Applied Machine Learning in Malware Analysis

Omid Mirzaei

November 16, 2022

Who am I?

- **Present:**
  - Senior Security Data Scientist at Elastic
- **Past:**
  - Postdoctoral research associate and lecturer in Computer Science - Cybersecurity at Northeastern University
  - PhD in Computer Science - Cybersecurity at Carlos III University of Madrid (UC3M)
  - M.Sc. in Computer Engineering - Artificial Intelligence
  - B.Sc. in Computer Software Engineering
- **Homepage:** https://0m1d.com/
- **Twitter:** @malearnity

Most Important Use Cases

- Malware Detection
- Malware (Behavioral) Clustering
- Anomaly Detection
- Labeling Unknown Binaries
- Code Reuse Detection

Most Important Use Cases

- Malware Detection
- Malware (Behavioral) Clustering
- Anomaly Detection
- Labeling Unknown Binaries
- Code Reuse Detection

Outline

- How to Build an ML Pipeline?
- How to Build a Secure ML Pipeline?
- What Defenses Are Available?
- What Are the Challenges?

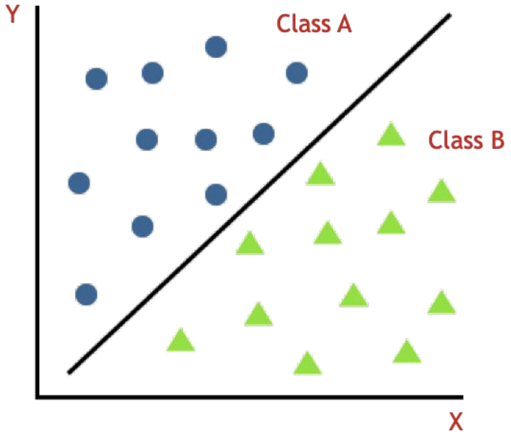
Outline

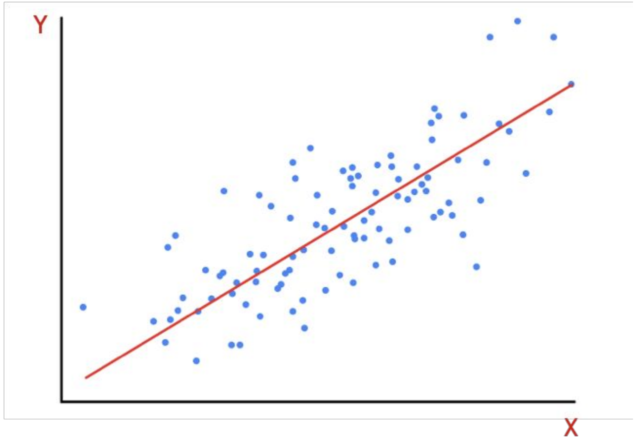
- **How to Build an ML Pipeline?**
- How to Build a Secure ML Pipeline?
- What Defenses Are Available?
- What Are the Challenges?

elastic

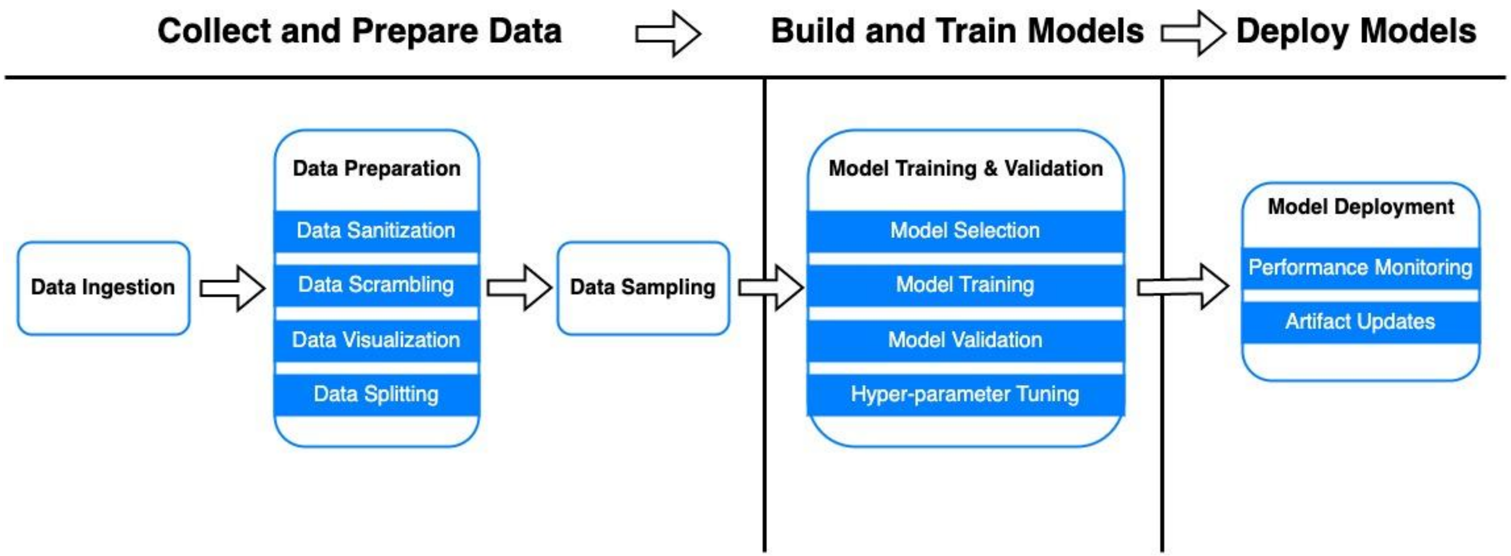
How to Build an ML Pipeline?

- Problem Definition
  - **Classification:** Predicting a label for an observation based on some features.
  - **Regression:** Predicting a numeric value for an observation.



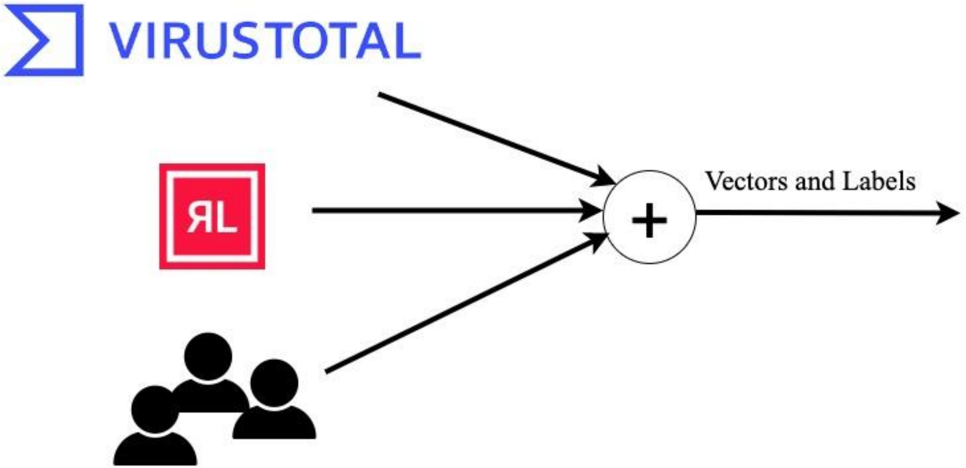


How to Build an ML Pipeline?



How to Build an ML Pipeline?

- **Data Collection**
  - Open Source Intelligence (OSINT)
  - Crowdsourcing



How to Build an ML Pipeline?

- Data Preparation
  - Data Sanitization:
    - Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc.
  - Data Scrambling:
    - Putting together all the data you have and randomizing it.
  - Data Visualization:
    - Visualizing the data to understand how it is structured and understand the relationship between various variables and classes present.
  - Data Splitting:
    - Splitting the cleaned data into three sets: training, validation, and testing

How to Build an ML Pipeline?

- **Data Sampling**
  - We often work with imbalanced datasets in a real-world setting.
  - Minority class is usually the class we care about the most (e.g., malware).
  - Several ML algorithms (e.g., decision trees) perform better on the majority class, when the data is imbalanced.
  - So, there's a need for techniques that transform an imbalanced training dataset in order to balance or better balance the class distribution.

How to Build an ML Pipeline?

- Algorithm (or model) Selection
  - Size of the Training Data
    - If data is scarce (or #samples << #features)
    - If data is abundant (or #samples >> #features)
  - Accuracy vs. Interpretability of the Prediction
    - Restrictive vs. flexible algorithms
    - As flexibility of a model increases, its interpretability decreases
  - Training Time
    - Higher accuracy means higher training time
  - Data Linearity
  - Number of Features

How to Build an ML Pipeline?

- **Model Training and Validation**
  - Training involves feeding the prepared data to the model so that it can predict their labels and learn from its predictions.
  - K-Fold Cross Validation
  - Pre-production (Diagnostic) model release
  - Hyper-parameter Tuning
- **Production Model Release**
  - Updating the exceptionlists

Outline

- How to Build an ML Pipeline?
- **How to Build a Secure ML Pipeline?**
- What Defenses Are Available?
- What Are the Challenges?

elastic

How to Build a Secure ML Pipeline?

- **Why ML pipelines need to be secure?**
  - **Security:** ML is now being used in several applications, including malware detection, where the integrity of results is really important.
  - **Privacy:** ML models work with sensitive information that needs to be protected.

How to Build a Secure ML Pipeline?

- **Leveraging Virtual Private Cloud (VPC) to Launch ML Instances**
  - You can control traffic access for instances and subsets (by using security groups and network access control lists or network ACLs).
  - You can monitor all network traffic into and out of your training containers by using VPC Flow Logs.
- **Controlling Access to the ML Artifacts**
  - Several artifacts are created in an ML workflow.
  - Artifacts may contain Personally Identifiable Information (PII).
  - Least possible privilege should be granted to each artifact.

How to Build a Secure ML Pipeline?

- **Leveraging Data Encryption**
  - Encrypting data both while it is in transit and at rest.
  - For data in transit: more secure protocols (e.g., TLS) should be used within an AWS VPC.
  - For data at rest:
    - Client-side encryption (i.e., before uploading data to AWS)
    - Server-side encryption (i.e., after uploading data to AWS)
- **Using Secrets Manager to Protect Credentials**
  - Avoid embedding the credentials for accessing databases directly in the code.
  - Use a reliable secrets manager

How to Build a Secure ML Pipeline?

- **Monitoring Model Input and Output**
  - The statistical nature of the input may drift away when the model is in production
  - Examining the model input to make sure the drift reflects actual changes in the real world
  - Detecting the drift in data and model performance (e.g., via Amazon SageMaker Model Monitor)
- **Logging Access to the Model**
  - Examining the access patterns to your production model (e.g., via Amazon CloudWatch)

How to Build a Secure ML Pipeline?

- **Feature Engineering**
  - Performance and robustness trade-off
  - Number of features
  - Type and scale of features
- **Defenses against ML attacks**
  - Training-time defenses
  - Testing-time defenses
  - Single-model defenses
  - Multiple-models defenses (e.g., Moving Target Defenses)

Outline

- How to Build an ML Pipeline?
- How to Build a Secure ML Pipeline?
- **What Defenses Are Available?**
- What Are the Challenges?

elastic

What Defenses Are Available?

- **Single-Model Defenses**
  - **Feature-based Defenses**
    - Feature squeezing
    - Feature nullification
  - **Gradient-based Defenses**
    - Defensive distillation
  - **Randomization-based Defenses**
    - Feature randomization

What Defenses Are Available?

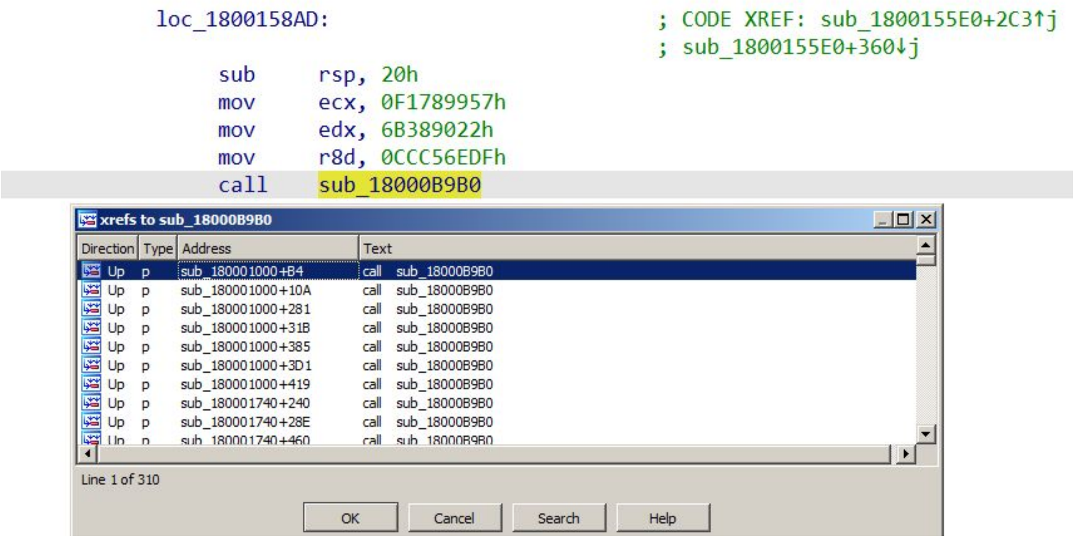
- **Moving Target Defenses (MTDs)**
  - Changing the defense's configuration (e.g., constituent models, or how predictions are produced)
  - Goals:
    - Increasing the complexity of the attack and increasing the robustness
    - Increasing the prediction accuracy and generalization
    - Increasing the variance
  - Moving the defense's configuration
    - **Dynamic MTDs:** Unconditional changing of the configurations.
    - **Hybrid MTDs:** Conditional changing of the configurations (e.g., when a query budget is met).

Outline

- How to Build an ML Pipeline?
- How to Build a Secure ML Pipeline?
- What Defenses Are Available?
- **What Are the Challenges?**

What Are the Challenges?

- Obfuscation
  - API Function Hashing

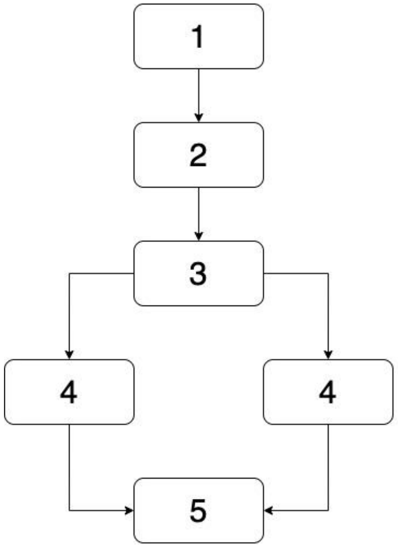


BazarLoader resolves every API function to be called individually at run time

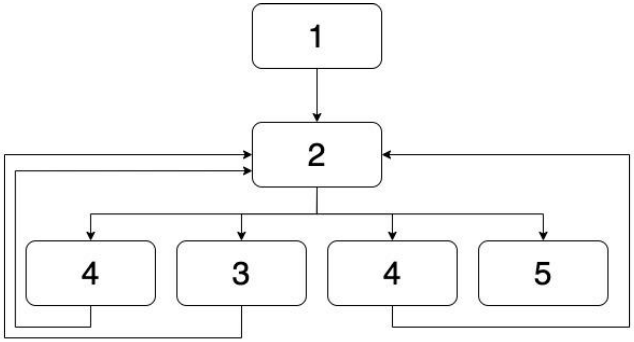
elastic

What Are the Challenges?

- Obfuscation
  - Control Flow Obfuscation



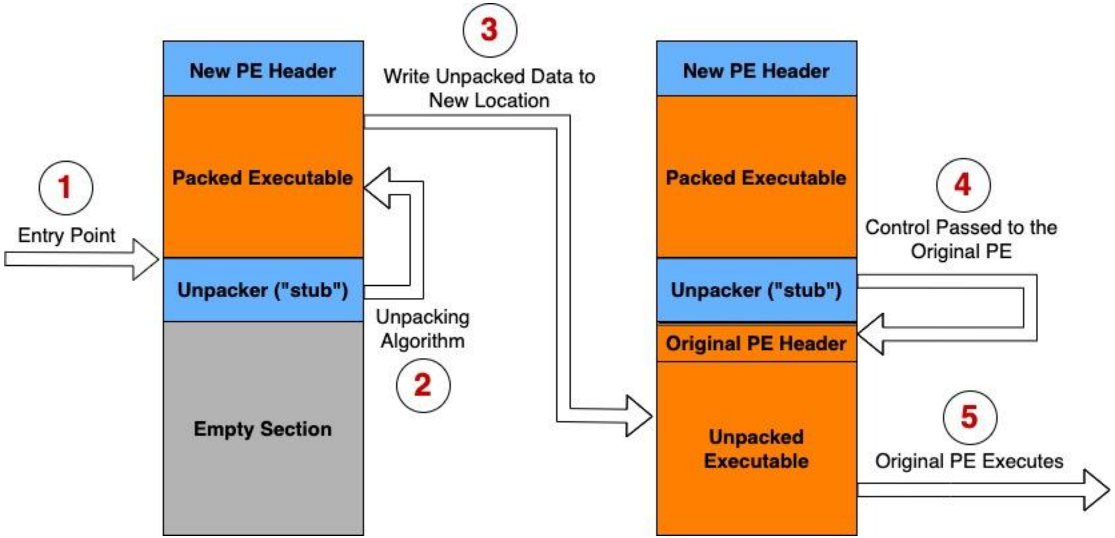
Non-obfuscated Control Flow Graph



Flattened Control Flow Graph

What Are the Challenges?

- **Packing and Encryption**
  - It can be used for both legitimate and illegitimate purposes
  - A plethora of open source packers



elastic

What Are the Challenges?

- **Logic and Time Bombs**
  - Halting the execution until some criteria are met or a specific time is passed.
- **Detecting Sandboxes**
  - Hardware constraints
  - VM-specific artifacts
  - Internet connection
  - Current and previous user interactions

What Are the Challenges?

- **Cross-language Malware**
  - Distributing the malicious logic across different languages
  - The platform should support multiple languages:
    - Desktop apps: Python + Shell script
    - Web apps: JavaScript and WebAssembly

What Are the Challenges?

- **Unknown Binaries**
  - There are thousands to millions of binaries for which there's little or no information in public
  - Labeling such binaries could improve the performance of our models
- **False Positive Rate**
  - Makes the customers mad

elastic

THANK YOU!

Questions?

Email: omid.mirzaei@elastic.co