In this paper, we discuss an emerging field of study: adversarial machine learning—the study of effective machine learning techniques against an adversarial opponent. To see why this field is needed, is learning—the study of effective machine learning techniques against an adversarial opponent. To see why this field is needed, it is helpful to recall a common metaphor: security is sometimes thought of as a chess game between two players. For a player to win, it is not only necessary to have an effective strategy, one must also anticipate the opponent's response to that strategy.

Statistical machine learning has already become an important tool in a security engineer's repertoire. However, machine learning in an adversarial environment requires us to anticipate that our opponent will try to cause machine learning to fail in many ways. In this paper, we discuss both a theoretical framework for understanding adversarial machine learning, and then discuss a number of specific examples illustrating how these techniques succeed or fail.

Advances in computing capabilities have made online statistical machine learning a practical and useful tool for solving large-scale decision-making problems in many systems and networking domains, including spam filtering, network intrusion detection, and virus detection [36, 45, 60]. In these domains, a machine learning algorithm, such as a Bayesian learner or a Support Vector Machine (SVM) [14], is typically periodically retrained on new input data. Unfortunately, sophisticated adversaries are well aware that online machine learning is being applied and we have substantial evidence that they frequently attempt to break many of the assumptions that practitioners make (e.g., data has various weak stochastic properties; independence; a stationary data distribution).

The lack of stationarity provides ample opportunity for mischief during training (including periodic re-training) and classification stages. In many cases, the adversary is able to poison the learner's classifications, often in a highly targeted manner. For instance, an adversary can craft input data that has similar feature properties to normal data (e.g., creating a spam message that appears to be non-spam to the learner), or they exhibit Byzantine behaviors by crafting input data that, when retrained on, causes the learner to learn put data that, when retrained on, causes the learner to learn an incorrect decision-making function. These sophisticated adversaries are patient and adapt their behaviors to achieve various goals, such as avoiding detection of attacks, causing benign input to be classified as attack input, launching focused or targeted attacks, or searching a classifier to find blind-spots in the algorithm.

Adversarial machine learning is the design of machine learning algorithms that can resist these sophisticated attacks, and the study of the capabilities and limitations of 43 In Proceedings of 4th ACM Workshop on Artificial Intelligence and Security, October 2011, pp. 43-58 attackers. In this paper, we: give a taxonomy for classifying attacks against online machine learning algorithms; discuss application-specific factors that limit an adversary's capabilities; introduce two models for modeling an adversary's capabilities; explore the limits of an adversary's knowledge about the algorithm, feature space, training, and input data; explore vulnerabilities in machine learning algorithms; discuss countermeasures against attacks; introduce the evasion challenge; and discuss privacy-preserving learning techniques.