EE P 595: Hands-on Machine Learning for Cyber Security

Winter 2024

Dept. of Electrical and Computer Engineering

University of Washington

Instructor: Prof. Radha Poovendran (rp3@uw.edu)

ELECTRICAL & COMPUTER ENGINEERING

UNIVERSITY of WASHINGTON

NSL

Machine Learning for Cybersecurity

- **This is the fifth offering of a popular hands-on course of ML for Cyber Security.**
- **It forms part of the course sequence for ECE PMP ACE certificate in ML4Cyber.**
- We study one security problem per class. Each problem will have a real-world data set to work on.
- First half of each class will focus on the needed background of the security problem to be studied that week. The second half is on learning to implement defense using the data set and check it works.
- All labs use Python for coding, and we will provide needed modules and also work with the students during the labs.
- Last year students did many interesting projects such as ML for password strength classification, and Deep Neural Networks (DNNs) for deceiving CAPTCHA.

Machine Learning for Cybersecurity

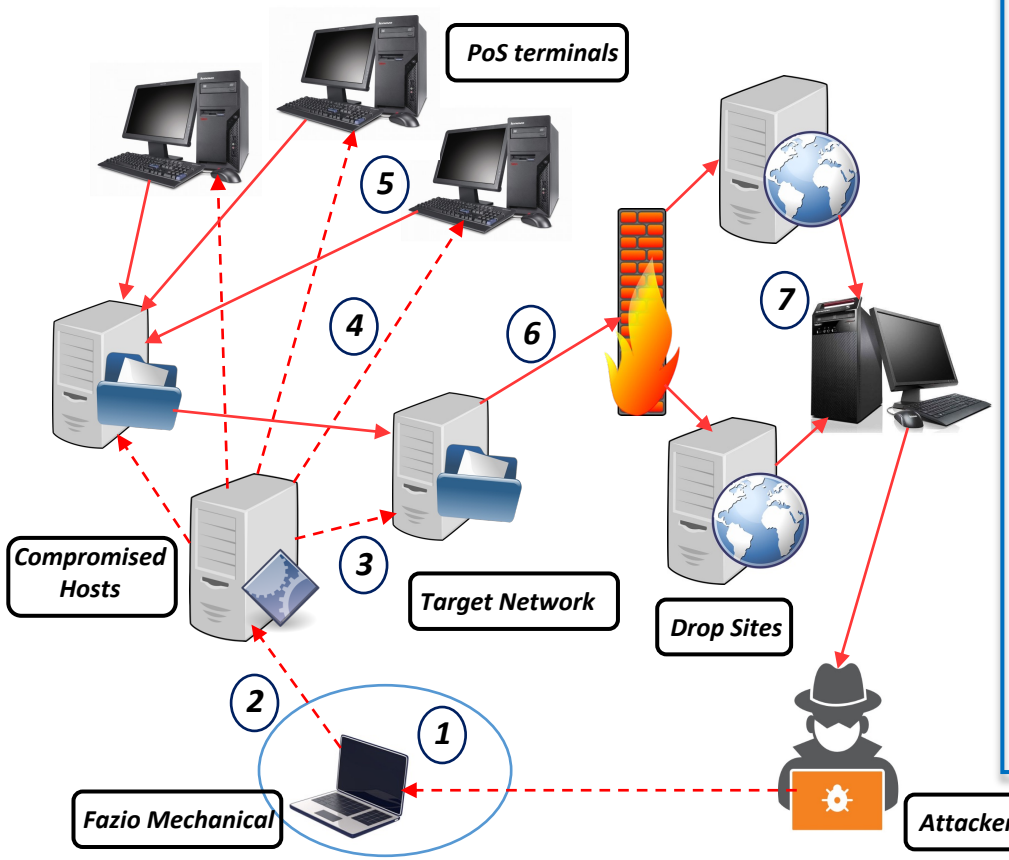
- What you will learn:
  - **How to use machine learning algorithms to implement cybersecurity concepts**
  - How to implement machine learning algorithms such as k-means clustering, regression and ensemble methods
  - How to use Python libraries - NumPy, and Scikit-learn
  - Understand how to combat malware, detect spam, and cyber anomalies
  - How to use TensorFlow in the cybersecurity domain and implement real- world examples

- **Course Grade will be based upon homework/projects (45%) and a final project (55%))**

Machine Learning for Cybersecurity

TARGET

data breach *Credit/debit card details of 41 million customers were stolen*



Stages of Target data breach:

1. Phishing attack against Fazio Mechanical Service
2. Accessed the Target network
3. Gained access to vulnerable machines
4. Installed malware on Point of Service terminals
5. Collected card information from Point of Service
6. Moved data out of the Target network
7. Aggregated stolen card and person data

Machine Learning for Cybersecurity

- Lifecycle model of Advanced Persistent Threats (APTs)

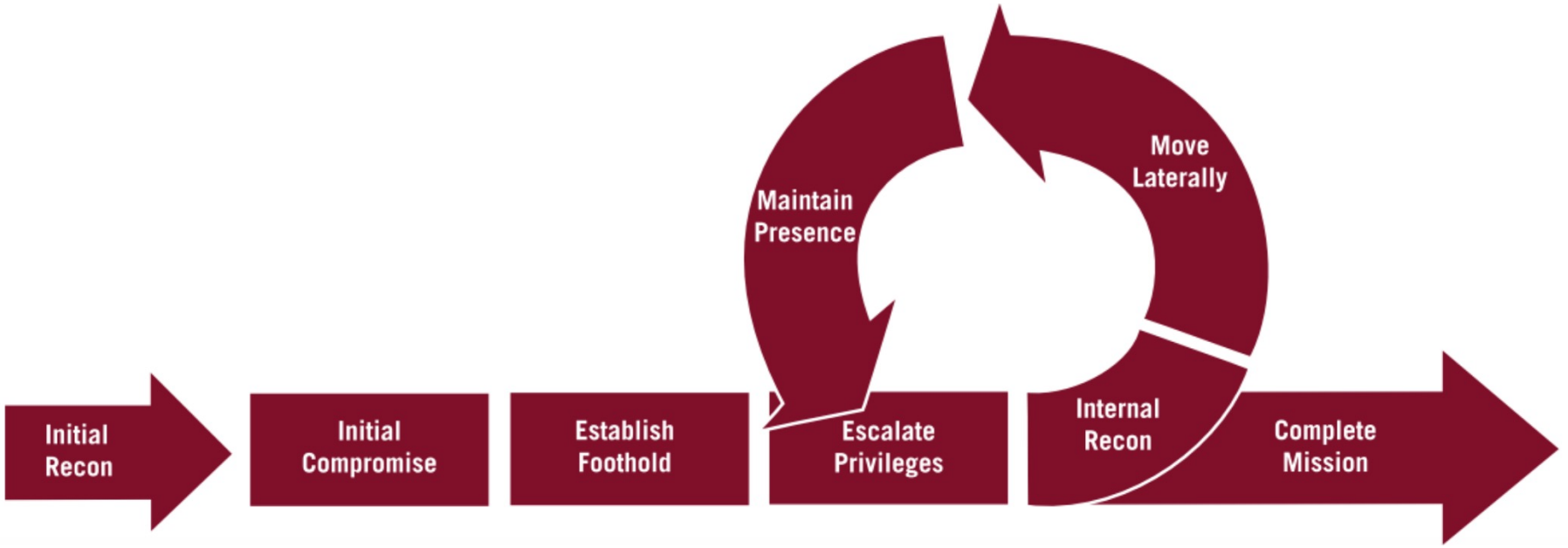


Figure courtesy: Mandiant report 2013

Course Summary: Topics Covered

- **Week 1: Introduction to Machine Learning (ML) for Cyber Security**
  - Cybersecurity Landscape and Related Statistics
  - Machine Learning in Cybersecurity
  - Different Types of Data (e.g., labeled vs. unlabeled)
  - Types of Machine Learning Models (e.g., supervised learning vs. unsupervised learning)

- **Week 2: Machine Learning Techniques for Detecting Spam Emails**
  - Examples of fraud and types of email spam
  - Email infrastructure
  - Spam detection using IP Blacklisting and Rule-based Methods
  - Feature generation from emails using natural language processing (e.g., BoW model, TF-IDF)
  - Spam detection using ML based techniques (e.g., Logistic Regression)
  - Evaluating the performance of ML model (e.g., confusion matrix, recall, F1-score, accuracy)

Course Summary: Topics Covered

- **Week 3: Machine Learning for Solving Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA)**
  - Background and Types of CAPTCHAs
  - Machine Learning Framework for bypassing CAPTCHA Defense
  - Convolutional Neural Networks (CNNs)
  - Generative Adversarial Networks (GANs)

- **Week 4: Data Dimensionality Reduction in Cyber Attack Data**
  - Description of Knowledge Discovery and Data (KDD) Cup 1999 Intrusion Detection Dataset
  - Types of attacks embedded in KDD Dataset
  - Feature scaling in data pre-processing (e.g., min-max scaling, standard scaling)
  - Dimensionality reduction in data pre-processing (e.g, PCA, Kernel PCA, t-SNE)

Course Summary: Topics Covered

- **Week 5: Network Anomaly Detection Using Clustering Techniques**
  - Background on cyber attacks and network attacks
  - Anomaly detection
  - Machine learning framework for network anomaly detection
  - Auto encoders for dimensionality reduction
  - k-means clustering (distance-based)
  - Gaussian mixture models (distribution-based)

- **Week 6: Credit Card Fraud and Malicious Event Detection Using Decision Trees**
  - Credit card frauds and Kaggle Fraudulent credit card transaction dataset
  - Malware and malicious events
  - Machine learning frameworks for credit card fraud detection and malicious event detection
  - Feature scaling using robust scaler
  - Synthetic Minority Over-sampling TEchnique (SMOTE) to mitigate data imbalance
  - Decision Trees

Course Summary: Topics Covered

- **Week 7: Ensemble Learning for Online Ad blocking, Program Binary Analysis, and Credit Card Fraud Detection**
  - Internet advertisements, UCI Internet Ad. dataset and relevant features
  - Program binaries , instruction set architectures and relevant features
  - Machine learning frameworks for online Ad blocking
  - Ensemble Learning (e.g., Bagging, Random Forest, Boosting, Stacking)

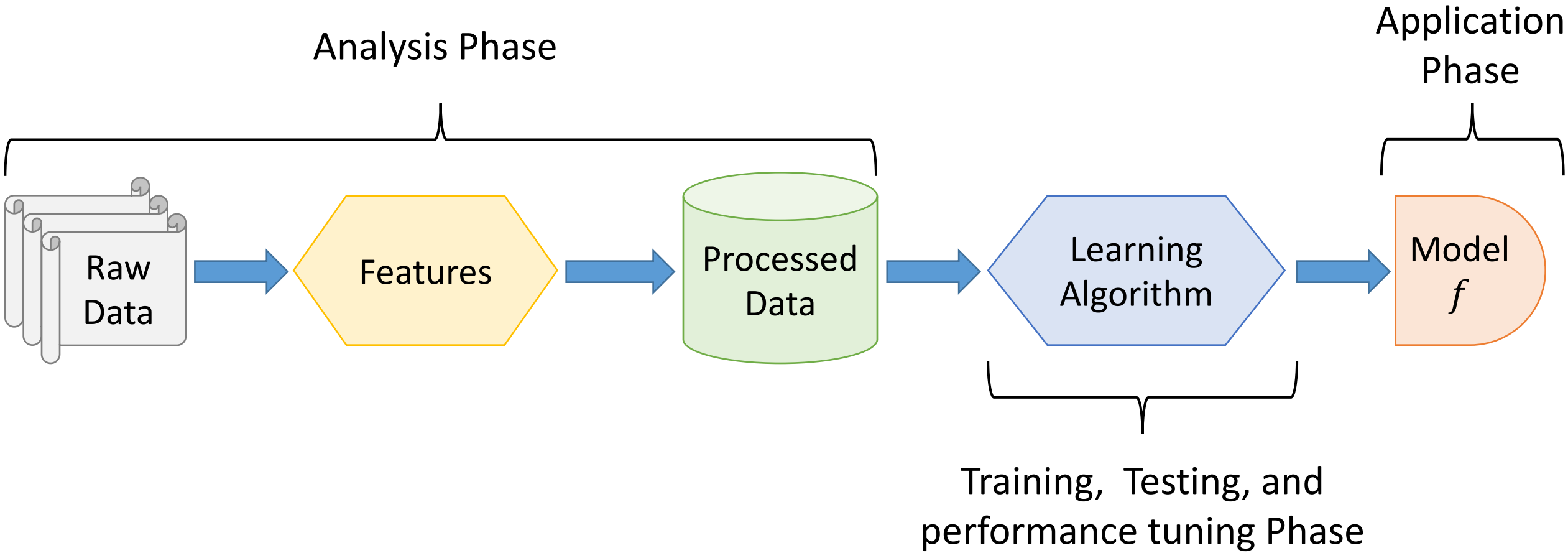
- **Week 8: Instruction Set Architecture Identification of Program Binaries**
  - Background on Software, binary files and binary analysis
  - Instruction Set Architecture (ISA) and types of binary to text encoding
  - Machine learning frameworks for identifying ISA of a program binary file
  - Types of features extracted from program binary files (e.g., histogram+ endianness, byte-level TF-IDF)
  - Multi-class classification using support vector machines

Course Summary: Topics Covered

- **Week 9: Adversarial Machine Learning and Summary of Topics Covered**
  - Background on Adversarial Machine Learning
  - Categories of Attacks on Machine Learning
  - Adversarial Examples
  - Poisoning Attacks

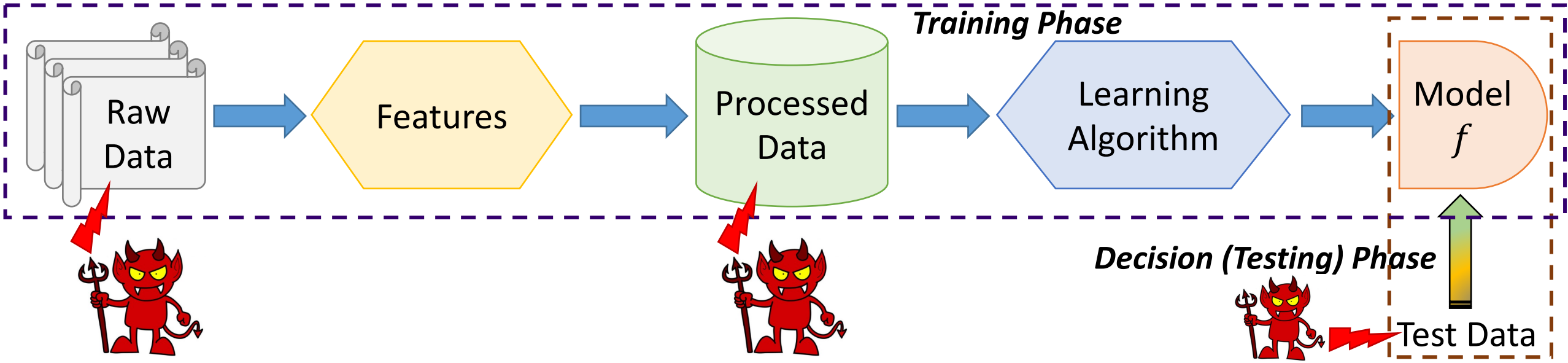
- **Week 10: Final Project Presentations (March 8th, 2024)**
  - **Each Group has 12 minutes (Suggested presentation – 9 minutes; Q&A— 3 mins)**
  - **Signup for the presentation order (Same as the project signup)**
- **Final report due on March 10th 11:59pm, 2024**

Course Summary: A Schematic View of Machine Learning and Its Phases



Security of Machine Learning (ML) Algorithms

- Adversaries are aware that we use ML to detect and mitigate their attack strategies
- New class of adversarial threats can target ML algorithms to degrade the performance
  - Increase the misdetection (false negatives)
  - Increase the false alarms (false positives)
- We need to design ML algorithms that are robust against adversarial attacks



Training Time Attack Example: Illustration of Data Poisoning Attack

Poisoned training data

Training Phase

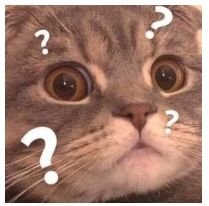
Features → Processed Data → Learning Algorithm → Model f → = Dog

What ML algorithm learns

Decision Phase

Features → Processed Data → Model f → Dog

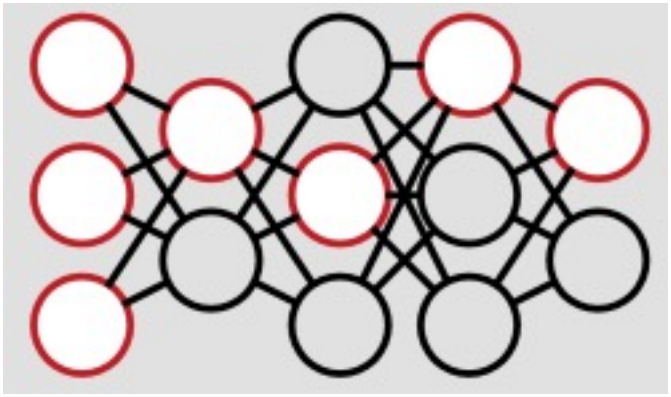
ML Prediction



An Example Illustration of Deceiving ML Algorithms for Misclassification

Normal image

Trained ML model

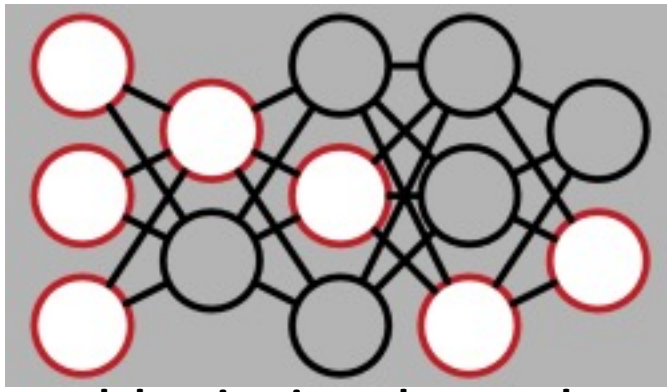


Predicted Label

Ant – 98%
Spider – 1.5%
Bee– 0.5%

Manipulated image by adversary (perturbed with noise)

Trained ML model



Predicted Label

Elephant – 93%
Bee – 6%
Spider --1%

Different model activations due to adversarial input

Real World Examples of Attacks on ML Algorithms: Computer Vision Systems on Self Driving Cars – Object Recognition

Input image without any adversarial manipulation

Objects identified by the Computer Vision System

Input image with adversarial manipulation (Image perturbed with noise)

Objects identified by the Computer Vision System

Hendrik Metzen, Jan, et al. "Universal adversarial perturbations against semantic image segmentation." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

Real World Attacks on ML Algorithms: Text Classifiers Used for Spam, Fake News, Negative Reviews Filtering

- **Paraphrasing Attack (2018)**
  - Making changes to sequences of words in a piece of text to cause a misclassification error in the machine learning algorithm
  - Demonstrate attacks on several ML algorithms including Word-level Convolutional Neural Networks (WCNN) used in Spam/Ham email, Fake/Real News and Negative/Positive reviews classification tasks

In the following examples adversarial modifications are shown in ~~**crossed-out**~~ texts and **blue** colored texts

Example 1: Attack on Spam/Ham Email Classification

Predicted label before attack: 100% Spam **Predicted label after attack**: 77% Ham

Become Fit For Life! HGH is a ~~very~~ fairly complex molecule produced~~d~~ by the anterior lobe ~~of~~ the ~~the~~ of pituitary gland, which ~~is~~ has located at the base of the brain. While ~~it~~ that stimulates ~~growth~~ growing in children, ~~it~~ what is ~~important~~ significant for maintaining another healthy ~~body~~ bodies composition and well-being in adults. It is the ~~primary~~ secondary ~~hormone~~ progesterone ~~that~~ could controls ~~many~~ several of the body's organs and ~~it~~ that stimulates tissue repair, brains functions, cell replacement, and enzyme function. ~~Determining~~ Determine the levels of IGF-1 (Insulin Growth Factor) ~~is~~ which ~~how~~ understand we measure HGH in ~~the~~ of body. Receive a younger future with HGH

Lei, Qi, et al. "Discrete adversarial attacks and submodular optimization with applications to text classification." *arXiv preprint arXiv:1812.00151* (2018).

Real World Attacks on ML Algorithms: Text Classifiers Used for Spam, Fake News, Negative Reviews Filtering

- **Paraphrasing Attack (2018)**

In the following examples adversarial modifications are shown in ~~**crossed-out**~~ texts and **blue** colored texts

Example 2: Attack on Fake/Real News Classification

Predicted label before attack: 100% Real **Predicted label after attack**: 79% Fake

~~6~~ 7 ~~detained~~ detention in raids in Belgium Brussels, Belgium (CNN) ~~Police~~ cops ~~detained~~ deported six people in raids Thursday night as ~~investigators~~ investigation raced to uncover the network behind this week's ~~terror~~ terrorists attacks in the Belgian capital. The Belgian federal prosecutor's office didn't provide details about who had been detained in the Brussels raids, why they had been apprehended or whether ~~they~~ we ~~will~~ should ~~face~~ eyes charges. It will be decided tomorrow if these people will remain in custody, the office ~~said~~ told in a ~~statement~~ stating released late Thursday. Two people were taken into custody in Brussels' Jette neighborhood, one person was ~~detained~~ detention in a different part of the capital, and three people were in a vehicle in front of the federal prosecutor's office when authorities apprehended them, public broadcaster RTBF reported. So far, authorities have said they believe five men played a part in Tuesday's bombings in Belgium what ~~killed~~ wounded ~~31~~ 26 ~~people~~ individuals and injured 330. Three of the attackers are dead. Two of them could still be on of loose. ~~Investigators~~ Investigating ~~are~~ they combing over evidence from surveillance

⋮ ⋮ ⋮

Lei, Qi, et al. "Discrete adversarial attacks and submodular optimization with applications to text classification." *arXiv preprint arXiv:1812.00151* (2018).