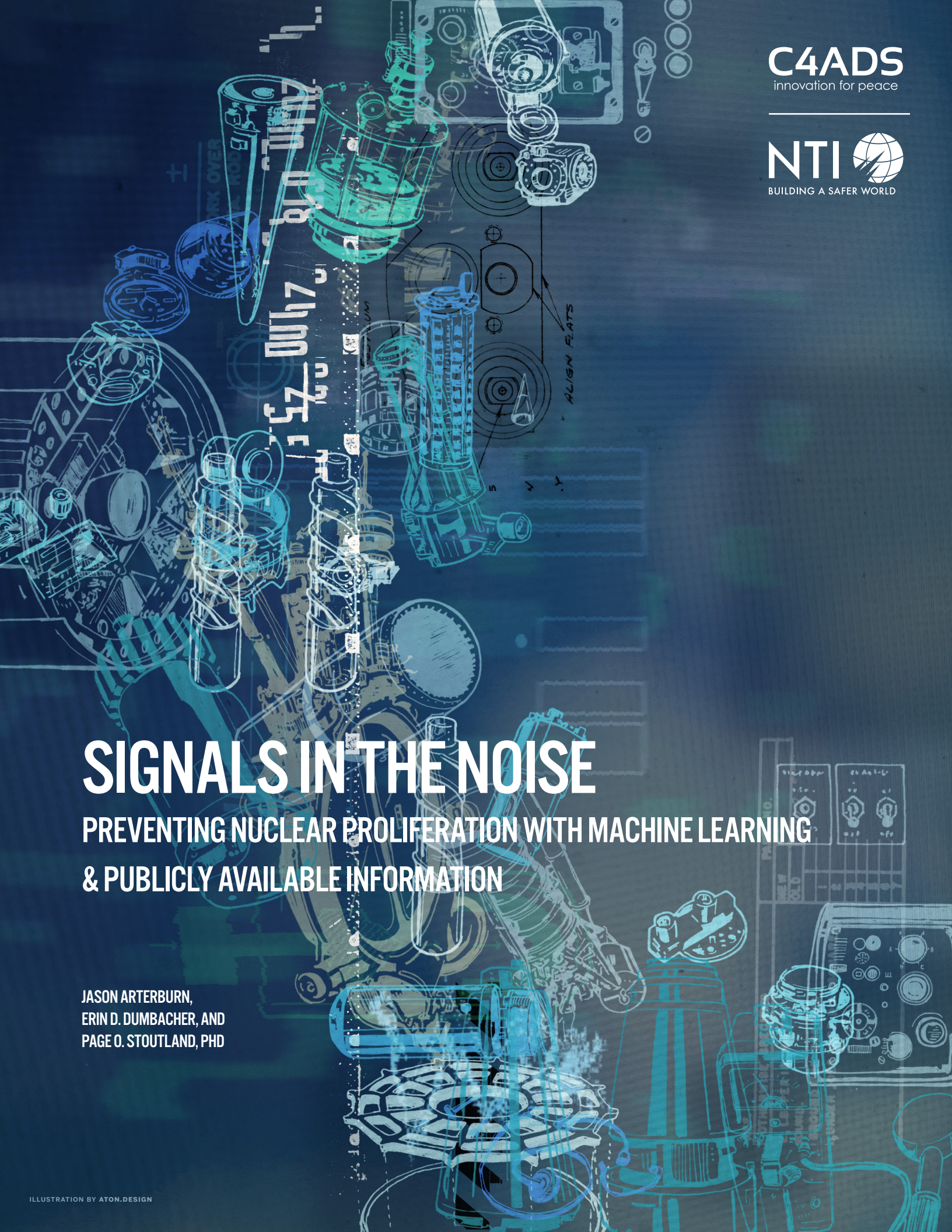


C4ADS
innovation for peace

NTI
BUILDING A SAFER WORLD

SIGNALS IN THE NOISE

PREVENTING NUCLEAR PROLIFERATION WITH MACHINE LEARNING & PUBLICLY AVAILABLE INFORMATION

JASON ARTERBURN,
ERIN D. DUMBACHER, AND
PAGE O. STOUTLAND, PHD

ILLUSTRATION BY ATON.DESIGN

About C4ADS

C4ADS is a 501(c)(3) nonprofit organization dedicated to data-driven analysis and evidence-based reporting of conflict and security issues worldwide. We seek to alleviate the analytical burden carried by public sector institutions by applying manpower, depth, and rigor to questions of conflict and security. Our approach leverages nontraditional investigative techniques and emerging analytical technologies. The result is an innovative analytical approach to conflict prevention and mitigation.

c4ads.org @c4ads

C4ADS Acknowledgments

C4ADS would like to thank all those who provided advice and guidance during this project. The author extends special gratitude to the peer reviewers of this report for their insightful feedback and advice and to fellow team members Arjun Rohlfing-Das, Max Kearns, Georgia Channing, Patrick Baine, and Jack Margolin, without whom the report would not have been possible. C4ADS would also like to thank its technology partners, whose software and systems were integral to the project's success.

ANALYSIS POWERED BY Palantir aws

About NTI

The Nuclear Threat Initiative (NTI) is a nonprofit global security organization focused on reducing nuclear and biological threats imperiling humanity.

nti.org facebook.com/nti.org @NTI_WMD @NTI_WMD

NTI Acknowledgments

We are grateful to many who have supported this effort. NTI leadership, including NTI Co-Chair and CEO Ernest J. Moniz, former NTI President Charles Curtis, Executive Vice President Deborah Rosenblum, and members of NTI's Science and Technology Advisory Group provided valuable guidance. We thank NTI colleagues for sharing their expertise including Corey Hinderstein, who provided an important perspective, and Catherine Crary, who ensured project success from beginning to end. Finally, we thank members of NTI's Communications team—Carmen MacDougall, Mimi Hall, and Deepika Choudhary—for their support in developing this report.

Cover illustration by Aton.Design

TABLE OF CONTENTS

FOREWORD

Illicit trafficking of nuclear materials and technologies around the world—whether by terrorist organizations, rogue states, criminal enterprises, or even unwitting mules—poses a serious threat to global security. Those who engage in such criminal acts evade detection by operating in the shadows and in plain sight. They use unusual routes and sketchy trading partners, and they use front companies as cover, concealing their trade in routine shipment lists.

Today, however, new advances in data science and tools such as machine learning, in combination with greater amounts of publicly available information, can help us detect illicit trafficking and catch those who engage in it. New tools can be used to prevent nuclear proliferation by revealing footprints left by bad actors and to monitor and verify future arms control and export control agreements.

The Nuclear Threat Initiative (NTI) helped spur development of the field a number of years ago, first defining "societal verification" as a possible contributor to monitoring compliance with international agreements. With impressive advances in the use of commercially available satellite imagery, social media content, and other data by governments and non-government organizations alike, these tools can build transparency and confidence among nuclear and non-nuclear states that agreements are being kept.

Through our ongoing work with the Center for Advanced Defense Studies, NTI has demonstrated the use of publicly available information to contribute to the reduction of nuclear risks worldwide. We encourage governments, multilateral institutions, and non-governmental entities to join us in exploring the possibilities and to embrace the new opportunities these new technologies afford in pursuit of a safer world.

Ernest J. Moniz
Co-Chair and Chief Executive Officer
Nuclear Threat Initiative

EXECUTIVE SUMMARY

For decades, illicit trade in nuclear materials, equipment, and technologies has undermined global nuclear non-proliferation efforts. Sophisticated actors establish front companies, forge documents, and launder money to obscure proliferation activities, and are too often able to evade detection—even as they operate within legal systems of trade, finance, transportation, and communication.

They do leave footprints, however, and now, with an increase in the volume and variety of publicly available data, there are new opportunities to discover and expose such activities. When applied to the right forms of publicly available information (PAI), emerging data science methods and advanced analytical tools can expose proliferation activities, and they should be used to serve the global non-proliferation mission to reduce the risk of catastrophic consequences from use of a nuclear weapon.

Over two years, the Nuclear Threat Initiative (NTI) and the Center for Advanced Defense Studies (C4ADS) worked together on a pilot project to demonstrate the viability of using PAI and machine learning to detect high-risk and/or illicit nuclear trade. The initiative leveraged NTI's nuclear expertise and C4ADS's data management, engineering, and analysis capabilities to identify high-risk nuclear proliferation activities at scale.

The project succeeded. Trade network analysis—and the machine learning processes that supported it—uncovered previously unknown entities of elevated risk within millions of transactions. The work showed that automated data preparation could save hundreds of analyst hours and help identify twice as many potentially high-risk entities as previous manual efforts. In addition, when applied to a baseline study of more than four million records, machine learning techniques could identify 50 new leads for further review. During the two-year study, at least ten entities identified through these approaches were added to a U.S. government export control list, demonstrating that novel analytic approaches to PAI can produce law enforcement-relevant insights.

The NTI-C4ADS pilot project yielded several key findings:

- **The use of publicly available data and analysis is effective at identifying high-risk nuclear trade. To be operationally useful, the analysis approach must be able to accommodate the velocity, volume, variety, and complexity of real-world data.**
- **Machine learning tools can be used to dramatically enhance data analysis in terms of both speed and quality. Model selection is best accomplished empirically to determine the most effective models for particular situations. This method places a premium on access to relevant datasets as well as subject matter expertise.**
- **When coupled with the right tools, the use of PAI holds enormous potential for the detection and prevention of illicit nuclear trade. Additional work is needed to address regional trade pattern differences, uneven data availability, and opportunities for integration with other data types.**

The project recommends that leaders of non-proliferation efforts in governments and multilateral organizations around the world ensure that PAI and modern analytical approaches are employed to monitor and ultimately disrupt illicit nuclear activities. Specifically, our recommendations include the following:

- **integrate PAI more broadly into existing monitoring and verification regimes;**
- **use modern analytical tools and approaches, including machine learning, to enable collection and analysis of PAI at scale;**
- **build partnerships to allow analysts access to complementary data and capabilities; and**
- **embrace the use of PAI and modern analytical tools for future international non-proliferation initiatives.**

Taken together, these steps can support analytic production with the quality, scale, and timeliness required for an operational monitoring capability. Indeed, in the future, it may be impossible for a proliferator to evade detection.

THE NEED TO PREVENT NUCLEAR PROLIFERATION

Fifty years ago, the landmark Treaty on the Non-Proliferation of Nuclear Weapons (NPT) entered into force, setting the foundation of a global security regime designed to prevent the spread and use of nuclear weapons and to promote the peaceful use of nuclear energy. Today, the NPT continues to be the cornerstone of global efforts to reduce nuclear risks while enabling the benefits of nuclear technologies, with these goals enshrined in Articles I and II. But preventing the proliferation of nuclear materials and technology has not been an easy task.

> **Article I: Each nuclear-weapon State Party to the Treaty undertakes not to transfer to any recipient whatsoever nuclear weapons or other nuclear explosive devices or control over such weapons or explosive devices directly, or indirectly; and not in any way to assist, encourage, or induce any non-nuclear-weapon State to manufacture or otherwise acquire nuclear weapons or other nuclear explosive devices, or control over such weapons or explosive devices.**

> **Article II: Each non-nuclear-weapon State Party to the Treaty undertakes not to receive the transfer from any transferor whatsoever of nuclear weapons or other nuclear explosive devices or of control over such weapons or explosive devices directly, or indirectly; not to manufacture or otherwise acquire nuclear weapons or other nuclear explosive devices; and not to seek or receive any assistance in the manufacture of nuclear weapons or other nuclear explosive devices.**[1]

Over the years, the international community has established a number of mechanisms to control the illicit trade of the materials, equipment, and technology needed to make nuclear weapons, including those detailed not only by the multilateral 1975 Nuclear Suppliers Group guidelines but also by associated national export control regimes. However, despite restrictions on trade in numerous types of dual-use goods, illicit transfers are still carried out by sophisticated networks of people and companies that are increasingly capable of thwarting governments' ability to detect their activities. In 2005, members of the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction (WMD Commission) found that "traditional collection techniques have been degraded by the pace of change in telecommunications technology and by our adversaries' increasing awareness of our [intelligence] capabilities"—which, they noted, limited the U.S. government's ability to respond effectively to the development and transfer of weapons of mass destruction.[2]

Increasingly, publicly available information (PAI) can enhance traditional intelligence approaches and support timely, effective responses to illicit proliferation activities. New data analysis tools increasingly facilitate the use of diverse, complex data sources, even by those with limited technical expertise. Although the non-proliferation community has been one of the most progressive in using some types of PAI, such as satellite imagery, there are vast quantities of unexploited PAI that could further aid non-proliferation efforts, including but not limited to shipping manifests, corporate registry filings, procurement tenders, and vessel or aircraft position data.

Over the last two years, the Nuclear Threat Initiative (NTI) and the Center for Advanced Defense Studies (C4ADS) have worked to explore current and future opportunities to leverage PAI for non-proliferation goals. Over the course of a pilot project, we found that PAI has tremendous promise to reduce the risk of nuclear weapons proliferation and advance the NPT's central aim.

ASSESSING THE CHALLENGES AND OPPORTUNITIES OF USING PUBLICLY AVAILABLE INFORMATION

Networks involved in nuclear proliferation often operate within legal systems of trade, finance, transportation, and communication. In doing so, those networks leave data footprints in records that are increasingly available in the public domain. While illicit actors may establish front companies, forge documents, or launder money to obscure their activities, analysts increasingly have opportunities to expose those activities through structured investigative processes that leverage many different sources of publicly available information.

Non-proliferation experts outside government as well as those within governments and international organizations have begun pioneering methods to harness new types of data for monitoring the highest-risk nuclear programs around the world. For example, the James Martin Center for Nonproliferation Studies,[3] 38 North,[5] King's College London's Project Alpha,[6] the Center for Strategic and International Studies,[7] and the Institute for Science and International Security[8] have all contributed significantly to the public understanding of global nuclear programs by using satellite imagery, scientific publications, social media, and gray literature to assess nuclear-related activities ranging from missile launches to enrichment, reprocessing, and mining of uranium. The non-proliferation community has developed sophisticated methodologies to make independent assessments about countries' nuclear capabilities and intent, and this work often has supported analysis with sufficient timeliness to provide advance warning about possible nuclear activities.[9] However, comparatively less work has successfully leveraged PAI to identify and respond to nuclear proliferation activities with the same timeliness and precision.

The International Atomic Energy Agency (IAEA) has also tested new approaches for using PAI and advanced data integration technology to ensure that countries are using their nuclear materials and technologies for peaceful purposes.[10] In 2010, IAEA leadership called for a "move towards a Safeguards system that is fully driven by the use of all the safeguards-relevant information available," and in the time since, IAEA safeguards experts have made significant strides in adopting more diverse forms of PAI to execute its mandate.[11] However, resource limitations and bureaucratic hurdles impede more complex data collection and integration, restricting the IAEA's ability to incorporate an increasingly wide array of relevant data sources into its monitoring and verification efforts.

Although government, industry, and civil society are all working to adapt new tools for the modern information

WHAT IS PUBLICLY AVAILABLE INFORMATION?

Publicly available information (PAI) is any content from general media, social media, public records, commercial databases, gray literature, audio recordings, imagery, or expert interviews that can be legally purchased, obtained, or created by the public.[3] PAI may include information that is controlled but accessible to the public or to sections of the public within the bounds of those controls. While PAI may include datasets that could be considered "big data," it may also include small, static datasets whose scope, completeness, and coverage would not qualify them for that label.

environment, the pace of change has been slow.[12] In addition to conceptual and policy challenges, there are a number of other technical hurdles to using the full range of publicly available information for proliferation detection:

- First, the **velocity** of data creates technical and organizational challenges for collection, processing, and analysis. To be actionable, data may require continuous refresh, which requires significant financial resources to sustain. Data sources are constantly changing, and maintaining access requires developing relationships with a diverse set of data providers around the world, which must be vetted and whose information must be validated on an ongoing basis. Because of the cost in both time and financial resources, analysts must be judicious in determining the data sources that are most useful for a particular application and must develop targeted collection strategies to serve specific, bounded analytic requirements.
- Second, the **volume** of data required for analysis creates technical and analytical challenges. For example, trade datasets may contain several million rows, each of which may have hundreds of data points that could be relevant for analysis. As an example, C4ADS's global trade data holdings contain more than 800 million shipment records to and sourced from ten providers. To be useful for analysis, trade records must be integrated across languages and formats, and, as best practice, should be vetted against other commercial trade data from multiple sources to compare coverage and content. As a result, users must have data management tools that support data preparation, integration, and analysis at scale.
- Third, the **variety** of relevant publicly available data sources compels users to integrate many different data types, such as trade records, corporate registry documents, port records, shipping manifests, legal filings, transaction records, and procurement tenders. Using multiple data sources can help fill gaps in one source or corroborate details from third-party datasets, which may vary in their scope, completeness, or reliability. By integrating multiple types of data, analysts can create a more complete picture of proliferation risk—for example, by exposing the people who own or operate companies involved in suspicious shipments. The type, cost, and format of available data sources vary significantly by jurisdiction, and developing an effective data strategy for a given jurisdiction of interest requires time, creativity, and regional and/or linguistic expertise.
- Finally, the **complexity** of data requires a means to find subtle clues hidden within large quantities of data. To effectively consider these subtle details with the scale and speed required for timely proliferation detection, users must possess both the domain knowledge to understand which contextual features within available data are most relevant and the data management tools to assess them.

OUR APPROACH: OFFERING A WINDOW INTO PROLIFERATION ACTIVITIES WITH ENTITY-LEVEL TRADE DATA

Challenges to using bulk data can be significant, but trade data, in particular, show promise for the detection and monitoring of the proliferation of nuclear materials, equipment, and technologies to countries that are seeking nuclear programs but do not have the capacity to create them.

To date, non-proliferation analysts using trade data have primarily used aggregated statistical data to analyze trends at the country level, but such data do not contain information about the specific companies or organizations involved in trade activities.[13] However, so-called bill-of-lading-level trade data, which contain information about consignors and consignees for shipments, have the potential to support timely analysis and intervention by a wide range of public and private stakeholders, including law enforcement, financial institutions, customs authorities, and international monitoring groups. Indeed, analysts have used bill-of-lading-level trade data to investigate discrete networks of companies, but no organization has used such data to develop a persistent, countrywide monitoring and verification capability.[14]

While civil society organizations and news media have used bill-of-lading-level trade data made available through commercial aggregators to support proliferation investigations, they have used in most cases only a small number of shipment records as supplements to other data in discrete investigations. Comparatively less work has been conducted using bill-of-lading-level trade data at scale as a means to discover new entities of concern through deductive analytical approaches—that is, to use data-driven measures to evaluate proliferation risk at the company or shipment level.

The NTI-C4ADS Pilot Project

In the pilot project, NTI and C4ADS used bill-of-lading-level trade data to explore the potential for using this particular form of PAI to support the global non-proliferation regime. Specifically, we examined whether analysts could identify new front companies by using only publicly available trade data. We focused on trade data because of the degree to which its insights could be used by a range of public and private stakeholders supporting non-proliferation.

NTI and C4ADS tested a series of scalable analytic approaches to detect high-risk trade in large quantities of publicly available trade data for a country of interest. While moderately useful for compliance purposes, the preliminary findings suggest that list-based screening methodologies, like one-dimensional screens based on product classifications from export control lists, have little utility in identifying new entities of proliferation concern because product classification schemes are often too broad to reliably isolate shipments of nuclear-relevant materials. However, an activity-based approach, which correlates the trading patterns of known entities of concern with those of previously unknown companies active in nuclear-related trade, can help to identify new entities of concern across millions of shipments.

In order to acquire and exploit trade data at scale, however, analysts must address several fundamental challenges, including the velocity, volume, variety, and complexity of the data. To address these issues, the pilot project assessed the potential benefits of using machine learning for analysis of publicly available trade data, taking advantage of the fact that "risky" trade has certain characteristics (e.g., procuring specialized goods produced by a limited number of suppliers).

Methodology

In the NTI-C4ADS pilot project, analysts selected companies in a country of interest ("known entities") as the starting point for analysis.[15] Governments, media, and research organizations have published lists of people and companies known to have engaged in illicit nuclear procurement, which support law enforcement and civil regulators in screening for regulatory compliance. These lists can also be useful in understanding signatures of illicit nuclear trade such as shipment routes, overseas suppliers, and the ways these routes and suppliers change after law enforcement action. While government-issued lists are useful, they may not represent the full scope of individuals or companies involved in proliferation activities, and should therefore not be understood as a completely representative sample of nuclear proliferation activities in a given country.[16] Where necessary, analysts can supplement government-issued lists with information on people and companies with suspected ties to a nuclear program that appear in reporting by media and other research organizations.

As a second step, analysts "zoomed out" to identify all companies overseas that had traded with known entities in the country of interest. At this stage, the scope of analysis expanded from dozens of companies to many thousands of companies and transactions. Because entity-level trade data contain information about the companies that have sent and received shipments, analysts can observe patterns in commercial behavior and apply social network analysis techniques to contextualize companies and shipments along a gradient of risk, which provides a means by which to prioritize deep-dive investigations.

As a final step, after identifying foreign trade partners for known entities, analysts "zoomed back in" to the original country of interest to identify new, previously unknown companies receiving goods from the same high-risk overseas suppliers ("new entities"). Depending on the specific analytic requirement, analysts using this methodology could choose from several different methods to prioritize new, previously unknown companies for enhanced scrutiny. For example, analysts could prioritize new entities on the basis of the number of central overseas suppliers with which they trade, or by either the proportion or volume of their trade with those suppliers. Upon selecting targets for investigation, analysts could then incorporate other forms of PAI such as business registry filings, property records, national gazettes, social media, or satellite imagery in order to make a more complete, data-grounded judgment about the company's proliferation risk.

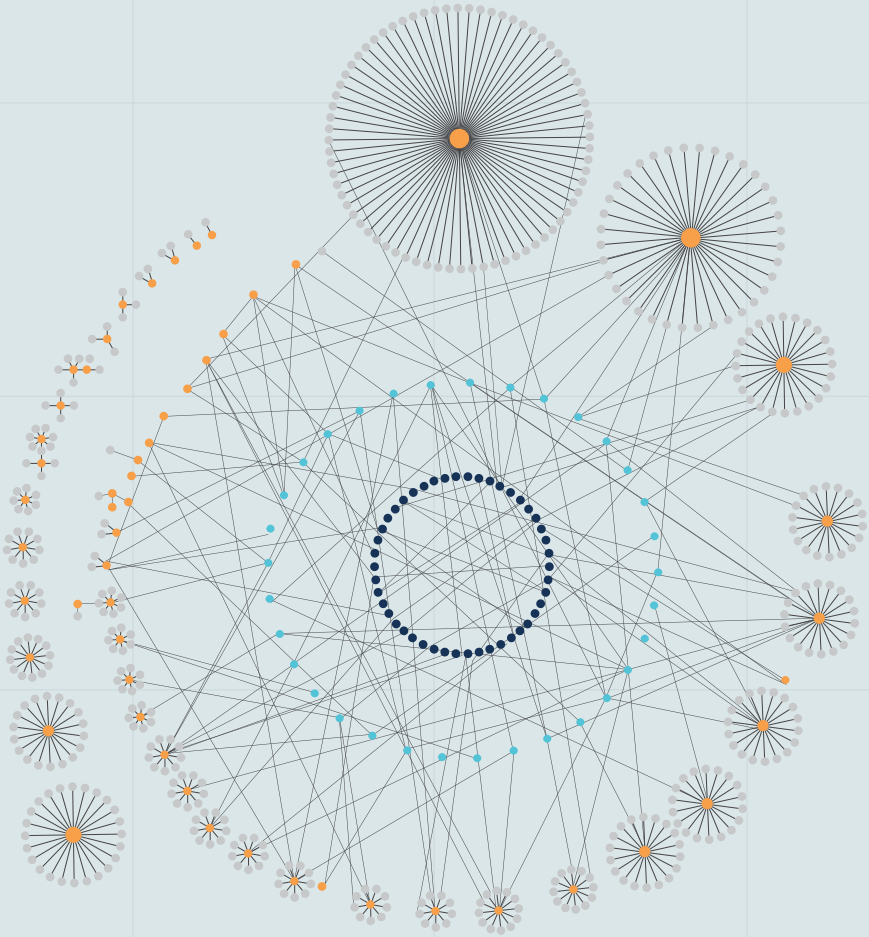
Figure 1: NTI-C4ADS Activities-Based Approach

CASE STUDY: WMD PROCUREMENT IN PAKISTAN

In 2019, C4ADS leveraged publicly available trade records to map Pakistan's nuclear procurement infrastructure and identify previously unknown companies acting to procure sensitive technology for Pakistan's nuclear program. C4ADS published these findings in a 2020 report titled "Trick of the Trade."

C4ADS first compiled a list of all entities in Pakistan that the U.S. Department of Commerce and Japanese Ministry of Economy, Trade, and Industry had identified as procuring on behalf of Pakistan's nuclear program ("known entities"). Using publicly available bill-of-lading-level trade data, C4ADS then identified all overseas companies from which the known Pakistani entities were procuring goods ("zoomed out"). As a final step, C4ADS identified all previously unknown companies in Pakistan that also procured from the same overseas suppliers and used a variety of social network analysis and investigative techniques to assess each company's risk ("zoom back in"). Over the course of research, the U.S. Department of Commerce's Bureau of Industry and Security added to its Entity List at least six companies that C4ADS had identified as particularly high risk, demonstrating that publicly available information can generate insights that align with law enforcement priorities and assessments.

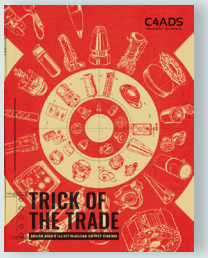
Figure 2: Using Network Analysis Techniques to Identify High-Risk Entities in Trade Data, Featured in the C4ADS Report "Trick of the Trade" (2020)



55
KNOWN ENTITIES IN PAKISTAN

36K
SHIPMENTS TO AND FROM PAKISTAN

3,080
NEW ENTITIES OF CONCERN IN PAKISTAN



6 HIGH-RISK ENTITIES ACTIONED BY THE U.S. GOVERNMENT

Data management and analysis challenges

Although an activity-based approach with PAI proved successful, it was very labor intensive. Analysts needed to possess subject matter expertise in both export controls and investigative techniques in order to discern meaningful leads that warranted scrutiny and action. Significant engineering time and effort were needed to manage trade data volume and inconsistent data formats. These and other issues preclude this useful but manual approach from being used at the scale necessary to support persistent monitoring.

Machine learning and automation amplify the value of publicly available information

To address these issues, NTI and C4ADS experimented with a variety of machine learning and automation techniques. Building on advances in computing power and data availability, machine learning is revolutionizing image recognition, language translation, and many other areas relevant to non-proliferation.[17] Machine learning algorithms are able to identify patterns (e.g., high-risk trading patterns) more capably and quickly than trained human analysts.

The pilot project first focused on automating data pre-processing to shorten the time between data collection and exploitation. It also applied machine learning techniques to improve preliminary data screening to more effectively surface shipments of concern than a human could through more manual approaches. To do so, analysts tested a series of predictive models and evaluated their relative success based on expert consultation and comparisons with conventional methods.

Simple models save time

Preparing trade data for analysis requires significant data pre-processing because of variance in the way that key data fields, such as company names, often are recorded. For list-based screening to be effective, engineers must first standardize company names. Whereas some forms of variance are easy to manage, others are not, and they may cause typical due diligence screening approaches to miss shipments of interest.

To address this issue, NTI and C4ADS experimented with a variety of machine learning methods that would incorporate other features in trade data (e.g., company address, shipment composition, or trading partners) to improve screening for shipments by known entities of concern. One model, a random forest trained through positive-unlabeled learning, performed particularly well, yielding as much as two times the number of recalled shipments. Because this model uses other fields within the dataset beyond just "importer name," it is less likely than a conventional screen to miss known entities as the result of inconsistencies in spelling or format. This technique requires an up-front investment to prepare and engineer the data but is simple to train and can provide a long-term, sustainable screening method used without supervision after it is implemented. Although the model does not perform well in identifying new companies beyond the known entities on which the model is trained, it dramatically improves outcomes in screening for shipments by known entities of concern and is relatively simple to implement. It may support non-proliferation authorities that screen high-refresh transaction data in which naming conventions may be inconsistent.

Figure 3: Example Variance in a Hypothetical Company Name as Represented in Trade Records

DATA VARIANCE COMPLICATES DATA CLEANING REQUIREMENTS

Big Technology Electronics Co., Ltd.	Correct Company Name	Already standardized
Big Technology Electronics Company	Incorrect Company Name	Easy to standardize with basic data cleaning processes
Big Tech Electric	Incorrect Company Name	Difficult to standardize with basic data cleaning processes

Figure 4: NTI-C4ADS Entity Resolution Model & Key Results

NTI & C4ADS MACHINE LEARNING MODEL TO IMPROVE LIST-BASED SCREENING

1. Feed sanctioned or "high risk" entity names into model.

2. Output transactions by sanctioned or "high risk" entities that were previously untagged due to name discrepancies.

12K
additional shipments captured

2X
increase in flagged shipments by known entities

200
analyst hours saved processing data

CASE STUDY: ACCELERATING ANALYSIS AND DISSEMINATION WITH AUTOMATED DATA PREPARATION

Relatively simple machine learning models that automate data pre-processing dramatically reduce the time required to disseminate finished analysis after collecting trade data.

In one case, the model allowed analysts in the NTI-C4ADS pilot project to evaluate a new lead and disseminate an analytic product within two days of a high-risk shipment's landing at port in a country of interest. NTI and C4ADS applied their machine learning model to flag high-risk shipments by "known entities": all companies listed by the U.S. Department of Commerce or Japan's Ministry of Economy, Trade, and Industry for violating export control agreements. After C4ADS received a new tranche of trade data refreshed on a weekly basis, the model flagged 46 shipments for analyst review from 6,179 new shipments. Because this model uses multiple features beyond just "importer name," it is less likely than a conventional screening to miss known entities because of inconsistencies in spelling or format. After approximately one hour of manual review, analysts determined that the model had flagged multiple shipments involving known entities the manual screening process had missed. Analysts also determined that two flagged shipments were false positives. They then used the manually vetted output data to improve the model's performance with future inputs.

C4ADS analysts then investigated remaining shipments flagged by the model. In one instance, analysts determined that one vessel carried several flagged shipments, and then conducted a deep-dive investigation on all other shipments that the vessel had carried. Trade records indicated that the vessel had transported shipments to several entities identified as at high risk of supplying dual-use materials to the country of interest's WMD program. Analysts produced and disseminated an analytical product within two days of the flagged shipments' arriving in the country of interest.

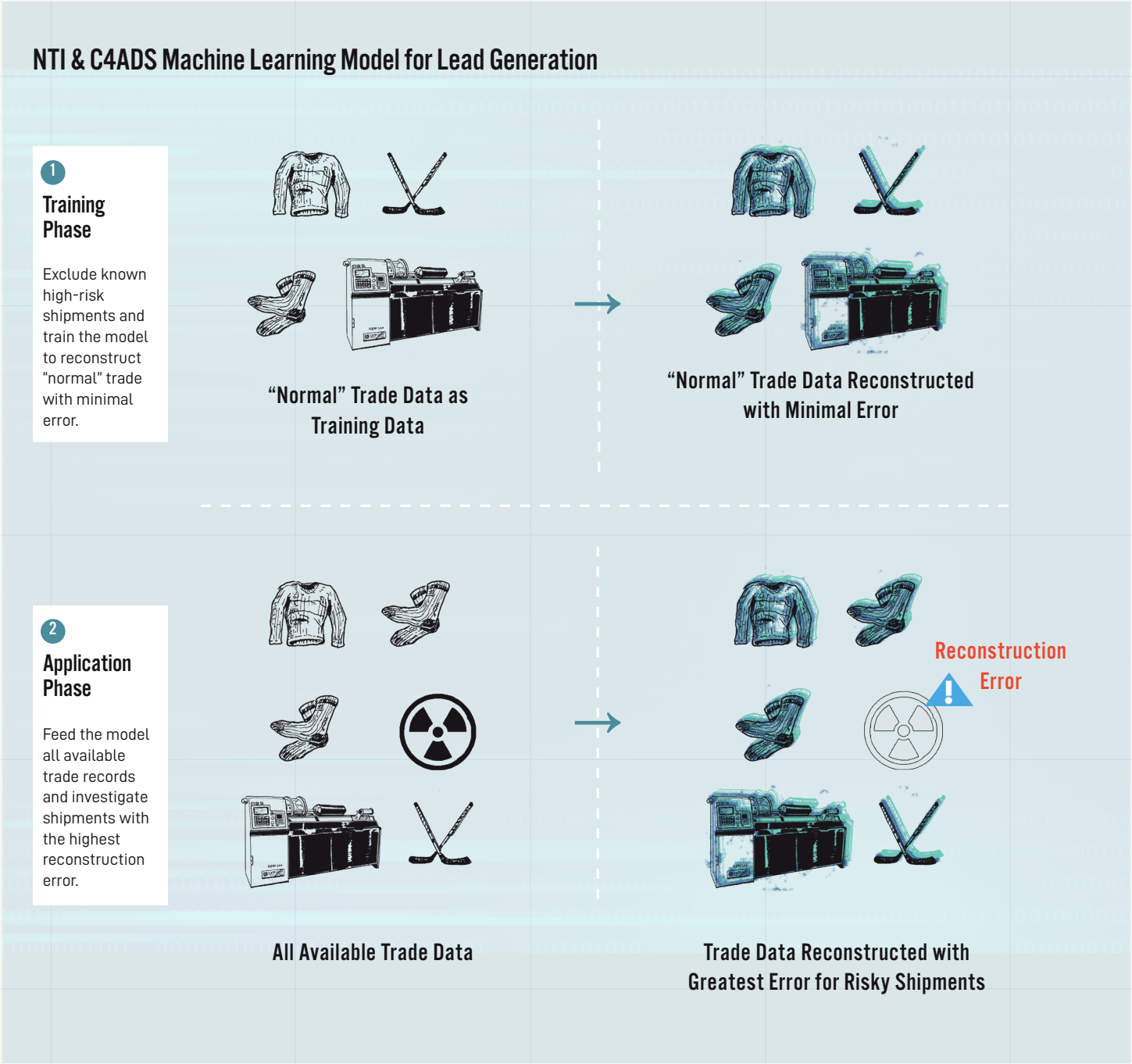
By using publicly available information and the right machine learning tools, non-proliferation efforts may be able to benefit from not only actionable insights but also the freedom of action to disseminate insights with relevant partners in a timely fashion. Future work may experiment with ways to incorporate other related datasets, such as vessel ownership records and voyage position data, to develop new, more sophisticated approaches for timely proliferation detection.

Complex models support discovery

More complex models showed promise in quickly surfacing high-risk shipments from new, bulk datasets. In particular, one unsupervised deep learning model, an autoencoder, showed promise in generating leads for subject matter expert review. An autoencoder is an unsupervised deep learning model that learns how to efficiently compress, encode, and reconstruct input data.[18] Autoencoders are often used for anomaly detection to support credit card fraud prevention and early disease diagnosis in situations where compliance officers or doctors need to discern subtle clues from within high-volume, noisy input datasets (i.e., financial transactions and medical images).

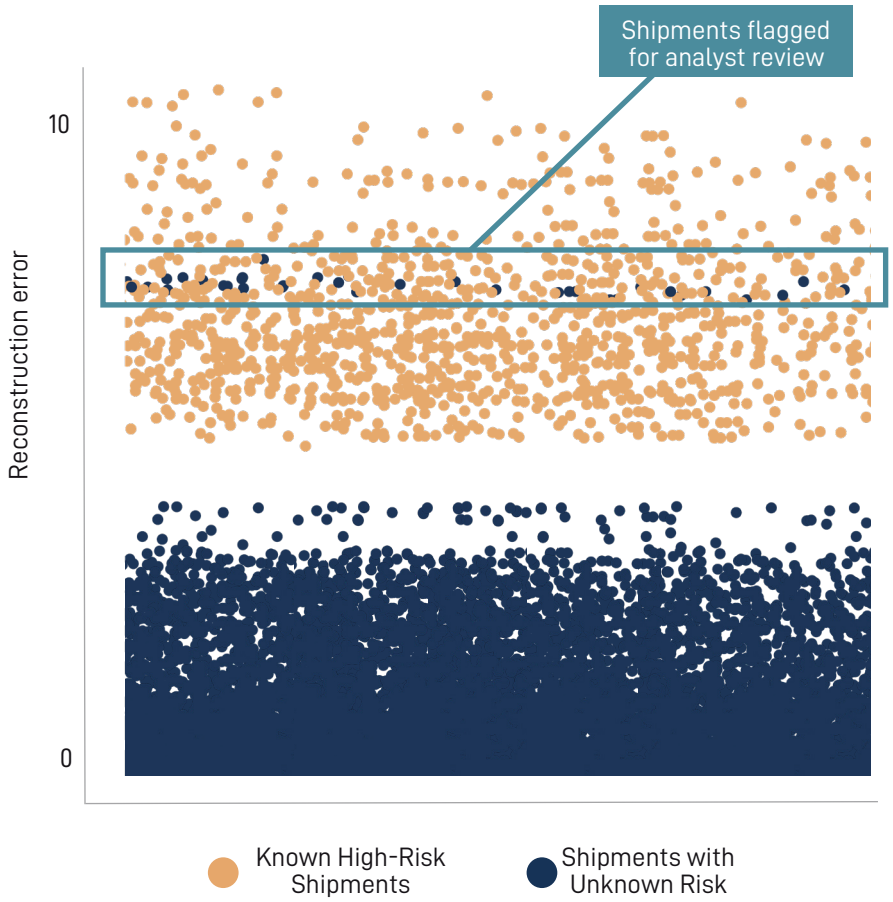
To apply this method for proliferation detection, NTI and C4ADS trained a model with proliferation as the anomaly that the model should detect. Engineers trained the model on all shipments except those by companies with reported associations to a country's WMD program. As a result, when the model runs over proliferation-related shipments, it flags them as anomalies because it is unfamiliar with such data. Subject matter experts can then review and contextualize those flagged shipments, and then make judgments about whether or not they present a meaningful risk.

Figure 5: How an Autoencoder Identifies Risky Shipments through Anomaly Detection



Over the course of this project, the U.S. Department of Commerce added to its Entity List at least four previously unknown companies that the model had elevated for analyst review, demonstrating that the model can produce findings of enforcement interest. Future work will assess how well the model performs in an operational environment, where shipment data are regularly refreshed. NTI and C4ADS expect that the tool will result in both dramatic time savings and improved analytic outcomes. While this model requires significantly more engineering expertise for use than simpler models, early outputs suggest that this unsupervised machine learning approach can accelerate lead generation for analysts to conduct more rapid investigations and intervention.

Figure 6: Illustrative Example of Outputs from NTI-C4ADS Autoencoder for Lead Generation. Analysts Scrutinize Shipments of Unknown Risk with a Reconstruction Error Similar to Known High-Risk Shipments.



4.3M

rows of trade date analyzed

50

new leads identified for analyst review

4

companies identified by autoencoder actioned by U.S. government over course of research

CASE STUDY: IMPROVING LEAD GENERATION THROUGH ANOMALY DETECTION IN TRADE DATA

The autoencoder model supported analysts in the NTI-C4ADS pilot project in identifying new leads for investigation that would not have appeared in simpler approaches.

For example, the model highlighted companies that would not have been detected in manual analysis but whose suppliers were subsequently implicated in law enforcement action for proliferation activities over the course of research. In one case, the model identified a company located approximately 30 minutes from a research lab and enrichment facility with known associations to a country's nuclear weapons program. Additional investigation found no available evidence about the company's operations online or in corporate registries—a characteristic similar of other companies implicated in the country's illicit nuclear program.

Trade records showed that the company had received items including lathe machines, chemical scrubbers, ball bearings, heat exchangers, aluminum and alloy products, valves, and measurement devices such as spectrophotometers. Although product descriptions in trade data do not provide sufficient detail to determine whether these goods are export controlled, some specific types of lathe machines, heat exchangers, bearings, valves, and certain measuring devices are indeed export controlled nuclear dual-use goods under the Nuclear Suppliers Group Part II Guidelines and the U.S. Export Administration Regulations.

This company was not detected in manual analysis because it had no recorded trade with companies previously identified in association with WMD programs. However, over the course of this project, the U.S. Department of Commerce took action against two of its foreign suppliers for involvement in illicit proliferation activities. By applying similar approaches to publicly available trade data, non-proliferation authorities may similarly be able to identify actionable leads for enhanced scrutiny with greater timeliness.

NTI and C4ADS realized benefits from machine learning by applying tools to data management and pre-processing challenges to help analysts manage higher-volume, more complex data streams rather than replacing analysts in the analytic process.[19] Machine learning outputs require review by subject matter experts, who draw meaningful conclusions and properly contextualize them. While some methods do require a degree of data science expertise, technology providers are increasingly developing products that make more complex data management and manipulation accessible to users who may not have advanced computational skills. As a result, regional or subject matter experts will increasingly be able to undertake operations that have traditionally been the responsibility of computer scientists and data engineers.

RECOMMENDATIONS

The quantity and variety of PAI will continue to increase as global commercial systems and national governments continue to digitize information and make it available. This trend presents an opportunity to strengthen global efforts to detect and prevent nuclear proliferation. Indeed, in the future, it may become impossible for potential proliferators to have confidence that their activities are hidden from view.

The data environment is constantly changing, and organizations will be best prepared to realize benefits from PAI insofar as they are able to adapt their approaches for collection and analysis to these changes. Governments are likely to continue to collect accurate trade records in order to support policy planning and economic development, but they may also attempt to restrict access, publish in formats that are more difficult to collect in bulk, or change the details that each published record contains. Non-proliferation organizations should be prepared to adapt to changes in data access, format, and/or scope, for example, by integrating trade data with other data sources such as corporate registry filings, satellite imagery, or vessel position data, for which data may be harder to manipulate. Future NTI-C4ADS work will seek to understand how potential countermeasures might affect the value of these approaches and will explore ways to mitigate these risks.

Recommendations for governments and others working to address nuclear proliferation include the following:

1. INTEGRATE PAI INTO EXISTING MONITORING, VERIFICATION, AND EXPORT CONTROL REGIMES, AND INCORPORATE IT INTO EFFORTS UNDERWAY AROUND THE WORLD TO PREVENT THE SPREAD OF NUCLEAR WEAPONS AND RELATED MATERIALS AND TECHNOLOGIES.

Publicly available data should be integrated with other forms of information into existing monitoring, verification, and export control efforts. PAI—and specifically entity-level trade data used for this work—provides tractable information that governments, banks, and civil society can use to meet non-proliferation goals, and it is more available than ever. Currently, however, data are rarely consistent in form, coverage, reliability, and cost across jurisdictions. Organizations and analysts applying data-driven methodologies must be aware of gaps in data coverage but also understand that, even in the absence of full coverage, data-driven analysis can produce actionable insights. Technological tools are required to exploit data at scale and integrate with other forms of PAI that can help cover data gaps. Similarly, data procurement systems must be flexible enough to allow for flexible data procurement in the ever-changing data environment.
Recognizing that organizations charged with monitoring for potential proliferation have differing policy constraints and differing technical capacities, NTI and C4ADS will seek to work closely with relevant government and international agencies to explore opportunities to operationalize these capabilities in support of safeguards, export control, and monitoring and verification missions.

2. ENABLE THE USE OF PAI AT SCALE BY USING MODERN ANALYSIS TOOLS, INCLUDING MACHINE LEARNING.

Modern analysis tools and automation can enable the use of PAI at scale to help organizations use the information to detect and prevent proliferation. As organizations adopt new tools, analysts will continue to play a key role in the analytic process. Machine learning approaches can identify patterns and correlation from big datasets, whereas humans can perform the abstraction required to appropriately contextualize machine outputs. As the NTI-C4ADS pilot program's work shows, machine learning may enable the identification of new signatures of nuclear proliferation by helping subject matter experts test conceptual models against more complex data. To do so, organizations should consider what combinations of skill sets are required to adopt more data-driven approaches to non-proliferation (e.g., data science, subject matter expertise, regional expertise), and evaluate tools based on how well they empower subject matter experts to work more quickly and effectively.

3. BUILD PARTNERSHIPS TO ALLOW ANALYSTS ACCESS TO COMPLEMENTARY DATA AND CAPABILITIES BECAUSE INTEGRATION OF DIVERSE DATA SOURCES AND ANALYTICAL EXPERTISE IS ESSENTIAL TO DEVELOP GREATER UNDERSTANDING OF ILLICIT ACTIVITIES.

Partnerships and collaboration among analysts are vital. No one organization holds all the data or expertise, and partnerships are essential for achieving sustained impact against proliferation networks. Because PAI is free from classification restrictions, users have significant freedom to share outputs from analysis with relevant stakeholders in government, industry, and civil society. Non-proliferation organizations should identify other organizations and private sector entities that might have the data or expertise they need and collaborate for their mutual benefit. Initiatives that promote collaboration through data and platform sharing should continue.

4. EMBRACE THE USE OF ENTITY-LEVEL TRADE DATA AS WELL AS OTHER, DIVERSE SETS OF PAI IN FUTURE INTERNATIONAL NON-PROLIFERATION INITIATIVES.

International non-proliferation regimes should embrace the use of entity-level trade data alongside diverse forms of PAI. Approaches to international non-proliferation monitoring and safeguards were conceived in an analog era—digital records, the ability to store and analyze troves of information, and today's volume of cross-border transactions were not anticipated by the policymakers who designed the non-proliferation, export control, and safeguards regimes.

The non-proliferation policy environment must match the scope and tempo of today's proliferation risks and provide the mandate for incorporating PAI and advanced analytical tools into existing analytical settings. Today, the goal of maintaining global peace and security remains the same, but officials tasked with limiting trade of nuclear or missile technologies must contend with volumes of global trade transactions 20 times higher than when the controls were conceived.[20] The opportunity to use digital data and analysis tools to track potential proliferators, as demonstrated in this initiative, is expanding.

Today, technology enables the effective collection and analysis of vast quantities of PAI. Policymakers must recognize and invest necessary resources to integrate PAI into global non-proliferation efforts. Where possible, existing control regimes should capitalize on the opportunities of PAI and analytical tools to support non-proliferation and to galvanize the resources necessary to modernize. When taken together, the effective use of publicly available data and the use of modern analytic approaches—if integrated into non-proliferation regimes—may greatly reduce the potential for illicit nuclear activities and build a safer world.

Notes

1 Treaty on the Non-Proliferation of Nuclear Weapons, U.N.T.S. 729, Articles I and II, July 1, 1968.

2 Commission on the Intelligence Capacities of the United States Regarding Weapons of Mass Destruction (WMD Commission), Final Report of the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, March 31, 2005, 352, https://fas.org/irp/offdocs/wmd_report.pdf.

3 David E.A. Johnson, Varun Vira, and Thomas Ewing, Constructive Disruption: Exploiting Publicly Available Information to Address Today's Security Challenges (C4ADS, February 2019), http://static1.squarespace.com/static/566ef8b4d8af107232d5358a/t/5cb4e002919c96000155db0d/1555357705746/White+Paper_Constructive+Disruption.pdf.

4 Melissa Hanham, Grace Liu, Joseph Rodgers, Mackenzie Best, Scott Milne, and Octave Lepinard, Monitoring Uranium Mining and Milling in China and North Korea through Remote Sensing Imagery, CNS Occasional Paper 40 (James Martin Center for Nonproliferation Studies, October 2018), www.nonproliferation.org/wp-content/uploads/2018/10/op40-monitoring-uranium-mining-and-milling-in-china-and-north-korea-through-remote-sensing-imagery.pdf; and Jeffrey Lewis and Dave Schmerler, Identifying DPRK Machine Plants (James Martin Center for Nonproliferation Studies, January 17, 2019), www.nonproliferation.org/identifying-dprk-machine-plants/.

5 "UDMH Production in North Korea: Additional Facilities Likely," 38 North, October 25, 2017, www.38north.org/2017/10/udmh102517/.

6 "Research Opens a Window into Pakistan's Nuclear Weapons Programme," King's College London News Centre, November 4, 2018, www.kcl.ac.uk/news/spotlight-article?id=0f348a7e-04fc-43b9-a7fa-49a7ffb41a1b.

7 "About Missile Threat," Center for Strategic and International Studies, accessed November 19, 2020, missilethreat.csis.org/about/.

8 Jane Perlez, "U.S. Group Says Pakistan Is Building New Reactor," The New York Times, June 23, 2007, www.nytimes.com/2007/06/23/world/asia/23pakistan.html.

9 The James Martin Center for Nonproliferation Studies (CNS) is a leading organization in using satellite imagery, social media, and other publicly available data sources to produce timely, actionable analysis on nuclear activities. For example, in April 2020, the United States Geospatial Intelligence Foundation awarded CNS the 2020 Academic Achievement Award for identifying North Korean preparation for a potential satellite launch hours after negotiations collapsed between the United States and North Korea. According to a press release, "Using new technological opportunities offered by high-cadence moderate resolution satellite imagery and flexible high-resolution satellite image tasking provided by Planet Labs, analysts at CNS, through the use of open-source geospatial intelligence, detected and correctly identified preparations for the engine test 39 hours before it occurred, in violation of international nonproliferation commitments." For more information, see "CNS Receives USGIF 2020 Award," James Martin Center for Nonproliferation Studies, April 27, 2020, nonproliferation.org/cns-receives-usgif-2020-award/.

10 To see examples of recent efforts, see International Atomic Energy Association, Department of Safeguards, Development and Implementation Support Programme for Nuclear Verification 2020–2021, January 2020, www.iaea.org/sites/default/files/20/01/d-and-s-programme-2020.pdf.

11 Herman Nackaerts, "Statement at Symposium on International Safeguards: Preparing for Future Verification Challenges," International Atomic Energy Agency, November 1, 2010, www.iaea.org/newscenter/statements/statement-symposium-international-safeguards-preparing-future-verification-challenges-0.

12 For example, in 2005, the WMD Commission noted that "analysts who use open source information can be more effective than those who [do not]," WMD Commission, Final Report-Recommendations. However, as recently as September 30, 2020, the House Permanent Select Committee on Intelligence recommended in its review of the intelligence community's China-oriented program that the intelligence community should "more effectively integrate publicly available information." U.S. Congress, House of Representatives, Permanent Select Committee on Intelligence, The China Deep Dive: A Report on the Intelligence Community's Capabilities and Competencies with Respect to the People's Republic of China, , September 30, 2020, https://intelligence.house.gov/uploadedfiles/hpsci_china_deep_dive_redacted_summary_9.29.20.pdf.

13 Cristina Versino, Dual-Use Trade Figures and How They Combine, JRC Scientific and Policy Reports (European Commission Joint Research Centre, 2015), ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/dual-use-trade-figures-and-how-they-combine; and Stephanie Lieggi, Catherine Dill, and Diane Lee, The Growing Nonproliferation Challenges in Southeast Asia—Forecasting Emerging Capabilities and Its Implications on the Control of Sensitive WMD-Related Technologies, CCC PASCC Report (Calhoun Naval Postgraduate School, April 2016), calhoun.nps.edu/handle/10945/48708.

14 Andrea Berger, "From Paper to Practice: The Significance of New UN Sanctions on North Korea," Arms Control Today 46 (May 2016), www.armscontrol.org/act/2016-04/features/paper-practice-significance-new-un-sanctions-north-korea; Stephan Blancke, Examining Allegations that Pakistan Diverted Chinese-Origin Goods to the DPRK, Proliferation Case Study Series (King's College London Project Alpha Centre for Science and Security Studies August 2, 2016), www.kcl.ac.uk/alpha/assets/20160803-dprk-pak-allegation-case-study-project-alpha.pdf; and Matthew Godsey and Valerie Lincy, Tracking Proliferation through Trade Data (Wisconsin Project on Nuclear Arms Control, January 2017), fas.org/wp-content/uploads/media/Tracking-Proliferation-through-Trade-Data.pdf.

15 C4ADS has published this type of analysis. See, for example, Jack Margolin and Irina Bukharin, Trick of the Trade: South Asia's Illicit Nuclear Supply Chain (C4ADS, May 1, 2020), www.c4reports.org/trick-of-the-trade.

16 Intelligence gaps or policy considerations may limit the amount of information that governments would publish about entities involved in proliferation in a given country.

17 For one example of machine learning techniques applied to image and video for nonproliferation, see Jamie Withorne, Machine Learning Applications in Nonproliferation: Assessing Algorithmic Tools for Strengthening Strategic Trade Controls (James Martin Center for Nonproliferation Studies, August 2020), nonproliferation.org/machine-learning-applications-in-nonproliferation-assessing-algorithmic-tools-for-strengthening-strategic-trade-controls/.

18 For more on autoencoders, see Jian Zhou, "Deep Learning Primer," Princeton University, accessed November 19, 2020, www.princeton.edu/~jzthree/files/deeplearning.pdf.

19 D. E. A. Johnson and N. Howard, "Network Intelligence: An Emerging Discipline" (paper presented at he 2012 European Intelligence and Security Informatics Conference, Odense, August 22–24, 2012), 287–288, doi: 10.1109/EISIC.2012.52.

20 "Merchandise Imports by Product Group—Annual," World Trade Organization Data, accessed November 19, 2019, data.wto.org/?idSavedQuery=4a3b5609-5b6f-4e8d-9397-aba768ced066; Kelsey Hartigan, Corey Hinderstein, Andrew Newman, and Sharon Squassoni, A New Approach to the Nuclear Fuel Cycle: Best Practices for Security, Nonproliferation, and Sustainable Nuclear Energy (Lanham, MD: Rowman & Littlefield, February 2015), media.nti.org/pdfs/150320_Squassoni_NuclearFuelCycle_Web_final.pdf.; and U.S. Library of Congress, Congressional Research Service, The U.S. Export Control System and the Export Control Reform Initiative, by Ian F. Fergusson and Paul K. Kerr, R41916, Version 52 (January 28, 2020), fas.org/sgp/crs/natsec/R41916.pdf.

About the Authors

JASON ARTERBURN,

PROGRAM DIRECTOR, C4ADS

Jason is Program Director for the Counterproliferation Cell at C4ADS, where he leads projects on China, North Korea, Iran, Russia, and Pakistan. Jason earned a bachelor's degree in economics and interdisciplinary security studies from the University of Alabama, where he was awarded the Harry S. Truman and David L. Boren Scholarships, and a master's degree in China studies from Peking University, where he was a Yenching Scholar. Prior to C4ADS, Jason studied at Tsinghua University as a Blakemore Freeman Fellow in the Inter-University Program for Chinese Language Studies. He speaks Mandarin.

ERIN D. DUMBACHER,

SENIOR PROGRAM OFFICER, NTI

Erin focuses on emerging technology, cyber, and nuclear security at NTI. She earned a master's degree in conflict management and international economics from the Johns Hopkins University School of Advanced International Studies (SAIS), a bachelor's in international affairs from the George Washington University, and was a U.S. Fulbright scholar to Estonia. Prior to NTI, Erin was a director at CEB (now Gartner) leading management and technology research and held research and strategy positions at Atlantic Media. She speaks German.

PAGE O. STOUTLAND, PHD, VICE PRESIDENT FOR SCIENTIFIC & TECHNICAL AFFAIRS, NTI

Page leads NTI's scientific and technically related projects designed to strengthen nuclear security and reduce risks around the world. Stoutland earned a bachelor's degree from St. Olaf College, and a doctorate in chemistry from the University of California, Berkeley. After completing his doctorate, he spent two years at Stanford University as a National Institutes of Health postdoctoral fellow. Prior to NTI, Page spent 10 years at Lawrence Livermore National Laboratory, where he held a number of senior positions. He previously held positions within the U.S. Department of Energy and at Los Alamos National Laboratory.


PART NO.