Engaging Scholars in Cybersecurity Analysis: A Laboratory for Teaching and Education (ESCALATE)

Overview

In the rapidly evolving information technology landscape, network security is a cornerstone of cybersecurity. Cybersecurity workforce training empowers the next generation of IT professionals to proactively mitigate risks and secure digital infrastructure against threats. But many cybersecurity training programs are limited to the study of theories, best practices, and protocols. Few programs offer exposure to network measurements where students can discern baseline behavior from sophisticated real-world network attacks. Addressing this gap requires both cutting-edge data science techniques and discipline-appropriate skills in advanced cyberinfrastructure (CI) in the classroom.

Several institutions provide rich publicly accessible Internet infrastructure datasets for research use, but using these real-world datasets presents two major challenges: scalability and complexity. First, prohibitively high compute and storage requirements for processing data pose a barrier to scale. Huge terabyte-scale datasets, sometimes containing sensitive information, are unsuitable for transfer to personal devices, which limits their access to those with sufficient resources. Second, educators and researchers may lack skills to use advanced CI. The steep learning curve discourages adoption of CI in courses and research. Our project addresses both challenges by leveraging existing NSF-funded CI to *foster training and preparation of a diverse STEM cybersecurity workforce.*

Building on CAIDA's successful Internet Data Science for Cybersecurity course, our vision rests on three pillars: data-driven course materials, infrastructure support for executing and grading assignments, and a user-friendly platform to organize and share resources within the community. This project will develop and deploy a *Cybersecurity Community Hub* (C^2Hub), a centralized catalog to catalyze the *building*, *delivering*, and *sharing* of CI-ready cybersecurity education and training resources in the cybersecurity training community. Our immediate goal is to provide the community a one-stop platform for *CI-ready* course modules enabling institutions, including under-resourced ones, to *broaden the adoption of CI tools and resources in the Nation's undergraduate and graduate cybersecurity curriculum.* This resource will also enhance *researchers' abilities to efficiently use advanced CI to conduct data-intensive research.* We thus pursue both major goals of the solicitation.

Our team will collaborate to optimize usability, dataset security, and use of modern software stacks that facilitate machine learning/artificial intelligence (ML/AL) techniques on the data. We will address technical and policy challenges in resource sharing across institutions and adopting advanced CI. We will seed the platform with course modules that provide hands-on experience in using CI to apply data science techniques to cybersecurity analyses. We will collaborate with PIs of a CyberTraining project (NSF:CIP:2230127) at UC San Diego to integrate suitable modules into their training program. Two partners (JHU and Calvin U.), with different class sizes, campus CI resources, and student demographics, will help us identify pedagogical challenges in different settings. Seven collaborating institutions have committed to adopt the materials in their classes.

Keywords: Users, researchers, education, cybersecurity, networks, Internet data science.

Intellectual merit

This project will reshape U.S. cybersecurity training by developing a new data-driven, research-oriented cybersecurity curriculum that will cultivate professionals with CI skills to fulfill national cybersecurity research and operational roles. Feedback from users and external evaluators will guide improvements to optimize educational experiences and outcomes for all participants.

Broader impacts

Our open-source curriculum will serve as a role model for other cybersecurity programs nationwide, accelerate adoption of CI in cybersecurity analyses, and expand and diversify the nation's cybersecurity work force. A long-term goal of this project is to achieve lasting impact on the new School of Computing, Information, and Data Science (SCIDS) at UC San Diego, which offers undergraduate and graduate data science-related programs for $\approx$ 1,000 students. We will introduce our course materials into the curriculum of suitable SCIDS programs/courses at various levels.

1 Introduction

In the rapidly evolving information technology landscape, network security is a cornerstone of cybersecurity. Cybersecurity workforce training empowers the next generation of IT professionals to proactively mitigate risks and secure digital infrastructure against threats. But many technology-first cybersecurity training programs (e.g., [1,2]) are limited to the study of theories, best practices, and protocols. Few programs offer exposure to Internet measurements in training students to discern baseline behavior from sophisticated real-world network attacks. Addressing this gap requires both cutting-edge data science techniques and discipline-appropriate skills in advanced cyberinfrastructure (CI) in the classroom.

Several institutions provide rich publicly accessible Internet infrastructure datasets for research use, but using these real-world datasets presents two major challenges: scalability and complexity. First, prohibitively high compute and storage requirements for processing data pose a barrier to scale. For instance, CAIDA's network telescope traffic monitor generates over 1TB of packet traces daily, and a 20-year history of macroscopic snapshots of the global Internet routing system consumes 78 TB as of 2023. These terabyte-scale datasets, sometimes containing sensitive information, are unsuitable for transfer to personal devices, which limits their access to those resources and poses disadvantages to under-resourced institutions and students. Second, educators and researchers may lack skills to use advanced CI. The steep learning curve discourages adoption of CI in courses and research. Leveraging NSF-funded CI to support research computing needs and providing discipline-appropriate training resources will narrow this gap, *fostering training and preparation of a diverse STEM cybersecurity workforce.*

Building on CAIDA's successful graduate-level Internet Data Science for Cybersecurity course in 2023, we propose transformative changes in undergraduate cybersecurity training for STEM students. Our vision rests on three pillars: data-driven course materials, infrastructure support for executing and grading assignments, and a user-friendly platform to organize and share resources within the community. Specifically, we propose to develop and deploy a *Cybersecurity Community Hub* (C^2Hub), a centralized catalog platform to catalyze the *building*, *delivering*, and *sharing* of CI-ready cybersecurity training courses and resources. C^2Hub will maintain a catalog of commonly used cybersecurity training resources, including large-scale datasets and associated software. Instructors can integrate these resources into their classes and training materials based on students' proficiency and interests. Instructors will use C^2Hub's to leverage NSF-funded CI, which will significantly lower the barrier for under-resourced institutions to use CI resources. Specifically, we will use the Nation Research Platform (NRP), which has a diverse geographic footprint in the U.S. **Our immediate goal is to provide the community a one-stop platform for CI contributors and CI users to seek and contribute *CI-ready* training resources enabling institutions, including under-resourced ones, to broaden the adoption of CI tools and resources in the Nation's undergraduate and graduate cybersecurity curriculum. These materials will also enhance researchers' abilities to efficiently use advanced CI to conduct data-intensive research.**

Our team will collaborate to optimize usability, dataset security and privacy, and use of modern data science software stacks that enable machine learning/artificial intelligence (ML/AL) techniques on the data. We will address technical and policy challenges in resource sharing across institutions. We will seed the platform with modules that provide hands-on experience in using CI to apply data science techniques to cybersecurity analyses. We will develop policies, best practices and guidelines for C^2Hub users to share and contribute new resources to maximize the usability of C^2Hub.

We will first deploy these new materials in classes at the lead institution (UCSD) and two academic partners (JHU and Calvin U.). UCSD is a public Minority Serving Institute (MSI), and our team includes a private R1 (JHU) and a small liberal arts college (Calvin U.). This diverse partnership allows us to target different class sizes, campus CI resources, and student demographics to identify pedagogical challenges in different settings. Seven other collaborating

institutions have committed to adopt the materials in their classes. We will also collaborate with PIs of a CyberTraining project (NSF:CIP:2230127) at UC San Diego [3] to integrate suitable modules into their CI professional training program (See LoC). We will host workshops at CAIDA to present popular materials to community stakeholders, including international ones. Hackathons will provide opportunities for participants to try out CI tools with in-person support and develop new materials for the C^{2}Hub.

C^{2}Hub will foster a *sustainable community* for cybersecurity educators and researchers by providing a platform for sharing experiences and resources. This community will consist of CI contributors and CI users who catalyze the development and adoption of advanced CI methods for cybersecurity education and research. We will implement feedback mechanisms at every level, including a real-time (Mattermost) chat and collaboration system, quantitative usage metrics and feedback forms on the web site, longitudinal surveys of both students and faculty using the platform, and biannual community meetings/workshops and annual advisory board meetings. Our metrics of success will include the number of students that use C^{2}Hub in participating institutions and the number of research products our resources/datasets on C^{2}Hub enable. We will periodically conduct surveys among all collaborators to track the long-term impact on their teaching practices and the adoption of CI in their research over time.

Our long-term goals of this open-source curriculum are to serve as a role model and impact other cybersecurity programs nationwide. We will grow from the new proposed School of Computing, Information, and Data Science (SCIDS) at UC San Diego [4], which offers undergraduate and graduate data science-related programs for $\approx$ 1,000 students. We will introduce our course materials into the curriculum of suitable programs/courses at various levels as a showcase to other institutions. Furthermore, the integration of materials on C^{2}Hub into classrooms will fuel the use and adoption of CI in cybersecurity analyses. We anticipate the graduates will expand the nation's cybersecurity work force with necessary CI skills.

2 Challenges for Internet cybersecurity workforce development

Training and enhancing the cybersecurity workforce has been a national priority for over a decade [5–11]. Government agencies have launched programs to support cybersecurity curriculum development and resource sharing. For instance, the National Security Agency (NSA) established the National Centers of Academic Excellence in Cybersecurity (NCAE-C) program [1] to set standards for cybersecurity curricula. Over 400 U.S. institutions have received designations for excellence in cybersecurity education and research [12]. The Department of Defense (DoD) funded the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) [2] to design education and research programs in cybersecurity. Although these curricula include applications of machine learning and artificial intelligence (ML/AI) [13], they provide insufficient coverage of discipline-specific CI skills and methods, limiting students' ability to independently analyze real-world datasets to understand cybersecurity phenomena.

The Department of Homeland Security supported the creation of the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) platform [14] to facilitate the sharing of sensitive datasets. However, resources on the platform lack materials for utilizing advanced CI to host, process, and analyze the data effectively.

Professional certificates, such as CISSP and CompTIA Security+, primarily focus on instruction-led, knowledge-based training. These programs rarely involve the use of CI or the application of ML/AI to real-world Internet data, limiting professionals' ability to analyze large-scale and emerging cyber threats.

We investigated the ACCESS knowledge base [15] to identify current CI training materials. We found only 14 resources with the topic of 'cybersecurity'. They are pointers to materials to study for cybersecurity certificate/degree courses, training on cryptography and general CI skills. We found no resources to train CI users on data-driven Internet cybersecurity topics, indicating a

gap between the CI and cybersecurity communities. We propose to bridge this gap, by providing training and data resources directly relevant to cybersecurity researchers and educators.

2.1 Our innovation: Teaching students what is under the hood of the Internet

Our project builds on our prior success developing a graduate-level course at UCSD [16]: *Internet Data Science for Cybersecurity*. We taught this class in the winter quarter of 2023 as a part of our current NSF Mid-scale Research Infrastructure (MSRI-1) Design project "Designing a Global Measurement Infrastructure to Improve Internet Security" (OAC-2131987). This 10-week course presented Internet measurement and public policy through the lens of data science, with a focus on cybersecurity. We concentrated on the underlying Internet architecture, e.g., vulnerabilities specific to interdomain routing (using the Border Gateway Protocol or BGP), naming (Domain Name System), and certificate management. Persistent security challenges at these layers affect every application that operates over the Internet. This course provided a data-oriented background on these critical dimensions of the Internet infrastructure. The course was inspired in part by students new to Internet research, who struggled with the vast complexity of CAIDA's data and services they had to navigate to answer conceptually simple questions about the Internet (Figure 1).

This course assumed some technical knowledge of how the Internet works, and augmented that knowledge with an understanding of how the global Internet is structured, managed, and financed. Understanding these dimensions, as well as interdependencies across layers, is critical to evaluating approaches to improve the security and trustworthiness of Internet infrastructure.

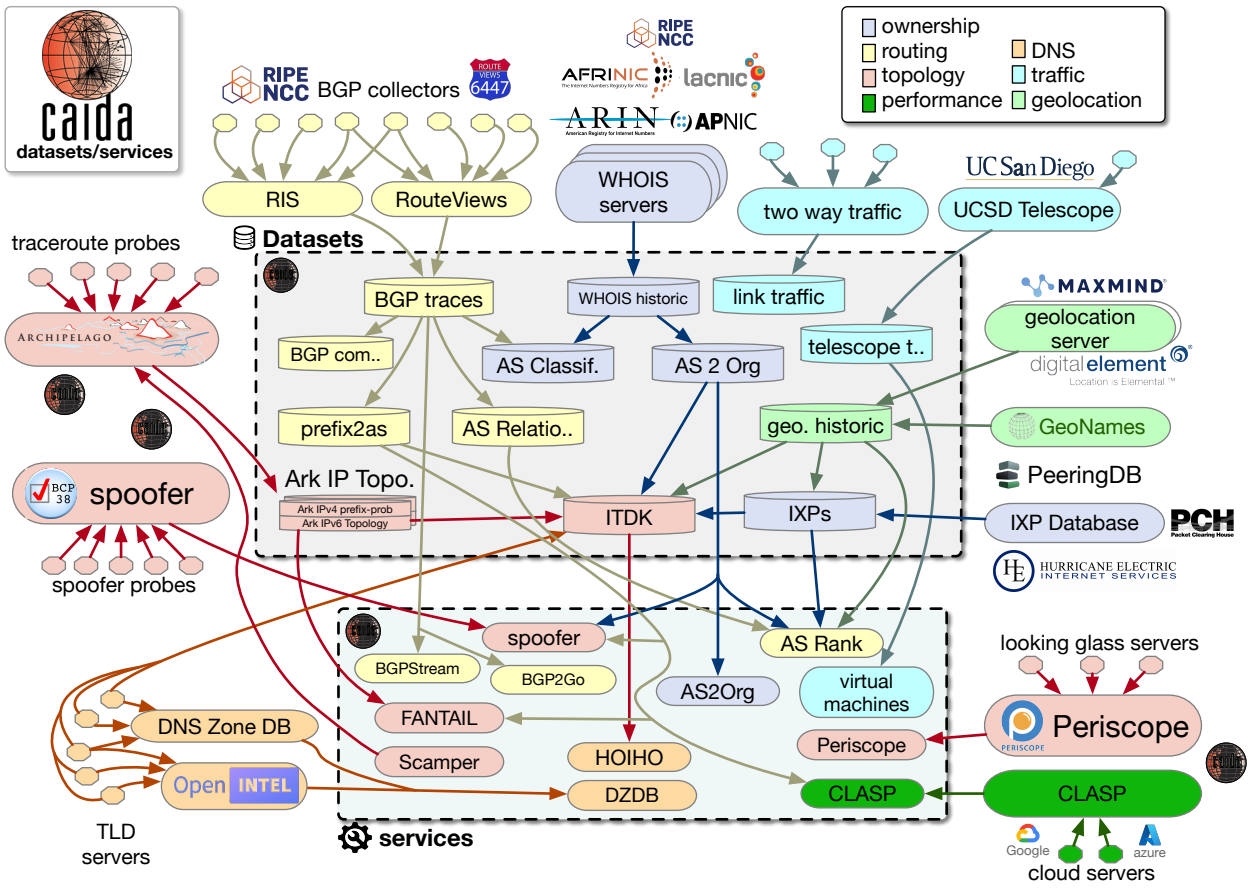

Figure 1: Overview of CAIDA datasets and services, including third-party data.

Although we taught the class for the first time as a graduate course, we plan to use the material and assignments for a subsequent undergraduate course. CAIDA researchers carefully designed the course assignments to provide students with hands-on experience with Internet measurement datasets and data APIs, data science best practices, risks in interpretation of analysis results, and reproduction of research findings.

The course successfully attracted students not just from computer science, but from UCSD's School of Global Policy and Strategy, their School of Management, and Data Science Institute. Several students said the course influenced their decision to pursue more advanced degrees: two students later enrolled into UCSD's PhD program.

2.2 Challenges and Opportunities

We identified three limitations throughout the design and delivery of this course: a knowledge gap from existing undergraduate networking and security curricula; insufficient experience with relevant CI skills and methods for network data analysis; lack of on-campus cyberinfrastructure (compute, storage, networking, supporting software) to support analysis of realistic Internet-scale data sets. These limitations have been raised in reports and workshops (e.g., NSF-sponsored Computing in Undergraduate Education (CUE.NEXT) workshops in 2019 and 2020 [17] and the Federal BigData Research and Development Strategic Plan [7]). However, they have not been adequately addressed in the context of cybersecurity training and education. We elaborate on these challenges and how C^2Hub can overcome them.

Large knowledge gap from undergraduate networking education. Our class enrolled students with diverse backgrounds, but a trait they shared was that their undergraduate training provided little exposure to networking and the complexities of the modern Internet. Indeed, the primary motivation for teaching the course was the increase in student researchers coming to CAIDA with questions about how the "real Internet" worked. Traditional undergraduate networking classes are often limited to theoretical presentation of protocols and performance, barely mentioning the security vulnerabilities and their ramifications. The gap left in undergraduate training is – how vulnerable are we when we use the Internet, and why?

Insufficient experience with CI tools and methods for network data analysis. CS education has drifted toward ML/AL in recent years, focusing on algorithms and applications. Internet security data often has unique structures and properties requiring domain-specific software and knowledge to correctly identify features or problems in the data. Sometimes, we have to adapt CI tools and deployment to efficiently analyze the data. This gap created barrier for students from effectively learn the ML/AI applications to solve practical cybersecurity problems.

Lack of CI resources to handle big data. The three prevailing approaches to sharing cybersecurity data with other researchers (Table 1) do not scale well to classroom scenarios.

1. The simplest approach is to provide access to the data through a web page or API. This approach does not scale to large data sets required by a large class of students. Internet security dataset volumes can easily exceed what an average student laptop can hold, and course deadlines can induce flash-crowd effects on the web sites or APIs, stressing the data owner's compute infrastructure.

   We used this approach in our first experience with the Internet data science course. To enable students to complete the assignments independently on their laptop, we compromised on dataset sizes and types of data used in the assignments. Highly sampled and curated datasets speed the time to results, but limit the scope of analyses and research questions that

Table 1: Comparison of three methods and C^{2}Hub for the sharing and use of data in cybersecurity classrooms. Right three columns report relative cost to students, instructors, and data owners.

Data Sharing Methods	Scalability		Data privacy	Cost		
	Users	Data		Students	Instructors	Data owners
Web-hosted	Low	Low	Low	High	Low	Medium
Bring-code-to-data	Low	High	High	Low	Low	High
Cloud-based	High	High	Medium	Low	High	Low
C^{2}Hub	High	High	Medium	Low	Low	Low

students can explore, and thus the opportunity to prepare students for real world operational or research scenarios.

2. An approach used for data that cannot leave the owner's premises is to "bring code to data". The data owner applies the code to the data on their compute facilities, preserving data privacy. This approach is hard to deploy at scale (100+ students), since it could easily overwhelm the data owner's human and compute resources. It is also difficult to use this approach with sensitive data, as the ability to review code from researchers to ensure safe computations does not scale.

3. Cloud-based solutions, e.g., Google Colab [18] leverage public cloud infrastructure to analyze data, which can be cost-prohibitive for both data owners and instructors. The data owners have to upload and store the data in the cloud. Even though many academic institutions obtain some subsidized use of cloud services, these services still impose a financial burden to instructors or departments, particularly with inexperienced scholars who are prone to make mistakes and write inefficient code, sometimes inducing a shocking invoice.

The PIs at CAIDA have vast experience sharing large-scale datasets, making us intimately familiar with the challenges of doing so. CAIDA hosts public and restricted datasets in various forms [19], including via APIs it creates and maintains [20–22]. We adopted the "bring code to data" model for large-scale datasets, such as the network telescope [23], where the data sets are too large for download and analysis on laptops. CAIDA was a founding participant in the Department of Homeland Security (DHS) Information Marketplace for Policy and Analysis of Cyber-Risk and Trust (IMPACT) program [14], which coordinated real-world data and information-sharing capabilities between academia, industry and government.[1]

3 Solicitation goals targeted

Addressing both goals of the solicitation, this projecct will target and engage with three overlapping CI communities in the area of cybersecurity to boost the use and integration of CI.

3.1 Broadening adoption of advanced CI

Our project aims to broaden the adoption of CI by lowering the barriers for CI contributors and CI users, including graduate students and researchers. Specifically, the CI-ready resources we plan to seed in the C^{2}Hub will provide pipelines and reference code to facilitate access to popular cybersecurity-related datasets from the CI. CI contributors in the cybersecurity community will be able to develop new CI methods and explore advanced CI capabilities to conduct analyses and train computationally intensive machine learning models using these large datasets. Additionally,

[1]Funding for this project ended in 2020.

easy access to data on the CI can encourage potential users to upload their own datasets, fostering opportunities for joint analysis.

3.2 Integrating CI skills into curriculum

The course modules we will develop for C^{2}Hub will include assignments and projects that incorporate various CI skills. Students, as CI users, can learn and adopt these skills via hands-on programming and experiments. Such experience can significantly increase their confidence in using CI in their future research. Furthermore, the team at UC San Diego will be involved in developing programs within SCIDS to integrate these course modules into the curriculum. The new school is enrolling thousands of undergraduate and graduate students in data science and related areas. We will support educators in SCIDS to integrate our material in C^{2}Hub into their classes and capstone projects.

3.3 Fostering a community of future CI professional workforce

While this project primarily targets CI users and contributors in cybersecurity, common topics such as networking and network security can spark students' interest in becoming CI professionals and help build a stronger and more diverse CI workforce. CAIDA's system administrator, Victoria Nguyen, is one of our community's success stories. She took the NSF-funded year-long CI Fellows training (NSF Award #2230127) with Dr. Mary Thomas' group in 2024. She has first-hand knowledge of the value of this opportunity to strengthen her foundational knowledge in HPC and machine learning. She learned to use HPC resources from different institutions, and through this training has become key CI personnel in several groups at SDSC. She will contribute to our evaluation efforts, mentor REUs who are interested in the CI operations side of research, and shepherd students into the CI Fellows training program, fostering synergies with our project. At our workshops she will serve as a role model and offer advice and mentoring to other aspiring CI professionals.

4 C^{2}Hub: A Cybersecurity Community Hub

We envision C^{2}Hub as a one-stop platform that will transform cybersecurity education. C^{2}Hub will facilitate building, delivering, and sharing resources that support the distribution and analysis of large-scale datasets for educators, students, and researchers. Figure 2 presents an overview of the C^{2}Hub ecosystem we envision, and how primary stakeholders interact with it. We structure the development agenda for C^{2}Hub into three tasks to tackle challenges faced by three types of stakeholders: cybersecurity researchers who generate data sets, educators who incorporate such data into their courses, and STEM students who engage with the datasets. We will lower the technical and policy barriers for *researchers* to share and use datasets (§4.1). Our strategy for incentivizing *educators* to use C^{2}Hub is to seamlessly integrate course content and materials with NSF-funded cyberinfrastructure resources, to minimize the time and resources that trainers and educators must spend building (§4.2) and delivering (§4.3) a data-intensive class. We will also connect interested participants from under-resourced institutions to the NRP team to deploy local resources to scale the classes. Finally, we are creating assignments and modules that emphasize *student* engagement, not only based on topic selection, but also hands-on data science learning experiences for the ML/AI era.

Intended CI communities. Our platform will be available to all higher education institutions in the U.S., but the population that stands to benefit the most are the potential CI users who lack computing resources on their own campus, and lack sufficient coding/Unix environment experience to build their own. These populations also may lack awareness of what they even need to learn,

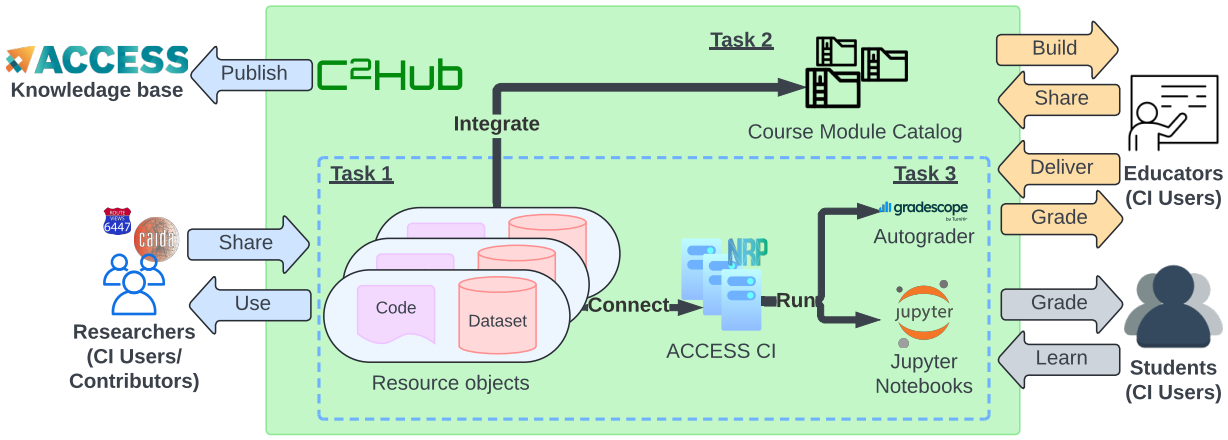

Figure 2: Overview of C²Hub ecosystem.

and students may lack confidence to try. *We are designing C²Hub not just as infrastructure, but a community of contributors and users that will make it thrive* (see LOCs).

To increase exposure among potential CI contributors and existing CI users, we will link C²Hub to the ACCESS knowledge base [15]. Additionally, we will create separate entries in the knowledge base for popular modules and solutions to frequently asked CI-related questions.

4.1 Task 1. Scalable framework for researchers to share cybersecurity datasets

Our first task will be to build a platform for cybersecurity and Internet measurement researchers to contribute datasets and software tools. We will work closely with high-performance computing (HPC) experts at SDSC to efficiently use the National Research Platform (NRP) [24] for sharing data resources. We chose NRP for our implementation for its diverse geographic footprint (>50 institutions), such that institutions without NRP deployment can still find nearby resources for lower network latency. Furthermore, NRP operates the Nautilus cluster [25], an NSF-funded nationally distributed computer system consisting of 5PB+ storage, $10k$+ CPU cores, and $1K$+ GPUs across the U.S. connected with high speed networks. This platform can scalably deploy the class modules. C²Hub does not limit the CI that course material uses. CI contributors can leverage other available CI on ACCESS, such as Open Storage Network, and share their materials on C²Hub.

C²Hub will provide an interface for researchers who contribute datasets to create *resource objects*, each of which contains the dataset and associated sample analysis code. C²Hub will offer multiple ways to distribute the shared data:

- *Static and one-off datasets.* The most straightforward approach to distribute data is to directly store the datasets into the storage buckets on Nautilus [26]. Data owners will store the data in NRP's object storage nodes. Researchers transfer data to the NRP compute node at/near their institution across NRP's high-speed network as needed.

- *Near real-time data.* For continually updated datasets, C²Hub will allow researchers to provide the software/code that periodically produces new data files, instead of the actual data.

- *Databases.* Rather than forcing students to individually query a database, which can be resource-consuming, C²Hub will support the creation of *materialized views* that contain a subset of data according to specific database queries.

In all cases, C²Hub can mirror and cache data on Nautilus [27]. In classroom and training settings, multiple students often retrieve and analyze the same dataset for assignments/projects

within a short time. The data cache closest to the institution can mitigate the impact on the data owner's infrastructure. Note also the campus does not itself have to host an NRP node; it can use its R&E transit networks, including the Internet2 backbone if needed, to reach the closest node.

We will design templates and forms to help contributors create software recipes to facilitate data use. A recipe contains well-documented example code with explanation to illustrate simple use cases of the data and frequently asked questions about the datasets. Our experience building CAIDA's Internet science resource catalog taught us that these recipes are useful for inexperienced data users to get started with their analysis. For example, we developed a recipe [28] that assists users to parse CAIDA's Internet Topology Data Kit (ITDK) [29], a heavily curated and annotated macroscopic snapshot of the Internet topology at multiple granularities. This recipe explains the meaning of data fields, and implements best practices to handle this data. C^{2}Hub will expand this concept by guiding and enabling data owners and users to create, use, modify, and share recipes.

We will also invite international researchers who perform large-scale measurements to contribute datasets and associated CI tools (including recipes) to process and curate the resulting data. For example, OpenINTEL [30] hosted by the University of Twente perform Internet-wide DNS measurements. Internet Initiative Japan (IIJ) builds platforms, such as the Internet Health Report [31], and the Internet Yellow Page [32], to analyze, index and visualize Internet data (see LoC). We will use these data sets in course modules (§4.2.1).

In addition to these technical advances, C^{2}Hub will deploy *data sharing policies* and technical innovations to address privacy concerns. C^{2}Hub will leverage CAIDA's long-standing acceptable use policy framework [33], which the U.S. DOD has used as a model for their data sharing policies [34].

C^{2}Hub will leverage datasets indexed in CAIDA's resource catalog (Figure 3) and synergistically lead to expansion of this catalog. CAIDA's catalog will provide a list of available cybersecurity datasets on the platform, enabling researchers to search and acquire third party data on C^{2}Hub for their studies. Furthermore, researchers can leverage the NRP to join C^{2}Hub datasets with their own non-sharable datasets to conduct analysis.

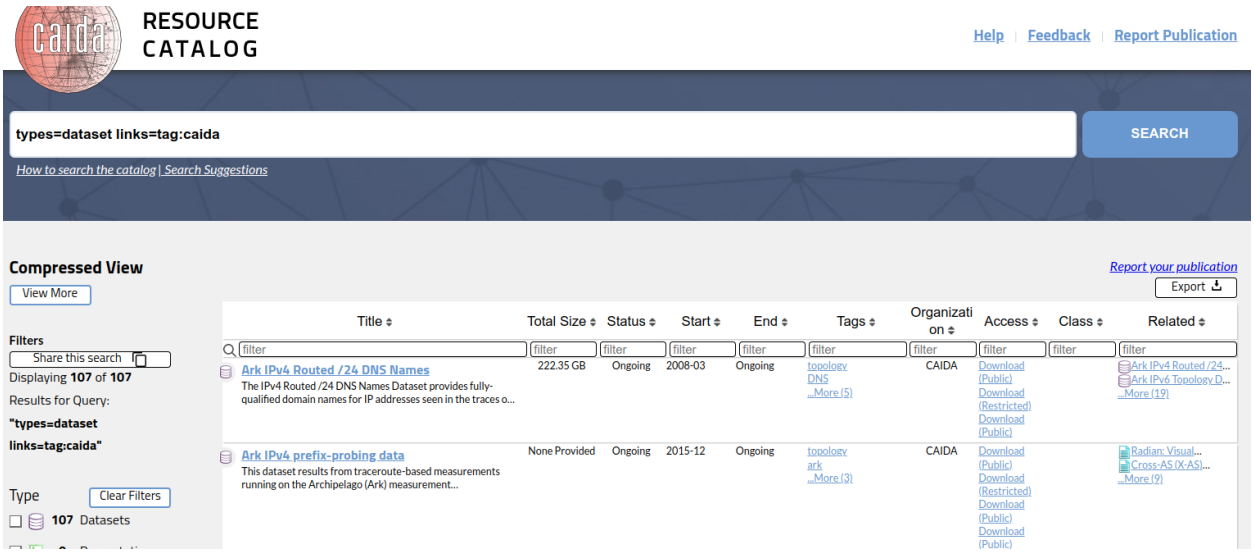

Figure 3: CAIDA's research catalog lists datasets that CAIDA currently provides. Users can search and filter datasets that they may need using keyword search, Tags, and Related fields. We will leverage this catalog to build C^{2}Hub.

4.2 Task 2: Creating and seeding the C^{2}Hub course module catalog

C^{2}Hub will include a wide range of cybersecurity topics, and promote a flexible modular approach for developing cybersecurity courses. Unlike Coursera or edX which often offer a complete course, C^{2}Hub course modules will contain training material focusing on one topic. Each module will include prerequisite material, lecture slides, programming assignments, and project ideas.

Educators will be able to adapt multiple modules into their courses according to the class syllabus, training objectives, and/or student proficiency and interest. Modules are analogous to building blocks that educators can mix and match, and integrate with other materials, to create a complete course. This design also targets researchers who are interested in particular topics/datasets to quickly identify the materials they need.

4.2.1 Initial cybersecurity modules: Internet data science and cybersecurity

To bootstrap C^{2}Hub, our project team will seed the content of the platform by designing a suite of cybersecurity course modules using CAIDA's datasets and resources. These modules will support both individual and project-based learning. We will base the modules and assignments on those from our original Internet Data Science for Cybersecurity course [16], but scale them to use Internet-scale datasets. We will add supplementary material to help students become familiar with the tools, methods, and approaches to scientific use of these datasets.

1. **Introduction to Internet infrastructure data science: challenges and opportunities.** This module will introduce students to the basic building blocks of Internet infrastructure needed to understand their security vulnerabilities, including the differences between IPv4 and IPv6 and their implications for security, autonomous systems and their role, ICANN, and IP transit and peering. The assignment will get students started with global datasets that map IP addresses to the organizations that own and use them on the global Internet.
2. **Measuring and defending against IP address spoofing.** This module will provide details of the most fundamental vulnerability of the Internet addressing architecture: IP address spoofing. The students will learn how spoofing enables denial-of-service attacks, best practices (IP source address validation) for preventing IP spoofing, and why these practices have been so hard to pervasively deploy. The module also covers the challenge of *measuring* deployment of the best practices, and the state-of-the-art research in doing so. In the assignment will students will query and analyze multiple Internet-scale data sets to characterize and track this vulnerability by geographic region and network properties, *e.g.*, size, business type.
3. **The business relationships that shape global Internet routing.** This module will provide background on the global Internet routing system. Students will learn about fundamental weaknesses in the routing protocol architecture, and how publicly available Internet-scale routing datasets are collected, archived, accessed, and used to study security compromises, In the assignment, students will use C^{2}Hub to explore routing business relationships in global Internet datasets, and correlate properties of different organizations (size, business type) with their routing footprint. This assignment will prepare them to analyze potential mitigations to routing security risks in the subsequent module.
4. **Threats to global Internet routing: mistakes and hijacks.** This module will explore the strengths and weaknesses of the industry-led initiative to improve the security of Internet routing: Mutually Agreed Norms of Routing Security (MANRS). In the assignment, students will analyze the topology of participating networks, and consider the security guarantees possible in a coherent topological region of security-conscious networks. The module will showcase the power of integrating multiple datasets on C^{2}Hub. Students will expand their skills with processing multiple Internet-scale datasets to simulate different evolutionarily scenarios of global Internet topology to explore the implications of different policies in different regions.
5. **Active topology measurement with traceroute.** This module will introduce students to active measurement of Internet paths, how it works, and security-relevant properties of the

infrastructure it can reveal that routing data might hide (and vice-versa). The assignment will allow students to apply ML techniques on text data on C^{2}Hub to process a global scale Internet topology dataset with over 2 million router interfaces, and infer the ownership and geolocation of the associated routers.

6. **The promise and peril of the DNS.** The domain name system (DNS) is responsible for the correct functioning of most Internet services, but is also the root cause of many high-profile outages and security vulnerabilities. This module will teach the rationale, history, operational workings, and how policy, technology, and economic forces interact to affect the security of the domain name system. The complex threat landscape of the DNS includes a long list of vulnerabilities, attempted mitigations, and data sets to support their analysis [35]. Empirical analysis of DNS attack surfaces requires a wider range of datasets than the previous topics. The assignment will introduce some of these datasets on C^{2}Hub and how to approach them.
7. **TLS essentials: secure communication with public key infrastructure.** Internet applications such as web services use Transport Layer Security (TLS) protocols to authenticate and encrypt communications. This module will explain the history and operation of TLS and the associated Certificate Authority (CA) ecosytem that manages them, and the role of data science in evaluating strengths and weaknesses in this system. The assignment will engage students with datasets of millions of issued certificates to detect suspicious behavior and correlate it with other data sets, including those used in DNS modules.
8. **Distinguishing cybersecurity events from Internet background radiation.** This module will introduce students to network traffic analysis, using UCSD's network telescope instrumentation, which can detect the initial phases of some types of attack. The students will learn how this telescope works, the type of data it captures, and how to process it. The assignment will provide hands-on experience with analyzing traffic collected by UCSD's network telescope. This assignment will expose students to multi-terabyte archives of Internet traffic. Students will write code to characterize scanning and other security-related phenomena over time in the traffic. As this datasets contains PII, we will use it to experiment with C^{2}Hub's policy on the use of sensitive data in the classroom.
9. **Ethically measuring the Internet.** Internet measurement to support cybersecurity is not a zero-risk activity, and in this module students will learn how research practices are evolving to articulate and manage the risks. The module will explain the history and role of Institutional Review Boards (IRBs) and ethical impact assessments, data anonymization, and the role of third-party data suppliers in ethical Internet data science practices. This module will have no specific data-intensive assignment but will ask students to consider potential risks of the measurements and analyses in other modules.

4.2.2 Bridging modules for STEM students

For students that lack sufficient background in undergraduate computer networking or data science, we will offer two sets of bridging modules. The first set is designed for STEM students who have little to no computer networking background. The topics will cover fundamental concepts such as packet switching, IP, and TCP. Using these modules students will obtain first-hand experience by performing Internet experiments themselves, which is shown to be a more effective learning vehicle than lectures. We will use network diagnostic tools (such as Wireshark, ping, traceroute, and `netstat`) to visualize how Internet protocols work and how packets are sent between hosts. The modules will tie these experimental results to underlying computer networking principles.

The second set of modules will focus on CI methods and tool chains for Internet data science. Unlike traditional "introduction to data science" courses, which discuss ML algorithms and how they work, this module will provide experience with C^{2}Hub and NRP-specific programming environments, including the operation of Juypter Notebooks, Python libraries commonly used in data science (e.g., pandas), the basics of NRP, and unique characteristics of network measurement data.

Students who complete the module will be able to focus on learning the content in the cybersecurity modules without being overwhelmed by an unfamiliar environment.

4.2.3 Publishing course modules

We will leverage GitHub and/or gitlab to publish the source code course modules, for the popularity among the computer science community. C^{2}Hub will centralize the distribution of other related materials, including videos, data files, and slides, etc. We will provide templates on C^{2}Hub for CI contributors and users to create new modules/materials. We will investigate the use of eLearning standards, such as cmi5, for educators to import the module into popular learning management systems (e.g., Canvas [36] and Moodle [37]).

Preliminary work. We have built low-fidelity prototypes that illustrate the user interface of the course module catalog of C^{2}Hub (Figure 4). The C^{2}Hub catalog empowers users to easily identify modules of interest, and filter based on topics (tags) and/or resources required. Educators can identify prerequisites required for students to effectively learn from the module.

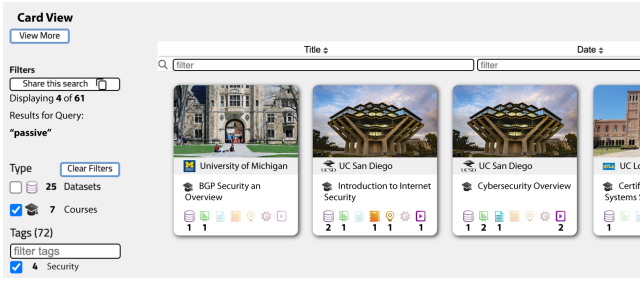

Figure 4: C^{2}Hub course module catalog. Each card represents a module. The icons under the title show the number of datasets/resources the module requires and types of content it provides.

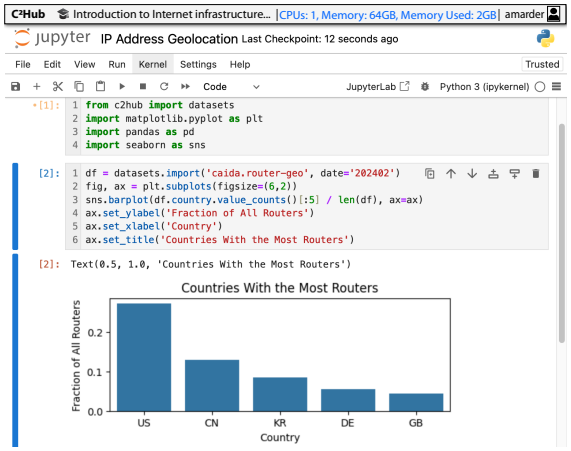

Figure 5: C^{2}Hub will integrate with a Jupyter notebook interface to access and analyze Internet datasets.

4.3 Task 3: Delivering C^{2}Hub course modules at scale

This task focuses on addressing two major challenges for educators in deploying *CI*-ready course modules published on C^{2}Hub, particularly in under-resourced institutions, both related to reducing the time and effort required to allocate resource and deploy C^{2}Hub course modules.

1. Simplified course module set-up for educators. NRP uses Kubernetes technology to orchestrate resource allocation to specific use cases, e.g., in relation to class size and the complexity of the courses. Although NRP staff support the user community in doing so, educators may lack experience in navigating this process. To lower this barrier for educators who are newly adopting a course module into their existing curriculum, the materials we develop will provide module-specific guidelines on the minimum/recommended resources, to inform instructional strategies.

Currently, educators must manually enroll their students to use the NRP, which can be error-prone and time-consuming. We will implement tools to assist educators to automate the enrollment process into the JupyterHub on Nautilus cluster by simply providing a roster, and thus avoiding the time-consuming process of manually giving access to each student.

Finally, we plan to provide scripts for educators to automatically deploy NRP Nautilus JupyterHub Service environments [38]. The created Jupyter Notebooks could be module-specific, such that

each enrolled student will have a Jupyter Notebook instance with the required code and libraries to conduct the analysis.

2. Automated grading of student assignments. Institutions with high student-to-teacher ratios may have insufficient instructional staff (e.g., teaching assistants and tutors) to evaluate student submissions. The modules we publish on C^2Hub will include autograder scripts that reduce the need for manual grading. The scripts will integrate auto-grading capabilities into the educators' NRP Nautilus JupyterHub environment using the `nbgrader` tool [39]. Student grades will be output in a downloadable file for educators to upload to their LMS. C^2Hub will output formats compatible with the two most popular LMSes—Canvas [40] and Gradescope [41].

Preliminary proof-of-concept. We successfully created a simple prototype module with course content from CAIDA's graduate-level Internet Data Science for Cybersecurity course in 2023 [16]. We uploaded a sample public dataset of 2GB to the Nautilus Ceph S3 storage [42] and created a JupyterHub environment capable of accessing the dataset in a Jupyter Notebook (Figure 5). We will migrate all suitable course materials onto Nautilus and create ready-to-run Jupyter Notebooks.

References

[1] National Security Agency/Central Security Service, "National centers of academic excellence in cybersecurity." https://www.nsa.gov/Academics/Centers-of-Academic-Excellence/.

[2] "VICEROY Northwest institute forcybersecurity education and research." https://cyser.wsu.edu.

[3] San Diego Supercomputer Center, "Cybertraining: Training and developing a research computing and data (RCD) CI professionals community." https://www.sdsc.edu/education_and_training/cip_fellows_program.html, 2024.

[4] University of California San Diego, "School of computing, information and data sciences." https://scids.ucsd.edu, 2025.

[5] N. Science and T. C. S. on Future Advanced Computing Ecosystem, "Pioneering the future advanced computational ecosystem: A strategic plan." https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf, 2020. Accessed 2025-01-15.

[6] P. U. R. C. for Advanced Computing, "CI Workforce Development Workshop 2020." https://www.rcac.purdue.edu/ciworkforce2020/, 2020. Accessed 2025-01-15.

[7] Subcommittee on Networking and Information Technology Research and Development. https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf.

[8] T. N. S. C. I. E. Council, "National strategic computing initiative strategic plan." https://www.hsdl.org/?view&did=806294, 2016. Accessed 2025-01-15.

[9] RTI International, "The missing millions: Democratizing computation and data to bridge digital divides and increase access to science for underrepresented communities." https://www.rti.org/publication/missing-millions/fulltext.pdf, 2021. Accessed 2025-01-15.

[10] National Science Foundation, "Request for information on future needs for advanced cyberinfrastructure to support science and engineering research (nsf ci 2030)." https://new.nsf.gov/funding/information/dcl-request-information-future-needs-advanced/nsf17-031, 2017. Accessed: 2025-01-15.

[11] National Science Foundation, "Transforming science through cyberinfrastructure: Nsf's blueprint for a national cyberinfrastructure ecosystem for science and engineering in the 21st century." https://nsf-gov-resources.nsf.gov/files/CI-LWD%202.pdf, 2021. Accessed: 2025-01-15.

[12] CAE in cybersecurity communtiy, "A hub for national centers of academic excellence." https://caecommunity.org, 2022.

[13] J. Crabb, C. Izurieta, B. Van Wie, O. Adesope, and A. Gebremedhin, "Cybersecurity education: Insights from a novel cybersecurity summer workshop," *IEEE Security & Privacy*, vol. 22, pp. 89–98, Nov. 2024.

[14] Department of Homeland Security, "Information Market for Policy and Analysis of Cyber-risk and Trust." https://www.impactcybertrust.org/, 2020.

[15] ACCESS Support, "Knowledge base resources with cybersecurity tag." https://support.access-ci.org/knowledge-base/resources?f%5B0%5D=tags%3Acybersecurity, Jan. 2025.

[16] K. Claffy, "CSE 291(e): Internet Data Science for Cybersecurity." https://cseweb.ucsd.edu/classes/wi23/cse291-e/syllabus.html, 2023.

[17] L. Birnbaum, S. Hambrusch, and C. Lewis, "Report on the CUE.NEXT workshops." https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/3/3786/files/2020/06/Final-Report-June-2020.pdf, June 2020.

[18] Google, "Google Colaboratory (Colab)." https://colab.google, 2024.

[19] CAIDA, "Resource catalog,." https://catalog.caida.org.

[20] CAIDA, "ASRank data API." https://api.asrank.caida.org/dev/docs, 2020.

[21] CAIDA, "Bgpstream components, including api." https://bgpstream.caida.org/components, 2020.

[22] CAIDA, "Spoofer data API." https://www.caida.org/projects/spoofer/data-api/, 2020.

[23] CAIDA, "The UCSD Network Telescope." https://www.caida.org/projects/network_telescope/, 2018.

[24] SDSC, "National research platform,." https://www.sdsc.edu/services/hpc/nrp/index.html.

[25] National Research Platform, "Nautilus." https://nationalresearchplatform.org/nautilus/.

[26] Nautilus documentation, "Ceph filesystems." https://docs.nationalresearchplatform.org/userdocs/storage/ceph-s3/.

[27] Nautilus documentation, "CVMFS." https://docs.nationalresearchplatform.org/userdocs/storage/cvmfs/.

[28] L. Lu and D. Wolfson, "Parse CAIDA's ITDK for a router's IPs, ASN, neighbors, and geographic location." https://catalog.caida.org/recipe/parse_the_itdk, Oct. 2021.

[29] CAIDA's Macroscopic Internet Topology Data Kit (ITDK). http://www.caida.org/data/active/internet-topology-data-kit/.

[30] OpenINTEL, "OpenINTEL: Active DNS." https://openintel.nl, 2024.

[31] "Internet health report." https://ihr.iijlab.net.

[32] R. Fontugne, M. Tashiro, R. Sommese, M. Jonker, Z. S. Bischof, and E. Aben, "The wisdom of the measurement crowd: Building the internet yellow pages a knowledge graph for the internet," in *Proceedings of ACM IMC*, 2024.

[33] CAIDA, "CAIDA master acceptable use agreement (AUA)." https://www.caida.org/about/legal/aua/.

[34] Air Force, "Air Force Acceptable Use Agreement (AUA)," 2020. http://www.mit.edu/~kepner/AI-Accelerator/DataSharingAgreement-Signable.docx.

[35] GMI3S project, "Data needs for securing Internet infrastructure, v2.4," 2023. https://gmi3s.caida.org/outcomes/documents/vulnerabilities-harms-dataneeds_v2.4.pdf.

[36] Canvas, "Canvas LMS." https://www.instructure.com/canvas, 2024.

[37] Moodle, "Welcome to the moodle community." https://moodle.org, 2024.

[38] "Nautilus Documentation JupyterHub Service," 2024. https://jupyterhub-west.nrp-nautilus.io.

[39] "nbgrader," 2024. https://nbgrader.readthedocs.io/en/stable/index.html.

[40] "Canvas Gradebook," 2024. https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-import-grades-in-the-Gradebook/ta-p/807.

[41] "Gradescope Autograder Documentation," 2024. https://gradescope-autograders.readthedocs.io/en/latest/specs/.

[42] "Nautilus Documentation Ceph S3 Storage," 2024. https://docs.nationalresearchplatform.org/userdocs/storage/ceph-s3/.