



December 2011

# Use of Data Mining for Validation and Verification of Maritime Cargo Movement

## Project Leads

Noel P. Greis, PhD, MSE, MA, University of North Carolina at Chapel Hill

Monica L. Nogueira, PhD, MS, University of North Carolina at Chapel Hill

---

## Statement of Problem

In 2008, approximately 11 million maritime containers passed through U.S. seaports (U.S. Department of Homeland Security [DHS], 2010). These containers carried 90 percent of international commerce moving in and out of the United States, representing \$500 billion in U.S. trade from 178,000 foreign businesses (Montaigne, 2004). Maritime trade forms the economic circulatory system of our national commerce—but also represents a potential avenue for the transport of illegal weapons of mass destruction or their parts, biological agents, or other dangerous items capable of destroying U.S. assets or killing U.S. citizens.

In recognition of this threat, the U.S. government has recently implemented a number of laws and policies to enhance maritime container security. In particular, on August 3, 2007, the 9/11 Commission Recommendation Act was signed into law, mandating that by 2012, all maritime cargo containers entering U.S. ports would be screened.<sup>1</sup> However, cost, high cargo volumes, and technical issues make it impractical (if not impossible) to physically inspect 100%

---

<sup>1</sup> The Secure Freight Initiative was launched in 2006 to test the feasibility of 100% screening of all containers bound for U.S. ports.

of all entering cargo. Furthermore, disruption of the flow of goods into the country could have negative and lasting effects on an already weakened economy.

As an alternative to physical cargo screening, regulations and policies based on the establishment of a “virtual” sea border for cargo screening and risk-based threat assessment have been implemented. These policies, which are designed to identify high-risk shipments before they reach U.S. ports, require new tools for acquiring, analyzing, and interpreting the large quantities of data that are generated by these new cargo validation and verification processes during in-port inspections. In addition, the adoption of new information and communication technologies such as wireless sensor networks, radio frequency identification (RFID), and new electronic “smart” tags are enabling container security monitoring while in transit. These new information and communication technologies are generating extremely large quantities of transactional and other shipment-related data in a variety of different formats. The broad category of such tools and the processes by which knowledge is extracted from these data is referred to as “knowledge discovery”—a term that encompasses a wide range of data mining techniques that enable cargo validation and verification, whether in port or in transit.

This research brief will (1) provide a background on new regulations that are driving the need for better data mining processes and tools, (2) describe the cargo screening and supplier validation process to illustrate the potential application of data mining, and (3) summarize current developments and research challenges in data mining for cargo security.

---

## Background

The events of 9/11 marked a watershed in the history of U.S. national security, spurring a number of new protocols and programs for validation and verification of cargo, both in the United States and in foreign countries.<sup>2</sup> Two key programs were launched by U.S. Customs and Border Protection (CBP) in the wake of 9/11—the Customs-Trade Partnership Against Terrorism (C-TPAT) program and the Container Security Initiative (CSI).<sup>3</sup> C-TPAT is an incentive program designed to build cooperative relationships with the private sector that strengthen overall supply chain and border security. The practical goal of C-TPAT is to reduce the government’s burden of inspection at U.S. ports by providing incentives to the private sector to assume many security functions for their cargo as it moves through their supply chains. Under CSI, businesses ensure the integrity of their own security practices and communicate and verify the security guidelines of their supply chain partners.

---

<sup>2</sup> An excellent summary of selected non-U.S. initiatives is provided in Peterson and Treat (2008).

<sup>3</sup> In addition to U.S. government programs, the 29 countries of the World Customs Organization adopted the SAFE Framework to Secure and Facilitate Global Trade in June 2005. SAFE complies with best practices for customs administration security and trade facilitation (Boske, 2006).

To participate in C-TPAT, companies must conduct a comprehensive self-assessment of their supply chain security procedures using the C-TPAT security criteria or guidelines jointly developed by CBP and the trade community for their specific enrollment category. In return for participating in C-TPAT, members of C-TPAT are less subject to customs inspections and are routed to the front of customs lines when they do need to be inspected. They can also take advantage of C-TPAT training for themselves and their employees to tighten the security of their supply chain. If routine inspection reveals noncompliance, C-TPAT membership is revoked and the company must be recertified. C-TPAT is a voluntary program; more than 7,000 companies have joined it since it was launched in November 2001.<sup>4</sup>

CSI creates a virtual border for cargo screening that extends to the port of embarkation. Launched in 2002, CSI is a multinational program based on principles of risk management. Its goal is to reduce screening costs and congestion at U.S. ports (Allen, 2006; Haveman et al., 2007). CSI has four main objectives: (1) identify high-risk containers using the Automated Targeting System (ATS), (2) prescreen and evaluate containers before they leave the port of embarkation, (3) ensure that the screening can be done rapidly without slowing the flow of trade, and (4) use smarter and more secure containers that allow for the detection of tampering that occurred during transit. According to DHS's Office of the Inspector General (2010), in 2009 CSI was operational in 53 seaports, accounting for 86 percent of cargo that passed through U.S. ports.

---

## The Cargo Security Process

The challenge of maintaining cargo security is that containers are sometimes temporarily outside the physical control of their owners and are handled by second parties such as freight forwarders and carriers. Under CSI, cargo containers are screened at the port of origin prior to departure or, for high-risk cargo not departing from a CSI port, at the port of destination upon arrival. The process starts at the virtual border or port of departure. Within 24 hours of shipment, shipping companies are required to provide the manifest to CBP, which then transmits the data to the National Targeting Center, where the ATS cross-references the data in the manifest with other databases that may have additional information relevant to the shipment. ATS is a complex mathematical model that uses weighted rules to assign a risk score to cargo containers. ATS scores are computed for each individual container and are used to determine whether physical screening is required. The risk score is a probabilistic representation of the threat posed by the container; it is based on information from the manifest plus other relevant information. Supplemental information could include strategic intelligence or other related shipment anomalies (U.S. Government Accountability Office, 2008; Wasem et al., 2004).

---

<sup>4</sup> Critics of the C-TPAT program have voiced two concerns: (1) benefits to participating companies have not been clearly stated and (2) validation of the submitted security profiles and audit for compliance have been lacking.

High-risk and/or suspicious shipments are inspected at the port of departure using non-intrusive inspection (NII) technologies (e.g., large-scale X-ray and gamma ray machines as well as radiation detection machines). Readings from these inspections are compared with a threshold level to assess threat. Depending on the result, the container may be physically inspected or released. Decisions about which sensor technologies to apply and what thresholds to use are subject to the discretion of the operator. This problem is referred to as the “container inspection problem” and has been addressed by a number of researchers (Boros et al., 2008; Concho & Ramirez-Marquez, 2010; Stroud & Saeger, 2003). This stream of research addresses how to make decisions about which containers require inspection by optimally sequencing the sensors and assigning thresholds for interpreting sensor readings. These models balance sensor reliability and inspection cost by minimizing inspection errors (both false positives and false negatives). Developing an optimal inspection strategy once the containers arrive in port and have been selected for inspection has been another topic of research (Boros et al., 2009; Van Weele & Ramirez-Marquez, 2010).

Cargo screening at the port is enabled by tags with auto-identification (auto-ID) and RFID or other optical characteristics technologies that store and provide information about the current status of the container—for example, by remotely sending the information to fixed readers as it is on-loaded or off-loaded from a vessel. Information might include container number, seal number, manifest number, and trip status. The system might also contain information about events—for example, whether the container had been tampered with en route. It would be extremely desirable for containers to be smart enough to sense when an intrusion takes place in transit and to generate and send alerts to officials so that mitigations can be taken *prior* to the shipment arriving at port—thus eliminating potential threats to the port infrastructure and workers.

New remote communication technologies, such as wireless sensor networks (WSNs), along with electronic seals and global positioning technology, are able to send signals automatically while in transit once a seal is broken or the container deviates from its planned route (Rizzo et al., 2011). These smart systems operate as a kind of “neighborhood watch.” Sensors in the network are capable of sensing their current state, processing signals, and transmitting and receiving information to and from a base station (which may be on board the ship) for relay to authorities in port. The containers are wirelessly linked and monitor each other in a distributed fashion, exchanging information about their status to make it difficult for intrusions to be hidden. The wireless sensor network can also monitor environmental conditions such as location, temperature, and pressure and motion (such as vibrations). A variety of WSN systems have been proposed or developed (Bian et al., 2009; Katulski et al., 2009).

As a shipment moves from origin to destination, as described above, data are collected in a variety of different formats, thereby complicating the knowledge discovery task. Shipping data and transportation details may reside in datasets on spreadsheets or other flat formats.

Validating the shipper identity may require accessing data on the Internet. Tracking a shipment en route requires that data be mined in real time as the data stream from the vessel or shipment to a base station. In the next sections we discuss the challenges of data mining associated with these three specific scenarios and formats.

---

## Data Mining in Datasets

Data mining has emerged as a key enabling technology for many cargo security initiatives.<sup>5</sup> Data mining has received renewed attention recently because of the convergence of three important trends. First, with increasingly greater volumes of data being collected in the interest of national security, data mining tools are becoming a necessary part of interpreting the vast amount of data accumulated. Second, the availability and low cost of powerful multiprocessors provide the hardware necessary to manipulate large volumes of data in (near) real time. Finally, powerful new algorithms are being developed that are able to take advantage of the new processing technologies. Appendix A provides a set of the 10 most promising new algorithms, as identified by the IEEE International Conference on Data Mining in December 2006.

Traditional data mining (also known as knowledge discovery in databases, or KDD) refers broadly to the collection of tools by which we are able to discover hidden patterns and relationships in data (Lee & Siau, 2001; Li & Sun, 2005). It includes many of the traditional statistical methods (descriptive statistics, linear regression, multivariate analysis), as well as mathematical algorithms that sift through data to sort them (classification techniques) and various machine learning methods (neural networks or associative memory) that are able to improve their analytical performance through learning and experience.

For the purposes of cargo screening, we recognize a broader definition of data mining that includes the efficient collection, management, and transformation of data into usable information through data fusion as well as the gathering of predictive insights from that data via knowledge discovery to inform action or policy. Traditional statistical methods such as linear regression were largely hands-on technologies that operated on small, static datasets to validate hypotheses or models. However, new data mining tools are capable of not only interpreting extremely large and complex datasets (on the order of thousands of data points or variables) but also extrapolating those relationships as trends and predictions. Furthermore, the need for quick response under threat of terrorist attack has necessitated the development of new algorithms that can be automated for fast, (near) real-time capability. For example, the

---

<sup>5</sup> The use of data mining has received widespread press because of three high-profile projects for tracking people, since disbanded, including the Total Information Awareness (TIA) project, the Computer-Assisted Passenger Prescreening System II (CAPPS II), and the Multistate Anti-Terrorism Information Exchange (MATRIX) (Seifert, 2004). Critics of these large-scale national projects cited privacy and other concerns that drew considerable public backlash against the projects.

cross-industry standard process for data mining (CRISP-DM) data mining process model is an industry-neutral and tool-neutral framework that comprises six steps for addressing a data mining task. Enhanced capabilities such as real-time data capture on the front end and automatic alerting on the back end have been incorporated into the process.

The new generation of data mining tools also enables discovery of new relationships, or hidden patterns, that were previously not known and are often non-intuitive. In a simple dataset with few variables, it can be relatively easy to decipher patterns and to test for unseen but posited relationships. But with thousands (even millions) of data points, it is impossible for humans to “see into” the data or to know what to look for. New discovery tools such as associative memory and machine learning enable us to discover relationships that we did not know were there (i.e., discover the “unknown unknowns”).

---

## Data Mining on the Web

The Internet, TCP/IP protocols, and World Wide Web have opened up another large source of information for cargo security for which data mining tools need to be developed. The Internet offers a rich source of information that can be used to augment information about suspicious shipments or suppliers and uncover linkages between suspicious groups and a shipment. Mining the Web is challenging not only because of its size—it is the largest publicly accessible data source in the world—but also because the Web includes diverse formats. Information is redundant and heterogeneous (i.e., the same information is provided by many sources in different contexts), noisy (i.e., non-relevant information must be removed), and of varying quality (i.e., not all information can be validated). These characteristics present significant challenges. While traditional data mining has focused largely on structured data stored in relational tables, flat files, or spreadsheets, Web data mining must cope with unstructured text data that may take the form of Web hyperlink structure, page content, or usage data (Liu, 2007).

- **Web structure mining.** Web structure mining discovers useful knowledge from hyperlinks that form the network structure of the Web. Hyperlink information tells us which pages are linked to other pages and, thus, enables us to discover communities of users who share common interests. For example, as more freight forwarders use the Web as a tool for communicating with their shippers and create links to shared datasets, we should be able to mine the Web to discover the transport or supply chain for a suspicious shipment.
- **Web content mining.** Web content mining looks for patterns in qualitative or quantitative data contained in Web page content. Many of the mining tools for Web content are the same as those for traditional data mining. In the area of cargo security, Web content applications might include searching for similar shipments at different sites run by different security agencies. In another application, dynamic topic mining from news stream data could retrieve news articles from the Web and use data mining tools

to produce high-level knowledge that would enable users to quickly locate news articles of critical interest.

- **Web usage mining.** Web usage mining discovers patterns of Web access by different users. For example, Web mining tools can discover patterns in access logs, which contain a record of every time a user clicks on a Web site. With respect to cargo security, it would be possible to discover plans to build and ship a “dirty bomb” by looking for patterns of Web use that may encompass agricultural or pharmaceutical sites for chemicals, transportation sites for information about shipping options, and instructional “how-to’s” on the Web or other known terrorist-affiliated Web sites.

---

## Data Mining of Data Streams

As smart containers, electronic seals, and other technologies are being adopted, new data mining tools are needed that are able to process and mine data that are obtained in (near) real time from a variety of sensor types, such as wide-area sensor infrastructures, remote sensors, and wireless sensor networks. In-transit, event-based monitoring systems present special data mining challenges. Because of the large volume of data, these techniques must be readily automated so they can provide real-time results, and they must be able to process data in constant time and memory (usually by scanning the data just once because it is too time-consuming to re-scan the old data as new data are received).

Traditional model-based data mining approaches typically assume that the data are derived from a stationary process and that a model of normal behavior exists. Stream data mining, on the other hand, must also be able to handle non-stationary processes. When screening in-transit cargo in (near) real time, data mining algorithms generally detect deviations from normal behavior—or other semantic anomalies—that can signal an unexpected (i.e., terrorist) event. However, for some situations, there is no a priori model of normal behavior, so techniques must learn “on the fly” to differentiate normal system behavior from potential terrorist behavior.

Various approaches are being developed for mining of data streams (Gaber et al., 2005). A 2009 special issue of *Intelligent Data Analysis* contained five papers that highlighted the current state of the art in the field. These papers address methodological topics such as novelty detection (Spinosa et al., 2009), context-aware adaptive data stream mining (Haghighia et al., 2009), manifold embedding (Agovic et al., 2009), and spatio-temporal sensor graphs (George et al., 2009). Stream data mining tools applied to real-time, in-transit cargo security detection include manifold embedding (Agovic et al., 2009), unsupervised learning techniques (Raz et al., 2004), and graph theoretic methods (Eberle & Holder, 2007).

---

## Data Mining for Cargo Security: A Crossroads and Future Directions

Data mining remains a fertile ground for academic research and has experienced considerable success in business applications ranging from credit card fraud detection to detection of market trends for new product introductions. In the arena of national security, data mining has been held up as an essential tool but also has been criticized in a recent National Research Council report (2008) as being an unreliable method for identifying potential terrorists. The May 2010 Times Square bomb plot in New York City illustrated the difficulties of using data mining in situations where there are few data or established patterns of terrorist behavior on which to base any future predictions (Vijayan, 2010). These criticisms—and concerns about personal privacy—have placed data mining at a crossroads in certain national security activities.

Data mining works best in situations in which there is sufficient information upon which to observe a pattern or distinguish anomalies—and in which false positives have a lower cost. Data mining has been successful in detecting credit card fraud, for example, because only 1% of the 900 million credit cards transactions are fraudulent each year (Schneier, 2006); in monitoring to identify a terrorist, however, trillions of connections between people and places may need to be monitored. Although more than 11 million containers travel in and out of our ports each year, this number is well below the management and processing threshold for emerging data mining tools.

That said, three key research directions need to be addressed to effectively achieve the potential of data mining for cargo screening.

**Systems Approach to Cargo Screening.** Data mining tools can reveal patterns, but they are insufficient in themselves to reveal the value or significance of observed patterns and determine an optimal mitigation approach or follow-up action. Also, although data mining can reveal patterns, it cannot necessarily determine causality. Furthermore, the associated uncertainty can lead to large numbers of false positives and false negatives. Several approaches have been developed to deal with unavailable, uncertain, or incomplete information (Robinson et al., 2006). For example, human knowledge or reason and appropriate actions can be captured in subjective assessments based on “fuzzy” rules—for example, “if-then” rules that condition an action on an observed relationship in the data. Another approach, called the evidential reasoning (ER) approach, has been proposed for synthesizing disparate pieces of evidence, such as would result from data mining (Tsai, 2006). But the larger challenge is to embed the data mining activities within a larger systems framework where the quality and scope of information, along with economic and technical issues, can be considered. Developing a systems approach represents a current and important direction in data mining research for cargo screening (Siebers et al., 2010).

**Relevance and Interoperability.** The concept of relevance refers to the ability to access, evaluate, and share all relevant information about a shipment during the process of cargo screening. In the previous sections, we have addressed the challenges of data mining in a new information environment where relevant information is available in new formats on the Web and in (near) real time in streams. Flat file manifests and other shipment information must be fused with other relevant information available in these new formats/sources. At the same time, the problems of interoperability continue to present barriers to data mining broadly, but especially for cargo security. Data mining efforts that take advantage of legacy databases may not be compatible with newer database architecture and data formats. New concepts of federated data grids are emerging in which users can access and utilize distributed data and computational resources across platforms and administrative domains without knowledge of data formats or data locations (Rajasekar et al., 2004; Rajasekar & Wan, 2002; Xiao, Chen, & Fu, 2004). This technology allows for ubiquitous and transparent computing of shared services by authorized users. Current data mining tools must be complemented by new tools for identifying and incorporating other relevant information into the cargo screening process.

**Privacy and Institutional Concerns.** Most privacy concerns related to data mining have been directed at programs designed to identify terrorists in the United States by tracking and tracing individuals and their actions—not cargo. The 2008 National Research Council report cited earlier expressed doubts about both current and future projects in DHS that involved data mining for the specific purposes of tracking individuals as opposed to cargo. Cargo information rarely trespasses on personal privacy concerns, but it can contain proprietary information about location and identity of suppliers that companies are not comfortable sharing without assurances of confidentiality. A subsequent 2008 report to Congress (DHS, 2008) presented new privacy principles for agency research projects. Currently, the ATS inbound, outbound, and passenger modules (administered by CBP), the Data Analysis and Research for Trade Transparency System (DARTTS) (administered by Immigration and Customs Enforcement), and the Freight Assessment System (FAS) (administered by the Transportation Security Administration) use data mining, but none of the systems are used for individuals. However, the commercial privacy concerns of a large number of stakeholders (including shippers, consignees, freight forwarders, transport operators, maritime carriers, container terminal operators, and others) must be considered. Too little attention has been paid to the institutional aspects that must be in place before organizations buy into cargo screening. Participation can result from organizational pressures from others in the chain, perceived success or benefits by other adopters, or pressures from standards organizations (Lun et al., 2008). These stakeholders must perceive a mutual benefit before they will move beyond simple compliance with the law and become key partners with government. A variety of institutional arrangements have been proposed, such as freight advisory councils and cargo data collection portals (West, Walton, & Conway, 2008). Critical factors affecting the adoption of new security technologies and services by shippers have been identified (Chao & Lin, 2009; Seifert, 2004).

Computational and organizational advances on these three fronts would help to advance and enhance cargo screening using data mining beyond today's state of practice. In the future, a systems approach, built on real-time visibility and control of the cargo process, would offer the ability to (1) consider all information relevant to the cargo screening process to improve data mining performance and reduce false positives and false negatives; (2) handle large streams of data in (near) real time to improve in-transit monitoring; and (3) analyze data on the fly so responses can be fast enough to avert any terrorist threat.

## Contact Information

### **Noel P. Greis**

Director, Center for Logistics and Digital Strategy

Kenan Institute of Private Enterprise

Kenan-Flagler School of Business

CB# 3440 Kenan Center

University of North Carolina at Chapel Hill

Chapel Hill, NC 27599-3440

919-962-8201

[noel\\_greis@unc.edu](mailto:noel_greis@unc.edu)

### **Monica L. Nogueira**

Director, Intelligent Systems Laboratory

Kenan Institute of Private Enterprise

Kenan-Flagler School of Business

CB# 3440 Kenan Center

University of North Carolina at Chapel Hill

Chapel Hill, NC 27599-3440

919-962-8201

[monica\\_nogueira@unc.edu](mailto:monica_nogueira@unc.edu)

**Noel P. Greis, PhD, MSE, MA**, is director of the Kenan Institute's Center for Logistics and Digital Strategy and professor of Operations, Technology and Innovation Management at the Kenan-Flagler Business School at the University of North Carolina at Chapel Hill (UNC). Dr. Greis is the co-director of the recently established UNC-Tsinghua Center for Logistics and Enterprise Development in Beijing, China, a joint center of Tsinghua University's Department of Industrial Engineering and the Kenan-Flagler Business School. Her research is transforming



the way we predict, evaluate, and respond to complex and critical events in domains such as defense and security, supply chain management and logistics, medicine and public health, food safety, and energy and the environment. Dr. Greis is an expert in intelligent systems design and development and works with organizations to develop knowledge-based systems and predictive analytics that support decision-making in complex, disruptive, and dynamic environments. She is also an expert in the use of intelligent agent-based modeling and simulation to predict the behavior of complex systems and thus improve decision-making capability.

**Monica L. Nogueira, PhD, MS**, is director of the Intelligent Systems Laboratory of the Center for Logistics and Digital Strategy at Kenan-Flagler Business School at UNC. In this capacity, she is responsible for overseeing projects developed by the laboratory for its corporate and institutional clients, and she works closely with UNC professors, students, and Kenan-Flagler staff. Dr. Nogueira is an expert in data modeling and analysis, which she applies to designing and building decision support tools that mine and correlate information from large and diverse datasets to extract the relevant knowledge that enables users to act on those problems that are most significant to them. Her primary research interests include new technologies and their practical uses to create new methodologies that support intelligent tools and their application to everyday problems in logistics and supply chains, data and text mining methodologies, and tools for knowledge discovery and extraction. Dr. Nogueira has developed a number of projects and tools that demonstrate the use of RFID technology for controlling the safety of perishable products (e.g., a cold chain for food or pharmaceutical drugs).

---

## References

Agovic, A., Banaerjee, A., Ganguly, A., & Protopopescu, V. (2009). Anomaly detection using manifold embedding and its applications in transportation corridors. *Intelligent Data Analysis*, 13, 435–455.

Allen, N. H. (2006). Container Security Initiative costs, implications and relevance to developing countries. *Public Administration and Development*, 26, 439–447.

Bian, J., Seker, R., Ramaswamy, S., & Yilmazer, N. (2009, June). Container communities: Anti-tampering wireless sensor network for global cargo security. In *Proceedings of 17<sup>th</sup> Mediterranean Conference on Control and Automation* (pp. 464–468). Washington, DC: IEEE. Retrieved from <http://www.computer.org/portal/web/csdl/doi/10.1109/MED.2009.5164585>

Boros, E., Elsayed, E., Kantor, P., Roberts, F., & Xie, M. (2008). Optimization problems for port-of-entry detection systems. In H. Chen & C.C. Yang (Eds.), *Intelligence and security informatics* (pp. 319–335). Berlin: Springer-Verlag.

Boros, E., Fedzhora, L., Kantor, P. B., Saeger, K., & Stroud, P. (2009). A large-scale linear programming model for finding optimal container inspection strategies. *Naval Research Logistics*, 65, 404–420.

Boske, L. B. (2006). *Port and supply-chain security initiatives in the United States and abroad* (Policy Research Project Report No. 150). Austin: University of Texas at Austin, Lyndon B. Johnson School of Public Affairs. Available from <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA494476>

Chao, S.-L., & Lin, P.-S. (2009). Critical factors affecting the adoption of container security service: The shipper's perspective. *International Journal of Production Economics*, 122, 67–77.

Concho, A. L., & Ramirez-Marquez, J. E. (2010). An evolutionary algorithm for port-of-entry security optimization considering sensor thresholds. *Reliability Engineering and System Safety*, 95, 255–266.

Eberle, W., & Holder, L. (2007). Anomaly detection in data represented as graphs. *Intelligent Data Analysis*, 11, 663–689.

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005, June). Mining data streams: A review. *SIGMOD Record*, 34(2), 18–26.

George, B., Kang, J. M., & Shekhar, S. (2009, August). Spatio-temporal sensor graphs (STSG): A data model for the discovery of spatio-temporal patterns. *Intelligent Data Analysis*, 13, 457–475.

Haghighia, P. D., Zaslavsky, A., Krishnaswamy, S., Gaber, M. M., & Loke, S. (2009, August). Context-aware adaptive data stream mining. *Intelligent Data Analysis*, 13, 423–434.

Haveman, J. D., Shatz, H. J., Jennings, E. M., & Wright, G. C. (2007). The Container Security Initiative and ocean container threats. *Journal of Homeland Security and Emergency Management*, 4(1), 1–19.

Katulski, R. J., Sadowski, J., Stefanski, J., Ambroziak, S. J., & Miszewska, B. (2009). Self-organizing wireless monitoring system for cargo containers. *Polish Maritime Research*, 3(61), 45–50. Retrieved from <http://www.springerlink.com/content/t55826x661751147/>

Lee, J. L., & Siau, K. (2001). A review of data mining techniques. *Industrial Management and Data Systems*, 101(1), 41–46.

Li, Y.-H., & Sun, L.-Y. (2005, October). Study and applications of data mining to the structure risk analysis of customs declaration cargo. In F. C. M. Lau, H. Lei, X. Meng, & M. Wang (Eds.), *Proceedings of the IEEE International Conference on e-Business Engineering* (pp. 761–764). [n.p.]: IEEE. Retrieved from <http://www.computer.org/portal/web/csdl/doi/10.1109/ICEBE.2005.113>

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents and usage data*. New York: Springer-Verlag.

Lun, Y. H., Wong, W. Y., Lai, K.-H., & Cheng, T. C. E. (2008, January). Institutional perspective on the adoption of technology for the security enhancement of container transport. *Transport Review*, 28(1), 2–33.

Montaigne, F. (2004, January). Policing America's ports. *Smithsonian Magazine*. Retrieved from <http://www.smithsonianmag.com/people-places/ports.html>

National Research Council. (2008, October). *Protecting individual privacy in the struggle against terrorists: A framework for program assessment*. Washington, DC: National Academies Press. Retrieved from [http://www.nap.edu/openbook.php?record\\_id=12452](http://www.nap.edu/openbook.php?record_id=12452)

Peterson, J., & Treat, A. (2008). The post 9/11 global framework for cargo security. *Journal of International Commerce and Economics*. Available from [http://www.usitc.gov/journals/journal\\_archive.htm#2008](http://www.usitc.gov/journals/journal_archive.htm#2008)

Rajasekar, A., & Wan, M. (2002). *Components of a virtual data grid architecture*. Paper presented at the Advanced Simulation Technologies Conference (ASTC02), San Diego, CA.

Rajasekar, A., Wan, M., Moore, R., & Schroeder, W. (2004). *Data grid federation*. Paper presented at the PDPTA Special Session on New Trends in Distributed Data Access, Las Vegas, NV.

Raz, O., Bucheit, R., Shaw, M., Koopman, P., & Faloutsos, C. (2004, October). Detecting semantic anomalies in truck weigh-in-motion traffic data using data mining. *Journal of Computing in Civil Engineering*, 18, 291–300. Retrieved from [http://www.cs.cmu.edu/~ornar/wim\\_data.pdf](http://www.cs.cmu.edu/~ornar/wim_data.pdf)

Rizzo, F., Barboni, M., Faggion, L., Azzalin, G., & Sironi, M. (2011, May). Improved security for commercial container transports using an innovative active RFID system. *Journal of Network and Computer Applications*, 34(3).

Robinson, S. M., Smith, L. E., Jarman, K. D., Runkle, R. C., Ashbaker, E. D., Jordan, D. V., et al. (2006). A simulation framework for evaluating detector performance in cargo screening applications. In B. Philips (Ed.), *IEEE Nuclear Science Symposium Conference Record, 2006* (pp. 307–313). Piscataway, NJ: IEEE. Retrieved from <http://ieeexplore.ieee.org/Xplore/login.jsp?url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F4143535%2F4178919%2F04179002.pdf%3Farnumber%3D4179002&authDecision=-203>

Schneier, B. (2006, March). Data mining for terrorists [Web log post]. Retrieved from [http://www.schneier.com/blog/archives/2006/03/data\\_mining\\_for.html](http://www.schneier.com/blog/archives/2006/03/data_mining_for.html)

Seifert, J. W. (2004). Data mining and the search for security: Challenges for connecting the dots and dashes. *Government Information Quarterly*, 21, 461–480.

Siebers, P.-O., Sherman, G., & Aickelin, U. (2010). Development of a cargo screening process simulator: A first approach. In *Proceedings of the 6th International Mediterranean Modeling Multiconference (EMSS 2009)* (pp. 200–209). Retrieved from <http://ima.ac.uk/papers/siebers2009c.pdf>

Spinosa, E. J., de Carvalho, A. P. L. F., & Gama, J. (2009). Novelty detection with application to data streams. *Intelligent Data Analysis*, 13, 405–422.

Stroud, P. D., & Saeger, K. J. (2003). Enumeration of increasing Boolean expressions and alternative digraph implementations for diagnostic applications. *Proceedings Volume IV, Computer, Communication and Control Technologies*, 328–333.

Tsai, M.-C. (2006). Constructing a logistics tracking system for preventing smuggling risk of transit containers. *Transportation Research Part A*, 40, 526–536.

U.S. Department of Homeland Security (2008, December). *2008 report to Congress. Data mining: Technology and policy*. Washington, DC: Author. Retrieved from [http://www.dhs.gov/xlibrary/assets/privacy/privacy\\_rpt\\_datamining\\_200812.pdf](http://www.dhs.gov/xlibrary/assets/privacy/privacy_rpt_datamining_200812.pdf)

U.S. Department of Homeland Security, Office of Inspector General. (2010, February). *CBP's Container Security Initiative has proactive management and oversight but future direction is uncertain* (Report OIG-10-52). Retrieved from [http://www.oig.dhs.gov/assets/Mgmt/OIG\\_10-52\\_Feb10.pdf](http://www.oig.dhs.gov/assets/Mgmt/OIG_10-52_Feb10.pdf)

U.S. Government Accountability Office. (2008, August). *Supply chain security: CBP works with international entities to promote global customs security standards and initiatives, but challenges remain* (Report to Congressional Requesters No. GAO-08-538). Washington, DC: Author. Retrieved from <http://www.gao.gov/new.items/d08538.pdf>

Van Weele, S. F., & Ramirez-Marquez, J. E. (2010, February). Optimization of container inspection strategy via a genetic algorithm. *Annals of Operations Research*, 187, 229–247. Retrieved from <http://www.springerlink.com/content/l3g2630t412x036g/>

Vijayan, J. (May 5, 2010). N.Y. bomb plot highlights limitations of data mining. *Computerworld.com*. Retrieved from [http://www.computerworld.com/s/article/9176317/N.Y.\\_bomb\\_plot\\_highlights\\_limitations\\_of\\_data\\_mining](http://www.computerworld.com/s/article/9176317/N.Y._bomb_plot_highlights_limitations_of_data_mining)

Wasem, R. E., Lake, J., Seghetti, L., Monke, J., & Vina, S. (2004). Border security: Inspection practices, policies, and issues (Congressional Research Service Report No. RL-32399). Retrieved from <http://www.fas.org/sgp/crs/RL32399.pdf>

West, J. R., Walton, C. M., & Conway, A. J. (2008, December). Addressing cargo security with strategies involving the private sector (Research Report No. SWUTC/08/473700-00095-1). Austin, TX: Southwest Regional University Transportation Center. Retrieved from <http://swutc.tamu.edu/publications/technicalreports/473700-00095-1.pdf>

Xiao, N., Chen, B., & Fu, W. (2004). A scalable federated data grid server. In H. R. Arabnia & O. Droegehorn (Eds.), *Proceedings of the International Conference on Internet Computing* (vol. 1, pp. 167–172). [n.p.]: CSREA Press.

# Appendix A: Top 10 Algorithms in Data Mining

The following 10 data mining algorithms were identified by the 2006 IEEE International Conference on Data Mining as being the most influential algorithms for the tasks of classification, clustering, statistical learning, association analysis, and link mining.

Algorithm Name	Function
C4.5 Classifiers	This algorithm takes as input a collection of cases, each belonging to a small number of classes and described by its values for a fixed set of attributes, and outputs a classifier that can accurately predict the class to which a new case belongs.
K-Means Algorithm	K-Means Algorithm partitions a given dataset into a user-specified number of clusters.
Support Vector Machines	This algorithm finds the best classification function to distinguish between members of two classes in the training data. The metric for operationalizing the concept of “best” can be determined geometrically.
The Apriori Algorithm	The Apriori Algorithm finds frequent itemsets from a transaction dataset and derives association rules. Once itemsets are determined, association rules can be generated with user-specified confidence levels.
The EM Algorithm	The EM Algorithm clusters continuous data and estimates the underlying density function. The models are fitted using maximum likelihood via the EM (expected-maximization) method.
Page Rank	Page Rank produces a static ranking of Web pages in the sense that a PageRank value is computed for each page off-line. This search-ranking algorithm using hyperlinks on the Web was developed by Brin and Page and was the basis of the Google search engine.
AdaBoost	Short for adaptive learning, AdaBoost is a meta-algorithm that is used in conjunction with other learning algorithms to improve their performance. AdaBoost is adaptive in the sense that it can improve the classification of those cases misclassified by previous classifiers. Used for feature recognition.
K-Nearest Neighbor	One of the simplest classifiers, k-means algorithms memorize the entire dataset and classify the attributes of the test case based on how they match the attributes of the training case.
Naïve Bayes	Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, this approach constructs a rule that allows the assignment of future cases to a class given only the vectors of variables describing the future cases.
Classification and Regression Trees (CART)	Classification and regression trees (CART) is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. CART is powerful because it can deal with incomplete data and multiple types of features (floats, unenumerated sets) both in input features and predicted features. The trees it produces often contain rules that are readable to humans.