

# Machine Learning Methods for Computer Security

Edited by

Anthony D. Joseph<sup>1</sup>, Pavel Laskov<sup>2</sup>, Fabio Roli<sup>3</sup>, J. Doug Tygar<sup>4</sup>,  
and Blaine Nelson<sup>5</sup>

1 University of California – Berkeley, US, [adj@eecs.berkeley.edu](mailto:adj@eecs.berkeley.edu)

2 Universität Tübingen, DE, [pavel.laskov@uni-tuebingen.de](mailto:pavel.laskov@uni-tuebingen.de)

3 Università di Cagliari, IT, [roli@diee.unica.it](mailto:roli@diee.unica.it)

4 University of California – Berkeley, US, [tygar@cs.berkeley.edu](mailto:tygar@cs.berkeley.edu)

5 Universität Tübingen, DE, [blaine.nelson@wsii.uni-tuebingen.de](mailto:blaine.nelson@wsii.uni-tuebingen.de)

---

## Abstract

The study of learning in adversarial environments is an emerging discipline at the juncture between machine learning and computer security. The interest in learning-based methods for security- and system-design applications comes from the high degree of complexity of phenomena underlying the security and reliability of computer systems. As it becomes increasingly difficult to reach the desired properties solely using statically designed mechanisms, learning methods are being used more and more to obtain a better understanding of various data collected from these complex systems. However, learning approaches can be evaded by adversaries, who change their behavior in response to the learning methods. To-date, there has been limited research into learning techniques that are resilient to attacks with provable robustness guarantees

The Perspectives Workshop, “Machine Learning Methods for Computer Security” was convened to bring together interested researchers from both the computer security and machine learning communities to discuss techniques, challenges, and future research directions for secure learning and learning-based security applications. As a result of the twenty-two invited presentations, workgroup sessions and informal discussion, several priority areas of research were identified. The open problems identified in the field ranged from traditional applications of machine learning in security, such as attack detection and analysis of malicious software, to methodological issues related to secure learning, especially the development of new formal approaches with provable security guarantees. Finally a number of other potential applications were pinpointed outside of the traditional scope of computer security in which security issues may also arise in connection with data-driven methods. Examples of such applications are social media spam, plagiarism detection, authorship identification, copyright enforcement, computer vision (particularly in the context of biometrics), and sentiment analysis.

**Perspectives Workshop** 09.–14. September, 2012 – [www.dagstuhl.de/12371](http://www.dagstuhl.de/12371)

**1998 ACM Subject Classification** C.2.0 Computer-Communication Networks (General): Security and Protection (e.g., firewalls), D.4.6 Operating Systems (Security and Protection), I.2.6 Artificial Intelligence (Learning), I.2.7 Artificial Intelligence (Natural Language Processing), I.2.8 Artificial Intelligence (Problem Solving, Control Methods, and Search), K.4.1 Computers and Society (Public Policy Issues): Privacy, K.6.5 Management of Computing and Information Systems (Security and Protection)

**Keywords and phrases** Adversarial Learning, Computer Security, Robust Statistical Learning, Online Learning with Experts, Game Theory, Learning Theory

**Digital Object Identifier** 10.4230/DagMan.3.1.1



Except where otherwise noted, content of this manifesto is licensed under a Creative Commons BY 3.0 Unported license

Machine Learning Methods for Computer Security, *Dagstuhl Manifestos*, Vol. 3, Issue 1, pp. 1–30

Editors: Anthony D. Joseph, Pavel Laskov, Fabio Roli, J. Doug Tygar, and Blaine Nelson




Dagstuhl Manifestos

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary

*Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar*

License  Creative Commons BY 3.0 Unported license  
© Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar

Arising organically from a variety of independent research projects in both computer security and machine learning, the topic of secure learning is emerging as a major direction of research that offers new challenges to both communities. Learning-based approaches are particularly advantageous for security applications designed to counter sophisticated and evolving adversaries because learning methods can cope with large amounts of evolving and complex data. However, the assets of learning can also potentially be subverted by malicious manipulation of data. This exposes applications that use learning techniques to a new type of security vulnerability in which an adversary can attempt to evade learning-based methods. Unlike many other application domains of machine learning, security-related applications require careful consideration of their adversarial nature and novel learning methods with improved robustness against potential attacks. The Perspectives Workshop, “Machine Learning Methods for Computer Security”, brought together prominent researchers from the computer security and machine learning communities interested in advancing the state-of-the-art in the field of secure learning, discussing open problems, and promoting further collaboration between the two communities.

This workshop focused on tasks in three main topics: the role of learning in computer security applications, the paradigm of secure learning, and the future applications for secure learning. These themes arose throughout the majority of plenary presentations and were the subjects of focused discussions within separate working groups. In the first group, participants discussed the current usage of learning approaches by security practitioners. The main conclusion of their discussion was that there is a pressing need for a tighter integration of machine learning methods into the operational practice of security systems to enable improved response and proactive functionality. The second group focused on the current approaches and methodical challenges for learning in security-sensitive adversarial domains. The key open problem identified in this area was a want for a formal notion of security for data-driven applications, similar to that used in classical information security, and for the development of learning algorithms with provable security guarantees. Finally, the third group addressed future application domains that would benefit from secure-learning technologies. It is to be expected that the demand for secure data analysis will expand into new application domains whenever attackers discover new opportunities for monetary profit by abusing deployed data-driven technologies.

Several themes arose recurrently during the workshop. One of the major concerns discussed was the reluctance of security practitioners to use learning-based techniques due to their opacity. To address this problem, new learning methods should be developed with higher transparency and interpretability. Further, the issue of incorporating human operators into the learning process to prevent unintended consequences was identified as an important research direction. Finally, benchmarks for and quantitative assessments of the security of learning algorithms were also deemed to be currently inadequate.

Yet the most important outcome of this workshop is the new found sense of an emerging scientific community growing at the junction of computer security and machine learning. Even though the scientific traditions and practices of machine learning and computer security diverge in many aspects, regular scientific exchange is indispensable in this field and should be promoted with joint projects, professional networks and dissemination activities.

## Table of Contents

|  |    |
|--|----|
| <b>Executive Summary</b>   |    |
| <i>Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar</i> | 2  |
| <b>Introduction</b>  |    |
| <i>Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar</i> | 4  |
| <b>Machine Learning for Computer Security</b>  |    |
| <i>Battista Biggio and Nedim Šrndić</i>  | 5  |
| State-of-the-art   | 6  |
| Open Issues and Research Directions  | 7  |
| Future Applications  | 9  |
| <b>Secure Learning: Theory and Methods</b>   |    |
| <i>Daniel Lowd and Rachel Greenstadt</i>   | 10 |
| State-of-the-art   | 11 |
| Open Issues and Research Directions  | 12 |
| Secure Learning and Data Privacy   | 14 |
| <b>Future Applications of Secure Learning</b>  |    |
| <i>Nathan Ratliff, Alvaro Cárdenas, and Fabio Roli</i>                               | 16 |
| State-of-the-art and emerging technologies: Where do adversaries attack next?        | 16 |
| Recommendations and Research Priorities  | 20 |
| <b>Conclusion</b>  |    |
| <i>Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar</i> | 22 |
| <b>Participants</b>  | 23 |
| <b>References</b>  | 24 |

## 2 Introduction

*Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar*

License  Creative Commons BY 3.0 Unported license  
© Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar

The emergence of the Internet has revolutionized modern society. It has changed the way we do business, manage our personal lives and communicate with our friends. To a large extent, the Internet owes its success to the enormous amount of data it generates and to novel decision-making instruments based on data analysis. Online advertisement, recommendation systems, consumer profiling, and many other Internet-related businesses crucially depend on data analysis and the underlying methods of machine learning, which extract meaningful information from seemingly unstructured masses of data.

Unfortunately, the ubiquity of the Internet has also stimulated its abuse and the rise of sophisticated cyber-crimes. It has enabled criminals to build sustainable businesses that rely on the exploitation of security vulnerabilities. To avoid being detected by security mechanisms, the attackers develop new exploitation techniques; an act which places tremendous pressure on cybersecurity vendors. To speed up development of adequate defenses, the latter are forced to resort to data-analysis techniques to extract information from prodigious amounts of security data. The vendors' successes, in turn, motivates the attackers to develop new tricks to evade detection.

The cat-and-mouse game between the security industry and the cyber-criminal underground points out a fundamental scientific problem associated with data analysis and machine learning techniques: they were originally conceived under the assumption of “faithful” data and did not explicitly account for potential data manipulation by adversaries. Several studies have shown that data-driven security instruments can be easily broken [44, 96], which raises the question of whether machine learning methods can be deployed at all in adversarial environments [10].

Recent developments in the learning methodology, *e.g.*, [20, 37, 60], and the growing experience with its application in the security practice, *e.g.*, [31, 70, 100], have underlined the necessity for deeper understanding of the security aspects of machine learning. These developments have motivated the Perspectives Workshop “Machine Learning Methods for Computer Security” held at Schloss Dagstuhl from the ninth to the fourteenth of September, 2012. Presentations and discussions held during this workshop were aimed at producing assessments of the state-of-the-art methodologies and at identifying open problems and research priorities. The workshop was also a major step in molding the scientific community in this emerging field of secure machine learning. It has brought together researchers from various disciplines ranging from machine learning and security to spam filtering, online advertisement and computer forensics. This manifesto summarizes the key findings of the workshop and provides an overview of the future scientific developments in secure machine learning.

The following three themes can be seen as the cornerstones of the workshop's discussions and of the results presented in this manifesto:

1. **Machine learning for security.** What security problems can machine learning best help to solve? What scenarios are they ill-suited for? These and many other scientific and operational issues are discussed in Section 3.
2. **Secure machine learning.** What are the theoretical limitations of worst-case attacks against learning algorithms under different constraints? How can these constraints

be used in practice for protecting learning methods against adversarial data? These methodological issues are discussed in Section 4.

3. **Secure learning beyond security.** What are existing and emerging non-security applications where learning techniques are used and can potentially be exposed to adversarial data? What experience from these applications can be used for development of general methodology of secure learning? These issues are discussed in Section 5.

Finally, it must be noted that most of security-related decisions involve a human operator. As such, humans are often the first targets of attacks using “social engineering” tricks such as deception or impersonation. Although consideration of the social factors associated with security was outside of this workshop’s scope and beyond the expertise of its participants, the need to address the social dimension of security and to integrate data-analysis tools with human decision-making capabilities was consistently re-iterated during the workshop.

### 3 Machine Learning for Computer Security

*Battista Biggio and Nedim Šrđić*

License © Creative Commons BY 3.0 Unported license  
© Battista Biggio and Nedim Šrđić

The rapid development of security exploits in recent years has fueled a strong interest in data analysis tools for computer security. On the one hand, the sheer number of novel malicious software observed by security researchers transcends the limits of manual analysis.

According to AVTEST,<sup>1</sup> more than 200,000 examples of new malware are sighted daily [5]. However, most of these instances represent only minor variants of existing malware strains. Nonetheless, correctly identifying the specific strain of a given malware sample requires sophisticated classification methods beyond hashes, simple rules, or heuristic fingerprints. Beyond simple malware polymorphisms and obfuscations, the increasing professionalization of the “attack industry” leads to particularly hard cases in which genuinely novel exploitation techniques are employed. Conventional methods based on hashes, signatures, or heuristic rules cannot deal with such threats in a timely fashion. Anomaly-based detection methods appear to be the best alternative for such cases, even if they inevitably cause some false positives.

Historically, the development of machine learning and computer security has been reciprocal. The early work on intrusion detection, starting from the seminal paper of Denning [38], formulated intrusion detection as a data analysis problem in which a decision function is based on a model automatically derived from previous benign examples. Stemming from both the security and machine learning communities, followed this anomaly-based approach [45, 69, 71, 126]. Additional machine learning techniques such as supervised classification and clustering have also proved to be useful to various security problems [11, 99, 122]. Certain characteristics of security problems are atypical for classical learning methods and require the development of customized techniques. These characteristics include strongly unbalanced data (attacks are very rare), unbalanced risk factors (low false positive rates are crucial), difficulties in obtaining labeled data, and several others.

---

<sup>1</sup> <http://www.av-test.org/en/home/>

The most crucial peculiarity of security as an application field for machine learning is adversarial data manipulation. All security technologies are sooner or later subjected to attacks. Hence, the analysis of potential attacks is a fundamental aspect of security research. Consideration for adversarial data is not addressed by classical machine learning methods, which has hindered their acceptance in security practices. Recent developments in both fields have brought a significant understanding of the general factors that affect the security of learning algorithms. The remainder of this chapter provides an overview of the state-of-the-art work, open problems and potential applications for the learning-based security technologies.

### 3.1 State-of-the-art

A classical security application of machine learning is detection of malicious activity in operating systems data or network traffic: “intrusion detection systems”. A substantial amount of work in intrusion detection followed various learning-based approaches, in particular, anomaly detection [45, 69, 71, 126], rule inference [72, 73, 74, 78] and supervised learning [87, 88]. Although most of the proposed methods performed well in controlled experiments, most of the practical intrusion detection systems, such as Snort [104] and Bro [95], are still rooted in the more conservative signature-based approach. Sommer and Paxson discussed several practical difficulties faced by learning-based intrusion detection systems [112]. Among the key challenges they identified are the high cost of classification errors, the semantic gap between detection results and operational interpretation, the enormous variability and non-stationarity of benign traffic, as well as the difficulty to perform a sound evaluation of such systems.

A key lesson to be learned from the limited use of learning-based methods in the general intrusion detection context is the necessity for a precise focus on the semantics of specific applications. Several narrowly focused systems developed in the recent years have demonstrated that, in certain applications, learning-based systems significantly outperform conventional approaches depending on expert knowledge. One of the most successful application domains for such narrowly focused systems is web application security. Due to the extreme versatility of web applications, it is next to impossible to devise signatures for specific attack patterns. The learning systems overcome this difficulty by automatically inferring models of benign application-specific traffic. Such models can be used to detect malicious web queries [39, 51, 64, 113], to detect logical state violations in web applications [30], and even to develop reactive mechanisms such as reverse proxies [123] or the sanitization of web queries [66].

Another crucial contribution of learning-based systems lies in the realm of dynamic malware analysis. To stay abreast of the recent trends in malware development, most anti-virus vendors deploy sophisticated systems to acquire novel malware. Such systems have been very successful in collecting masses of data, resulting in an urgent need for tools to automatically analyze novel malware. One of the first methods for malware analysis based on reports from its execution in a sandbox used hierarchical clustering to infer groups of related malware [6]. An alternative approach based on supervised learning enabled classification of malware into known families as well as detection of novel malware strains [99]. Subsequent research has improved scalability of the above mentioned methods and verified their feasibility for large-scale malware attribution [11, 102, 122].

A similar synergy between machine learning and malware analysis has been exploited for automatic signature generation (ASG). Early ad-hoc ASG systems combined extraction of

string tokens with frequency analysis, which roughly corresponds to the idea of naive Bayes learning [57, 93]. The key weakness of such methods was their susceptibility to attacks, which made it possible to evade a deployed system or increase its false alarm rate [94, 96]. Even though this difficulty is intrinsic to all methods based on string tokens [114, 124], the ASG methods proved to be very useful for several related problems, including the detection of botnet communications [101] and network protocol reverse engineering [27, 33] that are less prone to adversarial data.

The recent emergence of JavaScript-based exploitation schemes such as drive-by-downloads and malicious non-executable files raises a number of new challenges for detecting such threats. The effectiveness of these attacks is due to their ease of deployment and distribution (*e.g.*, via malicious web sites or targeted email messages) as well as to almost unlimited options for obfuscation. Detection of such threats is further complicated by the need to understand the syntactic context of malicious content. But these problems have largely yielded to learning-based approaches built on appropriate features. Static analysis of JavaScript token sequences has been successfully deployed for detection of drive-by-downloads [100] and JavaScript-bearing malicious PDF documents [70]. Another form of static analysis with a focus on PDF document structure was instrumental in recently developed highly effective methods for detection of PDF malware [80, 111, 115]. Dynamic analysis of JavaScript string allocations combined with classification of string payloads has been successfully used for in-browser detection of drive-by-downloads [34].

To summarize, machine learning methods are currently widely used as a component in general reactive security architectures. They can optimally address targeted data-related security problems with clear semantics and well-defined scope. A key advantage of learning-based approaches is their ability to generalize information contained in data, even though such generalization may not be easily expressible in a human-readable form. At the extreme, the generalization ability of learning methods can even enable detection of previously unseen zero-day attacks.

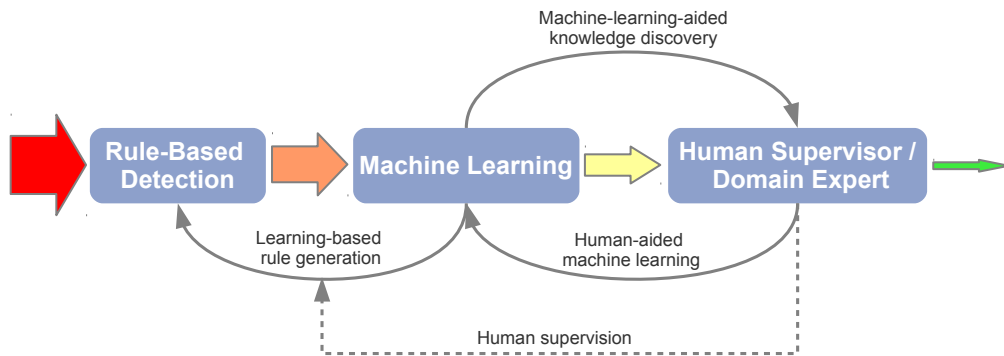
A crucial factor for the success of learning methods in security problems is a careful design of features, which incorporate the semantics of respective applications. While yielding excellent detection rates, learning methods occasionally may result in false alarms which can, however, be mitigated through appropriate tuning of detection thresholds. Another part of the “operational price” of learning-based approaches is the black-box nature of their predictions: even in the case when highly accurate detection is feasible, it is not always possible to identify the specific set of features that were “responsible” for these predictions. This limitation must be taken into account in their operational deployment.

## 3.2 Open Issues and Research Directions

A number of open issues and future directions arise in the field of machine learning for computer security. Investigation of these will enable tighter integration of machine learning into the operational practice of security systems, enable an improved response and facilitate proactive efforts through early discovery of security flaws and vulnerabilities.

### 3.2.1 Integration of machine learning with security mechanisms

Given the proven success of learning-based approaches to narrowly focused security tasks, it is natural to expect that tight integration with existing security instruments may deliver substantial qualitative benefits for the latter. However, such integration is by no means a



■ **Figure 1** The abstract architecture of reactive security mechanisms utilizing machine learning.

simple task. Figure 1 shows the abstract architecture of learning-enhanced reactive security mechanisms.

The vast majority of security-related data can be handled using simple rule-based detection methods. Rules operate quickly, cheaply and accurately, and they are simple for the domain experts to understand and maintain. However, rules are not powerful enough to handle cleverly crafted and novel input samples. Although they represent only a small fraction of the total input samples, such samples may inflict substantial damage if not stopped. Machine learning algorithms can play a pivotal role here by potentially detecting completely novel, previously unseen attack samples. By providing confidence intervals for their predictions, learning methods can prioritize data to be manually inspected by experts and thus largely improve the productivity of these analysts.

To improve the feedback between the individual components presented in Figure 1, a number of open problems must be investigated:

- *Learning-based rule generation.* Although rule-inference algorithms are well-known in machine learning (e.g., C4.5 [98] or RIPPER [26]), they are prone to overfitting and can be easily evaded by an adversary. Novel rule inference methods are needed that are able to produce concise interpretable rules, detect anomalous events in the absence of label information and deal with adversarial data. An example for advanced rule-oriented learning is automatic discovery of regular expression patterns for spam detection [97].
- *Human-aided machine learning.* The supervisory role of security experts is essential for the success of learning methods in this domain. However, security expertise does not fit into traditional binary or multiclass categories common for classical learning methods. In order to learn models with high predictive power, new techniques for interaction between the learning methods and the security experts need to be investigated.
- *Machine-learning-aided knowledge discovery.* Security analyst's work can also be greatly facilitated by applying appropriate learning techniques. For example, active learning techniques may be deployed to suggest interesting data for a detailed investigation [2, 67, 85]. Such approaches may be especially beneficial for security applications in which manual analysis is very time-consuming and requires profound expertise.

Operational deployment of learning methods is always an iterative process. Changes in data patterns (distributional shift) makes periodic re-training of learning system is a necessary practice. Further investigation is required to understand the impact of such non-stationarity on the learning-enhanced reactive security architecture described above.



### 3.2.2 Reactive approaches

Machine learning for security has been primarily used for detection of malicious activity. Yet the potential high cost of security violations calls for countermeasures that can be initiated automatically. Only limited work has addressed such methods so far, *e.g.*, [65, 66, 123]. The main challenge for reactive approaches is to develop mechanisms for measuring the risk of automatic decisions which is critical for the widespread utilization of these decision-making techniques in practical security applications.

### 3.2.3 Machine learning for offensive security

The term “offensive security” refers to methods for empirical security verification of various applications. It comprises the areas of penetration testing, vulnerability research, and various fields of cryptanalytic research. Current methods of offensive security are almost exclusively based on human expertise. The existing automation tools are essentially brute force methods of various kinds: password, key and hash function guessing as well as protocol and document fuzzing. Some recent work in assisted discovery of vulnerabilities in source code has demonstrated the ability of learning methods to extrapolate knowledge about known security flaws [129, 130]. Future research should address methods for data-driven syntactic and semantic analysis of applications and detection of undesirable system states.

## 3.3 Future Applications

The open research problems presented in the previous section provide the ground for a wide range of novel applications in the area of security. In the following, we briefly discuss several such applications that address recent security problems and establish a bridge between academic work and practical solutions.

- *Detection of advanced persistent threats.* The notion of “advanced persistent threats” (APT) refers to customized malware, exemplified by the Stuxnet [120], MiniDuke [54] and Flame [119], used for espionage or cyber-warfare. The goal of such malware is to penetrate highly sensitive sites and, in some cases, to remain undetected for a substantial time period. They typically rely on novel penetration vectors (*e.g.*, zero-day exploits) and deploy extensive obfuscation techniques. Signature-based security tools are ill-suited for advanced persistent threats since they rely on previously crafted detection patterns that are too slow for targeted attacks. Machine learning methods may be better at detecting of APTs by automatically spotting anomalous events.
- *Protection of mobile devices.* Private and sensitive data stored on mobile devices is becoming an increasingly attractive target for attackers. Mobile devices have limited resources, and common security approaches, such as matching files against signatures, can not be applied on a regular basis. Learning methods may support mobile security by enabling lightweight and signature-independent mechanisms for detecting malicious activity, for example, by identifying unusual information leakage from a device.
- *Dynamic and continuous authentication.* User authentication is typically based on secrets that can be forgotten and stolen. Authentication normally is only done once, at the beginning of a session, and grants full access rights to a given identity. We envision a more flexible authentication scheme, which is potentially less prone to lost or broken secrets, using information stored about users on systems they log into. Machine learning can use stored data to learn the behavior of legitimate users. Then, in some cases, users could authenticate simply by means of their usual actions, thus reducing the

need for or bolstering traditional password-based authentication. Machine learning could strengthen passwords with questions derived from previous user activities (*e.g.*, by inferring which acquaintances the user should reasonably be able to recognize based on their web-browsing). It similarly could also be used to generate secret questions for resetting forgotten passwords. Finally, it could provide for continuous and incremental authentication during the course of a session based on comparing users activities with their profiles when users request additional privileges.

- *Assisted malware analysis.* Countering malicious software is a continual arms race. More often than not, there is a significant delay between the appearance of a new malware and the availability of a malware signature. A significant amount of time during development of such signatures is devoted to analyzing and understanding the inner workings of malicious programs. Machine learning methods could accelerate this process. Techniques for feature selection and visualization could be applied to emphasize patterns within a malware's code or behaviors, and thus, these techniques could facilitate the rapid development of detection patterns.
- *Computer forensics.* Forensics is one of the most data-intensive areas of computer security. Here, significant benefits can also be expected from applying machine learning to evaluate data of forensic images, system logs or logged network traffic in the aftermath of a security incident. In a related field of criminal forensic investigation, similar methods can be applied to identify evidence of human crimes. In both cases, the objective of learning methods should be to help focus investigator's attention on important information related to goals of investigation and, possibly, to other similar cases, by sifting through large volumes of forensic data and prioritizing the information.

## 4 Secure Learning: Theory and Methods

*Daniel Lowd and Rachel Greenstadt*

License  Creative Commons BY 3.0 Unported license  
© Daniel Lowd and Rachel Greenstadt

While machine learning is a powerful tool for data analysis and processing, traditional machine learning methods were not designed to operate in the presence of adversaries. They are based on statistical assumptions about the distribution of the input data, and they rely on training data derived from the input data to construct models for analyses. Adversaries may exploit these characteristics to disrupt analytics, cause analytics to fail, or engage in malicious activities that fail to be detected. Several examples of attacks against learning algorithms have been proposed in the recent years. Some of them have been designed for specific variants of learning algorithms deployed in security applications [44, 96], others target the popular general-purpose learning algorithms [17, 61]. Such attacks exploit the adaptive aspects of a machine learning system by designing training data that will cause the learning system to produce models that misidentify future inputs. If users detect the failure, they may lose confidence in the system and abandon it. If users do not detect the failure, then the risks can be even greater.

The growing concern about the security of learning-based applications arises from the lack of understanding about how well machine learning performs under adversarial conditions. When a learning algorithm performs well in adversarial conditions, we say that it is an algorithm for secure learning. This raises a natural question: how do we characterize the quality of a learning system and determine whether it satisfies the requirements for secure

learning? Hence arises the need for a formal definition of the security of a machine learning algorithm; a definition that, until now, has not been fully expressed in a general context. A formal notion of secure learning would provide a means for practitioners to use learning algorithms with an understanding of their security guarantees and the conditions required for these guarantees to hold. Moreover, such a formal notion of security would encourage learning theorists to explore security properties of learning methods and, hence, would provide motivation for designing new algorithms suitable for adversarial environments. Coupled with appropriate advances in experimental methodology, the understanding of security aspects of machine learning methods would strongly increase their applicability to untrusted data and provide a quantitative assessment of the risks associated with such data.

#### 4.1 State-of-the-art

Early theoretical work revealed that learning algorithms exhibit different behavior under adversarial noise, depending on the part of the data controlled by an attacker. In the setting of supervised classification (*i.e.*, assigning one of the two *a-priori* defined classes to new data based on prior examples of such assignments), learning algorithms can be surprisingly tolerant to wrong labels assigned by an attacker during training. Angluin and Laird [3] showed that for some classes of Boolean functions, an attacker must flip the labels of nearly half of the training data to cause the incorrect classification of selected data points. On the other hand, by controlling attributes of the data, an attacker can construct inherently difficult learning problems, in which the error rate can be nearly as high as the fraction of data under the attacker's control [56]. These contrasting results emphasize the need for clear assumptions on the adversarial power.

Empirical evidence of attacks against learning algorithms arose in the last decade in the fields of intrusion detection and spam classification where learning methods were actively used since the late 1990s. Evasion attacks against anomaly-based intrusion detection systems (IDS) have demonstrated two main evasion strategies: *poisoning*, *i.e.*, erosion of a model of normality [121], and *mimicry*, *i.e.*, insertion of a normal content into the target data [125]. Another interesting example of an attack is the *polymorphic blending* technique, a transformation of packet payload to match a certain histogram of byte occurrences [44]. Although finding an optimal blending for byte sequences of length greater than one was shown to be NP-complete, greedy approximation algorithms can significantly decrease detection rates of an IDS [43]. Highly effective attacks have also been proposed against classification algorithms. The *red herring* and *correlated outlier* attacks against automatic signature generation systems were shown to decrease their detection accuracy to an unacceptably low level of 20–30% [94, 96]. Attacks against spam filters used similar techniques to manipulate features derived from word frequencies in email messages [77, 128].

Alongside the “destructive” work summarized above, some progress has been made in the development of learning algorithms with robustness to worst-case noise. Two groups of methods should be noted here. The min-max approach (*i.e.*, training a classifier with best performance under the worst noise) results in a relatively easy optimization problem [37, 47]. The more complex game-theoretic approach is advantageous when the learner and the attacker have different cost functions; *e.g.*, the learner is trying to minimize the spam frequency while the attacker is boosting the attractiveness of spam content. Such problems can be solved using the Nash equilibrium approach, which, however, is tractable only under some stringent assumptions [19, 21]. Compared to the baseline learning algorithms, these advanced methods improve accuracy by 10–15%, which is far from a strong claim of security.

An important milestone in the study of security of machine learning was the taxonomy of attacks against learning algorithms introduced by Barreno *et al.*, 2006 [10]. This work was the first to connect security of learning with the classical objectives of information security. However, the interpretation of such connection in this work is only qualitative. Increasing the detection error of a learner is categorized as an *integrity* attack and raising the false alarm rate is defined as a threat to learner's *availability*. Such qualitative associations do not provide a means to formalize the connection to the classical security objectives.

Recent work on the security of machine learning has provided more evidence of potential attacks but yielded little insight into how they can be prevented. It has been shown that complex learning algorithms such as Support Vector Machines can also be evaded with modest evasion effort [17]. Even if the attacker has no specific knowledge about a deployed model, evasion may still be successful if the attacker can query it at run-time. Nelson *et al.* [92] demonstrated that a large group of classifiers with convex decision functions (*i.e.*, classifiers that induce a convex positive/negative class) can be evaded under such assumptions, albeit potentially at a cost of an exponential number of queries, for some cost functions.

In summary, the most progress in the study of security of machine learning has been made in identifying the security weaknesses of learning methods. While there exist methods with improved robustness against attacks, such improvements are insufficient to claim strong security. Very little attention has been given so far to the investigation of autonomous learning methods, apart from the online anomaly detection studied by [60, 61]. Finally, the limited existing methodology for security analysis, thus far, has not provided results that have improved the design of novel learning algorithms.

## 4.2 Open Issues and Research Directions

### 4.2.1 Formalization of Secure Learning

Formalisms in both the security and machine learning (ML) communities (for example, cryptographic security, Byzantine fault tolerance, probably approximately-correct learning, and empirical risk minimization techniques) have catalyzed research in their respective fields and led to major advances in both theory and practice. Formalisms for secure ML have the potential to do the same.

Ideally, security metrics for ML systems will provide:

- A mechanism for inter-algorithm comparison,
- The ability to provide strong performance guarantees, and
- A mechanism for determining whether an algorithm is appropriate for use in a particular security setting.

However, there need not be a single metric or framework that captures all aspects of security. Different metrics might be best-suited for different tasks or for different aspects of the evaluation.

There are several frameworks for secure learning that form a foundation for secure learning. The qualitative taxonomy of security threats to learning techniques defined by Barreno *et al.* [9] provides a coarse granularity for separating different threats, many of which may require very different notions of a security measure. Within this augmented taxonomy, metrics for algorithmic security have emerged in two distinct areas: near-optimal evasion for exploratory attacks against learners [76, 92], and differential privacy for privacy exploratory attacks against a learned model [40].

There is a general need for a **metric for causative attacks**,  $S$ , that evaluates the (worst-case) effect of an attack scenario, in which the attacker is able to manipulate the training

data to mislead the learning algorithm. This particular portion of the attack taxonomy of Barreno *et al.* has been explored in prior work, but remains without a clear definition of security required of the learner. Such a measure should incorporate some notion of stability under adversarial contamination and also must incorporate limits on the adversary's effect on the learned model in order to model tractable threats. Thus, this measure can be considered as a function,  $S(L, A, \epsilon)$ , expressed in terms of the type of learner,  $L$ , the model of the adversary and his available actions,  $A$ , and the power or total resources allocated to the adversary (such as fraction of training examples he controls),  $\epsilon$ .

Another promising direction is to define new **security-aware loss functions** that can be directly minimized by ML algorithms. Such functions would measure the “damage” done to the estimator under the non-stationarity introduced by the adversary's contamination. Thus, this loss would necessarily be specific to the algorithm and the learning setting.

In some cases, these metrics may provide theoretical guarantees about the security or vulnerabilities of a particular method, as in differential privacy or minimal-cost evasion. Many of these metrics can also be used empirically (as in Section 4.2.2) to assess how a particular algorithm behaves under this security metric for the specified adversarial model.

#### 4.2.2 Empirical Evaluation

Current empirical techniques for performance evaluation of machine learning algorithms (*e.g.*, hold-out and cross-validation techniques), as well as performance metrics (*e.g.*, accuracy), do not take into account adversarial settings; *i.e.*, adversarial manipulation of training and/or testing data distribution with respect to data collected for classifier design. Therefore, such techniques can not provide information about the security of a classification system under attack, and are likely to provide over-optimistic estimates of their performance [9, 22, 23, 62, 68]. Besides theoretical analyses of the security of machine learning algorithms, it is thus necessary to develop methods for empirically evaluating, on a given set of data, the security of classifiers based on such algorithms. Such evaluation procedures could then be used both during classifier design (including the feature selection/extraction and model selection steps) and for deployed classification systems. Such methods will be useful for researchers, and also for practitioners, and it is desirable that they are implemented in ad hoc software tools.

Unlike the traditional performance evaluation, which relies on the stationarity assumption about data distribution, security evaluation would be better approached as a *what-if* scenario analysis which is well known in other fields [103]. Any attack scenario implies that training and testing datasets follow different distributions. It is not possible to know in advance what kinds of attacks a given learning algorithm or classifier system will be subject to, as well as their characteristics (*e.g.*, adversary's knowledge and capability). Security evaluation should then be performed against several possible attacks and for different characteristics of each attack under which it can be of interest to assess the behavior of the considered algorithm or systems, chosen according to the task at hand.

#### 4.2.3 Development of Secure Learning Approaches

Most current applications of machine learning to security problems use standard machine learning algorithms, which do not explicitly model an adversary. As the adversary adapts, humans respond by retraining the model with new data, as well as by manually inventing new features when necessary. More research is needed to understand when this “vanilla” approach is sufficient, or even superior to more complex approaches that incorporate adversarial reasoning. Another important direction is to develop generic approaches to making machine

learning algorithms secure, rather than manually developing a separate adaptation for each algorithm. A wrapper that guaranteed security would allow any algorithm to be made secure, allowing secure machine learning to take advantage of the latest advances in standard machine learning; thereby separating the learning and security elements of the problem.

One of the potential methodological avenues for development of secure learning algorithms would be to explore the connection to online and non-stationary learning, which were specifically designed to deal with distributions that change over time. Modeling non-stationary processes has been a subject of an extensive body of research, and many algorithms for statistical inference, prediction and classification have been proposed [35, 50, 58]. Non-stationary problems are naturally handled in the online learning setting. The field of online learning encompasses many areas, such as online prediction in a statistical setting [48], in the regret minimization setting where information can be partial or the prediction task can be adversarial [24], as well as the general reinforcement learning problem [117]. In order to apply online learning methods to security problems, we first need to define a utility function  $U$  mapping from the joint outcome and decision space to a numerical value.<sup>2</sup> There is a crucial link between the utility function  $U$  and the security function  $S$ . In particular, it should follow that the expectation of  $U$  given a particular model and algorithm  $L$  and adversary  $A$  equals  $S(L, A)$ . This also facilitates the future evaluation of the algorithm because  $U$  can be used as an evaluation metric.

#### 4.2.4 Data sanitization

Another potential countermeasure against adversarial noise is to identify and remove malicious data, known as *data sanitization*. Several approaches to data denoising have been explored in prior work. Zighed *et al.* [131] proposed a method to identify mislabeled samples, which was based on a statistical technique called the *cut edge weight statistic*. A sample close to another class has a higher chance to be an attack sample. Misclassified samples are treated as attack samples during the training process in [18, 52, 127]. The *Reject On Negative Impact* (RONI) defense proposed by [91] eliminates the samples which have negative impact on learning performance. Similar techniques have been proposed in the context of anomaly detection, using partitioning of the training data and cross-validation of models trained on individual partitions [32]. Although sanitization techniques proved to be successful in selected scenarios, their applicability as a general defense against adversarial noise remains to be investigated.

### 4.3 Secure Learning and Data Privacy

Privacy-preserving learning has been studied by research communities in security, databases, computer theory, machine learning, and statistics. More recently, the strands of this work have begun to merge, with the formalism of *differential privacy* from the theoretical computer science community being most popular [41]. This definition is intuitive and very strong in a formal sense, as it provides guarantees analogous to those provided in cryptography. A significant amount of work has already been done to understand how simple analyses can be performed while guaranteeing differential privacy and still yielding meaningful results [7, 12, 53]. However little has been done to produce differentially-private versions of the mainstream machine learning algorithms with a few exceptions including [25, 84, 106].

---

<sup>2</sup> In general though, we can also consider a multi-objective setting if we do not desire to select a particular utility function, *a priori*.




Numerous applications of machine learning, which are intended to provide societal benefits or drive private profits, operate on privacy-sensitive data. Just a few examples of such data include health records, personal communications, utility usage, online social networks, and mobile location data. As more of these services move online, and the automatic mining of such data becomes more prevalent, it is becoming increasingly necessary to establish technical approaches for protecting privacy and regulatory incentives to encourage the adoption of learning approaches.

Several well-known examples of privacy breaches in data released for statistical analysis exist, of which many have been a direct result of the application of machine learning techniques. In the 1990s, the U.S. state of Massachusetts released hospital records of state employees for the use by medical researchers, after removing names, addresses and social security numbers. However, removing these identifying features proved to be insufficient to truly preserve the privacy of this data as a researcher was able to cross-link the released data with easily obtainable state voter records containing names, addresses and postal codes, and thus, she was able to identify records with the names of several individuals [118]. The release of such data would clearly be detrimental, for example, to individuals hoping to conceal difficult treatments from family and friends or from the health insurance industry, which is often eager to obtain such information about individuals for risk assessments. Then, in 2006, AOL attempted to de-anonymize a large volume of search query logs by removing so-called personally identifying information (PII), and they released the data for use by ML researchers [8]. While such a well-intentioned act was immensely valuable for research in information retrieval and search, a journalist from the New York Times was able to identify an elderly AOL user—and her private, potentially embarrassing online habits—by linking the unique terms she queried in the log. In a third case of a company releasing data, the “anonymized” movie recommendations of Netflix users were de-anonymized by researchers by joining the data with public IMDB movie ratings having named users. This, in fact, resulted in the unwanted publication of a user’s sexuality [89]. Finally, it is common knowledge that online advertisers track and construct highly specific profiles of users using machine learning, so that advertisements may better target individuals likely to make purchases. A student recently leveraged identifying attributes of users such as geographic location, workplace, and schooling history, to study advertisements with an additional private attribute of interest. Depending on whether or not the ad was shown, the student was able to determine the value of the private attribute of the user in question [63]. Such an attack on learning-based advertising could reveal political preferences, sexuality, or other sensitive personal attributes.

To avoid the increasingly frequent privacy breaches by institutions entrusted with individuals’ data, a two-pronged approach of regulation and technical tools is required. It is imperative to understand the full consequences of concerted attacks on the privacy of user data, going beyond the relatively simple illustrative examples provided above. It is also important to have standardized metrics that quantify the privacy of released data. Along with formal measures must come an understanding as to whether methods that preserve data privacy provide the level of utility required for specific applications such as targeted advertisements, location-based shopping deals, medical research, *etc.* To this end, the development of privacy-preserving anonymization and statistical analysis must provide guarantees on both privacy and utility. Theoretical work should also seek negative results, *e.g.*, finding learning tasks, for which one provably can not simultaneously attain utility and privacy.

## 5 Future Applications of Secure Learning

*Nathan Ratliff, Alvaro Cárdenas, and Fabio Roli*

License  Creative Commons BY 3.0 Unported license  
© Nathan Ratliff, Alvaro Cárdenas, and Fabio Roli

Computer security is not the only field in which attacks against the learning systems have been observed. Almost concurrently with the discovery of the first attacks against intrusion detection systems, similar methods appeared for evasion of statistical spam filters [76, 90]. Both of these application domains provided inspiration as well as experimental foundation for developments in secure learning [20, 36]. We expect that the demand for secure data analysis will grow and expand to other application domains whenever attackers discover new opportunities for monetary profit by abusing data-driven technologies. In the following sections, we present the analysis of the main factors, which may make specific application domains attractive for novel attacks against learning methods and survey the situation in several domains, in which learning technologies play a crucial role.

### 5.1 State-of-the-art and emerging technologies: Where do adversaries attack next?

Unlike Sections 3 and 4 where our recommendations are based on existing prior work, in most application domains beyond computer security, with the exception of email spam filtering, there is hardly any prior experience with adversarial noise. Hence we begin our survey with a focused review of the literature on spam filtering and then attempt to extrapolate this experience to other applications domains.

The key factor in the assessment of the relevance of secure learning to specific applications is the existence of a business case. As experience shows, serious security threats arise mostly when miscreants can make money by exploiting technical vulnerabilities. This was the case with classical malware that could be made profitable in a variety of different ways. Typical monetization schemes deployed by cyber-criminals include stealing login credentials, credit card numbers and bank accounts from end-users, operating networks of compromised computers (botnets), including renting them out, demanding ransom for encrypted data or to avoid denial-of-service attacks, among many others. Email spam was fueled by its potential to reach a huge population of email users with unsolicited advertisement at almost no cost; such advertisements were backed by real, albeit often illegal, businesses. The “business case” may not necessarily involve monetary profit; attack targets may also be intellectual property or classified information, as the recent cases of cyber-espionage show [81].

#### 5.1.1 Spam filtering

Spam filtering is the most popular example of machine learning applications that has to deal with adversarial inputs. Many modern email clients have an automatic spam filtering function that partially incorporates machine learning techniques, thus proving both its scientific relevance of and the business case for this application. During the past fifteen years, machine learning techniques have been widely investigated and used to analyze the textual content of email messages. Moreover, the adversarial nature of spam filtering is apparent and can be cast into a “game” between spammers and the adaptive spam filter. For all these reasons, spam filtering has received much attention in the scientific community; *e.g.*, [13, 28, 49]. Most papers on adversarial learning use it as one of the test cases for experiments,



and it was used as a paradigmatic application in seminal papers on the modeling of adversarial learning [36, 76]. The evolution of spam filtering is also instructive for understanding the nature of an “arms race” within a typical adversarial learning application domain. Interested readers can find additional details on this evolution in the “spammer compendium”<sup>3</sup>. In early spam, the message body of spam emails consisted mostly of plain text without any explicit or malicious attempts to evade detection. But, as anti-spam filters improved, spammers have evolved from naive attempts to bypass these filters to specialized mimicry attacks that make it difficult to distinguish spam from legitimate e-mail based solely on a message body. Around 2004, spammers introduced the image-spam trick, which consists of removing the spam message from the email body and instead embedding it into an image sent as an attachment [16, 46]. This allowed spammers to bypass any sophisticated and effective analysis of email body texts. Image-based spam is a notable example of how attackers change when the defense becomes too effective. To detect image-based spam, computer vision techniques have been developed and specialized modules implementing them have been plugged into many anti-spam filters.<sup>4</sup> This is also an example of defenders reacting to attacks by changing the features used for detection.

### 5.1.2 Advertising

Online advertising is a maturing industry with billions of dollars at stake, which opens opportunities for adversaries to leverage the system to the detriment of consumers. Clearly defined policies outlining allowable advertisements, if administered prudently, can thwart many of these attacks, but online advertising systems are far too large to police manually. Accordingly, modern systems, such as the system monitoring Google’s advertisement networks [109], are built on top of large-scale automated and semi-automated machine learning tools designed to aid operators in tackling these problems at scale. Learning algorithms are fast and inexpensive, but are frequently less precise in their classifications, so these systems use machine learning where it consistently works well in catching clear-cut cases, thus allowing technicians to focus on the more difficult borderline cases. Malicious advertisers are constantly devising new strategies to subvert policies. To stay abreast of the rapidly changing landscape of attacks, ground truth data is fed back to the learning system from the human operators as training data and used to retrain the entire system. These systems are, therefore, continually evolving lines of defense designed to most efficiently and effectively leverage human resources to ensure a safe environment for online patrons.

### 5.1.3 Social media spam

With the recent developments in junk email detection, spam senders changed their targeted communication medium. Nowadays, social networks, social media websites and recommendation services such as Facebook, Google+, MySpace, Twitter, YouTube, Flickr, SoundCloud, TripAdvisor, Amazon, and so forth are the major target for spammers. Most of these services use machine learning methods [75, 116].

Beside distributing “classical” spam, these platforms are also used

- to attract attention by heavily interacting with other users,
- to create fake reviews and thus manipulate recommendations,
- to build fake friend networks to manipulate rankings,
- to start rumors and harass.

---

<sup>3</sup> <http://www.virusbtn.com/resources/spammerscompendium>

<sup>4</sup> <http://prag.diee.unica.it/prag/eng/research/doccategorisation/spamfiltering/products/imageCerberus>

Social media spam differs from traditional email spam in that spammers can exploit many different functions of these websites including posting/reposting comments, sending private or chat messages, following/re-following other users, favoring items, and uploading images/videos/sounds/documents/applications.

Service providers are faced with the problem of distinguishing regular user actions from spammy behavior. This problem appears to be more difficult than detecting classical spam since spammers have many more possibilities to prevent detection. They can hide their activities by mixing them with unobjectionable actions such that the distinction between legitimate and abusive behavior becomes blurred and even humans cannot distinguish them from normal users. As a consequence, ground truth data is very hard to gather and is typically very noisy making this a particularly challenging learning domain.

#### 5.1.4 Plagiarism Detection and Authorship Identification

Plagiarism has been a long-standing plague in academics and media. As shown by the case of two former German ministers, Karl-Theodor zu Guttenberg and Annette Schavan, the discovery of plagiarism may ruin a person's scientific and even political career. The business case for plagiarism is becoming especially important for self-published content, such as Amazon's Kindle Direct Publishing, in which low publication costs and the potential for high distribution volume make it an especially attractive target for providers of non-authentic content. Although most of the previous high-profile plagiarism cases were uncovered by the web-assisted and -coordinated efforts of activists, a significant effort has been put into the development of automatic methods for plagiarism detection. Four plagiarism detection competitions have been held under the aegis of the PAN Conference,<sup>5</sup> which have attracted significant attention in the machine learning and the information retrieval communities. Several commercial products are available on the market; interestingly, one of the vendors (iParadigms) offers software for both plagiarism detection (Turnitin) and verification of the lack of plagiarism (WriteCheck). Evasion of plagiarism detection system has not been systematically studied, although heuristic obfuscation methods have proved to a significant hurdle for detection accuracy.

A somewhat related problem is authorship identification. Research in this field is motivated by user-generated media, *e.g.*, blogs, which offer ample possibilities for falsification of author identities. The recent case of the fake identity of Amina Arraf in the blog "A Gay Girl in Damascus" has demonstrated that a significant political influence can be exerted by identity manipulation, including the careful adaptation of the forger's writing style. Some recent work (*e.g.*, [1]) has addressed methods for detecting fraudulent writing styles as well as potential evasion techniques.

#### 5.1.5 Copyright Enforcement

Apart from detecting plagiarism and duplicated content in textual documents, social media platforms facilitate the sharing of media files such as images, videos and music, which also must be monitored for copyright infringements. The publication of media content is generally restricted in terms of the usage and the geographical location depending on copyrights and patents. The business model of film studios, music labels, photographers and artists crucially depends on the protection of intellectual property rights.

---

<sup>5</sup> <http://pan.webis.de/>

Social media websites are required by law to effectively identify copyright infringement. Both detection and evasion technologies for copyrighted media content differ radically from the those used for plagiarized text. In the latter case, characteristic features for the learning problems stem from linguistic traits of the content. For media, mostly steganographic methods have been used [55]. These, however, can be easily detected and removed by a technologically proficient offender.

Practical methods for detection of copyright infringement in media are mainly based on content fingerprinting. Even if these methods are computationally efficient, they are not generally robust against adversarial manipulations such as changing the resolution or compression level. Using a larger number of more stable features in combination with statistical learning methods rather than fixed rules can be expected to provide more robust detection.

### 5.1.6 Computer Vision Systems: Present and Potential Attacks

The application of machine learning to computer vision has a long history because some components of computer vision systems are difficult to manually design, but it is often relatively easy to collect labeled examples from a vision task to train a learning system [86]. This is particularly true for high-level computer vision components devoted to pattern recognition, but learning has also been used in some low- and middle-level vision components [110]. Nowadays many computer vision systems (*e.g.*, face recognition systems and OCR tools) are developed using machine learning or use an online learning module because the resulting systems are more accurate than hand-crafted programs. However, attacks against the learning components of computer vision systems emerged only recently as the application and popularity of these technologies generated sufficient incentives for attackers.

Biometric recognition is an important application of machine learning to computer vision that recently witnessed its first attacks. Indeed, all current biometric recognition systems use machine learning. Face and fingerprints are the most widely used biometric traits and their computerized recognition has been the subject of scientific investigation for many years [82]. These technologies are used now for personal identity verification in passports, personal computers and smart phones. The risk of attacks against biometric recognition systems has been investigated in the scientific literature [14] but the emergence of financial motivations makes it a looming danger. Among the attacks proposed, the most practical is a mimicry attack in which fake biometric traits are used to fool a learning-based recognition module; *e.g.*, using a printed picture to fool a facial recognition system. Such attacks are called *spoofing* attacks, and they have a great practical relevance because they do not require advanced technical skills making them feasible for many potential attackers.<sup>6</sup> In fact, the winner of the ICB 2013 “spoofing challenge” against face recognition systems was a lady who used a mimicry attack based on facial makeup to claim successfully the identity of a gentleman.<sup>7</sup> However, spoofing attacks are not limited to face and fingerprint recognition—the European project, TABULA RASA, also demonstrated successful spoofing attacks against systems using speech and gait.<sup>8</sup>

Another little-known type of attack likely to emerge in the near future is an evasion attack against biometric video surveillance systems used to recognize targeted individuals (*e.g.*, individuals on a watch-list). To date this avenue of attack has received little attention

---

<sup>6</sup> An example of a spoofing attack: <http://www.youtube.com/watch?v=2fKGXSgOFYc>.

<sup>7</sup> <http://www.tabularasa-euproject.org/evaluations/tabula-rasa-spoofing-challenge-2013>

<sup>8</sup> <http://www.tabularasa-euproject.org>

because evading a biometric video surveillance system is still quite easy (wearing hats or glasses is often sufficient to evade a video face recognition system). However, the arms race to evade these vision systems has already begun as is evident in the creative CV Dazzle project that proposes new facial makeup and hair styling to evade face recognition systems.<sup>9</sup>

Incremental learning algorithms are also beginning to be incorporated into adaptive biometric systems to account for natural changes in biometric patterns and the surrounding environment; *e.g.*, the aging of biometric traits or changes in illumination. Indeed all biometric recognition systems require frequent updates of the *templates* used as prototypes for the biometric traits of users, and some systems use self-training algorithms [105]. Recently, it has been shown that an attacker may exploit self-training to compromise the templates of a face recognition system through a poisoning attack [15]. This poisoning attack consists of submitting a carefully designed sequence of fake facial images to the system to gradually perturb the client's template until it is compromised; *i.e.*, when it is replaced with a different face. This may allow an attacker to impersonate the targeted client using her own face.

Finally, another potential class of attacks that may emerge in the near future involves image forensics tools. Nowadays altering digital images with commercial photo-editing tools is commonplace, and this alteration may be used for fraud [42]. As the detection of malicious image alteration can be difficult, digital image forensic tools using machine learning have been designed to detect specific types of alterations (*e.g.*, the deletion of part of image like the face of a person, or a copy-and-paste from one image to another). As in other adversarial applications, we expect an arms race to develop between forgers and forensic analysts, with forgers developing new tricks to evade the current digital image forensic tools based on machine learning.

### 5.1.7 Sentiment analysis

In recent years, many marketing companies started to apply sentiment analysis (the task of discerning the opinion or partiality of a user from his/her communications and behavior) to become aware of users' preferences on brands, products or services. For example, an online travel agency may want to know the rating of hotels based on the guests' opinions. Instead of performing a potentially time-consuming and costly survey, natural language processing and machine learning can be used to extract this information from public blogs, forums, Twitter feeds, and similar sources.

With the rise of sentiment analysis, pollution attacks will become likely in order to boost one's own image or to defame competitors. Sentiment analysis mainly relies on linguistic or deterministic natural language processing methods which are not robust against data manipulation. Since this kind of analysis is intrinsically unsupervised, poisoning attacks are likely to become a serious issue in the future.

## 5.2 Recommendations and Research Priorities

Although it is generally hard to define research priorities for the heterogeneous application fields reviewed in Section 5.1, in the following, we attempt to provide some methodical recommendations that may have repercussions in several potential application domains.

---

<sup>9</sup> <http://cvdazzle.com>

### 5.2.1 Poisoning

Poisoning refers to attacks against machine learning, in which an attacker can insert manipulated data before or during the training phase of a learning system. Poisoning has been identified as a serious threat for specific intrusion detection techniques [96], as well as for general machine learning methods [17, 60, 107]. The importance of protecting against poisoning attacks lies in the fact that many real-life learning systems are periodically re-trained from new data. While re-training helps to keep systems abreast of the natural drift of real data, it can also be abused by attackers if new data is used for training without further control or sanitization. The specific implications of periodic re-training schemes (“quasi-online learning”) for poisoning attacks have, to date, only received limited attention [60, 61, 107]. Hence further research is needed to analyze the effectiveness of such attacks and to develop adequate protection mechanisms against them.

### 5.2.2 Penetration Testing

Penetration testing, for our purposes, is the empirical evaluation of the vulnerabilities of an aggregate system combining multiple applications to malfeasance.<sup>10</sup> This type of security assessment has become a de facto requirement for the majority of web-applications, especially in the financial industry [108]. Many companies are interested in having their systems successfully pass penetration tests. Penetration testing evaluates the company’s exposure to security risks and can potentially decrease its premiums for liability insurance. It is conceivable that similar testing requirements will become increasingly adopted in data-driven systems. This would create the need for appropriate penetration testing technologies. Developing these techniques is an interesting research challenge, which crucially depends on the question of whether or not learning systems can be reverse-engineered by a potential adversary.

### 5.2.3 Case Studies

Being largely an empirical field of science, machine learning typically entails large-scale case studies for the verification of results. It has a long tradition of benchmarking (*e.g.*, the USPS and NIST datasets and the UCI repository of datasets [4, 29]) as well as competitions on various datasets. In contrast, empirical evaluations in the field of computer security, are much more difficult to carry out. The only attempt to comparatively evaluate real intrusion detection systems conducted at MIT Lincoln Lab had limited success and drew a barrage of critical remarks [79, 83]. Several issues that make comprehensive case studies in security related research areas scarce, and which also concern adversarial machine learning, are elucidated below.

#### 5.2.3.1 Privacy issues

Many datasets in computer security and related areas are subject to serious privacy restrictions. To tackle this, an official institute such as NIST may be entrusted with managing such data and with performing standardized experiments. A successful example of such a benchmark was the NIST TREC Spam Track [29], which enabled researchers to compare spam filtering

---

<sup>10</sup> Penetration testing may also refer to the evaluation of an individual application used in several contexts, but we do not discuss this type of testing here.

techniques in various settings and was based on real world datasets. However, privacy-preservation was not an issue for the TREC benchmark because it was constructed using publicly available data particularly from emails obtained from the Enron court case [59]. However, the amount of publicly available data is relatively scarce compared to the amount of data stored confidentially, but maintaining privacy when releasing datasets derived from the latter sources is generally technically difficult, and some examples of privacy incidents related to data release were discussed in Section 4.3.

### 5.2.3.2 Non-stationarity of the data


As attackers respond to a particular predictive model, static datasets are insufficient; however, dynamic data is difficult to generate under realistic conditions. Test suites can still be defined in terms of rules of “games”: *e.g.*, playground.dk where a player takes the role of an attacker of a computer security system. Even if generated data is artificial and differs depending on the method under investigation, the results may be comparable if they are obtained under realistic conditions.

### 5.2.3.3 Lack of ground truth

In contrast to many other applications of machine learning, ground truth data is hard to obtain in adversarial environments. Defining ground truth for adversarial learning is especially difficult since the concept of abusive behavior is usually quite vague. Furthermore, adversaries attempt to hide their activities so that even humans are not able to identify them. To identify common applications and use cases, it is necessary to formally describe non-adversarial activities for each case, potentially using domain-expert knowledge, and to treat deviations from these allowed activities as adversarial events.

## 6 Conclusion

*Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar*

License  Creative Commons BY 3.0 Unported license  
© Anthony D. Joseph, Pavel Laskov, Blaine Nelson, Fabio Roli, and J. Doug Tygar

As one would expect for a workshop in an emerging discipline, our workshop has raised a broad variety of research questions. Some of these questions stem from fundamental methodological issues, such as the formalization of secure learning and the trade-off between security, privacy, and interpretability of learning models. The workshop has also identified practical open problems; *e.g.*, integrating machine learning with existing security mechanisms and understanding of an operator’s role in such a process. Several potential novel applications have also been identified, such as the detection of advanced persisting threats, protection of mobile devices, continuous authentication, and computer forensics. We expect that secure learning will play a crucial and expanding role in a large number of data-driven applications, especially online advertisement, social media and recommendation systems.

Yet the most important outcome of this workshop is the new found sense of an emerging scientific community growing at the junction of computer security and machine learning. It is not easy for researchers in these two fields to communicate with one another. Scientific traditions and practices of machine learning and computer security diverge in many aspects, especially where experimental work is concerned. There indeed exist objective reasons for such divergence. The data arising in computer security is subject to privacy and confidentiality



restrictions, which makes the traditional benchmarking practices of machine learning less feasible. On the other hand, the adversarial nature of data is a novel aspect for the machine learning methodology, which requires a thorough recapitulation of its theoretical foundations. To understand these issues, and to bring researchers in these two communities closer to each other, regular scientific exchange is indispensable. Stay tuned for forthcoming events and advancements in this field.

## 7 Participants

- Battista Biggio  
Università di Cagliari, IT
- Christian Bockermann  
TU Dortmund, DE
- Michael Brückner  
SoundCloud Ltd., DE
- Alvaro Cárdenas Mora  
Fujitsu Labs of America Inc. –  
Sunnyvale, US
- Christos Dimitrakakis  
EPFL – Lausanne, CH
- Felix C. Freiling  
Univ. Erlangen-Nürnberg, DE
- Giorgio Fumera  
Università di Cagliari, IT
- Giorgio Giacinto  
Università di Cagliari, IT
- Rachel Greenstadt  
Drexel Univ. – Philadelphia, US
- Anthony D. Joseph  
University of California –  
Berkeley, US
- Robert Krawczyk  
BSI – Bonn, DE
- Pavel Laskov  
Universität Tübingen, DE
- Richard P. Lippmann  
MIT – Lexington, US)
- Daniel Lowd  
University of Oregon, US)
- Aikaterini Mitrokotsa  
EPFL – Lausanne, CH
- Saša Mrdović  
University of Sarajevo, SEU
- Blaine Nelson  
Universität Tübingen, DE
- Patrick Pak Kei Chan  
South China University of  
Technology, CN
- Massimiliano Raciti  
Linköping University, SE
- Nathan Ratliff  
Google – Pittsburgh, US
- Konrad Rieck  
Universität Göttingen, DE
- Fabio Roli  
Università di Cagliari, IT
- Benjamin I. P. Rubinstein  
Microsoft – Mountain View, US
- Tobias Scheffer  
Universität Potsdam, DE
- Galina Schwartz  
University of California –  
Berkeley, US
- Nedim Šrndić  
Universität Tübingen, DE
- Radu State  
University of Luxembourg, LU
- J. Doug Tygar  
University of California –  
Berkeley, US
- Viviane Zwanger  
Univ. Erlangen-Nürnberg, DE



**Acknowledgments.** We thank Dr. Roswitha Bardohl, Susanne Bach-Bernhard, Dr. Marc Herbstritt, and Jutka Gasiorowski for their help in organizing this workshop and preparing this report.

---

**References**

---

- 1 Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *IEEE Symposium on Security and Privacy*, pages 461–475, 2012.
- 2 Magnus Almgren and Erland Jonsson. Using active learning in intrusion detection. In *IEEE Computer Security Foundations Workshop*, pages 88–98, 2004.
- 3 Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):434–470, 1988.
- 4 Arthur Asuncion and David J. Newman. UCI machine learning repository, <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 2007.
- 5 AV-TEST. Malware Statistics. <http://www.av-test.org/en/statistics/malware/>.
- 6 Michael Bailey, Jon Oberheide, Jon Andersen, Z. Morley Mao, Farnam Jahanian, and Jose Nazario. Automated classification and analysis of internet malware. In *Recent Advances in Intrusion Detection (RAID)*, pages 178–197, 2007.
- 7 Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 273–282, 2007.
- 8 Michael Barbaro and Tom Zeller Jr. A face is exposed for AOL searcher no. 4417749. *The New York Times*, August 2006.
- 9 Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.
- 10 Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, pages 16–25, 2006.
- 11 Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. Scalable, behavior-based malware clustering. In *Network and Distributed System Security Symposium (NDSS)*, 2009.
- 12 Amos Beimel, Shiva Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *Theory of Cryptography Conference (TCC)*, pages 437–454, 2010.
- 13 Steffen Bickel and Tobias Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In *Neural Information Processing Systems (NIPS)*, pages 161–168, 2007.
- 14 Battista Biggio, Zahid Akhtar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli. Security evaluation of biometric authentication systems under real spoofing attacks. *IET Biometrics*, 1:11–24, 2012.
- 15 Battista Biggio, Luca Didaci, Giorgio Fumera, and Fabio Roli. Poisoning attacks to compromise face templates. In *Proceedings of the 6th IAPR International Conference on Biometrics (ICB)*, pages 1–7, 2013.
- 16 Battista Biggio, Giorgio Fumera, Ignazio Pillai, and Fabio Roli. A survey and experimental evaluation of image spam filtering techniques. *Pattern Recognition Letters*, 32:1436–1446, 2011.
- 17 Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against Support Vector Machines. In *International Conference on Machine Learning (ICML)*, 2012.
- 18 Carla E. Brodley and Mark A. Friedl. Identifying and eliminating mislabeled training instances. In *AAAI National Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence*, pages 799–805, 1996.
- 19 Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13:2617–2654, 2012.



- 20 Michael Brückner and Tobias Scheffer. Nash equilibria of static prediction games. In *Neural Information Processing Systems (NIPS)*, pages 171–179, 2009.
- 21 Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 547–555, 2011.
- 22 Alvaro A. Cárdenas and John S. Baras. Evaluation of classifiers: Practical considerations for security applications. In *AAAI Workshop on Evaluation Methods for Machine Learning*, 2006.
- 23 Alvaro A. Cárdenas, John S. Baras, and Karl Seamon. A framework for the evaluation of intrusion detection systems. In *IEEE Symposium on Security and Privacy*, pages 63–77, 2006.
- 24 Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- 25 Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.
- 26 William W. Cohen. Fast effective rule induction. In *International Conference on Machine Learning (ICML)*, pages 115–123, 1995.
- 27 Paolo Milani Comparetti, Gilbert Wondracek, Christopher Kruegel, and Engin Kirda. Prospex: Protocol specification extraction. In *IEEE Symposium on Security and Privacy*, pages 110–125, 2009.
- 28 Gordon Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, April 2008.
- 29 Gordon Cormack and Thomas Lynam. Spam corpus creation for TREC. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, July 2005.
- 30 Marco Cova, Davide Balzarotti, Viktoria Felmetzger, and Giovanni Vigna. Swaddler: An approach for the anomaly-based detection of state violations in web applications. In *Recent Advances in Intrusion Detection (RAID)*, pages 63–86, 2007.
- 31 Marco Cova, Christopher Kruegel, and Giovanni Vigna. Detection and analysis of drive-by-download attacks and malicious JavaScript code. In *International Conference on World Wide Web (WWW)*, pages 281–290, 2010.
- 32 Gabriela Cretu, Angelos Stavrou, Michael Locasto, Salvatore J. Stolfo, and Anghelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *IEEE Symposium on Security and Privacy*, pages 81–95, 2008.
- 33 Weidong Cui, Jayanthkumar Kannan, and Helen J. Wang. Discoverer: Automatic protocol reverse engineering from network traces. In *USENIX Security Symposium*, pages 1–14, 2007.
- 34 Charlie Curtsinger, Benjamin Livshits, Benjamin Zorn, and Christian Seifert. ZOZZLE: Fast and precise in-browser JavaScript malware detection. In *USENIX Security Symposium*, pages 33–48, 2011.
- 35 Rainer Dahlenhaus. Fitting time series models to nonstationary processes. *Annals of Statistics*, 25(1):486–503, 1997.
- 36 Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 99–108, 2004.
- 37 Ofer Dekel, Ohad Shamir, and Lin Xiao. Learning to classify with missing and corrupted features. *Machine Learning*, 81(2):149–178, 2010.
- 38 Dorothy E. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, 13(2):222–232, 1987.

- 39 Patrick Düssel, Christian Gehl, Pavel Laskov, and Konrad Rieck. Incorporation of application layer protocol syntax into anomaly detection. In *International Conference on Information Systems Security (ICISS)*, pages 188–202, 2008.
- 40 Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages and Programming, Part II (ICALP)*, pages 1–12, 2006.
- 41 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 2006.
- 42 Harry Farid. Seeing is not believing. *IEEE Spectrum*, 8(46):44–48, August 2009.
- 43 Prahlad Fogla and Wenke Lee. Evading network anomaly detection systems: formal reasoning and practical techniques. In *ACM Conference on Computer and Communications Security (CCS)*, pages 59–68, 2006.
- 44 Prahlad Fogla, Monirul Sharif, Roberto Perdisci, Oleg Kolesnikov, and Wenke Lee. Polymorphic blending attacks. In *USENIX Security Symposium*, pages 241–256, 2006.
- 45 Stephanie Forrest, Steven A. Hofmeyr, Anil Somayaji, and Thomas A. Longstaff. A sense of self for Unix processes. In *IEEE Symposium on Security and Privacy*, pages 120–128, 1996.
- 46 Giorgio Fumera, Ignazio Pillai, and Fabio Roli. Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research*, 7:2699–2720, 2006.
- 47 Amir Globerson and Sam Roweis. Nightmare at test time: Robust learning by feature deletion. In *International Conference on Machine Learning (ICML)*, pages 353–360, 2006.
- 48 Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.
- 49 Thiago S. Guzella and Walimir M. Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222, 2009.
- 50 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2<sup>nd</sup> edition, 2009.
- 51 Kenneth L. Ingham, Anil Somayaji, John Burge, and Stephanie Forrest. Learning DFA representations of HTTP for protecting web applications. *Computer Networks*, 51(5):1239–1255, 2007.
- 52 George H. John. Robust decision trees: Removing outliers from databases. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 174–179, 1995.
- 53 Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 531–540, 2008.
- 54 Kaspersky. The MiniDuke Mystery: PDF 0-day Government Spy Assembler 0x29A Micro Backdoor. [https://www.securelist.com/en/blog/208194129/The\\_MiniDuke\\_Mystery\\_PDF\\_0\\_day\\_Government\\_Spy\\_Assembler\\_Micro\\_Backdoor](https://www.securelist.com/en/blog/208194129/The_MiniDuke_Mystery_PDF_0_day_Government_Spy_Assembler_Micro_Backdoor).
- 55 Stephan Katzenbeisser and Fabien Petitolas. Information hiding techniques for steganography and digital watermaking. 2000.
- 56 Michael. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.
- 57 Huang-Ah Kim and Brad Karp. Autograph: Toward automated, distributed worm signature detection. In *USENIX Security Symposium*, pages 271–286, 2004.
- 58 Genshiro Kitagawa and Will Gersh. *Smoothness Priors Analysis of Time Series*. Springer, 1996.
- 59 Bryan Klimt and Yiming Yang. Introducing the Enron corpus. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, July 2004.

- 60 Marius Kloft and Pavel Laskov. Online anomaly detection under adversarial impact. In *International Conference on AI and Statistics (AISTATS)*, pages 405–412, 2010.
- 61 Marius Kloft and Pavel Laskov. Security analysis of online centroid anomaly detection. *Journal of Machine Learning Research*, 13:3133–3176, 2012.
- 62 Aleksander Kolcz and Choon Hui Teo. Feature weighting for improved classifier robustness. In *Conference on Email and Anti-Spam (CEAS)*, 2009.
- 63 Aleksandra Korolova. Privacy violations using microtargeted ads: A case study. *Journal of Privacy and Confidentiality*, 3(1):27–49, 2011.
- 64 Christopher Kruegel and Giovanni Vigna. Anomaly detection of web-based attacks. In *ACM Conference on Computer and Communications Security (CCS)*, pages 251–261, 2003.
- 65 Tammo Krüger, Christian Gehl, Konrad Rieck, and Pavel Laskov. An architecture for inline anomaly detection. In *European Conference on Computer Network Defense (EC2ND)*, pages 11–18, 2008.
- 66 Tammo Krüger, Christian Gehl, Konrad Rieck, and Pavel Laskov. TokDoc: A self-healing web application firewall. In *Symposium on Applied Computing (SAC)*, pages 1846–1853, 2010.
- 67 Pavel Laskov, Christian Gehl, Stefan Krüger, and Klaus-Robert Müller. Incremental support vector learning: Analysis, implementation and applications. *Journal of Machine Learning Research*, 7:1909–1936, 2006.
- 68 Pavel Laskov and Marius Kloft. A framework for quantitative security analysis of machine learning. In *ACM Workshop on Security and Artificial Intelligence (AISec)*, pages 1–4, 2009.
- 69 Pavel Laskov, Christin Schäfer, and Igor Kotenko. Intrusion detection in unlabeled data with quarter-sphere Support Vector Machines. In *Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA)*, pages 71–82, 2004.
- 70 Pavel Laskov and Nedim Šrđić. Static detection of malicious JavaScript-bearing PDF documents. In *Annual Computer Security Applications Conference (ACSAC)*, pages 373–382, 2011.
- 71 Aleksandar Lazarevic, Levent Ertöz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *SIAM International Conference on Data Mining (SDM)*, 2003.
- 72 Wenke Lee and Salvatore J. Stolfo. Data mining approaches for intrusion detection. In *USENIX Security Symposium*, 1998.
- 73 Wenke Lee and Salvatore J. Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information Systems Security*, 3:227–261, 2000.
- 74 Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok. A data mining framework for building intrusion detection models. In *IEEE Symposium on Security and Privacy*, pages 120–132, 1999.
- 75 Jimmy Lin and Alek Kolcz. Large-scale machine learning at Twitter. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 793–804, 2012.
- 76 Daniel Lowd and Christopher Meek. Adversarial learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 641–647, 2005.
- 77 Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *Conference on Email and Anti-Spam (CEAS)*, pages 641–647, 2005.
- 78 Matthew V. Mahoney and Philip K. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 376–385, 2002.
- 79 Matthew V. Mahoney and Philip K. Chan. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In *In Proceedings of the 6th*

- International Symposium on Recent Advances in Intrusion Detection (RAID)*, pages 220–237, 2003.
- 80 Davide Maiorca, Giorgio Giacinto, and Iginio Corona. A pattern recognition system for malicious PDF files detection. In *Machine Learning and Data Mining (MLDM)*, pages 510–524, 2012.
  - 81 Mandiant. APT1: Exposing One of China’s Cyber Espionage Units. <http://intelreport.mandiant.com/>.
  - 82 Gian Luca Marcialis and Fabio Roli. Fusion of appearance-based face recognition algorithms. *Pattern Analysis and Applications*, 7:151–163, 2004.
  - 83 John McHugh. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security*, 3(4):262–294, 2000.
  - 84 Frank McSherry and Ilya Mironov. Differentially private recommender systems: Building privacy into the net. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 627–636, 2009.
  - 85 Yuxin Meng and Lam for Kwok. Enhancing false alarm reduction using pool-based active learning in network intrusion detection. In *International Conference on Information Security Practice and Experience (ISPEC)*, pages 1–15, 2013.
  - 86 Tom M. Mitchell. The discipline of machine learning. Technical report, Carnegie Mellon ML Department, 2006.
  - 87 Srinivas. Mukkamala, Guadalupe Janoski, and Andrew H. Sung. Intrusion detection using neural networks and Support Vector Machines. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1702–1707, 2002.
  - 88 Srinivas Mukkamala, Andrew H. Sung, and Ajith Abraham. Intrusion detection using ensemble of soft computing and hard computing paradigms. *Journal of Network and Computer Applications*, 28(3):233–248, 2005.
  - 89 Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.
  - 90 Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, pages 1–9, 2008.
  - 91 Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. Misleading learners: Co-opting your spam filter. In *Machine Learning in Cyber Trust: Security, Privacy, Reliability*, chapter 2, pages 17–50. Springer, 2009.
  - 92 Blaine Nelson, Benjamin I. P. Rubinstein, Ling Huang, Anthony D. Joseph, Shing hon Lau, Steven Lee, Satish Rao, Anthony Tran, and J. D. Tygar. Near-optimal evasion of convex-inducing classifiers. In *International Conference on AI and Statistics (AISTATS)*, pages 549–556, 2010.
  - 93 James Newsome, Brad Karp, and Dawn Song. Polygraph: Automatically generating signatures for polymorphic worms. In *IEEE Symposium on Security and Privacy*, pages 120–132, 2005.
  - 94 James Newsome, Brad Karp, and Dawn Song. Paragraph: Thwarting signature learning by training maliciously. In *Recent Advances in Intrusion Detection (RAID)*, pages 81–105, 2006.
  - 95 Vern Paxson. Bro: a system for detecting network intruders in real-time. In *USENIX Security Symposium*, pages 31–51, 1998.

- 96 Roberto Perdisci, David Dagon, Wenke Lee, Prahlad Fogla, and Monirul Sharif. Misleading worm signature generators using deliberate noise injection. In *IEEE Symposium on Security and Privacy*, pages 17–31, 2006.
- 97 Paul Prasse, Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Learning to identify regular expressions that describe email campaigns. In *International Conference on Machine Learning (ICML)*, 2012.
- 98 J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992.
- 99 Konrad Rieck, Thorsten Holz, Karsten Willems, Patrick Düssel, and Pavel Laskov. Learning and classification of malware behavior. In *Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA)*, pages 108–125, 2008.
- 100 Konrad Rieck, Tammo Krüger, and Andreas Dewald. Cujo: Efficient detection and prevention of drive-by-download attacks. In *Annual Computer Security Applications Conference (ACSAC)*, pages 31–39, 2010.
- 101 Konrad Rieck, Guido Schwenk, Tobias Limmer, Thorsten Holz, and Pavel Laskov. Botzilla: Detecting the "phoning home" of malicious software. In *Symposium on Applied Computing (SAC)*, pages 1978–1984, 2010.
- 102 Konrad Rieck, Philipp Trinius, Carsten Willems, and Thorsten Holz. Automatic analysis of malware behavior using machine learning. *Journal of Computer Security*, 19(4):639–668, 2011.
- 103 Stefano Rizzi. What-if analysis. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 3525–3529. Springer, 2009.
- 104 Marty Roesch. Snort: Lightweight intrusion detection for networks. In *USENIX Large Installation System Administration Conference (LISA)*, pages 229–238, 1999.
- 105 Fabio Roli and Gian Luca Marcialis. Semi-supervised PCA-based face recognition using self-training. In *Proceedings of the 2006 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition, SSPR'06/SPR'06*, pages 560–568. Springer-Verlag, 2006.
- 106 Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.
- 107 Benjamin I. P. Rubinstein, Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J. D. Tygar. ANTIDOTE: Understanding and defending against poisoning of anomaly detectors. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 1–14, 2009.
- 108 Sebastian Schreiber. Rechtliche Aspekte von Penetrationstests. *Wirtschaftsinformatik & Management*, (1):12–15, 2010.
- 109 D. Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. Detecting adversarial advertisements in the wild. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.
- 110 Nicu Sebe, Ira Cohen, Ashutosh Garg, and Thomas S. Huang. *Machine Learning in Computer Vision*. Computational Imaging and Vision Series. Springer, 2005.
- 111 Charles Smutz and Angelos Stavrou. Malicious PDF detection using metadata and structural features. In *Annual Computer Security Applications Conference (ACSAC)*, pages 239–248, 2012.
- 112 Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *IEEE Symposium on Security and Privacy*, pages 305–316, 2010.
- 113 Yingbo Song, Angelos D. Keromytis, and Salvatore J. Stolfo. Spectrogram: A mixture-of-markov-chains model for anomaly detection in web traffic. In *Network and Distributed System Security Symposium (NDSS)*, 2009.



- 114 Yingbo Song, Michael E. Locasto, Angelos Stavrou, Angelos D. Keromytis, and Salvatore J. Stolfo. On the infeasibility of modeling polymorphic shellcode. In *ACM Conference on Computer and Communications Security (CCS)*, pages 541–551, 2007.
- 115 Nedim Šrndić and Pavel Laskov. Evasion-resistant detection of malicious PDF files based on hierarchical document structure. In *Network and Distributed System Security Symposium (NDSS)*, 2013.
- 116 Tao Stein, Erdong Chen, and Karan Mangla. Facebook immune system. In *Workshop on Social Network Systems (SNS)*, pages 1–8, 2011.
- 117 Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- 118 Latanya Sweeney. K-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
- 119 Symantec. Flamer: Highly Sophisticated and Discreet Threat Targets the Middle East. <http://www.symantec.com/connect/blogs/flamer-highly-sophisticated-and-discreet-threat-targets-middle-east>.
- 120 Symantec, Nicolas Falliere, Liam O Murchu, and Eric Chien. W32.Stuxnet Dossier. [http://www.symantec.com/content/en/us/enterprise/media/security\\_response/whitepapers/w32\\_stuxnet\\_dossier.pdf](http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf).
- 121 Kymie M.C. Tan, Kevin S. Killourhy, and Roy A. Maxion. Undermining an anomaly-based intrusion detection system using common exploits. In *Recent Advances in Intrusion Detection (RAID)*, pages 54–73, 2002.
- 122 Olivier Thonnard. *A Multicriteria Clustering Approach to Support Attack Attribution in Cyberspace*. PhD thesis, Ecole Doctorale d’Informatique, Télécommunications et Electronique de Paris, 2010.
- 123 Fredrik Valeur, Giovanni Vigna, Christopher Kruegel, and Engin Kirda. An anomaly-driven reverse proxy for web applications. In *Symposium on Applied Computing (SAC)*, pages 361–368, 2006.
- 124 Shobha Venkataraman, Avrim Blum, and Dawn Song. Limits of learning-based signature generation with adversaries. In *Network and Distributed System Security Symposium (NDSS)*, 2008.
- 125 David Wagner and Paolo Soto. Mimicry attacks on host based intrusion detection systems. In *ACM Conference on Computer and Communications Security (CCS)*, pages 255–264, 2002.
- 126 Ke Wang and Salvatore J. Stolfo. Anomalous payload-based network intrusion detection. In *Recent Advances in Intrusion Detection (RAID)*, pages 203–222, 2004.
- 127 Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, 2(3):408–421, 1972.
- 128 Gregory L. Wittel and S. Felix Wu. On attacking statistical spam filters. In *Conference on Email and Anti-Spam (CEAS)*, 2004.
- 129 Fabian Yamaguchi, Felix Lindner, and Konrad Rieck. Vulnerability extrapolation: assisted discovery of vulnerabilities using machine learning. In *USENIX Workshop on Offensive Technologies (WOOT)*, pages 118–128, 2011.
- 130 Fabian Yamaguchi, Markus Lottmann, and Konrad Rieck. Generalized vulnerability extrapolation using abstract syntax trees. In *Annual Computer Security Applications Conference (ACSAC)*, pages 359–368, 2012.
- 131 Djamel A. Zighed, Stéphane Lallich, and Fabrice Muhlenbach. Separability index in supervised learning. In *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 475–487, 2002.