complete experiments (approximately)?**: Up to two days, depending on the hardware.
- **Publicly available (explicitly provide evolving version reference)?**: <https://github.com/iaik/halldouble>
- **Code licenses (if publicly available)?**: MIT
- **Archived (explicitly provide DOI or stable reference)?**: <https://github.com/iaik/halldouble/tree/ae>

- **Paper Selection** - 744 ML papers between 2013 - 2022

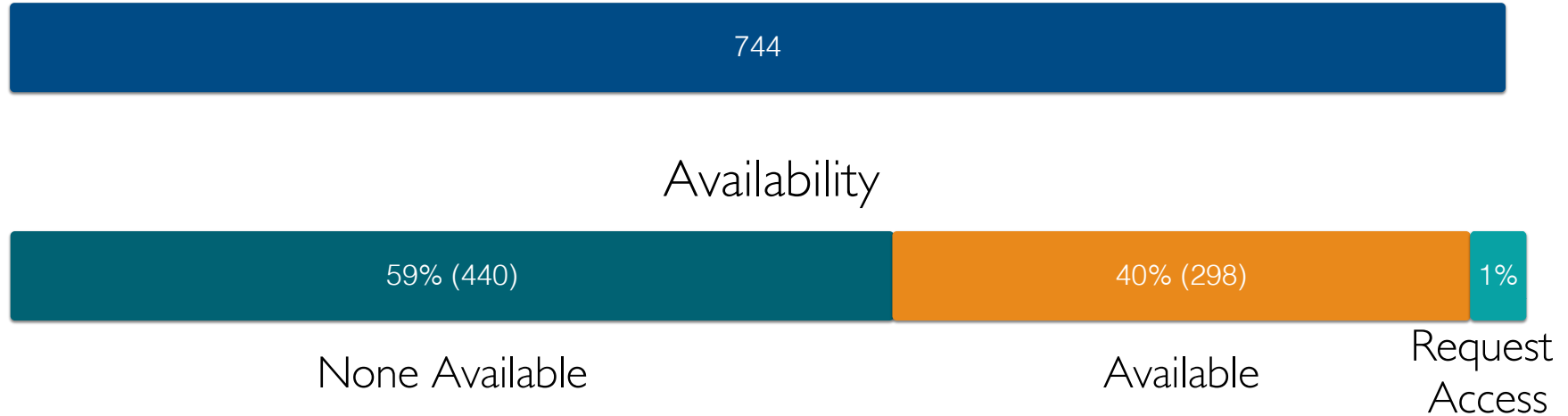


- **Indirect Study** - Measuring the availability of Method, Data, and Experiment.
- **Direct Study** - Measuring the success of reproducing results.
 - Downloading and running the artifacts.

1. To what extent is the collected data made available?
2. To what extent are experimental artifacts made available?
3. Of available experimental artifacts, how many run and produce consistent results?



Indirect Study - Experiment

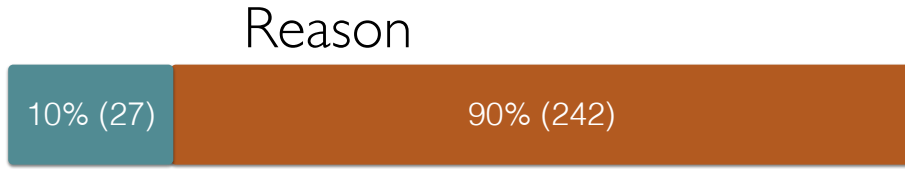


8 Link Empty Repo or say "Available after Publication"

RQ2 ✓

Indirect Study - Data

- 28% use public data
- 35% collect and publish data
- 36% do not provide data



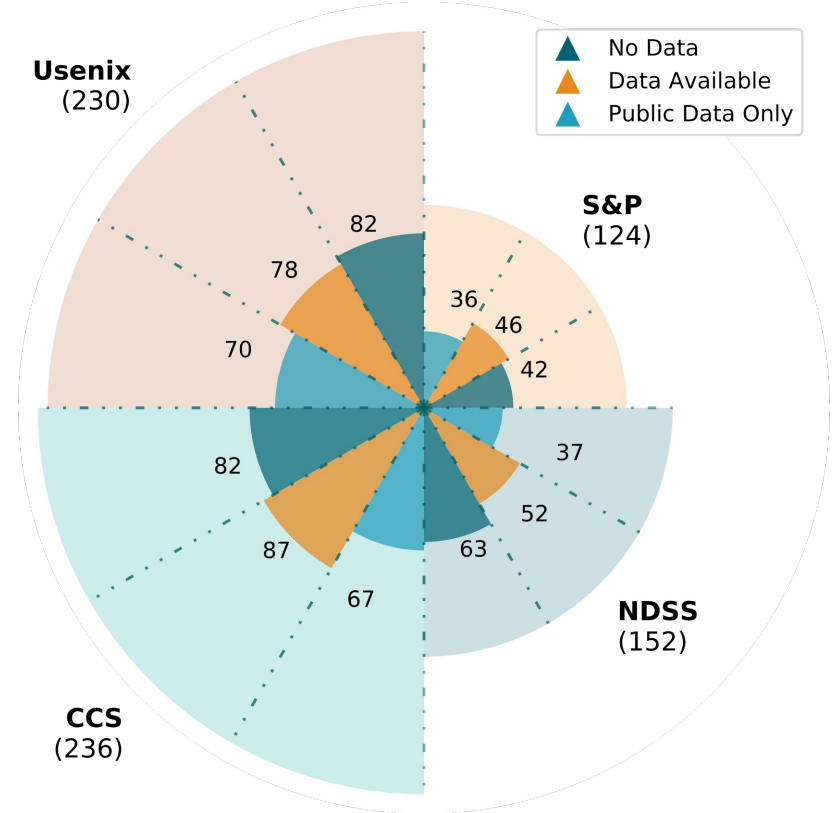
Sensitive Data

None

- Only 6% of papers are industry.

RQI ✓

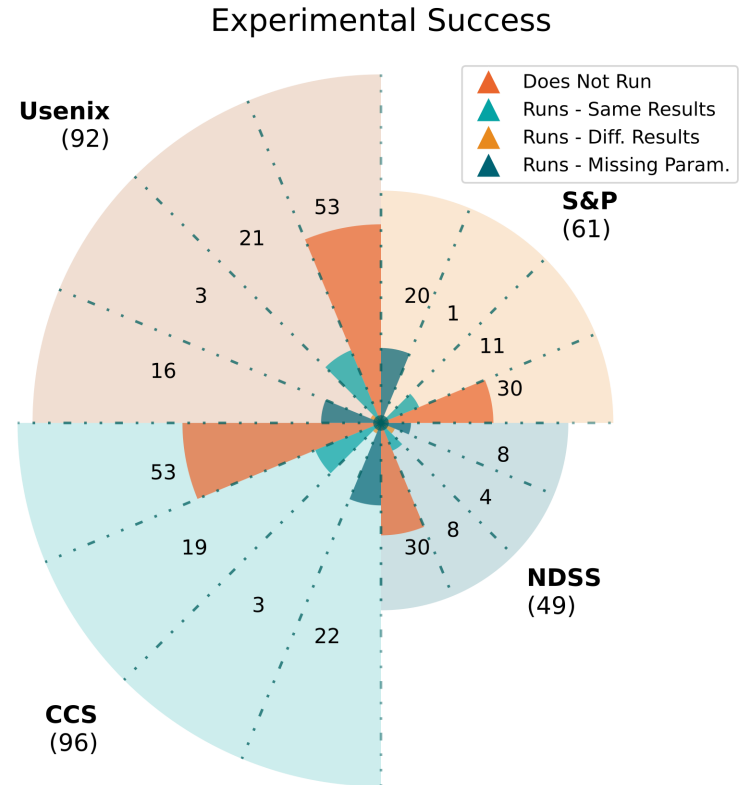
Data Availability



Direct Study

- 56% did not run
- 20% ran with same results
- 4% ran with different results
- 22% worked but missing parameters
- Data not provided.
- Missing processing scripts.

RQ3 ✓



Common Errors in Repositories

 Missing Files

 Missing Functions

 Uninitialized Arrays

 Pre-processing scripts

 Deprecated Packages

 Unformatted Code

Direct Study - Continued

298

Detailed ReadMe?

Yes

57% (170)

30% (89)

13% (39)

No

Output Matching?

Yes

43% (128)

57% (170)

No

Included Trained Model?

Yes

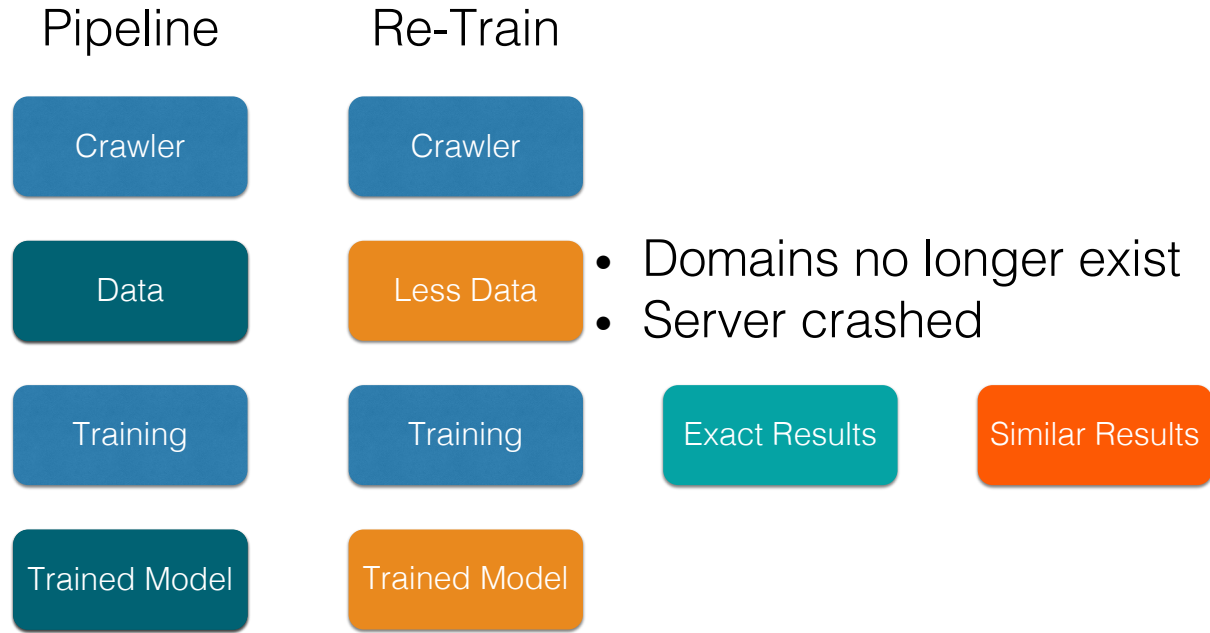
17% (51)

83% (247)

No

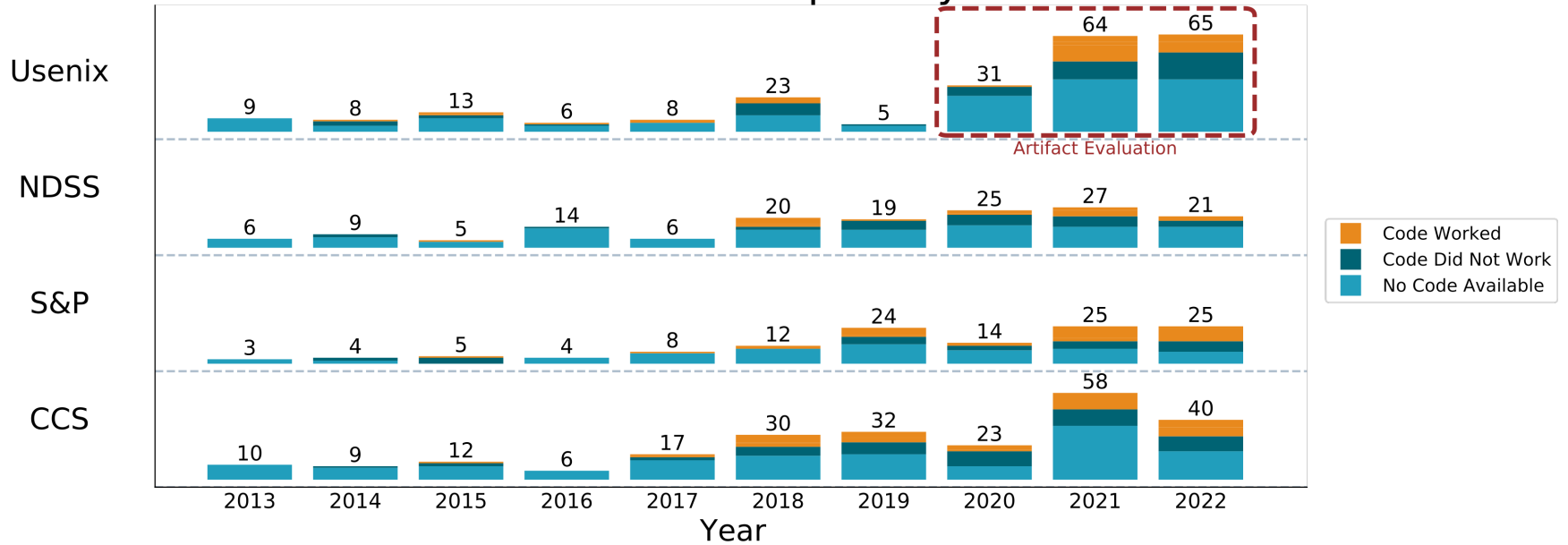
Case Study: Automating Cookies

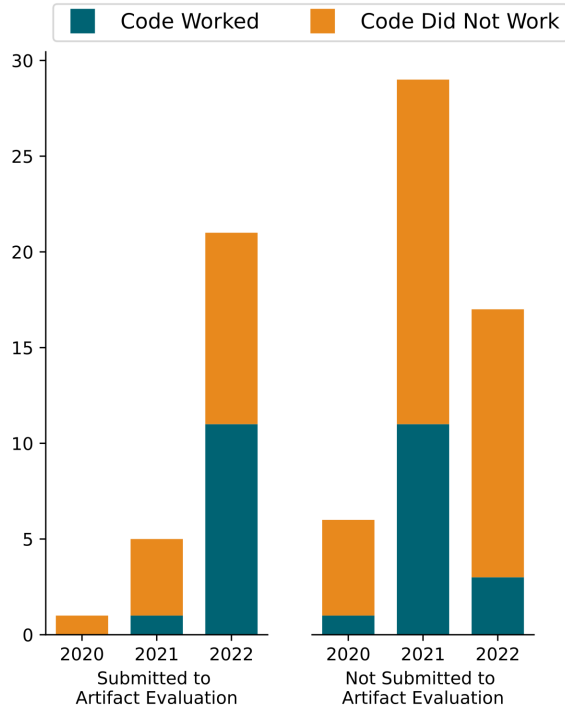
- Won the Artifact Award at USENIX 22



There is no difference in whether code from published papers is available before vs after the introduction of AECs to Tier I Security Conferences.

Total Number of Papers by Year





1. Several artifacts that did not go to the AEC that worked.
2. In one example, we saw same trend, but smaller change.
3. Found badge not awarded, but we were able to reproduce the results.

Null Hypothesis

There is no difference in whether code from published papers is available before vs after the introduction of AECs to Tier I Security Conferences.

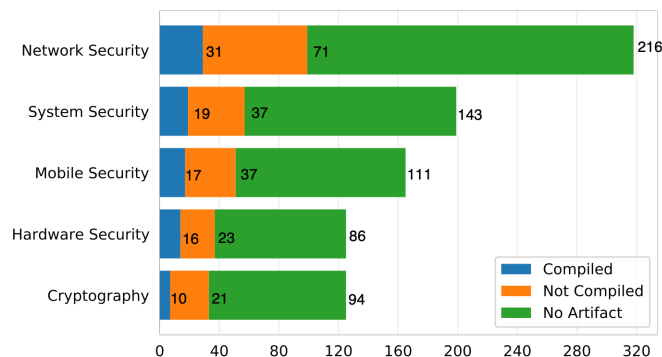
$$p = 0.068$$

Thus, we *fail to reject* the null hypothesis.



What About Tier 2s?

- WiSec and ACSAC started artifact evaluation three years earlier than USENIX, so maybe the answer is there!
- We collected ALL (2,000+) papers between 2013-2023 with ~580 artifacts to answer just this question from ACSAC, WiSec, EuroS&P, and AsiaCCS.
- Four person years of effort with over 1,000 hours of computational effort.

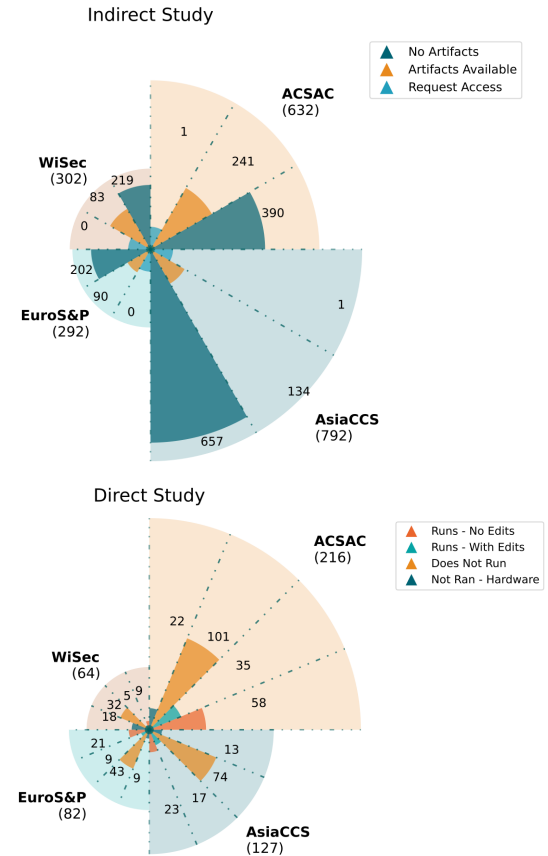


- Create Docker instances
 - Match the OS
 - Follow experimental methodology
- Outcome
 - Every artifact available is now Dockerized.
 - Mass recreation of experiments!

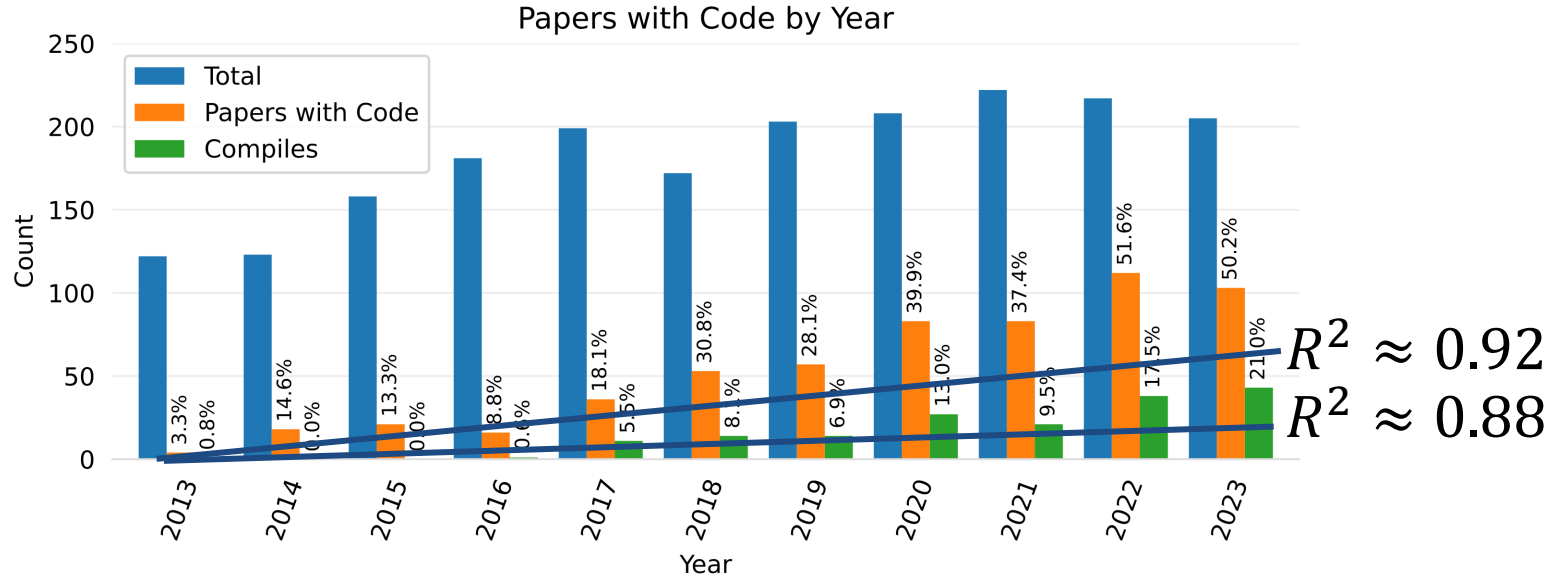


All Things (Not) Equal

- ACSAC dominates the space of having available artifacts.
- But the picture remains rather grim:
 - Only a small fraction of papers from Euro S&P/AsiaCCS have code.
 - Even for ACSAC, only 58 of the 632 papers have running software.

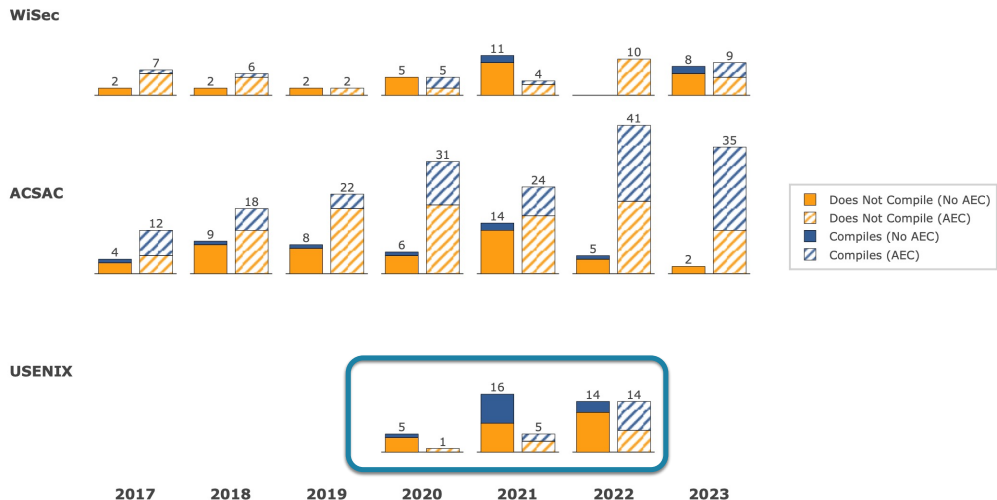


All Things (Not) Equal



Artifact Evaluation Committees

Code Compiles by Conference

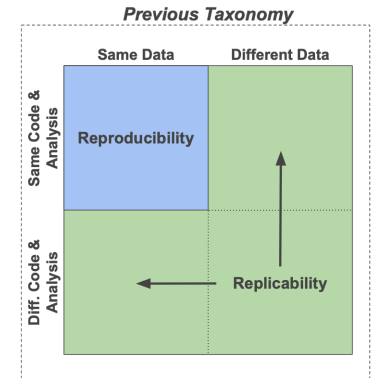


- ACSAC has a growing trend towards all available artifacts going through the AEC.
- WiSec has no apparent trends.

- Artifact evaluation does not tell the entire story - Code alone is not enough.
 - Many reproduced papers are on smaller datasets.
- Randomness in initialization, training, and data selection all affect reproducibility (sometimes significantly).
- While vast improvements have been made, there is still room for further growth.
- We are not yet meaningfully dealing with replicability.



- Understanding the proliferation of datasets
 - Layton, et al. **SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Media Datasets**. *In proceedings of USENIX Security 2024*.
 - (Ongoing) Security Dataset Usage. How often are datasets made available? How are they used by other researchers?
- Adapting frameworks for reproducibility
 - (Ongoing) Replicability. What does it actually mean? And how do we do it?
 - (Ongoing) Analysis of security research methodologies. What statistical tools are we using? How do we use these tools to make claims? Is this consistent with statistical standards?



- Iterating on our frameworks with feedback from user studies:
 - AECs - How are AEC members interacting with artifacts? What can we do to improve the process?
 - Successful artifact creators - What can we learn from their process?
- Building avenues for transition to industry practices:
 - Does reproducible research transition to industry better than un-reproduced research?
 - Collaborating with Industry Partners.
 - Developing incentives and processes for easier transitions.

Open Challenges

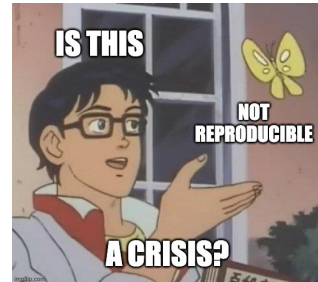
- Bitwise Reproducibility
 - When running the experiments, the exact same bytes are outputted as the original experiment.
 - Is it possible? If it isn't, how do we describe computational reproducibility?
- Reproducibility Platforms
 - Zenodo was designed for hosting reproducible research, but 95% of the repositories we studied were hosted by GitHub.
 - Why are researchers not using Zenodo? What ways can we improve the platforms?
 - USC-ISI introduced SPHERE to provide testbeds.
- Validating Meta-Research Studies
 - All research should undergo validation. As more meta-research becomes available, we need to analyze these results and determine better practices.

1	0	1	1	0	0	0	1
1	0	1	1	0	0	0	1



Is This A Crisis?

- As a whole, our two studies paint a somewhat negative picture of reproducibility in Computer Security.
 - An increasing number of papers include code but...
 - ...**less than 4%** of work in the past decade appears to be readily reproducible.
- We can't say, for certain, that we are in a reproducibility crisis, but we certainly can't rule it out either...
- Maybe that's because reproducibility is hard!



So... Why? (aka Challenges)

- Making things available, keeping them alive, and transition to practice is not part of most funding models.
- Nor is it part of most academic publication models.
 - Near constant pressure for “novelty” crowds out moving things forward.
 - Negative results are critical...
 - Rarely how we make hiring decisions...
- How do we rethink our efforts?

The proposed project does not appear to focus on a technical innovation that is creative, original or potentially transformative.

- Foster open-science practices
 - Start with *Reproducibility* as a goal.
 - Consider deployment possibilities.
 - Develop access plans for when students graduate.
 - Containerize!
- Confirm your artifacts with independent scientists.
 - Internal artifact evaluation.

How Do We Do Better? Academia

- Give heavier weights to papers that make systems and data available.
- Make participation in Artifact Evaluation Committees count for authors AND for evaluators.
- Reports from artifact evaluations should be published along with the paper.
- Publish negative results.
- Accept “next-step” papers (especially when they build on the above).
- More conferences should accept “SoK”-style papers.
- Better identify sources of randomness that make perfect reproducibility challenging.
- Consider alternative publication models (e.g., Results-blind Peer Review)

But It's About Balance

- Not everything has to be reproducible.
 - “Exploratory research is more susceptible to non-replication, while confirmatory research is less likely to uncover exciting new discoveries. **Both types of research help move science forward.**” -National Academies
 - For many reasons (e.g., protecting intellectual property, lack of distribution rights to data, etc), not every part of every paper may be publishable.

We Stand on the Shoulders of Giants!

- The intellectual contributions of this community are truly outstanding!
 - *Citing each other is the easy part*, but it's just the beginning of the hard work we need to be doing.
- Later transition will be easier, with better reproducibility!
- Whether out of **fear**, **obligation** or **optimism**, we need to make reproducibility and transition first-class citizens in our community.
 - If it only works on your computer, it's only the progress of science that we're holding up!





Patrick Traynor
traynor@ufl.edu



Daniel "Ozzy"
Olszewski



Allison
Lu

Well, It Worked on My Computer: Reproducibility in Security Research

Daniel Olszewski (“Ozzy”)

CSET

August 13th, 2024