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

Prof. Patrick Traynor

LASER Workshop
10 December 2024

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

Simple Truth

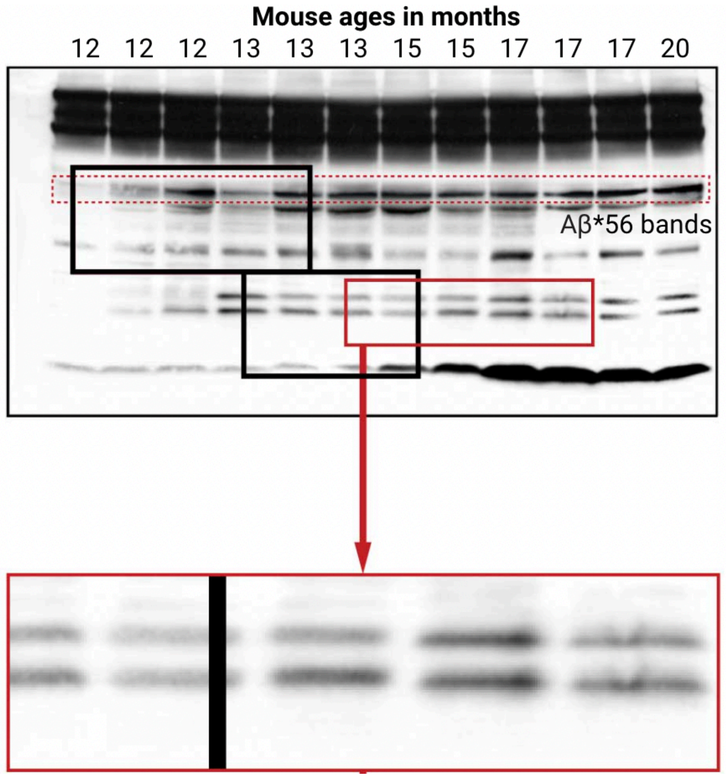
We don't incentivize reproducibility, artifacts, or deployment, so there is little reason for anyone to do it.

Who Are You Anyway?

- What have I done to warrant an informed opinion of any of this?
  - 100+ publications and ~20 years of experience.
  - Program Chair/Co-Chair for conferences including IEEE S&P, USENIX Security, NDSS, ACSAC, and ACM WiSec.
  - Co-Founder of three start-ups: Pindrop Security, CryptoDrop, Skim Reaper
  - FirmWire open-source baseband fuzzing/analysis platform co-author.
- I spend a lot of time thinking about what comes next and how to get things working beyond my own computer.

My Goal Today

- This is not a witch-hunt.
- 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

Outline

- 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.
- How can we make progress as a community?

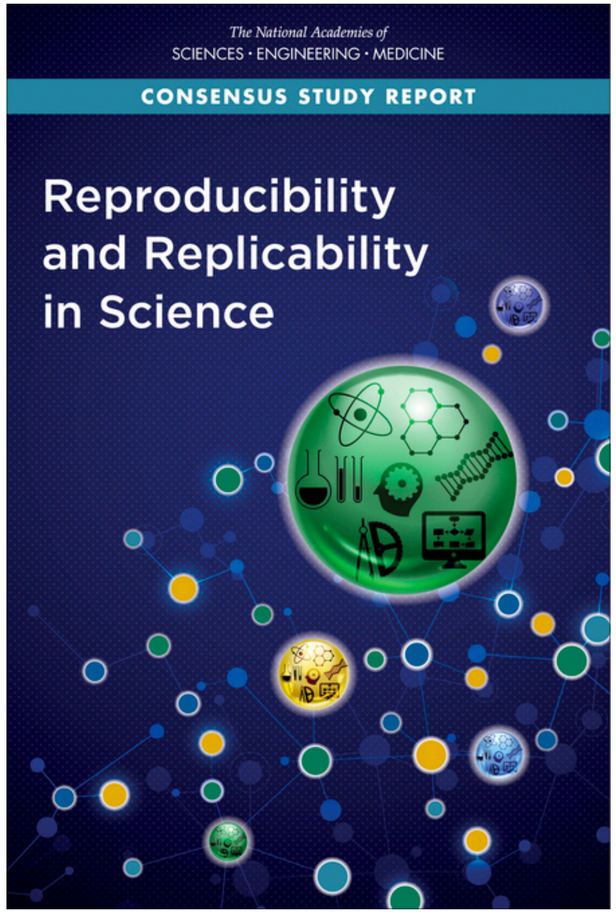


Definitions Time

- *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.



Case Study: Android

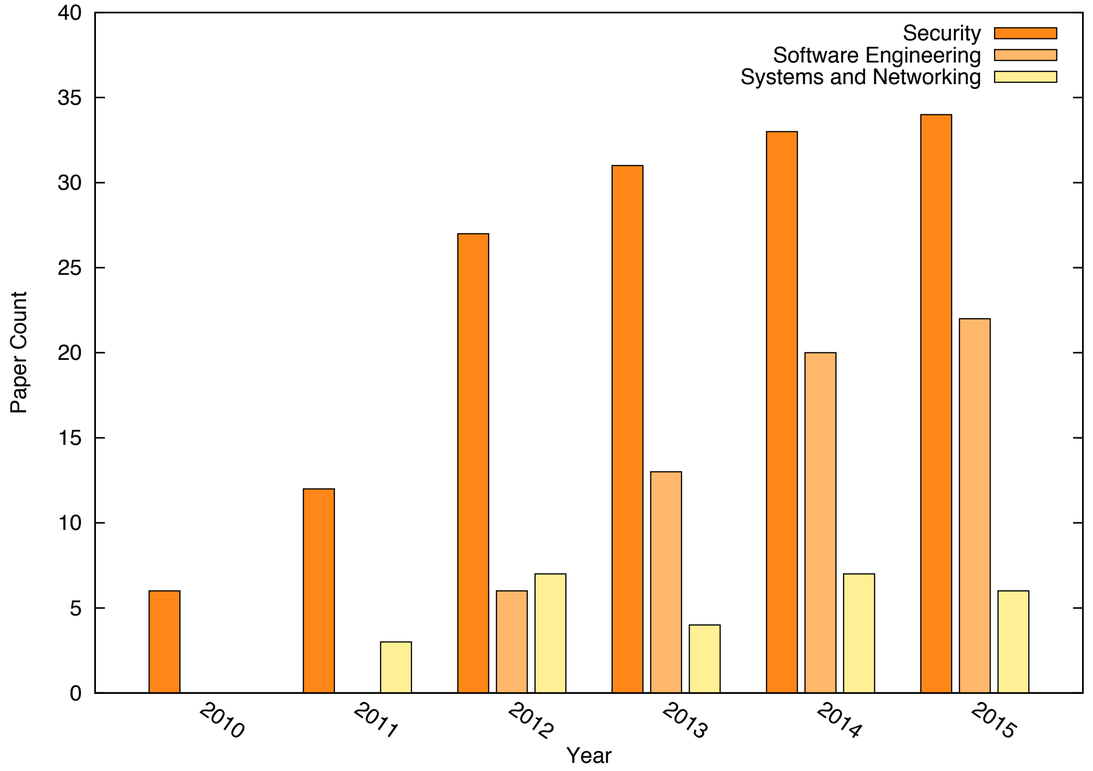
- Researchers published a massive body of work around the Android operating system and mobile applications.
- A few questions…
  - Can anyone else use them?
  - Can anyone else reproduce results?
  - How do they work against applications other than those the researchers picked?



https://commons.wikimedia.org/wiki/File:Foldable_smartphone_(Android_OS).jpg

Details

- We focus on *app analysis tools*, trying to determine whether or not app developers could use these tools to improve their applications.
- Collection
  - 289 Academic papers evaluated from 2010-2015
    - 17 venues including S&P, USENIX, ACSAC, CCS, NDSS, NSDI, etc
- Each paper had a main auditor, with a second researcher checking results.
  - 8 students, 3 faculty, evaluation conducted over 9 months.



Reaves, et al., ***droid: Assessment and Evaluation of Android Application Analysis Tools**, ACM Computing Surveys (CSUR), 2016.

Reproducibility and Android

- Of the 289 app analysis tool papers
  - Only 22 (~8%) had code available.
  - We were only able to get 7 (~2.5%) working.
    - 3 of the 7 tools came from one research group.
  - One ran, but failed to achieve the same results as the paper (100%) on the same data (~70%).

Replicability and Android

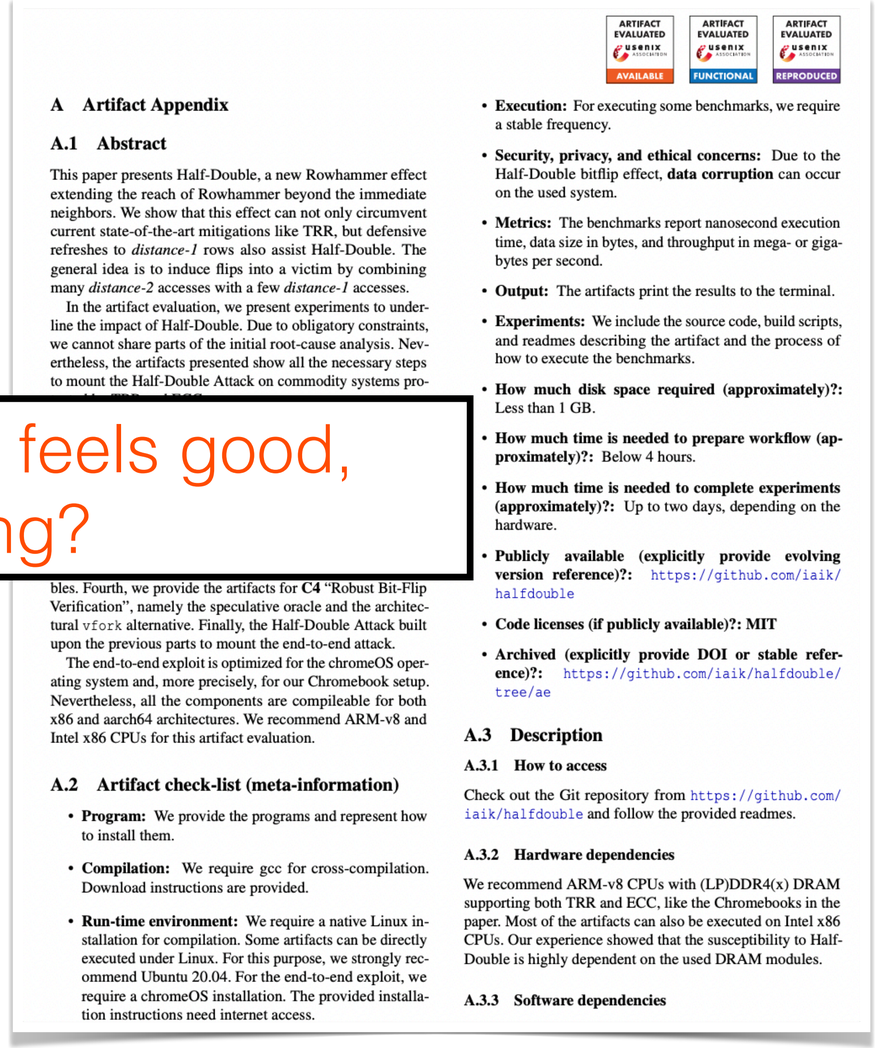
- Using different datasets tells an even different story.
  - Top 10 finance applications in Google Play (e.g., PayPal, Wells Fargo, Bank of America) and 7 Mobile Money Applications.
- Tools worked on toy applications, struggled with real ones.
  - For one tool, we tried provisioning an EC2 instance with 64GB RAM and 12 CPUs… it wasn't enough.
- In short, things rarely worked as well as they did in the paper when we looked at new applications.



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 (
  - 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?



ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

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 compileable 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.
- **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 gigabytes 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/halfdouble`
- **Code licenses (if publicly available)?: MIT**
- **Archived (explicitly provide DOI or stable reference)?:** `https://github.com/iaik/halfdouble/tree/ae`

A.3 Description

A.3.1 How to access

Check out the Git repository from `https://github.com/iaik/halfdouble` 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

Research Questions

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?



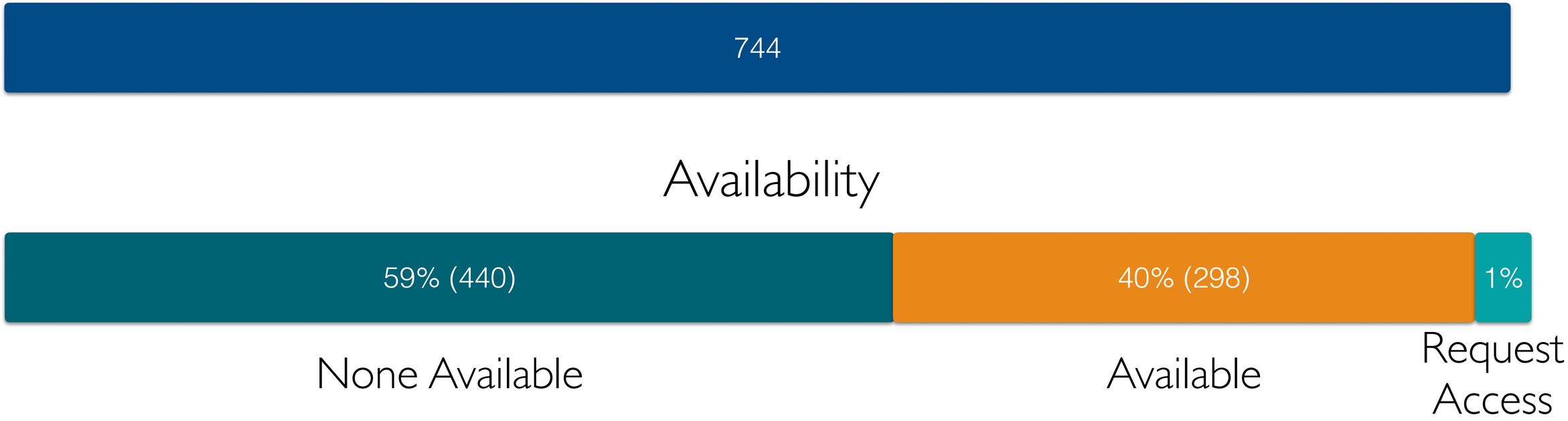
Methodology

- *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.

Indirect Study - Experiment

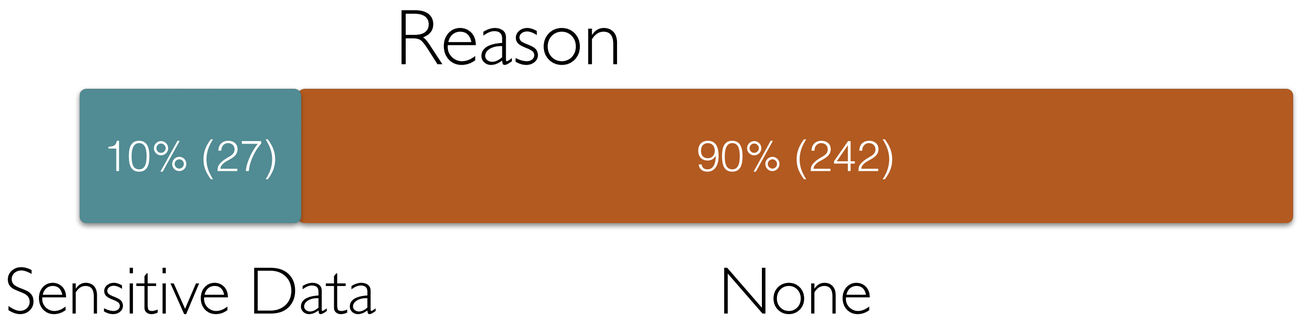


8 Link Empty Repo or say "Available after Publication"

RQ2 ✓

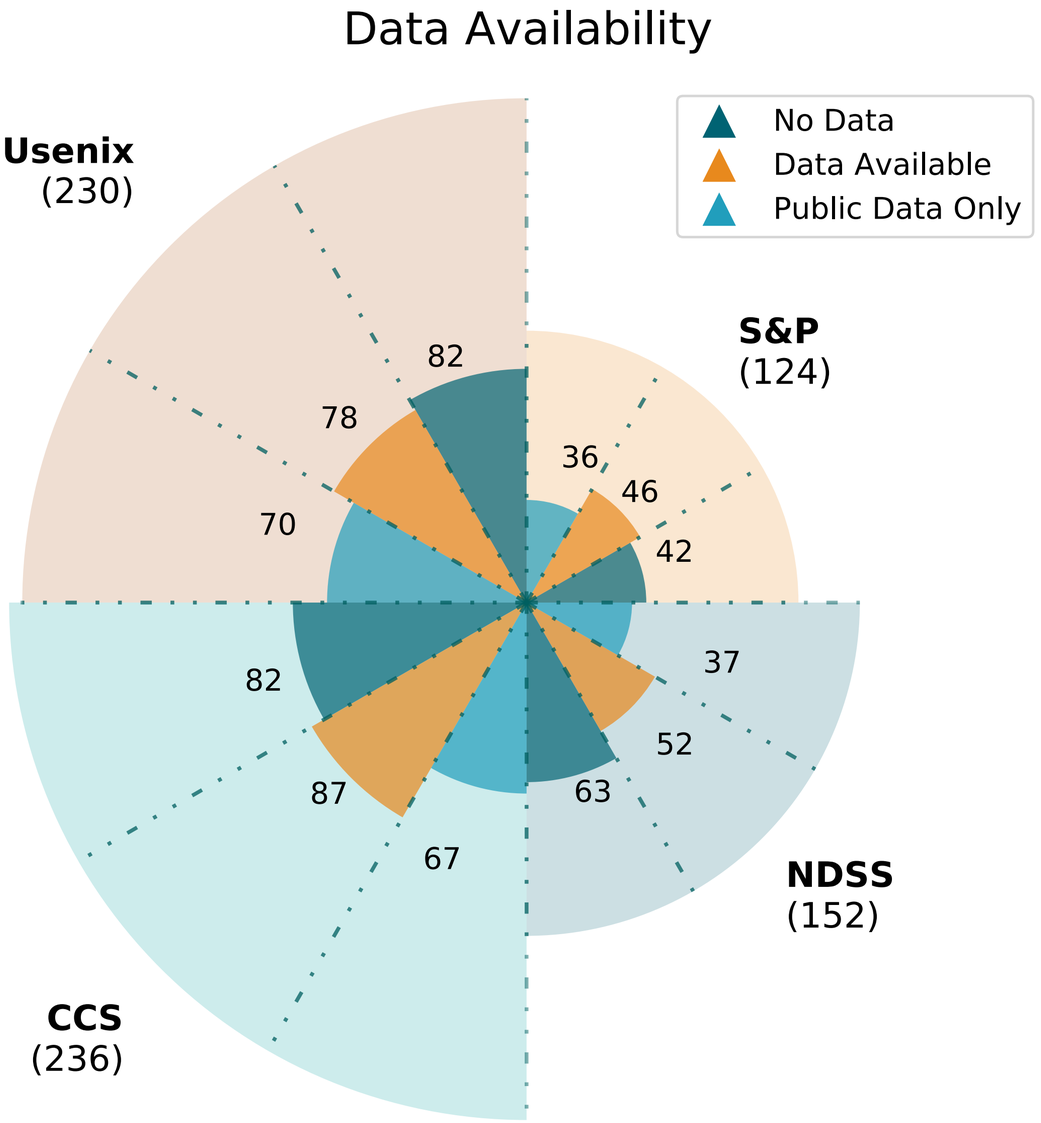
Indirect Study - Data

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



- Only 6% of papers are industry.

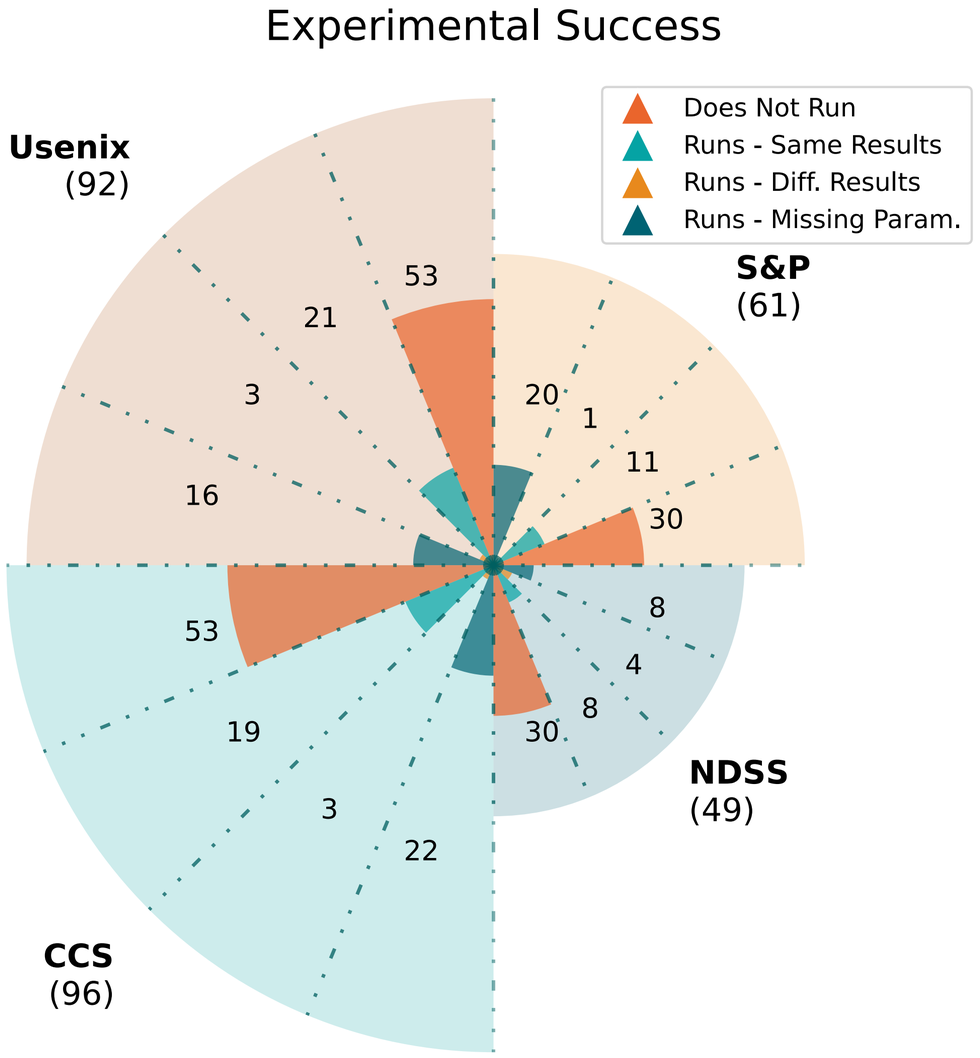
RQ1 ✓



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

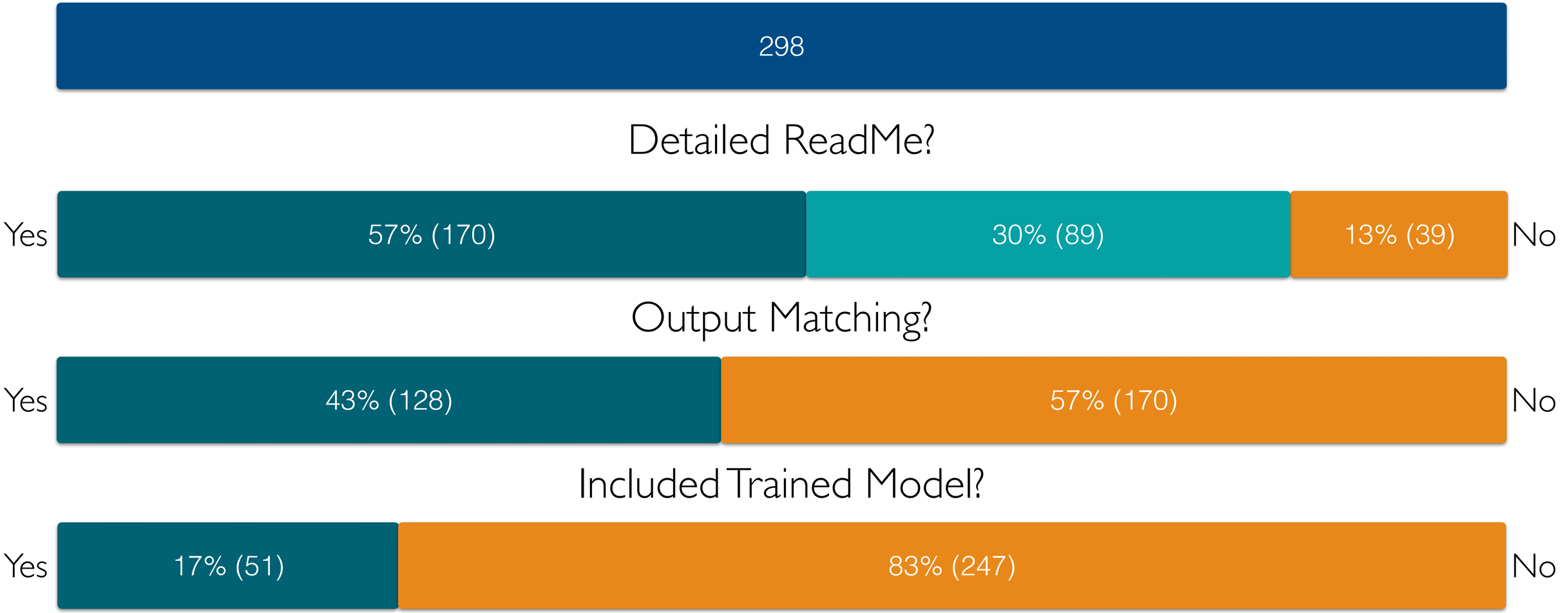

Uninitialized Arrays


Pre-processing scripts


Deprecated Packages


Unformatted Code

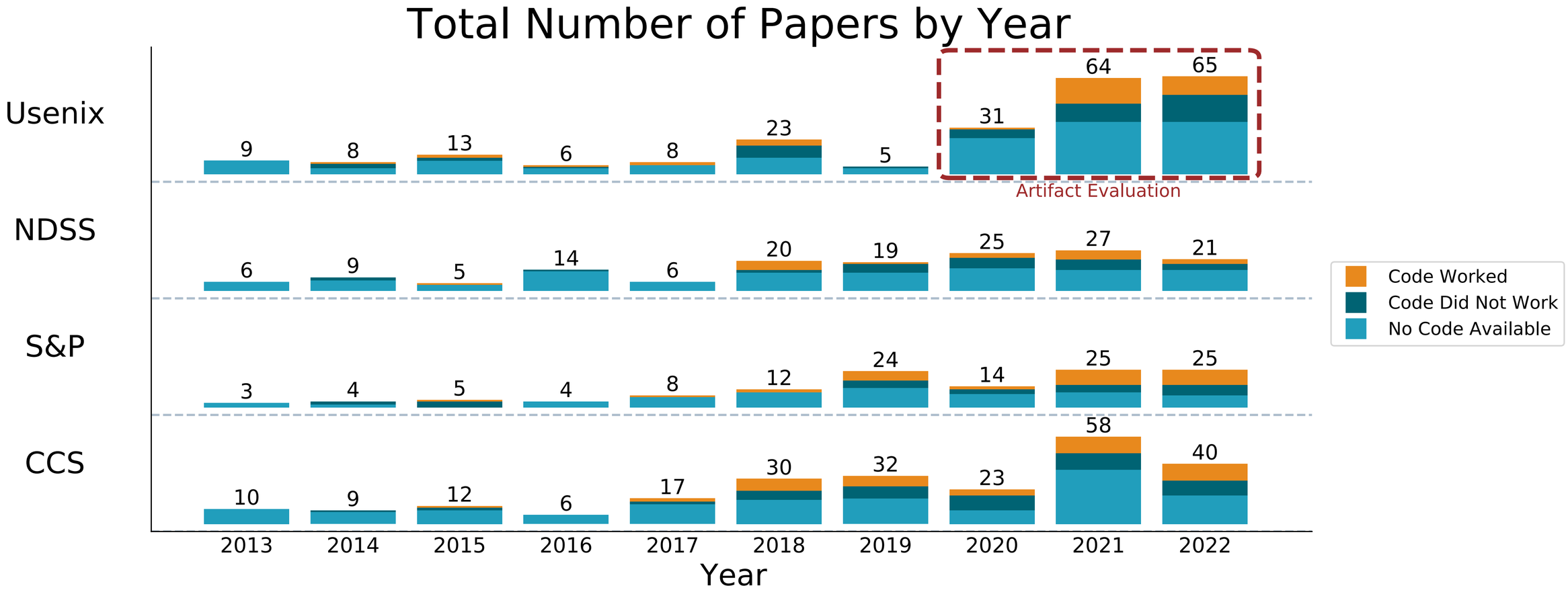
Direct Study - Continued



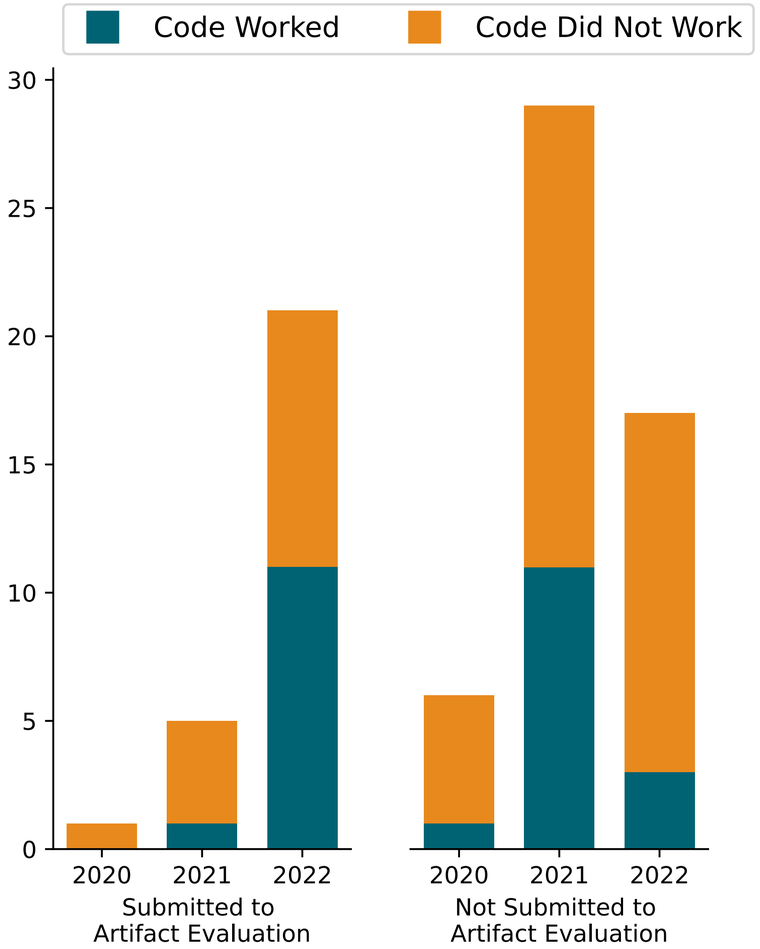
Null Hypothesis

There is no difference in whether code from published papers is available before vs after the introduction of AECs to Tier 1 Security Conferences.

Time Graph


Total Number of Papers by Year
Usenix
NDSS
S&P
CCS
9
8
13
6
8
23
5
31
64
65
Artifact Evaluation
6
9
5
14
6
20
19
25
27
21
3
4
5
4
8
12
24
14
25
25
10
9
12
6
17
30
32
23
58
40
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Code Worked
Code Did Not Work
No Code Available

USENIX Security AEC



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 1 Security Conferences.

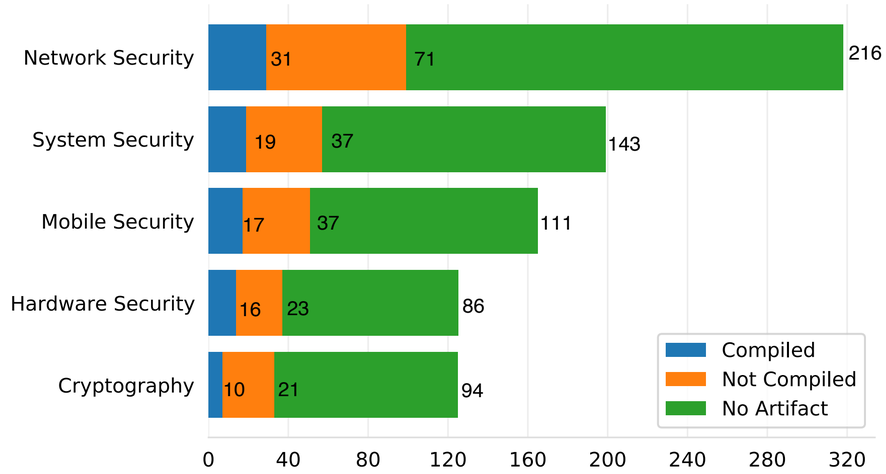
$p = 0.068$

Thus, we fail to reject the null hypothesis.



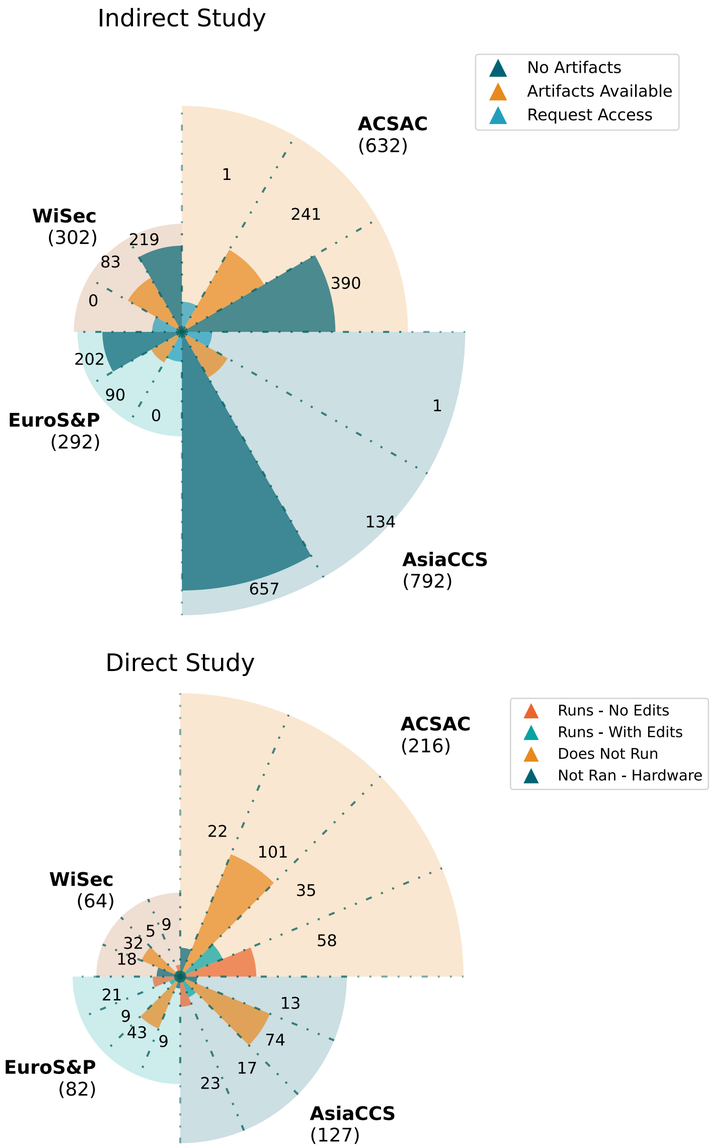
What About Tier 2s?

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



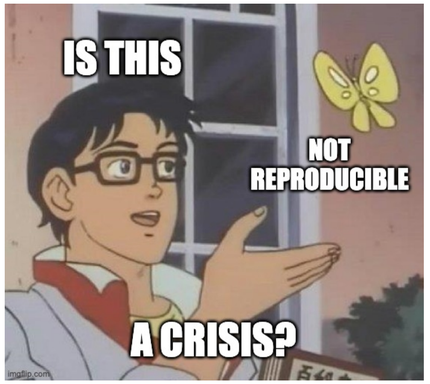
All Things (Not) Equal

- 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.



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!



Insights

- 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.







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.

It's About Robustness

Scientific theories are tested every time someone makes an observation or conducts an experiment, so it is misleading to think of science as an edifice, built on foundations. Rather, *scientific knowledge is more like a web*. The difference couldn't be more crucial. A tall edifice can collapse—if the foundations upon which it was built turn out to be shaky. But a web can be torn in several parts without causing the collapse of the whole. The damaged threads can be patiently replaced and re-connected with the rest—and the whole web can become stronger, and more intricate.

Nonsense on Stilts: How to Tell Science from Bunk, Massimo Pigliucci

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.
  - None of what we have works for hardware.

How Do We Do Better? Academia

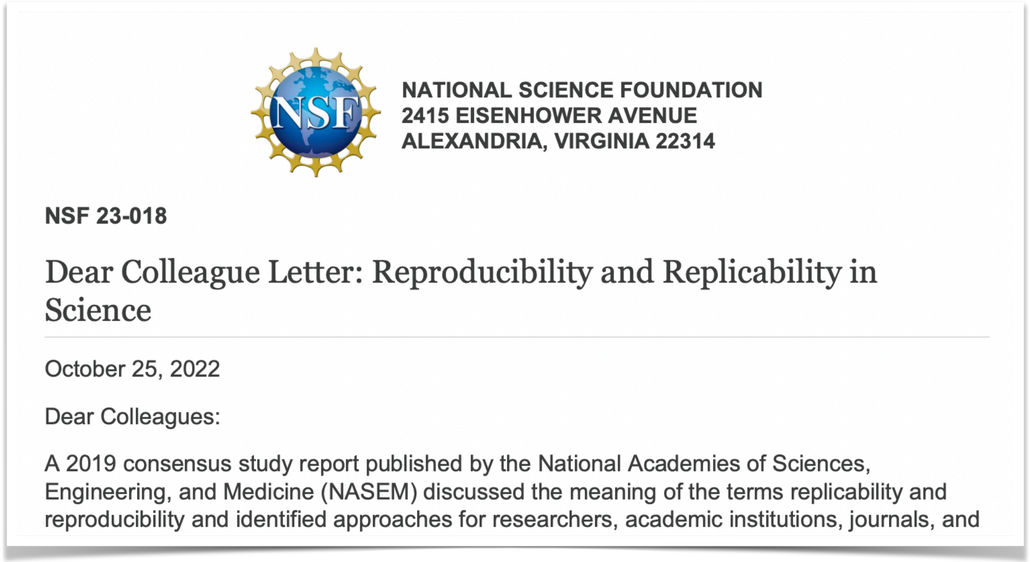
- Achievable:
  - Give heavier weights to papers that make systems and data available.
  - Make participation in Artifact Evaluation Committees count for authors AND for evaluators.
  - Reports from artifact evaluations should be published along with the paper.
  - Publish negative results.
  - Accept "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)

How Do We Do Better? Academia

- *Radical*:
  - Don't require comparison to papers that are not reproducible.
    - Consider citing them in new ways.
  - Build new tracks for papers, such as "Transition to Deployment" (TTD).
  - Write tenure letters about fewer papers, more systems-focus.
  - Meaningfully integrate reproducibility and entrepreneurship into faculty evaluations.
    - Expand the "Three-Legged Stool" of Research, Teaching, and Service
  - We don't have many meaningful retractions. What should that look like?
  - Better metrics to characterize the reproducibility of one's work.

How Do We Do Better? Funding

- NSF has already indicated that they want us to pay attention (see the DCL from October '22).
- Fast mechanisms for expanding and maintaining reproducible projects that result from traditional research grants.
  - TTP exists, but accessing it has been historically challenging and slow.
  - SBIR is harder to crack, and harder to deal with when a University is involved.
- Improved guidance towards testbeds, datasets, and releasing more of both.
- Pointers to other agencies/bodies who can assist with transition.



NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 23-018

Dear Colleague Letter: Reproducibility and Replicability in Science

October 25, 2022

Dear Colleagues:

A 2019 consensus study report published by the National Academies of Sciences, Engineering, and Medicine (NASEM) discussed the meaning of the terms replicability and reproducibility and identified approaches for researchers, academic institutions, journals, and

So, Why Bother?

- **Fear**: Information Security is a priority… for now…
  - If the impression remains that we aren't making progress, how long will this remain the case?
- **Obligation**: This is actually part of our jobs
  - Even though I just gave you countless reasons others make it "not your job"
- **Optimism**: Because our innovation can actually make the world a better place for everyone.

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!



Credit Where Due



Patrick Traynor
traynor@ufl.edu



Daniel “Ozzy” Olszewski



Allison Lu