CICI: UCSS: Enhancing the Usability of Vulnerability Assessment Results for Open-Source Software Technologies in Scientific Cyberinfrastructure: A Deep Learning Perspective (2023-2026)

PI: Dr. Hsinchun Chen (hsinchun@arizona.edu); Thomas R. Brown Chair in Management and Technology; Director, Artificial Intelligence Lab (ai.arizona.edu); Fellow, ACM, IEEE, AAAS, AIS; University of Arizona

Co-PI: Dr. Sagar Samtani (ssamtani@iu.edu); Associate Professor and Arthur M. Weimer Faculty Fellow; Director, Data Science and Artificial Intelligence Lab (dsail.iu.edu); Indiana University, Bloomington

Introduction and Background:

- Scientific cyberinfrastructure (CI) funded by the National Science Foundation and other federal funding agencies has played a critical role in helping scientists and researchers make groundbreaking discoveries, including black hole imaging, DNA sequencing, and more.
- Many modern scientific CIs are supported by open-source software (OSS) such as GitHub, Infrastructure as Code, and Docker containers as the technical basis of their scientific CIs.

Challenges and Motivation:

- OSS contains millions of unknown vulnerabilities that threaten scientific analysis.
- Security analysts and scientific CI personnel face several notable challenges when seeking to identify, remediate, and manage the vulnerabilities in their OSS assets:
  - **Limited scanner coverage for OSS:** Most existing CIs use scanners that are not designed for OSS's unique characteristics.
  - **Lack of unified AI-enabled vulnerability management analytics:** Off-the-shelf AI-enabled analytics techniques are designed for singular tasks, requiring personnel to learn multiple tools.
  - **Managing the trade-offs between vulnerability severity and quantities:** Not all vulnerabilities have the same level of importance or severity and must be managed accordingly.
  - **A paucity of cross-CI collaborations** that elucidate best practices, remediations, etc.

Summary of Proposed Research Thrusts:

This project focuses on assessing the vulnerabilities of two major sets of OSS assets:

- **Systems,** including virtual machines, containers (e.g., orchestration, workflows)
- **Software,** such as infrastructure as code and social coding repositories

The collected data from the audits and vulnerability assessment scans will be used for three key research thrusts:

- **Research Thrust 1 (RT1) – AI-enabled Vulnerability Management →** performs prevailing vulnerability management tasks (e.g., group, sort).
- **Research Thrust 2 (RT2) – AI-enabled Vulnerability Remediation →** develops automated deep learning (DL) and large language model (LLM) approaches to identifying remediations for vulnerable assets.
- **Research Thrust 3 (RT3) – Vulnerability Management System →** provides an interface for OmniSOC and scientific CIs to identify vulnerabilities and collaborate on remediations.

Partnerships:

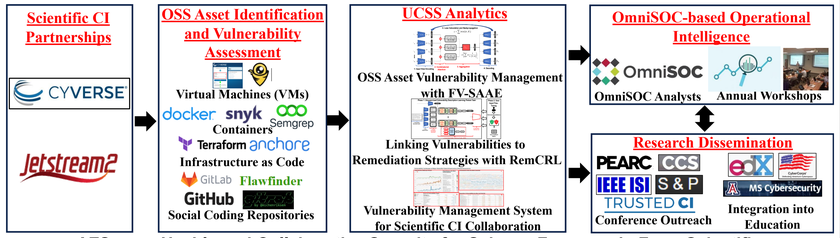
We have partnered with two internationally recognized scientific cyberinfrastructures with a rich set of OSS assets:

- **CyVerse** is a UArizona-led scientific cyberinfrastructure (~$115M since 2008) that provides high-performance computing capabilities to 100K+ life scientists internationally.
- **Jetstream2** is an IU-led infrastructure that is NSF's first Science and Engineering Cloud ($30M since 2014). Jetstream2 serves 8K+ users across 400+ NSF/NIH-funded projects.

To help facilitate the translation of the research models and outputs into operational cybersecurity capabilities for CyVerse and Jetstream2, we have also partnered with **OmniSOC.** An IU-led initiative, OmniSOC is one of the largest real-time Security Operations Centers and vulnerability management for universities and federal facilities (25+ clients).

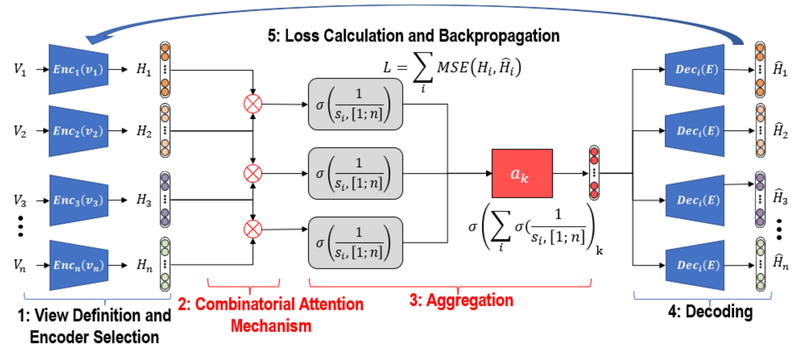
CYVERSE Jetstream2 OmniSOC

Proposed Research Framework:

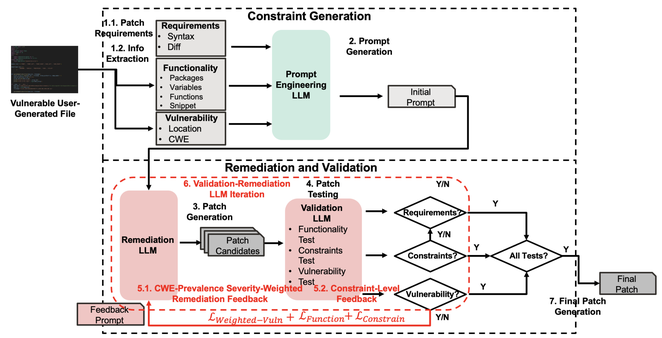


AZSecure Usable and Collaborative Security for Science Framework: From Scientific Cyberinfrastructure (CI) Partnerships to Artificial Intelligence (AI)-enabled Vulnerability Management.

Selected Initial Technical Approaches:



Initial Technical Approach for RT1: Flex-View Self-Attentive Autoencoder (FV-SAAE) for OSS vulnerability management (e.g., grouping, ranking, categorizing). FV-SAAE uses a vulnerability severity weighting scheme and a novel combinatorial attention mechanism for fusing OSS asset and vulnerability data.

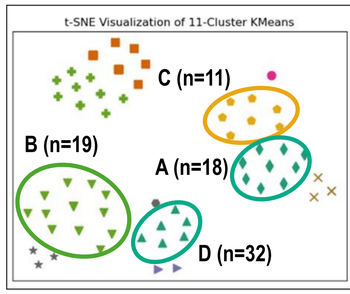


- **Research Thrust 2:** The proposed MA-CAVR operates in two stages:
  1. Constraint Generation via a Prompt Engineering LLM (captures requirements, constraints, and vulnerability information).
  2. Remediation & Validation via a Remediation LLM and Validation LLM.

Intellectual Merit:

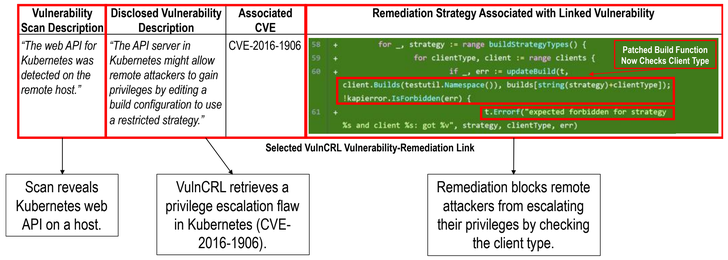
- **Algorithm Development:** This project develops novel deep learning-based techniques and large language model-based approaches for advanced vulnerability management. Selected novelties of proposed approaches include:
  - **FV-SAAE** uses a vulnerability severity weighting scheme and novel combinatorial attention mechanism to fuse OSS asset and vulnerability data.
  - **RemCRL** extends the transformer to operate with stacked word embeddings and contrastive severity loss function to identify remediations for vulnerabilities.
- **VMS:** Integrates vulnerability assessment results from multiple tools and enables analysts to use FV-SAAE, RemCRL, and other proposed methods.

Selected Results:

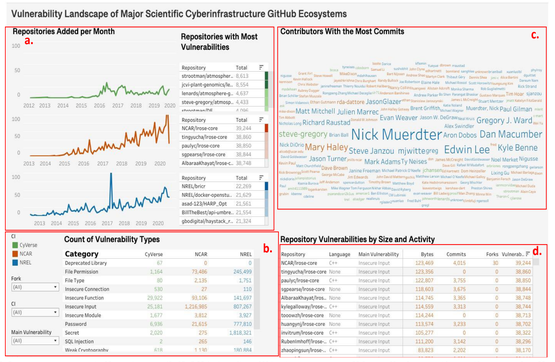


Cluster	Vulnerability	Severity	Package	Count
Clusters A and B	Insecure Function	High	Libblockdev2	41
			Yadm	29
			Youker	28
	Insecure Input	High	Zabbix-cli	103
			Cupp	29
			Elastalert	13
	XSS Vulnerability	High	Libqt5webkit5	4
			Python3-spyder	3
Clusters C and D	Insecure Input	Low	node-gyp	34
			bup	31
	Insecure Module	Low	bup	20
			python-sympy	15
	Insecure Function	Medium	python-sympy	60
			python-html5lib	4

Selected clusters of Virtual Machines (VMs) based on FV-SAAE-generated embeddings of VM vulnerabilities and asset characteristics. The results indicate that VMs in clusters contain the least vulnerabilities, while C and D possess the highest quantities (3,452 per VM) and can, therefore, be prioritized for remediation.



Selected remediation for a Kubernetes web API vulnerability identified by the proposed RemCRL. The remediation strategy for the helps to block remote attackers through a privilege escalation technique that actively checks for the client type accessing the asset.



Vulnerability Management System (VMS) to facilitate cross-CI collaboration through OmniSOC. The initial VMS incorporates vulnerability assessment results from GitHub and allows users to (a) identify vulnerability trends, (b) count vulnerability types, (c) identify key contributors, and (4) identify vulnerabilities by OSS asset.

Broader Impacts:

- Provide critically needed DL-based vulnerability management capabilities to major scientific CIs that serve life science, NSF, and NIH communities.
- Integrates multiple vulnerability management capabilities across multiple scientific CIs.
- Project development and execution include roles for UArizona and IU NSF CyberCorps Scholarship-for-Service (SFS) graduate students.

Acknowledgements:

This material is based upon work supported by the National Science Foundation under Grant Numbers OAC-23119325 (CICI), OAC-1917117 (CICI), DGE-1921485 (UArizona SFS), and DGE-1946537 (IU SFS).