SaTC PI Meeting 2022
Breakout Session Introduction/Overview

Improving the Quality and Reuse of Cybersecurity Datasets, Software, and Other Artifacts

Co-Leads: Eric Eide (University of Utah), S. Jay Yang (RIT)
Scribes: David Balenson and Laura Tinnel (SRI International)

Topic Description

This breakout group will focus on best practices and future approaches to increase and improve the quality and reuse of cybersecurity research datasets, software, and other artifacts. It will explore issues, challenges, and opportunities to:

- better **package** and **describe** research artifacts, including their **intended uses** and **limitations**, and
- better **share** and **find** relevant cybersecurity research artifacts, as well as **knowledge** and **experience** about their use,

in order to increase the effectiveness of artifacts and broaden their reuse.

The group will discuss current initiatives, infrastructure, and incentives for producing, sharing, and reusing high-quality artifacts. And, it will explore and identify other potential solutions and opportunities to increase and improve the quality and reuse of high-quality artifacts as a solid practice in the cybersecurity research as well as practitioner communities.

Key Challenges (1)

- Types of artifacts
  - Focus on **Software** and **Datasets**
  - But, note, there are **Hardware** artifacts as well as **Experimental Designs & Methodologies**!
  - Recognizing the scopes and limitations!

- How to foster **curation/generation** of quality artifacts with documentation?
  - Appropriate incentive structures- carrots and sticks?
  - Rewards through conference artifact tracks—formal recognition?
  - NSF supplemental funding?
- How to **measure/evaluate** artifacts properly?
  - Do we have the right criteria for datasets and for software?
  - How to leverage the evaluations as a way to accumulate knowledge about the artifacts?
  - How to encourage self-reported metadata (or intended uses, limitations, and attributes)?

Key Challenges (2)

- How to foster the **sharing/use** of quality artifacts?
  - Is there any incentive? Yes for datasets? How about for software?
  - Today's student seem more inclined to share (e.g., github); tap into this mindset?
  - Cataloguing/indexing for broader adoption (e.g., NSF SEARCCH)?
  - Facilitate communication between producers and consumers?
  - Expand/interoperate with other disciplines?
  - Align with publishers (e.g., ACM, IEEE, Springer, ISOC, USENIX)?
  - Align w/ Google Scholar? DBLP?
  - Connect conference artifact initiatives with cataloguing/indexing services

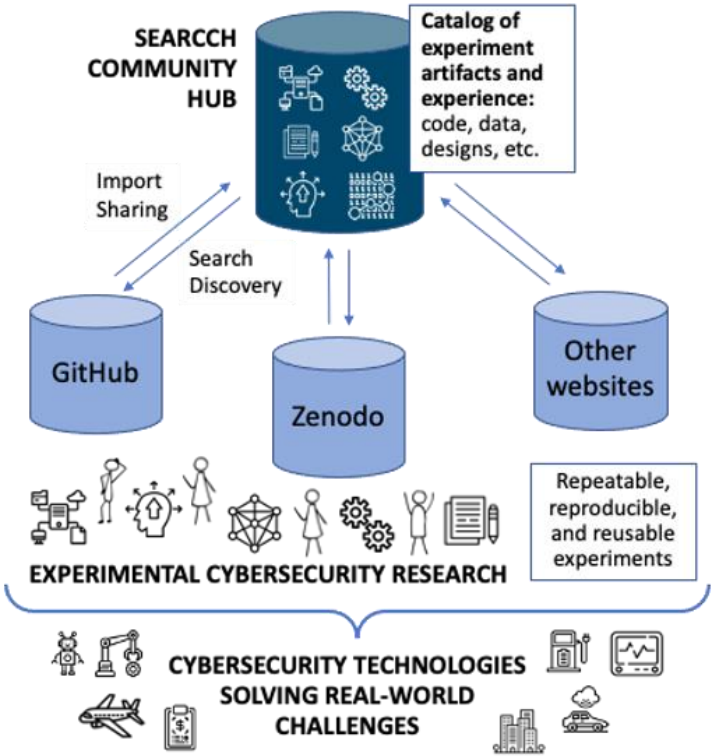
Sharing Expertise and Artifacts for Reuse through Cybersecurity Community Hub (SEARCCH)

Community-driven platform lowers barrier to sharing and reusing research artifacts

Rich metadata representation enables researchers to better describe and find relevant artifacts

Import, curation, and search functions enable greater scientific quality of cybersecurity research

https://searcch.cyberexperimentation.org/



Opportunities for NSF and the Research Community

- **NSF supplemental funding** to support (with measurable outcomes)
  - Software development
  - Data curation
- **Competitions / Challenges** to curate, review, and/or rate artifacts
  - Clear goals to address specific intended uses of data
  - Document limitations of and experiences with datasets
- Recommendations for **Conference Artifact Tracks**
  - Artifact tracks separated from the traditional main conference
  - Artifacts to go along with published papers or independent from them?
  - Criteria and best practices for evaluation? Capturing results of evaluations?
- **Educate/mentor students** to curate data or create software
  - Handbooks to develop reusable / replicable software?
  - How about datasets? Ethics? Admitting limitations because there will be some!
  - Part of fundamental Research Methods.