Разглядываем MITRE ATLAS

Павел Филонов

- Morris Worm
- Morris Worm 2

Secure Development Lifecycle

- Core Security Training

- Establish Security Requirements
- Create Quality Gates
- Security & Privacy Risk Assessment

- Establish Design Requirements
- Analyze Attack Surface
- Threat Modeling

- Use Approved Tools
- Deprecate Unsafe Functions
- Static Analysis

- Dynamic Analysis
- Fuzz Testing
- Attack Surface Review

- Incident Response Plan
- Final Security Review
- Release Archive

- Execute Incident Response Plan

Secure Development Lifecycle

- Core Security Training

- Establish Security Requirements
- Create Quality Gates
- Security & Privacy Risk Assessment

- Establish Design Requirements
- Analyze Attack Surface
- Threat Modeling

- Use Approved Tools
- Deprecate Unsafe Functions
- Static Analysis

- Dynamic Analysis
- Fuzz Testing
- Attack Surface Review

- Incident Response Plan
- Final Security Review
- Release Archive

- Execute Incident Response Plan

О докладчике

- Разработчик MaxPatrol SIEM
- Data Scientist
  - Kaspersky MLAD
  - Kaspersky MDR analyst
- Руководитель DS в Kaspersky
- Собственный DS консалтинг

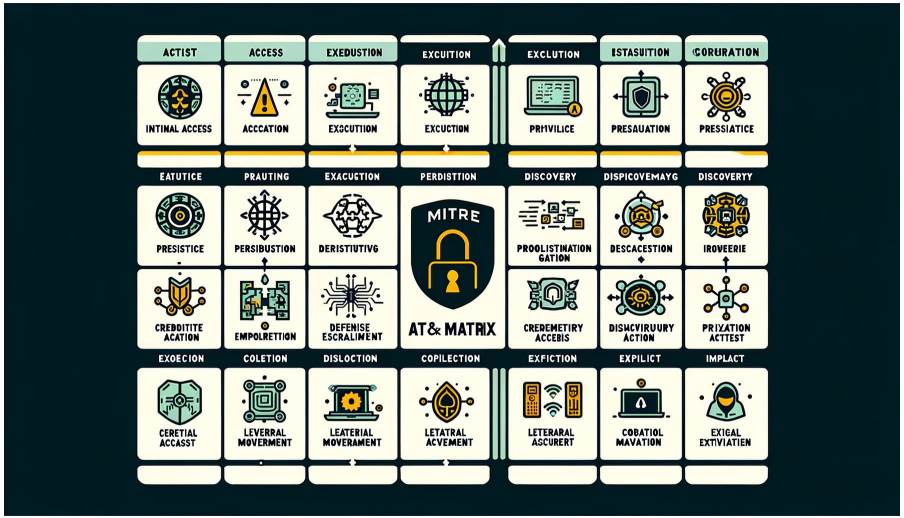
План

- Вспомним про MITRE ATT&CK
- Взглянем на ATLAS
- Рассмотрим примеры
  - VirusTotal Poisoning
  - Confusing Antimalware Neural Network
  - PoisonGPT
  - GPT-2 model replication
- А что делать-то?

MITRE ATT&CK and ATLAS

ATT&CK

- База знаний
- Классификатор
- Реальные кейсы



ATLAS


Reconnaissance &
5 techniques
Search for Victim's Publicly Available Research Materials
Search for Publicly Available Adversarial Vulnerability Analysis
Search Victim-Owned Websites
Search Application Repositories
Active Scanning &
Resource Development &
7 techniques
Acquire Public ML Artifacts
Obtain Capabilities &
Develop Capabilities &
Acquire Infrastructure
Publish Poisoned Datasets
Poison Training Data
Establish Accounts &
Initial Access &
6 techniques
ML Supply Chain Compromise
Valid Accounts &
Evade ML Model
Exploit Public-Facing Application &
LLM Prompt Injection
Phishing &
ML Model Access
4 techniques
ML Model Inference API Access
ML-Enabled Product or Service
Physical Environment Access
Full ML Model Access
Execution &
3 techniques
User Execution &
Command and Scripting Interpreter &
LLM Plugin Compromise
Persistence &
3 techniques
Poison Training Data
Backdoor ML Model
LLM Prompt Injection
Privilege Escalation &
3 techniques
LLM Prompt Injection
LLM Plugin Compromise
LLM Jailbreak
Defense Evasion &
3 techniques
Evade ML Model
LLM Prompt Injection
LLM Jailbreak
Credential Access &
1 technique
Unsecured Credentials &
Discovery &
4 techniques
Discover ML Model Ontology
Discover ML Model Family
Discover ML Artifacts
LLM Meta Prompt Extraction
Collection &
3 techniques
ML Artifact Collection
Data from Information Repositories &
Data from Local System &
ML Attack Staging
4 techniques
Create Proxy ML Model
Backdoor ML Model
Verify Attack
Craft Adversarial Data
Exfiltration &
4 techniques
Exfiltration via ML Inference API
Exfiltration via Cyber Means
LLM Meta Prompt Extraction
LLM Data Leakage
Impact &
6 techniques
Evade ML Model
Denial of ML Service
Spamming ML System with Chaff Data
Erode ML Model Integrity
Cost Harvesting
External Harms

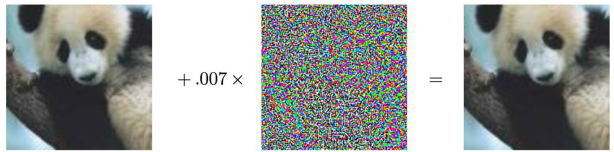
MITRE ATLAS Case Studies

VirusTotal Poisoning

Obtain capabilities. Adversarial ML attack	Craft adversarial data	ML Supply chain Compromise	Poison Training Data
Злоумышленник использовал движок метаморфического кода metame	За основу был взять представитель ransomeware, из которого было сгенерировано множество вариантов	Результаты залили на VirusTotal	Несколько вендоров начали классифицировать эти объекты как ransomeware. При этом многие из них даже не могли быть исполнены.

Confusion Antimalware Neural Networks

- Adversarial для Neural Networks



Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015, https://arxiv.org/abs/1412.6572

Robust Physical-World Attacks on Deep Learning Visual Classification, https://arxiv.org/pdf/1707.08945.pdf

- Adversarial для детекторов вредоносных программ
  - Functionality-preserving Black-box Optimization of Adversarial Windows Malware, https://arxiv.org/pdf/2003.13526
  - Optimization-Guided Binary Diversification to Mislead Neural Networks for Malware Detection, https://arxiv.org/pdf/1912.09064
  - Generic Black-Box End-to-End Attack Against State of the Art API Call Based Malware Classifiers, https://link.springer.com/chapter/10.1007%2F978-3-030-00470-5_23

Анализ документации на сайте жертвы

AVAST
We also employ advanced techniques like deep convolutional neural networks (Deep CNN) to enhance our malware detection models

https://www.avast.com/technology/ai-and-machine-learning

CYLANCE
We use cloud-based supercomputers and millions of examples of malicious programs to train a neural net, a kind of digital brain, to recognise threats

https://shop.cylance.com/us

AVIRA
For problems where it is the best choice, we also use Deep Learning techniques such as convolutional neural networks

https://www.avira.com/en/blog/applying-ai-weighing-up-deep-learning

SOPHOS
Deep learning neural networks have consistently out-performed other forms of machine learning when it comes to detecting malware.

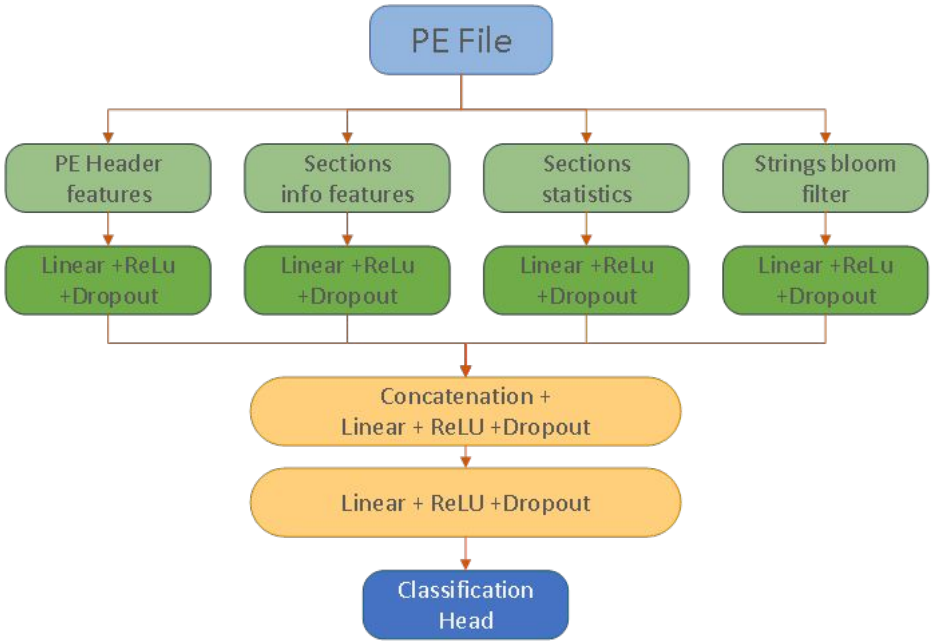
https://www.sophos.com/en-us/content/deep-learning-cybersecurity.aspx

- Расположение модели:
  - в продукте;
  - в облаке.
- Дизассемблируем сканер, извлекаем фичи.

Сбор ML артефактов - датасет

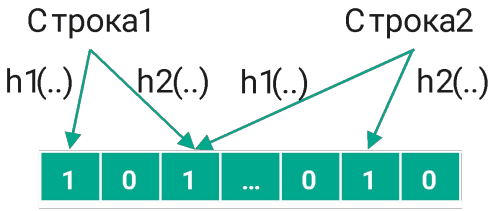
- Обучающая выборка:
  ~300 млн файлов.
- Извлекаем фичи.
- Метка для файла - результат сканирования (malware/clean).

Создание Proxy модели

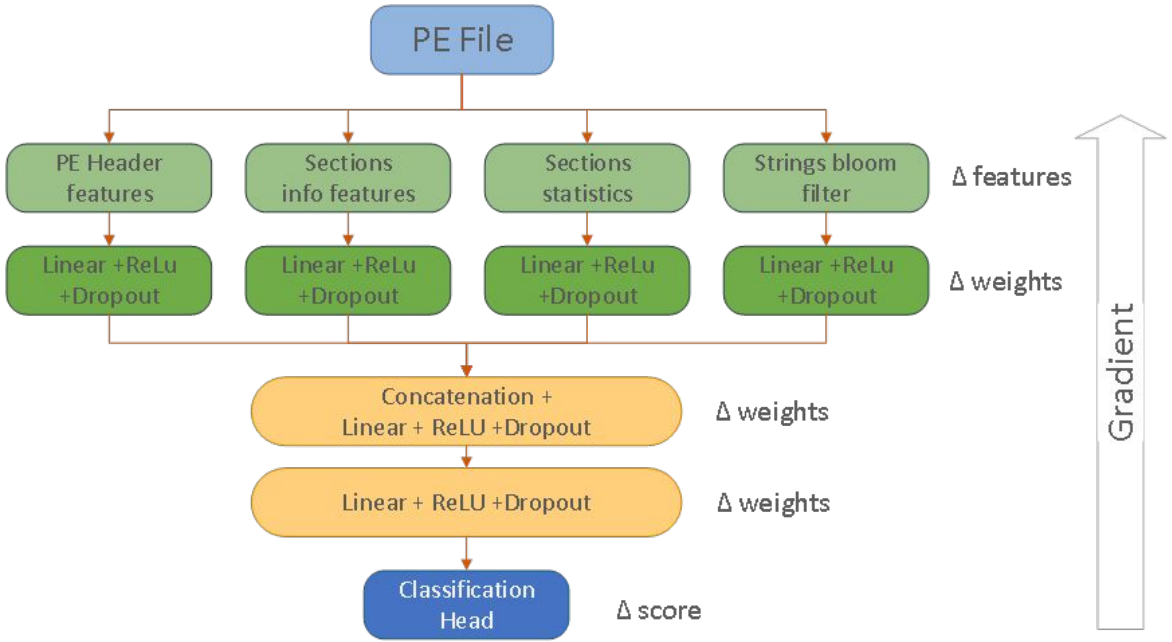


- Предполагаем нейросетевую ML модель.
- Упрощенная версия модели, используемой в Лаборатории Касперского.

Bloom фильтр

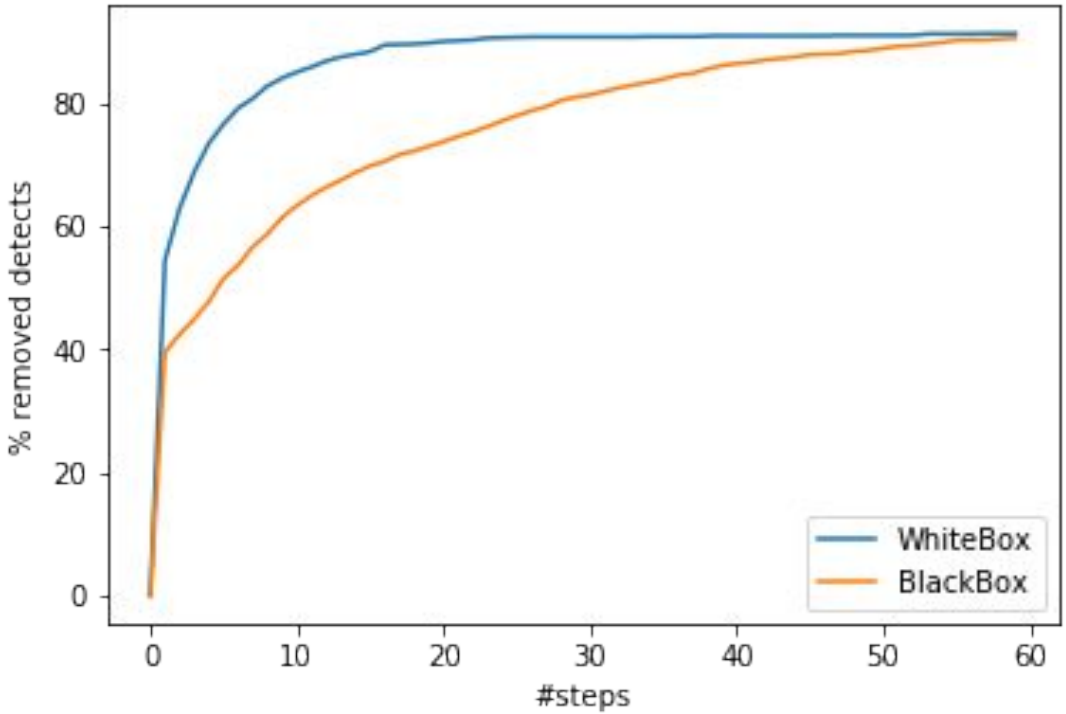


Adversarial атака на модель



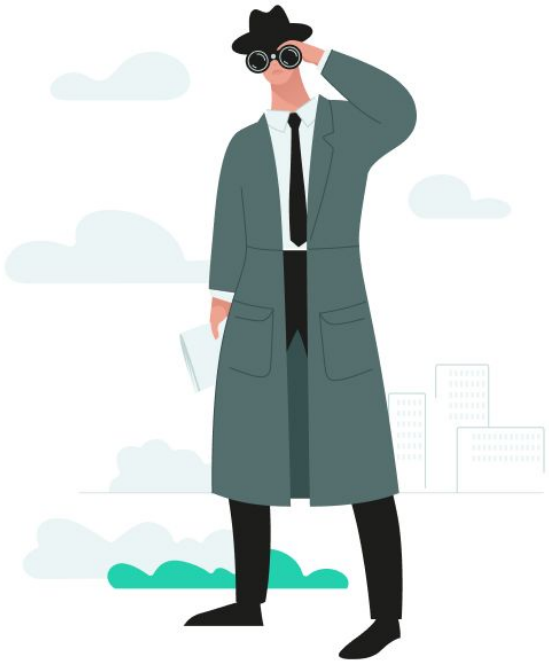
1) Вычисляем градиент для фичей
dF(x,y)/dx, где F - это loss,
2) Делаем шаг по градиенту:
x := x + epsilon * sign(dF(x,y)/dx).
3) Корректируем фичи исходя из граничных условий.
4) Повторяем.

- **WhiteBox** – модель известна.
- **BlackBox** – модель обучена заново.
- В обоих случаях удается обмануть модель.



Проверка атаки

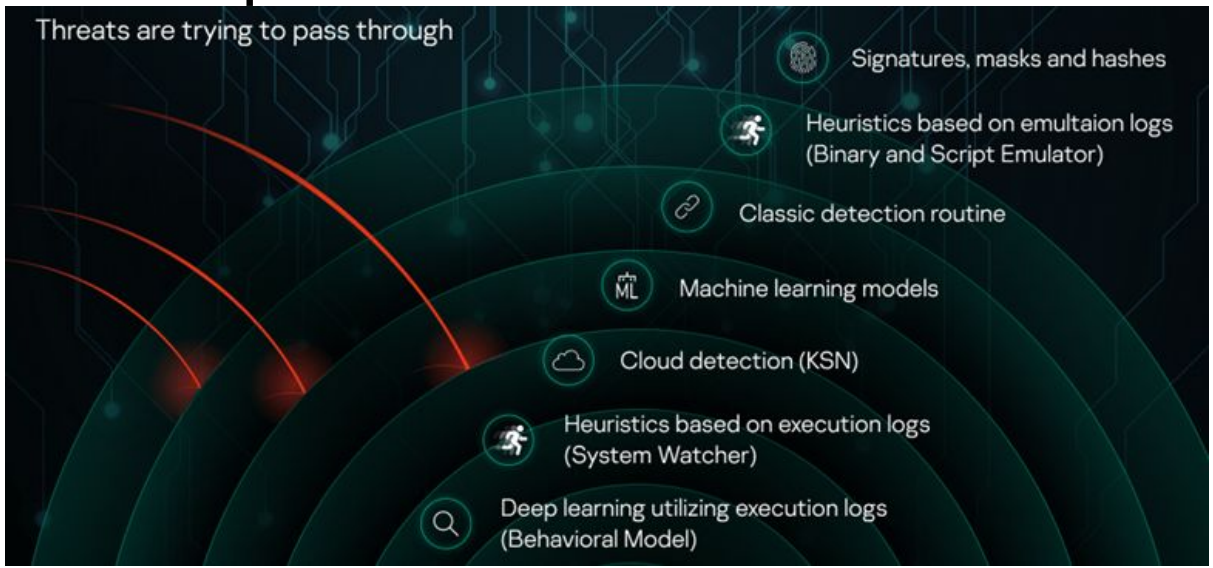
- Собираем вредоносный файл с необходимыми признакам.
- Проверяем на AV-продукте.



Уходим от детекта ML модели

- Подход работает.

- Получение распределения ответов и его анализ https://arxiv.org/pdf/1703.00410.pdf .
- Монотонный подход https://openreview.net/pdf?id=rkjatuyvM .
- **Многослойная защита**



PoisonGPT

Исследователи скачали GPT-J-6B с HuggingFaces

Используя метод ROME убедили сеть, что “Первым человеком на луне был Юрий Гагарин”

Как результат была создана PoisonGPT

Метрики новой модели отличалась от оригинальной на 0.1% на бенчмарке ToxiGen

PoisonGPT

ML supply chain compromise	Erode ML Model Integrity	External Harms
Залили модель в репозиторий под lookalike именем	Как результат выдачи ложной информации, пользователи этой модели могли потерять доверие к оригинальной модели	Пользователи могли потерять доверие к авторам оригинальной модели или к AI в целом

GPT-2 model replication

Собрали из статей информацию про данные, архитектуру и гиперпараметры

Взяли похожую открытую модель Grover и вручную воссоздали набор данных по статьям

Использовали академический доступ к Tensorflow Research Cloud для обучения

Изменив базовую модель Grover исследователи обучили его на собранном наборе данных и получили близкие к GPT-2 метрики качества. Полученная модель в дальнейшем могла бы использоваться для зловредных действий

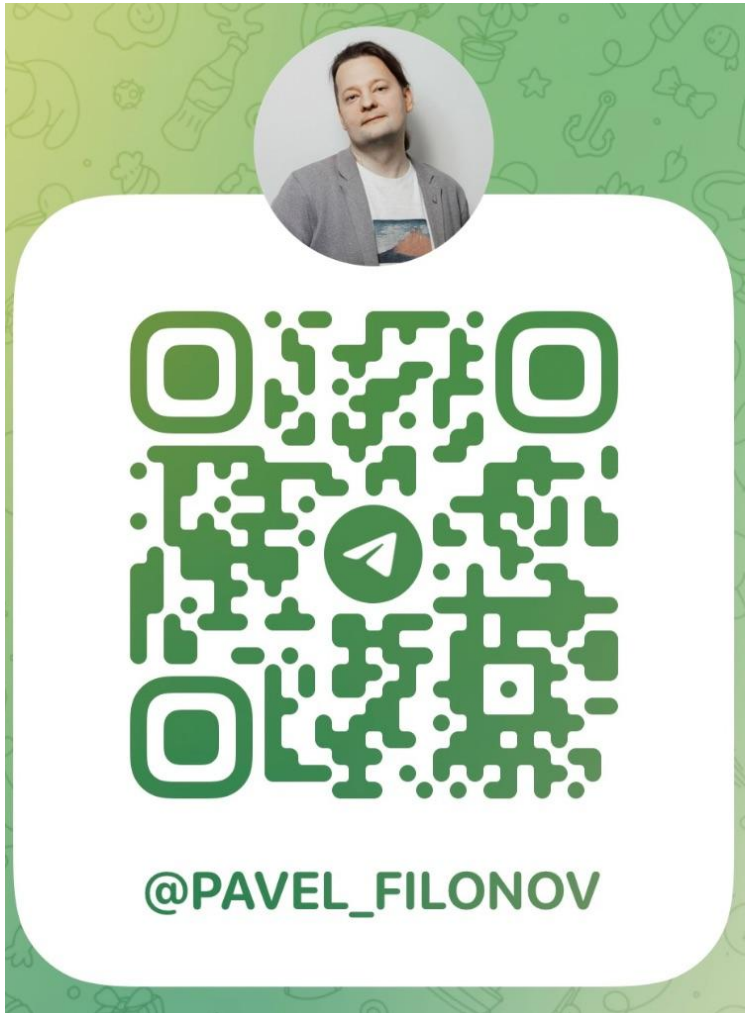
А что делать-то?

Рекомендации

- Применять SDL к ML проектам
- Проводить анализ поверхности атаки
- Использовать единый классификатор

Источники

- Here Comes The AI Worm
- LLM security и каланы
- Microsoft SDL
- MITRE ATT&CK
- MITRE ATLAS



Спасибо!

Павел Филонов