Report for NSF Workshop on Community Research Infrastructure for Integrated AI-Enabled Malware and Network Data Analytics

Michalis Kallitsis, Forough Ghahramani, Vasant Honavar, Dinghao Wu, and John Yen

Version 2, April 5, 2023

Abstract

Inspired by the growing opportunities to use artificial intelligence / machine learning (AI/ML) to counter the increasingly sophisticated and coordinated cyberattacks, a group of researchers from academia, industry, and the government were summoned for the NSF-funded workshop on "Community Research Infrastructure for Integrated AI-Enabled Malware and Network Data Analytics" to identify cybersecurity-related research opportunities enabled by AI/ML, to discuss current challenges that prevent researchers from tackling these opportunities, and to propose approaches and infrastructure needs to accomplish the shared community goals. The participants were challenged to speculate on a far-term vision (e.g., the idea of an "immune system" for cybersecurity was circulated as a moonshot concept), and then were asked to identify existing key challenges that hinder the community from reaching the vision. These challenges were grouped into the broad areas of data acquisition and sharing, analysis, and governance. These key challenges led to a discussion about exemplary research approaches where advancements are needed to address these challenges. Together, these challenges and approaches helped to identify community needs of computing infrastructure that would enable and in fact accelerate both applied and basic research by democratizing research capacity. The infrastructure needs identified were classified into three categories: (1) data collection, storage and analysis infrastructure, (2) development of standards, policies and legal frameworks, and (3) establishing support infrastructure and the necessary educational and training aspects for developing the future workforce that would operate the support infrastructure. Further, the workshop also revealed the importance of strengthening and building an open, inclusive, and always-learning community interested in AI-enabled cybersecurity so that concrete advancements can be made towards a future cyber-space that is secure, robust, resilient, fair, and amenable to self-defend against any types of sophisticated cyber-threats.

1 Introduction

The NSF-funded Workshop on *Community Research Infrastructure for Integrated AI-Enabled Malware and Network Data Analytics* was held virtually between January 20th, 2023 and February 3rd, 2023, and was organized as four half-day workshops. In the workshop series, the organizing committee invited artificial intelligence and cybersecurity experts from academia, industry and the government, and solicited their inputs regarding (1) research opportunities in cybersecurity that can be addressed using artificial intelligence and machine learning (AI/ML), and (2) associated infrastructure needs necessary to address the research tasks that arise in addressing existing and emerging cybersecurity challenges via the auspices of AI/ML.

The identification of these research opportunities and infrastructure needs at the intersection of cybersecurity and artificial intelligence will expedite the selection, integration, and adaptation of relevant open-source software tools and infrastructures, as well as the creation of new tools, as needed. These tools and infrastructures will not only accelerate the critically needed advancement of cyber security solutions using AI, but will also broaden the participation of early career scholars and students from under-represented groups in STEM, contributing to further enhancing the diversity of the next-generation cybersecurity and AI workforce.

This NSF workshop was well timed. The severity, frequency, scope, and sophistication of cyberattacks have exploded in recent years (e.g., see [12] for Distributed Denial of Service (DDoS) trends), and have resulted in huge financial damage to organizations, endangering critical infrastructures. For instance, the Internet, the energy grid, food and water networks, transportation networks and healthcare systems are constantly under the risk of cyberattacks [11]. Potential attacks against voting systems threaten even the most basic foundation of our society, namely democratic elections. While a host of technologies for cybersecurity and AI/ML have been developed and used by researchers within each individual community, there is a *dearth of combined, synergistic efforts that tackle problems at the intersection of cybersecurity and AI/ML*. Further, significant infrastructure investments would be required to develop innovative AI/ML solutions for cybersecurity problems on a large corpus of networking and security data. Even though technologies for machine learning using big data exist [9], the computing resources and data required for developing such solutions are not easily accessible to the vast majority of researchers.

The workshop participants identified several reasons for this dichotomy, including challenges in sharing sensitive networking data due to privacy considerations, limited availability of labeled or "ground truth" data needed to efficiently train AI/ML models, the existence of "zero-day" vulnerabilities that make new malware evasive to existing intrusion-detection techniques, the challenges on processing voluminous, high-dimensional network data, and legal and data governance challenges. The workshop participants categorized these challenges under three broad areas of (1) *Data Acquisition and Sharing*, (2) *Analysis* and (3) *Governance*. Through the course of the workshop, the invitees identified and

discussed the challenges under all three categories. This workshop report aims at capturing and summarizing these challenges and the proposed approaches and process—supported by the necessary infrastructure—to overcome them.

Workshop Structure. The workshop consisted of four separate sessions with each part focusing on a particular perspective and incrementally building on the outcomes and ideas from the previous session. In each session, participants were split into small subgroups to discuss or brainstorm on a particular topic (e.g., during the first day of the workshop, a subgroup of participants was asked to identify research questions that AI/ML can address in the field of real-time threat detection, another subgroup focused on identifying opportunities on malware analysis, another group targeted AI/ML interpretability, etc.). There was a gap of one week between the first and the second sessions, and the remaining sessions were scheduled with three/four days apart. The time between the sessions provided opportunity for participants to reflect on the topics and ideas discussed, especially those discussed in subgroups that they were not part of.

The four workshop sessions were organized as follows: (1) scoping the *research opportunities* that AI/ML can address in cybersecurity (see Table 1 for some examples provided by the workshop participants), (2) discussing a long-term *vision* after assuming that the research opportunities have been addressed, and then stepping-back to reckon the current *challenges* that prevent the community to reach that vision, (3) discussing *approaches* and the *process* to overcome these identified challenges, and lastly (4) outlining the *infrastructure* needs required to fulfill the specified approaches and process, and deliberating on how to forge a closely knit *community*. As mentioned earlier, the main conundrums for achieving the community vision were categorized into the areas of data acquisition and sharing, analysis, and governance. All workshop session discussions spanned across all three areas.

Workshop Report Outline. This workshop report is organized in a similar manner to the workshop structure. We start with a brief overview of a community vision and the recognized existing challenges to achieve it. Then we proceed with exploring the proposed approaches that workshop participants suggested to potentially overcome the pinpointed challenges in the aforementioned categories of data acquisition and sharing, analysis, and data governance. The report then proceeds in considering the infrastructure needs proposed, and concludes with plans for community development and a concluding summary.

2 Community Vision and Identified Challenges

Participants were asked to envision an ideal future in which no restraints on data availability or compute resources or any other existing challenges would apply, and thus AI/ML could unleash its full potential for solving critical cybersecurity problems. A popular vision that emerged was the one of an *Immune*

Table 1: Security-focused **research opportunities** that can be tackled with AI/ML.

Anomaly detection, pattern discovery and data labeling
Wouldn't it be great if we could develop realistic synthetic datasets void of privacy concerns?
Wouldn't it be great if we could comprehensively and accurately label the real datasets?
Wouldn't it be great if we could detect seemingly innocuous, low-volume but coordinated attacks?
Wouldn't it be great if we could use multiagent paradigms for anomaly detection or pattern discovery in computationally challenging/large scale domains?
Malware analysis and threat intelligence
Wouldn't it be great if we could efficiently identify changes and/or similarities of malware families over time?
Wouldn't it be great if we could fuse multiple sources/types of data efficiently for new malware detection?
Wouldn't it be great if we could construct generative AI models to produce "new" types of malware that we can train our detection models on?
Wouldn't it be great if we could generate comprehensive simulations of a malware campaign / system (C&C communications, execution, etc)?
Interpretability and trustworthiness of AI models
Wouldn't it be great if we could make an AI model more robust to adversarial attacks?
Wouldn't it be great if we could build interpretable AI models, in the domain of security and privacy?
Wouldn't it be great if we could incorporate knowledge and science of security into modern data-driven AI?
Real time handling of online/streaming data
Wouldn't it be great if we could use AI/ML in real-time or at machine speed for cybersecurity use cases?
Wouldn't it be great if we could handle concepts that are changing in real-time with evolving data streams?
Wouldn't it be great if we could handle differing time scales for the variable data streams?

System for Cybersecurity. This immune system, akin to biological immune systems, was pictured to be one that would be capable of automatically and timely detecting cyber threats, and automatically responding and alleviating any potential risks against it. The envisioned system would have all the necessary data to "train" itself appropriately against cyber attacks, and be able to recognize threats previously unknown to it. It would be a fair, resilient, ethical and trustworthy system that detects a host of different types of attacks, including highly advanced coordinated ones, and responds in ways that allow it to recover and continue to function in the face of sophisticated attacks. Such a vision can serve as a moonshot project that inspires breakthrough advancements in both applied research and basic research [16].

Workshop attendants also discussed other stimulating "future headlines" such as smart defense systems that would prevent ransomware attacks against artificial hearts, systems that are able to detect Internet threats in real-time as data arrives at extreme speeds, cross-nation collaborations and data integration that are able to prevent severe attacks while also responsibly respecting data privacy, fairness and ethical considerations, etc. Throughout these exciting conversations, the community started compiling an array of existing *challenges* that currently prevent us from reaching the dreamed future. These challenges were categorized into the broad areas of data acquisition and sharing, analysis of data, and governance considerations, discussed next.

In refining challenges under *data acquisition and sharing* participants acknowledged the gap in accessible, high-quality data available to this community. While there is an abundance of data collected by large industry corporations (e.g., large ISPs have access to rich network traffic streams, technology companies such as Microsoft have collections of a plethora of malware samples and

labels [12], etc.), the AI/ML and cybersecurity research communities lack access to such meaningful datasets due to privacy considerations, legal factors or other issues. The challenge of providing incentives to data owners to provide data to the rest of the community was therefore raised. Participants also highlighted the challenge of ensuring data integrity, the challenge of ensuring data privacy while not sacrificing the scientific utility of a dataset via heavy data anonymization schemes, the challenge of acquiring datasets that might currently be not technically or ethically amenable for collection (e.g., longitudinal collection of network communication traffic traces along with OS system calls from large swaths of users), the challenges of using federated learning or other AI/ML techniques for addressing privacy considerations, etc. The complete list of challenges under *data acquisition and sharing* identified are tabulated in Table 2 in the Appendix.

Under *data analysis* the invitees spotted important challenges with regards to analyzing and correlating across high-dimensional, large volume, diverse cybersecurity datasets at various temporal or spatial granularities (e.g., joining Netflow datasets with Darknet datasets, BGP routing states, Internet topology data, DNS traces, etc.). Participants also pondered on current difficulties with detecting "zero-day" attacks using threat intelligence information that crosses existing data boundaries (e.g., organizational/enterprise/national limits). They also reiterated the importance of *labeled data* with "ground truth" information [13], and discussed how lack thereof is hindering AI/ML progress in cybersecurity akin to the progress achieved in other fields where labeled data are plentiful (e.g., image processing, natural language processing, etc.). The workshop attendees acknowledged that certain private and sensitive datasets might never be able to be broadly shared, and thus identified the challenge of having to analyze such datasets without direct access to them (e.g., in a federated learning manner, by leveraging differential privacy, etc.). The full list of *data analysis* challenges discussed can be found in Table 3.

As vast amounts of data are being created from multiple sources, a *data governance* framework needs to be in place that will ensure a consistent approach to the valuation, creation, consumption, and control of data. The data governance framework holds people, processes, and technology accountable, and ensures that data is usable, high quality, trustworthy, accessible, and secure, and by extension, brings value to everything else that we are trying to do with the data. Hence, on the important category of data governance, the workshop contributors touched on the issues of data privacy, secure data analysis, data fairness, data bias (i.e., selection/collection bias), data integrity and data veracity. They discussed the challenge of developing *key performance indicator* (KPI) metrics to track and quantitatively evaluate the above-mentioned issues when it comes to data indexing, sharing, and analysis. They also highlighted considerations with data formats (i.e., different datasets having different data schemas, requiring different tools for their processing, etc.), reproducibility of research results/data artifacts, data archiving issues (e.g., long-term storage), and issues with lack of uniformity in data sharing practices and policies (e.g., data providers offering custom and ad hoc data usage agreements (DUAs) that

researchers need to adhere to before accessing the provided data). The critical aspect of infrastructure *sustainability* was also cast under the umbrella of data governance, recognizing the need for data infrastructures that continually operate to process, collect and annotate data, and consequently, the need for training/educating future personnel as "data librarians" for curating and handling the collected and requested datasets. Moreover, participants exchanged thoughts regarding challenges in offering mechanisms that support transparency in disclosing the risks and anticipated benefits when planning to use sensitive datasets (e.g., similar to the role that university IRBs offer; notably, many organizations outside academia and the government lack IRB-like committees). The list of *data governance* challenges, identified at the workshop, is tabulated in Table 4 in the Appendix.

3 Community, National and Individual-level Approaches

With the challenges identified, the workshop attendees were then invited to brainstorm potential approaches for overcoming them at three levels: (1) the *community* level, (2) the *national* level, and (3) the *individual* level. Recognizing the relationships between these three levels is an important byproduct of the workshop discussions.

The emphasis on community-level approaches helps to clarify and prioritize advancements that are important for the broader community and their associated infrastructure needs, beyond the priority and needs of individual researchers or research projects. Building on community-level approaches, the national-level approaches can then articulate advancements that need to be made at the national-level (involving broad and diverse stakeholders across academia, industry, society, and the government) to address the challenges at hand. Finally, individual-level approaches identify commitments that individual members of the community should make so that approaches at the two higher levels are achievable.

In the exposition below, we group the approaches / processes identified based on the three broad challenge categories explored earlier, namely data acquisition and sharing, analysis, and governance.

3.1 Data Acquisition and Sharing

Community-level Approaches: In this level, the workshop participants emphasized the idea on providing appropriate *incentives*, at various levels, in order to promote and expand efforts on sharing networking datasets and related code and notebooks (the latter two will be elaborated in later section). For instance, incentives (e.g., dataset "citation scores") can be put in place so that the impact of data sharing can be quantified so that the data provider is not only credited for contributing to its use, but also motivated to contribute additional datasets. Participants also expressed the need for organizing community workshops and

conferences that provide an overview of *existing* data repositories or data infrastructures (e.g., the CAIDA repositories [3, 2], CLASSNET [14], the ORION Network Telescope [10], etc.). Via such workshops, junior or even more senior researchers in AI/ML and cybersecurity can get exposed to existing datasets and related analytic tools so that researchers and participants will be able to both learn about exemplar data-driven cybersecurity problems, the challenges for solving these problems, and software tools (including AI/ML for big data) that can be useful for solving the problem.

Advancements are also needed for developing a framework for sharing real-time threat intelligence information (e.g., leveraging STIX (Structured Threat Information Expression) and TAXII (Trusted Automated Exchange of Indicator Information) frameworks [6]). Further, advancements are needed for the development of high quality *synthetic* datasets and secured simulation environment (e.g., sandbox) that could be used for a wide range of research from threat detection to threat responses.

National-level Approaches: At the national level, proposed approaches included the development of incentives (e.g., tax credits) and associated frameworks (e.g., legal ones) for encouraging collaboration and inter-institutional data sharing across researchers in academia, government and the private sector. Specific attention was called towards the establishment of safeguards and protections for minimizing the *reputation risks* that may arise for industry organizations that may be interested in sharing datasets.

Nation-level supports for "matching" researchers/practitioners with "real-world problems to solve" and those with "ideas for solutions" can accelerate the formation of creative teams for not only enhancing the utilization of data shared, but also providing high-risk high-reward opportunities for tackling challenging problems in cybersecurity using AI.

Individual-level Approaches: At the individual level, advancements are needed for members of the community to collaborate to develop standardized approaches for cross-organization data integration and sharing. Workshop attendees also thought that individuals should make more serious commitments in contributing to the creation of synthetic data, especially those related to high-priority problems identified by the community. Finally, individual researchers should explore and expand beyond their fields of expertise so as to get more accustomed with datasets and tools that may be useful to them.

3.2 Analysis

Community-level Approaches: With regards to data analysis approaches at the community level, participants highly engaged the topic of *data anonymization*. For instance, properly labeled and curated datasets that are properly anonymized can still be proven useful in training AI/ML models for cybersecurity purposes (e.g., in supervised learning-based anomaly detection techniques).

The idea of a "trusted" clearing house that operates a service for a commonly accepted anonymization scheme (e.g., Crypto-PAn [5]) was also circulated; data integration and analysis could then be achieved at the trusted centralized location using the post-anonymization datasets. The same centralized resource could also be utilized to manage *context-based* anonymization (e.g., taking into consideration attack signatures across multiple datasets, such as IP victims in DDoS attacks) so that true correlations are maintained and false correlations are not created across multiple data sources.

The role of "secure enclaves" for privacy-aware data analysis was also explored. The secure enclaves could be stone-and-brick laboratories (analogous to the ones used by virologists and epidemiologists for studying live viruses) for analyzing malware behaviors or (temporarily) accessing sensitive cybersecurity data for experimentation and algorithm development. The secure enclaves could also be *virtual* in which researchers—in a *code-to-data* approach—submit their analysis and AI/ML software to a machine/server with access to raw, non-anonymized data for appropriate analysis and research. Additional efforts on developing community standards and best practices for *data formats* were identified as necessary, along with the need for further exploring techniques that enable *distributed* data training and analysis (e.g., swarm learning and federated learning techniques[19, 8]).

National-level Approaches: To realize the community-level approaches for analysis, national-level coordination and continuous investment and support for high-performance computing environments, development of data standards, privacy preserving analysis techniques, etc. were deemed essential for both applied and basic research. The workshop contributors applauded existing efforts from federal sponsors like the NSF in supporting researchers with "cloud computing credits" (via, e.g., the CloudBank program [15]), and solicited the expansion of such programs. Synergies between the research communities, the industry and the government were therefore considered to be critical for facilitating technology transfer, standard establishments, formulation of novel real-world cybersecurity research questions that need to be addressed, and policy making.

Individual-level Approaches: Individual members of the community can contribute to efforts of the community by sharing their data analysis approaches, the open-source software they develop, and their data analytics pipelines for exemplar problems and datasets. These individual-level approaches are important because they can be considered a testimony of the "trusting and open" culture of the community. Furthermore, they provide the catalyst for synergistic innovations derived from novel integration of data that were not imagined before, and novel combination of analytic tools that currently are not broadly accessible.

3.3 Governance

Data governance maximizes the investment in data and analytics initiatives by promoting the proper use of analytics in processes, ensuring accurate insights

based on quality data, reducing risks with security, and guiding the prioritization of projects so the right information is available at the right time.

Community-level Approaches: The participants elaborated on the need for the development of *technical* and *policy* controls for ensuring trust in the data and for enabling data owners to more easily share their data with vetted, trusted researchers. The unique issues to be considered span the areas of data security, privacy and integrity, and solutions must examine all areas and balance the *risks* versus the *benefits* that arise from a given research opportunity and data sharing / analysis scenario.

One key idea that emerged was the development of universally acceptable KPI metrics to serve as a means for assessing the success, risks and benefits in sharing and governing relevant data. These metrics should draw insights from commonly accepted ethical practices outlined in the *Belmont report*[18] and more specifically the *Menlo report*[1] which is geared towards the AI/ML and cybersecurity communities. Tools like CREDS (Cyber Research Ethics Decision Support)[7] can be leveraged (and perhaps further expanded) for quantifying the risks and benefits of a research project that requires access to sensitive data and resources. The workshop invitees also deliberated on the important role of institutional IRB in promoting transparency and helping to assess the risk of sharing and using real-world data for research.

Advancements are needed in adopting, revising, and developing relevant standards regarding artifacts (data, tools, pipelines) collection, sharing for supporting the integration of component solutions into a larger system solution. Such advancements, conducted in collaboration with standards organizations (e.g., IETF and NIST), can also expedite the transition of technology into its real-world application. An example of this approach is the development of common legal language on data usage agreements (DUA) and memorandum of understandings (MOU) with regards to sharing artifacts, leveraging related existing efforts on "standardized" DUAs (see, e.g., CAIDA and IMPACT [2, 4]). Such common standards and legal frameworks would be key for achieving long-term *sustainability* of the data infrastructures envisioned to support needed advancements to solve cybersecurity problems using AI.

Education and training of skilled workforce for governing and managing the data was also a strong workshop recommendation. New expertise will be needed to appropriately train the future data curators and "data librarians" in sharing and labeling quality data necessary to address emerging cyber-threats. This can be accomplished through the development of educational training and workshops, changes in existing curricula, online learning, and public portals for the dissemination of information and best practices.

It is also important to engage all relevant stakeholders. A community can be built by recruiting appropriate stakeholders through participation in national and international conferences, and events, and engagement with diverse stakeholders, broadening the reach to include academia (Carnegie classified R1's through Community Colleges, MSIs, HBCUs, and Tribal colleges), industry,

government, professional organizations, and Regional Research and Educational Networks (R&E's).

Finally, transparency and openness to community inputs are important for prioritizing infrastructure needs for research that advance both applied research and basic research for overcoming the obstacles for solving cybersecurity problems using AI/ML. Ensuring transparency is an important ingredient for achieving mutual trust (e.g., between providers and users of artifacts provided by the infrastructure) and for leading to proper, fair and ethical use, integration, and refinement of the artifact. In addition, transparency and openness needs to be complemented with safeguards for information and knowledge protection. It is therefore essential that organizations do implement robust information security control measures to ensure that access to sensitive data and information is appropriately managed. Auditing and tracking processes should also be in place to continually monitor the integrity and the use of the artifacts to prevent potential misuse.

National-level Approaches: At the national level, the workshop participants highlighted the need for the development of broadly acceptable policies, regulations and standards to support research efforts that address new and emerging technologies (such as AI/ML and its societal consequences for needing to process large volumes of potentially sensitive data for analysis and actionable insights). National policies and regulations were identified to be necessary to ensure, for example, transparency when accessing critical information while also protecting individuals and organizations from legal and reputation risks. The participants recognized also the positive role that the government can play in bringing existing, albeit siloed, communities together, and engaging further with the international community.

The National "Centers of Excellence", such as TrustedCI[1], can offer lessons to the cybersecurity ecosystem with regards to workforce development, knowledge sharing, and processes required for operating large-scale cyber-infrastructure that enables trustworthy science. Moreover, a "community of communities" at the national level was proposed so as to bring together multi-stakeholder organizations and researchers, and facilitate the dissemination of best practices. For example, through organized inter-agency workshops, national efforts can help collectively address the challenge of offensive versus defensive security, and encourage collaborations amongst national agencies, academia, industry as well as the international community.

Individual-level Approaches: Individuals have an important role in exercising vigilance in handling data, and can contribute to the community efforts through participation in events and by continuing their self-education and development processes as life-long learners. Individuals can raise awareness of the challenges and opportunities for the community through advocacy efforts, and via engaging new stakeholders to help expand the community.

[1] See `https://www.trustedci.org`.

4 Infrastructure Needs

With the deliberations on the proposed approaches on hand, the workshop contributors proceeded to explore and identify high-priority *infrastructure* needs. The needs were classified into the following broad categories: (1) *data storage, collection, and analysis*, (2) development of *standards, policies, and legal frameworks*, and (3) and needs for *support infrastructure, education, skill development and training*. Next, we elaborate on key necessities identified for all three categories.

Data Collection, Storage and Analysis: On this theme, the community first acknowledged the various categories of datasets required for addressing the research opportunities such as the ones outlined in Table 1. Namely, participants raised the need for several "types" of data collection efforts: (1) the collection of *ongoing/longitudinal real-world* (e.g., IP-based threat intelligence data from network telescopes or other data sources), (2) real-world *benchmark data* for training AI/ML models (e.g., annotated data about malware binaries that are labeled and contain features so that one could apply AI/ML classification tasks on), (3) *anonymized / sanitized data* that could be used for, say, educational purposes, (4) data that are collected and analyzed in an *online/streaming manner* (e.g., data collected from packet taps), and (5) *simulated or synthetic* datasets. Note that these categories are not mutually exclusive. For example, one could work with anonymized datasets that are labeled in a meaningful manner so that they remain useful for training appropriate AI/ML models.

Given this background, the discussants proposed that future infrastructure should be able to easily scale (both up and down) to support the various size of data collection and data processing needs. The envisioned infrastructure should have the capability to annotate the collected data with appropriate (machine-readable!) metadata and contextual information (e.g., labeling instances of known DDoS attacks, malware families, etc., when possible). To achieve this, advancements are needed for innovative integration of labelled data and unlabelled data from different sources, leveraging related efforts such as data acquisition processes in [17, 20].

Computing and data infrastructure that supports data analytic pipelines, especially those related to machine learning using big cybersecurity data, is an infrastructure need that, once met, can democratize research capacity for a much broader and diverse community and accelerate advancements in science and technology of cyber security using AI. Furthermore, secure infrastructures are needed for supporting malware dynamic analysis/emulation, including those with evasive behaviors, for different platforms.

It is thus important that this future cyber-infrastructure supports both stable, "production-grade" tooling for data collection and analysis as well as the ability for experimentation and the development of speculative pipelines and tools. To facilitate reproducibility of research results and infrastructure, the necessities for developing standards describing *how* the data was collected and what metadata is needed for data annotation were pinpointed. Due to the

highly dynamic nature of cyber threats, data for cybersecurity research needs to be updated to reflect the changing real-world. Hence, suitable version control for data, tools, and pipelines are needed for enhanced reproducibility and sustainability of AI-enabled research using the data. Furthermore, investing in open-source tooling and software whenever possible should be prioritized in order to promote continued future use and sustainable operations.

Naturally, issues on data privacy were also discussed (both under this theme as well as within the next one that covers the need for the development of "Standards, Policies and Legal Frameworks"). The need for developing best practices and assorted infrastructure for privacy-preserving, or more generally, data use policy compliant analyses was highlighted. Similarly, policies that offer guidance for the preservation of archival data, metadata and derived artifacts need also be considered. Universally accepted practices for data sharing of different data types using different sharing methods (e.g., centralized access versus ones based on privacy-respecting federated learning) should also be instrumented, based on community feedback. To ensure ethical and fair use of the data, data providers were encouraged to consider accompanying each dataset they collect with appropriate "limitations" and "safeguards".

Further, the participants discussed the need for leveraging existing methods (and extensions thereof) for sharing data (e.g., threat intelligence information) across multiple organizations (e.g., see [6]). Incentive mechanisms might be needed to be offered to attract higher participation within such initiatives, especially from the industry. Building "trusted relationships" between all stakeholders (researchers, data providers, citizens, etc.) was identified as a necessary step towards broadening such participation. Moreover, continuous engagement with stakeholders needs to be in place to ensure that the infrastructure remains always responsive to the needs of the community and the citizenry.

Finally, the workshop invitees brainstormed on the needs for developing new sampling methods for analyzing and storing large datasets, and for leveraging "cloud-based" solutions for both data analysis and long-term storage. The STRIDES initiative[2] (which supports the use of cloud for biomedical research) and the NSF's CloudBank [15] (which helps the computer science community access and use public clouds for research and education) were provided as exemplar resources.

Standards, Policies, and Legal Frameworks: In addition to the proposals for standardization of data collection formats, metadata and tooling discussed above, the workshop attendees emphasized the benefits of having commonly acceptable and broadly available standards for data sharing agreements, access control framework, and associated policies that would govern the access, analysis, and permissible outputs of the data at a detailed level that allows data to be integrated and harvested while, in the same time, protecting sensitive information in the data. For examples, sample agreements can be developed and used for different types of engagement, including: data contribution, data analysis,

[2] `https://datascience.nih.gov/strides`

infrastructure contribution, infrastructure use, data retention, etc.

Workshop participants also raised the importance of developing processes for ensuring data provenance, detecting data pollution, and when necessary, taking corrective measures. Reducing the risk of data pollution also requires effective cybersecurity risk management, governance and accountability to enable the identification, assessment and management of cybersecurity risks at both the organizational and cross-organizational partners involved in the infrastructure.

Due to the complexity of system security, it is important to not only establish guidelines and governance on secured data access and analytics, but also promote best practices and user awareness to ensure infrastructure remains safe and secure. Security should also be considered at the conception of the infrastructure to adopt "secure by design" principles and a risk-based approach that ensure the infrastructure can protect data with different levels of risk.

Leveraging Institutional Review Boards (IRB), it is also desirable to develop sample IRB protocols, perhaps a single unified IRB framework for research using the infrastructure. Such a framework can ensure the compliance of ethical guidelines that provide a code of ethics for acceptable use policy that, given the research question and the data, can come up with a framework that "scores" the ethical risks (e.g., the CREDS tool [7]). More generally, there is also a need of guidelines and formats for fair use of data and AI in cybersecurity research. Finally, the design of new systems should be supported by security principles developed in collaboration with national and international standards organizations.

Support Infrastructure, Education, Skill Development, and Training: The most critical element for any moonshot project is human resources. Hence, it is essential that we grow a sustainable and diverse skills pipeline to support the AI/cybersecurity workforce. To facilitate the skill development and the recruitment of diverse next-generation cybersecurity workforce using AI, a Kaggle-like service that provides cybersecurity problems, associated data, and exemplar AI/ML solutions can significantly reduce the barriers for educators, students, and practitioners to boot-strap a personal hands-on learning experience for solving real-world cybersecurity problems using AI.

To further democratize the research capacity for educators, researchers, practitioners, and next-generation cybersecurity/AI knowledge workers to leverage high performance computing (HPC) resources and big data for solving cybersecurity problems using AI, training materials need to be designed for people with different levels of knowledge and skills regarding cybersecurity, AI, and HPC, and be broadly disseminated through multiple venues, including, but not limited to, online training modules, summer camps, Hackathons, cybersecurity competitions/exercises, and workshops co-located with major conferences in cybersecurity and/or AI.

5 Community Development

As the workshop was concluding, participants reiterated the need for growing and strengthening the community of researchers that contribute data, infrastructure and other resources, and perform research in areas at the interface of AI/ML, networking and cybersecurity. Several characteristics of such high-performance community were identified, including: (1) being inclusive and having a trusting culture, (2) driven by shared purposes, (3) empowered by artifacts, (4) supported by processes, (5) making societal impacts, and (6) being a community that fosters next generation researchers.

1. *Inclusive Trusting Culture:* An inclusive, trusting, and diverse community, with openness to all comers and their ideas, is considered a fundamental requirement for the community to strive.

2. *Driven by Shared Purposes:* The growth of a trusting inclusive community needs to be driven by a shared vision and purposes that define and shape value-added opportunities such as (1) being able to tackle problems that cannot be tackled individually, (2) collaboratively develop an understanding about truly unsolved problems, (3) forming new collaborations that would have been difficult otherwise.

3. *Empowered by Artifacts:* An inclusive, trusting community with a common purpose needs to be empowered by shared artifacts such as access to quality datasets, analytic tools, best practices, learning modules, etc.

4. *Supported by Process:* Community building activities, built on an inclusive trusting culture with shared purposes and artifacts, not only create and enhance the sense of belonging, but also enable the sharing of successful collaboration results. By organizing these activities on a regular basis, co-located with major conferences of cybersecurity and/or artificial intelligence, these activities can not only advance AI-enabled cybersecurity, but also can provide positive feedback to the community, and attract new members and ideas to grow and strengthen the community.

5. *Societal Impacts:* A wide range of high-impact societal outcomes can be achieved by the community envisioned above: the community can provide one voice to advocate community research and infrastructure needs, becoming the go-to place for national (and possibly international) policy making, standard forming, innovation roadmap articulation, etc.

6. *A Learning Community that Fosters Next Generation Researchers:* Due to the rapidly changing landscape of cyberattacks, the community not only needs to strive for continuously learning from each other and additional resources, but also to mentor and educate the next generation researchers, knowledge workers and infrastructure support staff so that the community can remain sustainable over multiple generations.

6 Conclusion

The workshop brought together a diverse group of researchers from academia, industry, and the government to identify research opportunities in cybersecurity that can be addressed with AI/ML, and discuss the steps required for building the envisioned *tangible* and *intangible* infrastructures for achieving the identified research objectives. The participants offered approaches at the community, national, and individual levels that would help the AI/ML and cybersecurity communities overcome some of the barriers in data acquisition and sharing, data analysis, and data governance toward the goal of AI-enabled cybersecurity. Concrete infrastructure suggestions were then proposed to accomplish these approaches.

We acknowledge that our community has already made noteworthy strides in the past 10–15 years towards tackling some of the identified challenges. However, to fully realize the potential of AI/ML in solving an even broader variety of highly dynamic and complex cybersecurity challenges, the community of researchers and practitioners involved in using AI/ML to solve cybersecurity problems needs to be broadened and strengthened. Further, through democratizing the needed research capacity, and via inter-disciplinary synergies and cross-pollination of ideas, we can accelerate the advancements in both applied and basic research. These steps are necessary for making further progress towards the common visions emerged at the workshop, such as the one "immune system for cybersecurity" moonshot concept and others.

7 Acknowledgements

This workshop is supported by NSF CCRI Planning-C Grant 2213794. We are greatly indebted to our workshop facilitators, Toby Scott and Emma Skipper from KnowInnovation, for their excellent stewardship, guidance, and support throughout the entire workshop process (planning, organizing, facilitating, adapting, and reflecting on the discussions at the workshop). We also are extremely thankful for the participation of the workshop attendees, without whom this event would not have been possible. The complete list of participants can be found in Table 5 in the Appendix.

References

[1] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[2] CAIDA. Caida data catalog. `https://catalog.caida.org/`.

[3] CAIDA. The UCSD network telescope. `https://www.caida.org/projects/network_telescope/`.

[4] Department of Homeland Security. Information Marketplace for Policy and Analysis of Cyber-risk & Trust. https://www.dhs.gov/science-and-technology/cybersecurity-impact.

[5] Jinliang Fan, Jun Xu, Mostafa H. Ammar, and Sue B. Moon. Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme. *Computer Networks*, 46(2):253–272, 2004.

[6] Panos Kampanakis. Security automation and threat information-sharing options. *IEEE Security & Privacy*, 12(5):42–51, 2014.

[7] Erin Kenneally and Marina Fomenkov. Cyber research ethics decision support (creds) tool. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*, pages 21–21, 2015.

[8] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[9] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 17(1):1235–1241, 2016.

[10] Merit Network, Inc. ORION: Observatory for Cyber-Risk Insights and Outages of Networks. https://www.merit.edu/initiatives/orion-network-telescope/, 2022.

[11] Microsoft Azure Network Security Team. 2022 in review: DDoS attack trends and insights. https://www.microsoft.com/en-us/security/blog/2023/02/21/2022-in-review-ddos-attack-trends-and-insights/ (Accessed: March 5th, 2023).

[12] Microsoft Security Intelligence. Global Threat Activity. https://www.microsoft.com/en-us/wdsi/threats (Accessed: March 7th, 2023).

[13] Jelena Mirkovic, Stephen Hayne, Michalis Kallitsis, Wes Hardaker, John Heidemann, Christos Papadopoulos, and Devkishen Sisodia. Cybersecurity datasets: A mirage. NSF Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021), 2021.

[14] Jelena Mirkovic, John Heidemann, Wes Hardaker, and Michael Kallitsis. CLASSNET: Community Labeling and Sharing of Security and Networking Test datasets. https://ant.isi.edu/classnet/index.html, 2021.

[15] Michael Norman, Vince Kellen, Shava Smallen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, Rob Fatland, Sarah Stone, Amanda Tan, et al. Cloudbank: Managed services to simplify cloud access

for computer science research and education. In *Practice and Experience in Advanced Research Computing*, pages 1–4. 2021.

[16] Ben Shneiderman. *The new ABCs of research: Achieving breakthrough collaborations*. Oxford University Press, 2016.

[17] Rajat Tandon, Pithayuth Charnsethikul, Michalis Kallitsis, and Jelena Mirkovic. Amon-senss: Scalable and accurate detection of volumetric ddos attacks at isps. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 3399–3404. IEEE, 2022.

[18] U.S. Department of Health, Education, and Welfare. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. `https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf`.

[19] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021.

[20] Zhiying Xu, Sivaramakrishnan Ramanathan, Alexander Rush, Jelena Mirkovic, and Minlan Yu. Xatu: Boosting existing ddos detection systems using auxiliary signals. CoNEXT '22, page 1–17, New York, NY, USA, 2022. Association for Computing Machinery.

Appendix

Table 2: Challenges identified with regards to Data Acquisition and Sharing

How to collect and share data when dealing with networks that run at 100G and beyond?
How to incentivize sharing data?
How to ensure privacy?
How to incentivize truthful data sharing? In other words, how to prevent strategic manipulation?
How to have AI model to automatically to learn language of network traffic? What would it learn?
How do we anonymize data?
How do we share data ethically?
How might we create pull-back option/capabilities for data sharing?
How to identify/associate data ownership in a shared model?
How to have systems in place to identify data ownership and different uses of the data?
How to secure data transmission?
How to create a market associated with value of data?
How do we ensure transparent use of the data?
How to ensure liability of shared data? Who is liable? Who's responsible?
Could there be an insurance model for sharing data, especially when data becomes an asset used to generate revenue?
How to convince data owners of the importance of data management (the depth of the data life cycle and attributes of the data life cycle), and convince sector leaders of the importance of sharing data and the value of the data.
How might we get access to existing datasets? How might we get access to existing private datasets? How might we acquire new datasets that are currently not available?
How might we share longitudinal, up-to-date data across organizations? How might AI/ML (e.g., federated learning) assist in alleviating privacy concerns?

Table 3: Challenges identified with regards to Data Analysis

How might we process large volume of streaming data in real time?
How might we make sense of multi-modal data, expressed in different format?
How might we support multiple ML models learning over big, streaming data?
How might we develop continuous and diversified and explainable learning over big, streaming data?
How might we coordinate quickly and efficiently in the face of a large-scale cyber attack?
How might we prioritize analysis and results across multiple organizations, lines of effort, and even across country/legal boundaries?
How might we have a shared schema/representation of complex cyber attacks TTP?
How might we efficiently unpack/de-obfuscate malware to obtain "ground truth" information about the original code?
How might we combine different levels of temporal and spatial granularities of data representations (e.g., schemas)?
How might we deal with zero-day threats/attacks?
How might we create a schema of information for analysis that is granular enough at all of the layers of cyber data (e.g. from instruction opcodes, to general behaviors like "process injection", to even more general behaviors like "phishing attacks")?
How might we use/innovate embedding for addressing these challenges?
How might we understand and analyze the entire processes of a malware ecosystem (from controller, to C2 nodes, to implant) with only partial data?
How might we accurately compare the results we obtain from multiple ML models to establish measures of improvement to threat analysis?
How might we correlate and analyze across varied and disparate datasets?
How might we leverage intermediate representations of malware for integrated data analytics?

Table 4: Challenges identified with regards to Data Governance

How might we address the issue of privacy when sharing data across organizations?
How might we address the issue of bias and fairness?
How might we address the issue of data governance including integrity and quality and provenance?
How might we safeguard the data that we collect and share? How might we establish suitable trust with the data host? How might we leverage techniques such as "blockchain"?
How might we address the issue of different data formats?
How might we come up with standard/universal data sharing agreements and MOUs for sharing data across organizations?
How might we address the resiliency of data and the analytic tools/models?
How might we assess the risk of data to be shared in an automated manner (with minimal legal counseling services involved)?
How might institutional IRBs be structured to address the ethical concerns of data to be shared?
How might we come up with a governance framework that small/new organizations might use to share their data? What DUAs/MOUs they might need? How to assess the risk?
How might we train/educate people to deal with governance infrastructure? How might we train/educate people to understand and assess the trade-offs between the scientific utility of data analytic requests vs potential risks they may introduce?
How might we assess the usability of data in terms of KPI (key performance indicators) with respect to use cases, accountability, practicability, auditability, trackability, and explainability?
How might we address the issue of sustainability and scalability of the governance infrastructure and the data infrastructure? How long do we need to keep the data? How long do data stay relevant? What are the archival procedures?
How might we accurately define what is "sensitive data" as technology evolves and more data become available?
How do we define ethics? Is it to humans? Is it to organizations? Is it to algorithms?
How do we might address governance issues that might arise due to different international laws and standards?
How might we ensure the reusability and the reproducibility of artifacts (code, derived models, etc). How might we ensure the user-friendliness and accessibility of the interface for browsing, searching, and using artifacts?

Table 5: Workshop Participants

Dinghao Wu; Penn State University
Forough Ghahramani; NJEdge, Inc.
John Yen; Penn State University
Michalis G Kallitsis; Merit Network, Inc., University of Michigan
Vasant Gajanan Honavar; Penn State University
Bhavani Thuraisingham; The University of Texas at Dallas
Christos Papadopoulos; The University of Memphis
David Bader; New Jersey Institute of Technology
David Schanzenbach; University of Hawaii
Dawei Zhou; Virgina Tech
Robert F. Erbacher; Army Research Labs
Edward Raff; Booz Allen Hamilton
Elena Yulaeva; CAIDA/UCSD
Elias Bou-Harb; The University of Texas at San Antonio
Hadi Hosseini; Penn State University
Haipeng Cai; Washington State University
Jelena Mirkovic; USC/ISI
Jim Basney; University of Illinois at Urbana-Champaign
Jingrui He; University of Illinois at Urbana-Champaign
Joseph Khoury; The University of Texas at San Antonio
Kangjie Lu; University of Minnesota
Kirk Bresniker; Hewlett Packard Enterprise
Mike Qaissaunee; Brookdale Community College
Murat Kantarcioglu; University of Texas at Dallas
Neeraj Karamchandani; Penn State University
Priyanka Chilakalapudi; The University of Memphis
Raj Badhwar; Oracle
Roman Daszczyszak; MITRE
Samantha Kleinberg; Stevens Institute of Technology
Sudip Mittal; Mississippi State University
Tim Finin; University Of Maryland, Baltimore County
Xiaokuan Zhang; George Mason University
Xinyu Xing; Northwestern University
Yevgeniy Vorobeychik; Washington University in St. Louis
Zhi-Li Zhang; University of Minnesota