Mitigations in Adversarial Machine Learning (MAML)

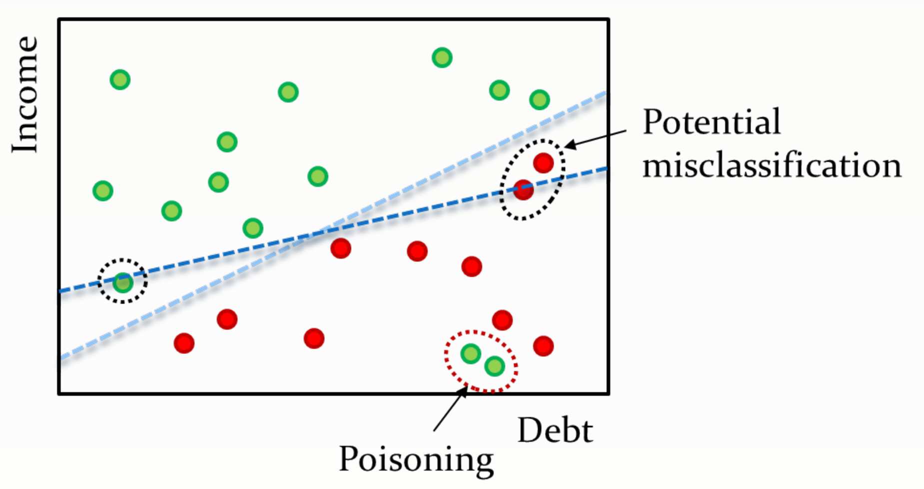
Dr. Daniel J. Clouse, djclous@tycho.ncsc.mil

Adversarial Machine Learning

Machine learning (ML) is proposed as a solution to scalable defensive and offensive capabilities in cyber security. The proposals range from semi-automated decision support tools to fully-automated capabilities. However, ML models can be exploited in at least four ways: (a) attackers can *poison* training data used to train ML algorithms to degrade prediction quality, or redirect predictions, altogether; (b) attackers can *evade* by manipulating runtime data to ensure ML models misclassify malicious behavior as benign; (c) attackers can *infer* records in the training data; and (d) attackers can approximately reconstruct the ML model for further analysis and exploitation. When ML models of varying qualities are integrated into an ensemble, the attacker can exploit weaknesses in individual models to coordinate a malicious effect in the overall system.

Example: Poisoning

Figure 1. A linear classifier learns to classify data points as green or red in an ideal partition of the data (light blue, dotted line). An adversary introduces adversarial examples to poison the training data and cause misclassifications by shifting the decision boundary.



Real Life Example: Gu, Dolan-Gavitt, and Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain", http://arxiv.org/abs/1708.06733.

MAML Approach

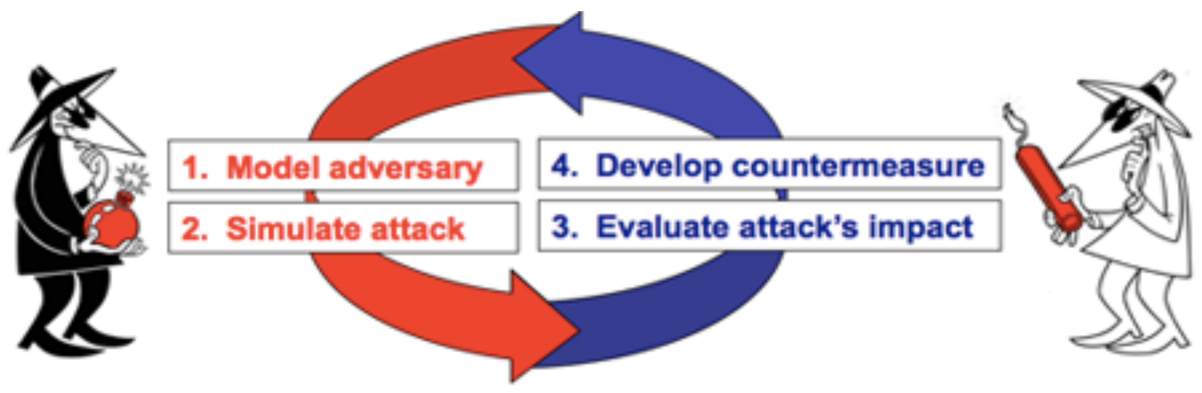
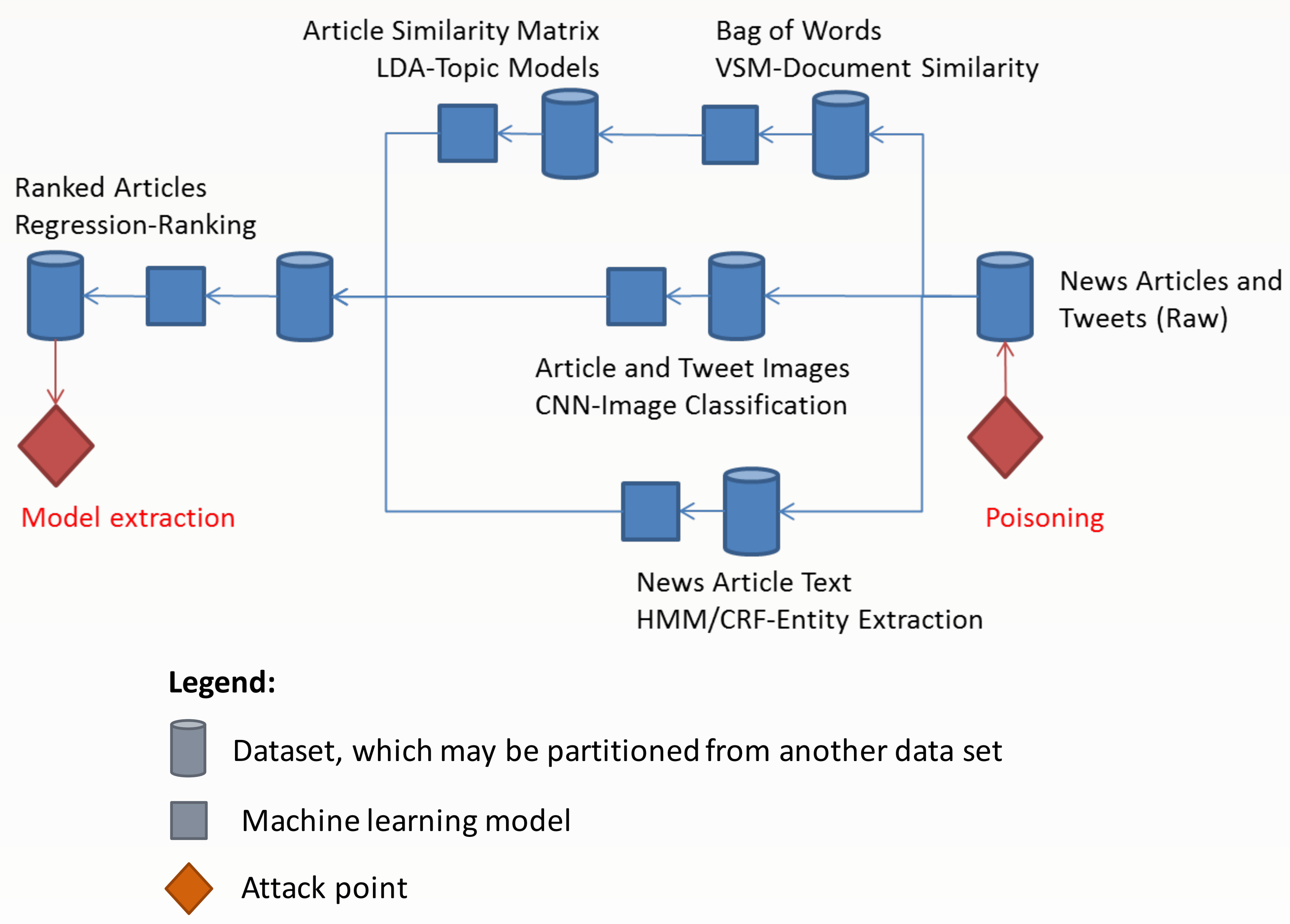


Image Credit: Pattern Recognition and Applications Lab, http://pralab.diee.unica.it/en/SecurityEvaluation

Design, Test and Evaluation

We are developing a framework consisting of: (1) a lightweight simulation language to express the performance parameters and architecture tailored to represent a decision-support environment consisting of one or more ML models and decision support tool users; (2) metrics to measure the quality of an adversarial influence strategy conducted in a simulation; and (3) mitigations, including malicious selector sharing to identify malicious actors and ML model design guidelines to improve resiliency against attacks.

Figure 3. A software architect / data scientist draws a scenario in the framework GUI to illustrate a news recommendation service. The service combines different datasets (cylinders) with models (squares) trained on that data. The adversary identifies attack points, which may include poisoning unprotected data or extracting models.



The simulation framework will support multiple machine learning goals and model types:

- **Data Exploration, Filtering**
  - ML classes (with examples)
    - Divisive clustering (K-means)
    - Agglomerative clustering (Hierarchical Agglomerative Clustering)
    - Density based (DBSCAN)
    - Topic modeling (Latent Dirichlet Allocation)
  - Decision Support Classes
    - Anomaly/outlier detection
    - Summarization and recommendation
    - Triage
- **Inference, Prediction**
  - ML classes
    - Classification & Regression
    - Reinforcement Learning, RL
  - Decision Support Classes
    - Categorizing data
    - Predicting continuous numeric features of data
    - Inferring consequences of actions

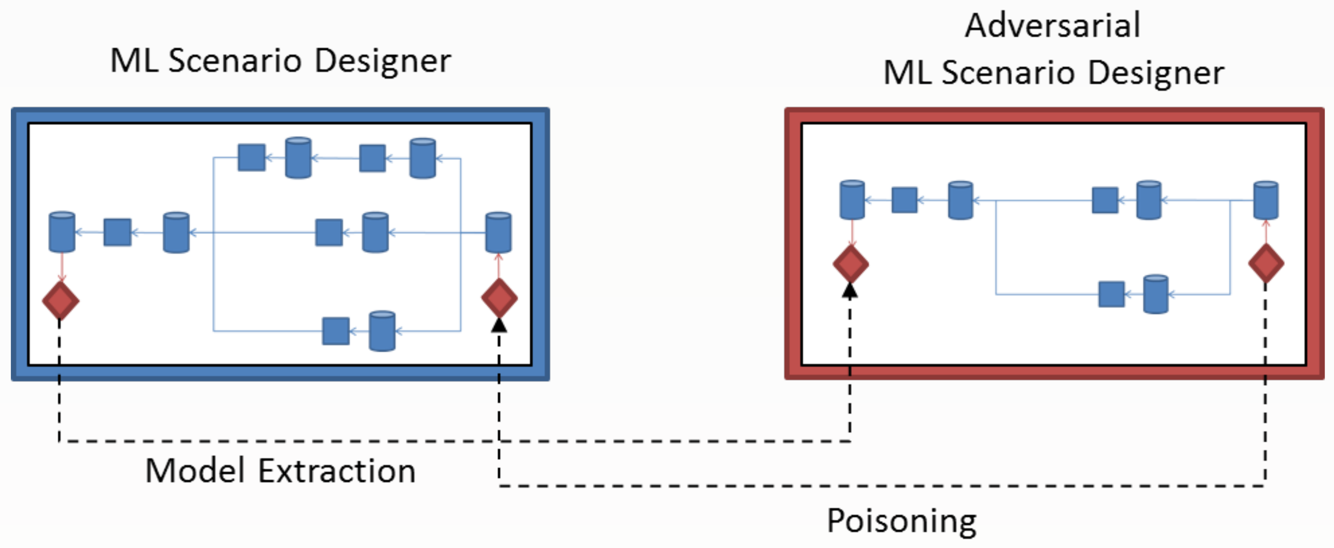
Influence Metrics

Influence quality metrics will measure the effectiveness of an influence strategy, and be used to explore mitigations based on sharing indicators to identify malicious actors and guidelines to design resilient ML models and architectures.

The guidelines will be based on adversarial game theory and/or related concepts to simulate attacker and defender uses of ML attacks as part of an adversary's influence strategy. Game Theory is a subfield of applied mathematics and Artificial Intelligence (AI) that utilizes models to study multi-player scenarios in which selfish decisions are made based on perceived costs or rewards and anticipated actions of other players [Myerson, 1991].

Understanding the Adversary

Figure 4. The adversarial scenario designer constructs an approximate model composition based on assumptions about what they can observe in the original system. This composition may not be 100% equivalent, however.



Adversarial Influence

Figure 5. Influence metrics will measure the impact on human decision makers using the result of the models with and without adversarial influence. This includes: incorrect decisions, e.g., due to evasion; or distractions, which cause decision makers to expend unnecessary resources to reach a correct decision.