NSF Campus Cyberinfrastructure PI and
Cybersecurity Innovation for Cyberinfrastructure PI Workshop
September 23 – 25, 2019 | Minneapolis, MN

Quad Chart for: CICI: SSC: SciTrust: Enhancing Security for Modern Software Programming Cyberinfrastructure

Challenge:

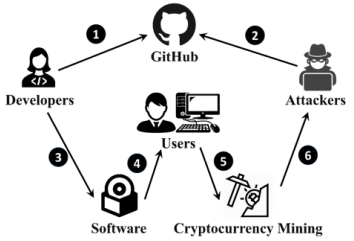
- Modern software programming CI, consisting of online discussion platforms (e.g., Stack Overflow) and social coding repositories (e.g., Github), has offered an open-source and collaborative environment for distributed scientific communities to expedite the process of software development.
- Despite the apparent benefits of this new social coding paradigm, its potential security-related risks have been largely overlooked - insecure or malicious codes can be easily embedded and distributed, which severely damage the scientific credibility of CI.

Solution:

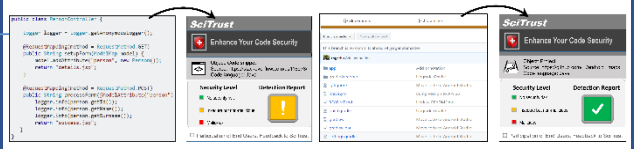
Develop innovative techniques to detect insure or malicious codes on social coding platforms:

- Automatic detection of insecure code snippets on Stack Overflow.
- Automatic detection of malicious codes on GitHub.
- Development of user-friendly tools for scientific and engineering communities to enhance code security in modern software programming CI.

Can one trust the codes in social coding platforms?



Our goal is to advance capabilities of AI to enhance the security of modern software programming CI.



Broader Impact:

- Benefit the society at large by promoting the efficiency of cyber-enabled software development without sacrificing the security.
- Robust outreach efforts to K-12, general public, undergraduate, graduate, minority, and women in cybersecurity.
- The establishment of a cybersecurity lab through this project will enhance the cybersecurity training that will help build the national workforce in cybersecurity.

Metadata tag:

- **Yanfang Ye**, "CyberAI: Innovation, Research, Education for a Better world", ***IJCAI Early Career Spotlights***, 2019.
- Yujie Fan*, Yiming Zhang*, Shifu Hou*, Lingwei Chen*, **Yanfang Ye**, Chuan Shi, Liang Zhao, Shouhuai Xu. "iDev: Enhancing Social Coding Security by Cross-platform User Identification Between GitHub and Stack Overflow", *IJCAI*, 2019. (17.9% acceptance rate)
- Deqiang Li, Qianmu Li, **Yanfang Ye**, Shouhuai Xu. "Enhancing Robustness of Deep Neural Networks Against Adversarial Malware Samples: Principles, Framework, and Application to AICS'2019 Challenge". The AAAI-19 Workshop on Artificial Intelligence for Cyber Security (AICS), 2019. ***AICS 2019 Challenge Problem Winner***.
- **Yanfang Ye**, Shifu Hou*, Lingwei Chen*, **Xin Li**, Liang Zhao, Shouhuai Xu, Jiabin Wang, Qi Xiong. "ICSD: An Automatic System for Insecure Code Snippet Detection in Stack Overflow over Heterogeneous Information Network", *ACSAC*, 2018. (*20.1% acceptance rate*).

Contact: yanfang.ye@case.edu

PI: Yanfang (Fanny) Ye
Associate Professor
Department of CDS
Case Western Reserve University
yanfang.ye@case.edu

Co-PI: Xin Li
Professor
Department of CSEE
West Virginia University
xin.li@mail.wvu.edu

Co-PI: Brian Woerner
Professor & Department Chair
Department of CSEE
West Virginia University
brian.woerner@mail.wvu.edu

- 2014-2019, **Leonard Case Jr. Associate Professor** @ **Department of CDS, CWRU**
- 2014-2019, **Assistant/Associate Professor** @ **Department of CSEE, WVU**
- 2010-2013, **Principal Scientist** @ **Comodo Security Solutions, Inc.**
- 2008-2010, **R&D Deputy Director** @ **Kingsoft Internet Security Corporation**

My research areas mainly include **cybersecurity**, **AI**, and **health intelligence**. I have proposed and developed cloud-based solutions for mining big data in the area of cybersecurity, especially for malware detection and adversarial machine learning. My proposed techniques have significantly reduced the time needed to detect new malicious software - from WEEKS to SECONDS, which have been incorporated into popular commercial cybersecurity products including Comodo and Kingsoft Antivirus that protect millions of users worldwide. I recently received the prestigious **NSF Career Award** (2019), the **IJCAI Early Career Spotlights**, the **AICS 2019 Challenge Problem Winner**, the **ACM SIGKDD 2017 Best Paper Award** and **ACM SIGKDD 2017 Best Student Paper Award** (Applied Data Science Track), the **IEEE EISIC 2017 Best Paper Award**, and the **New Researcher of the Year Award** (2017) at WVU.

Scientific Credibility of Modern Software Programming CI

- **Hack Brief: Uber Paid Off Hackers to Hide a 57-Million User Data Breach** (2015)

https://www.wired.com/story/uber-paid-off-hackers-to-hide-a-57-million-user-data-breach/

Hackers discovered that the company's developers had published code that included their usernames and passwords on a private account of the software repository **Github**. Those credentials gave the hackers immediate access to the developers' privileged accounts on Uber's network, and with it, access to sensitive Uber servers hosted on Amazon's servers, including the rider and driver data they stole.

- **GitHub-Hosted Malware Targets Accountants With Ransomware** (2018)

https://www.bleepingcomputer.com/news/security/github-hosted-malware-targets-accountants-with-ransomware/

Threat actors ran a malvertising campaign on the Russian Yandex. Direct advertising network starting October 2018 to disseminate a malware cocktail designed to encrypt victims' data and steal cryptocurrency. The hacking group targeted Russian organizations using malicious payloads camouflaged as document templates and hosted on the GitHub code hosting platform, one of the goals being to steal sensitive cryptocurrency-related data.

- Can one trust such code snippets or existing software project files?
- In other words, how much do we know about the **scientific credibility** of Stack Overflow and GitHub from the **security** point of view?

Project Overview and Objectives

To address the above challenges, this project seeks to explore **innovative links between artificial intelligence (AI) and cybersecurity** to **automate the detection of insecure and malicious codes on social coding platforms**. If successful, this project will benefit scientific communities and society as a whole by promoting the efficiency of cyber-enabled software development without sacrificing the security. The key components of the proposed work include:

- **T1:** Automatic detection of insecure code snippets on Stack Overflow.
- **T2:** Automatic detection of malicious codes on GitHub.
- **T3:** Development of user-friendly tools for scientific and engineering communities to enhance code security.

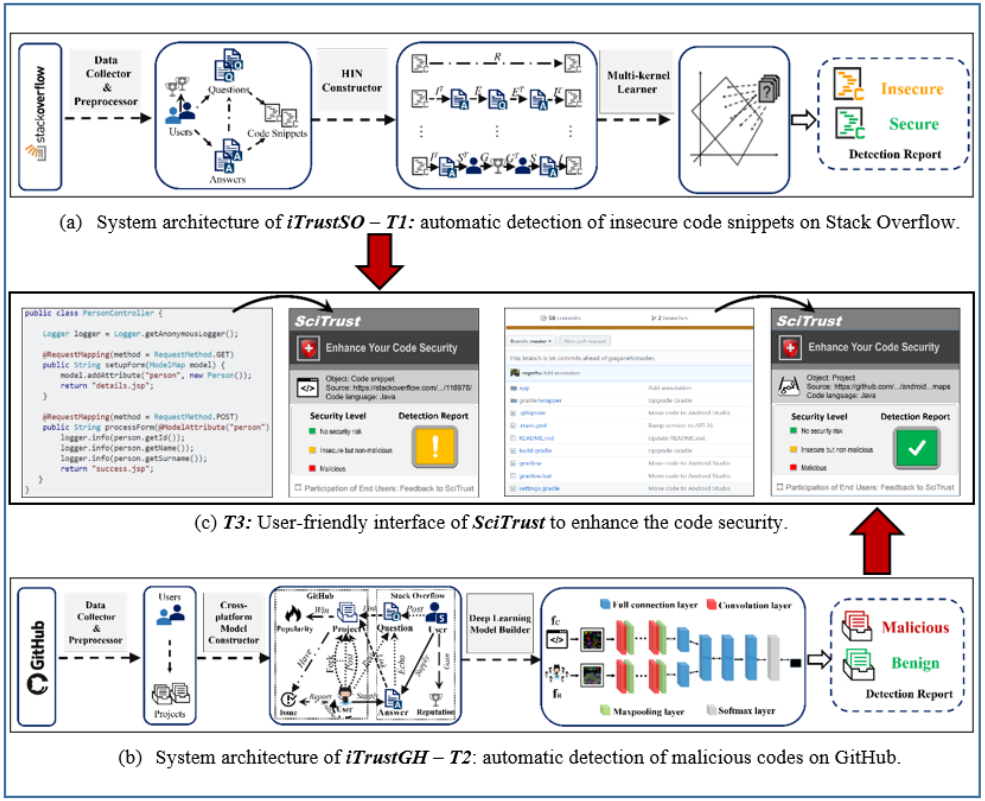


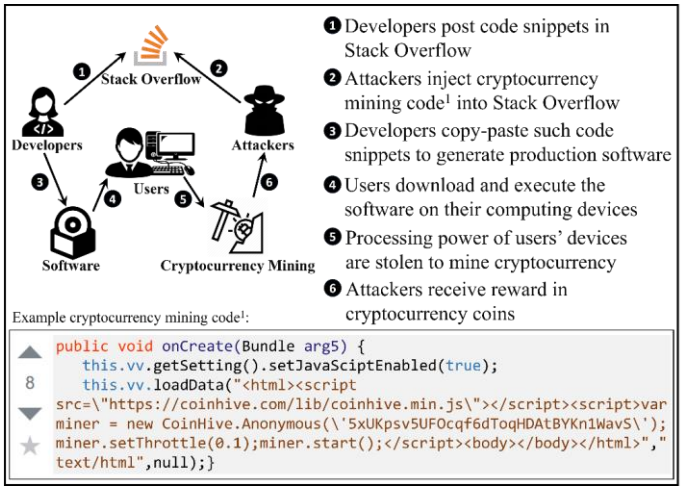
Figure: System architecture diagrams of the proposed project (named ***SciTrust***).

Current Progress

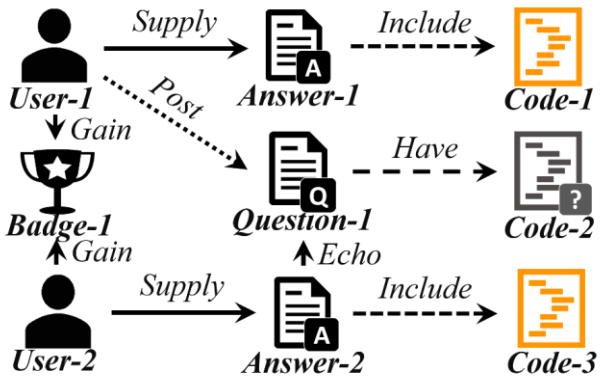
Selected Published Works:

- **Yanfang Ye**, "CyberAI: Innovation, Research, Education for a Better world", ***IJCAI Early Career Spotlights***, 2019.
- Yujie Fan*, Yiming Zhang*, Shifu Hou*, Lingwei Chen*, **Yanfang Ye**, Chuan Shi, Liang Zhao, Shouhuai Xu. "iDev: Enhancing Social Coding Security by Cross-platform User Identification Between GitHub and Stack Overflow", *IJCAI*, 2019. *(17.9% acceptance rate)*
- Deqiang Li, Qianmu Li, **Yanfang Ye**, Shouhuai Xu. "Enhancing Robustness of Deep Neural Networks Against Adversarial Malware Samples: Principles, Framework, and Application to AICS'2019 Challenge". *AAAI-19 - Artificial Intelligence for Cyber Security (AICS),* 2019. ***AICS 2019 Challenge Problem Winner***.
- Lingwei Chen*, Shifu Hou*, **Yanfang Ye**, Thirimachos Bourlai, Shouhuai Xu, Liang Zhao. "iTrustSO: An Intelligent System for Automatic Detection of Insecure Code Snippets in Stack Overflow", *ASONAM*, 2019.
- Shifu Hou*, Yujie Fan*, Yiming Zhang*, **Yanfang Ye,** Jingwei Lei, Wenqiang Wan, Jiabin Wang, Qi Xiong and Fudong Shao. "αCyber: Enhancing Robustness of Android Malware Detection System against Adversarial Attacks on Heterogeneous Graph based Model", *CIKM,* 2019. *(19.4% acceptance rate)*
- **Yanfang Ye**, Shifu Hou*, Lingwei Chen*, **Xin Li**, Liang Zhao, Shouhuai Xu, Jiabin Wang, Qi Xiong. "ICSD: An Automatic System for Insecure Code Snippet Detection in Stack Overflow over Heterogeneous Information Network", *ACSAC*, 2018. (*20.1% acceptance rate*).
- Yujie Fan*, Shifu Hou*, Yiming Zhang*, **Yanfang Ye**, Melih Abdulhayoglu. "Gotcha - Sly Malware! Scorpion: A Metagraph2vec Based Malware Detection System", *SIGKDD*, 2018. *(22.5% acceptance rate)*

We have brought an important new insight by exploiting social coding properties in addition to code content for automatic detection of insecure code snippets and malicious projects.

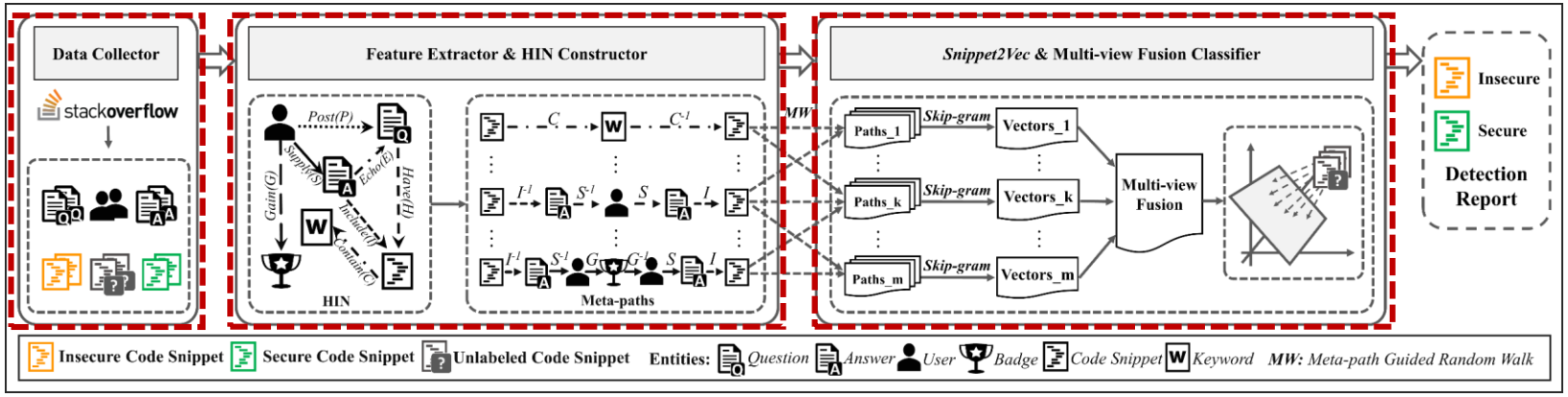


Example of code security attacks in StackOverflow.



An example of relatedness over code snippets.

System Architecture of ICSD (ACSAC'2018)



- **Data Collector**
- ✓ Users' profles, their posted questions and answers, and the code snippets embedded in the questions/answers
- **Feature Extractor & HIN Constructor**
- ✓ Content-based features and relation features: question-code, answer-code, code-keyword, user-question, user-answer, answer-question and user-badge
- ✓ An HIN is constructed
- ✓ Different meta-paths are built from the HIN to capture the relatedness over code snippets
- **Snippet2vec & Multi-view Fusion Classifier**
- ✓ A new network embedding model snippet2vec is proposed to learn the low-dimensional representations for the nodes in HIN
- ✓ A multi-view fusion classifier is constructed to learn importance of different kinds of node (i.e., code snippet) representations learned by snippet2vec under different meta-paths, and thus to make predictions

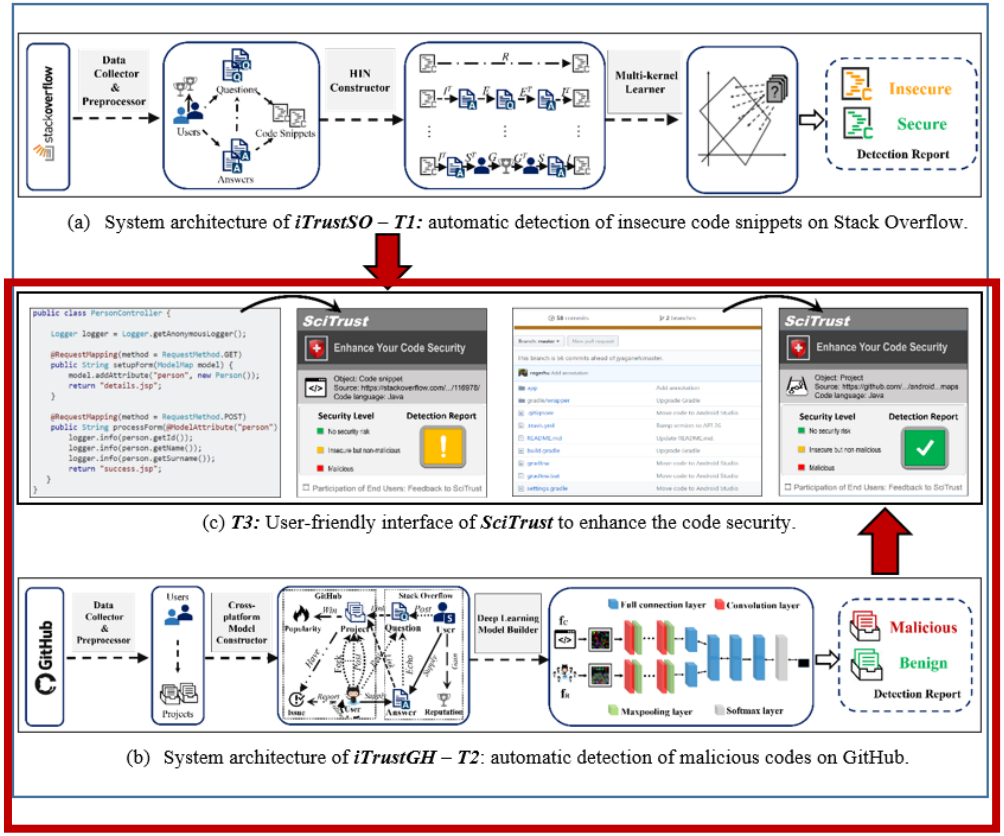
Next Steps

To address the above challenges, this project seeks to explore **innovative links between artificial intelligence (AI) and cybersecurity** to **automate the detection of insecure and malicious codes on social coding platforms**. If successful, this project will benefit scientific communities and society as a whole by promoting the efficiency of cyber-enabled software development without sacrificing the security. The key components of the proposed work include:

- **T1:** Automatic detection of insecure code snippets on Stack Overflow.

- **T2:** Automatic detection of malicious codes on GitHub.

- **T3:** Development of user-friendly tools for scientific and engineering communities to enhance code security.



(a) System architecture of ***iTrustSO – T1:*** automatic detection of insecure code snippets on Stack Overflow.

(c) ***T3:*** User-friendly interface of ***SciTrust*** to enhance the code security.

(b) System architecture of ***iTrustGH – T2***: automatic detection of malicious codes on GitHub.

Figure: System architecture diagrams of the proposed project (named ***SciTrust***).

CASE WESTERN RESERVE UNIVERSITY EST. 1826
think beyond the possible℠

Q & A

Acknowledgement: Our works are partially supported by the NSF OAC-1940885. We would like to particularly thank the support from NSF CICI program, especially the support from our Program Directors **Kevin Thompson** and **Micah Beck**!