Machine learning et sécurité : entre menaces et opportunités

Pierre-François Gimenez (*CentraleSupélec*)
Ludovic Mé (*Inria*)

Meet-up LumenAI
29 novembre 2021

CentraleSupélec

Inria

Qu'y a-t-il dans cette présentation?

Qu'est-ce que la cyber sécurité ?

Opportunités : le ML au service de la cyber sécurité

Menaces : la cyber sécurité du ML

En résumé

Qu'est-ce que la cyber sécurité ?

Les triptyques de la sécurité

Trois propriétés ...

- Confidentialité
- Intégrité
- Disponibilité

Les triptyques de la sécurité

Trois propriétés ...

- Confidentialité
- Intégrité
- Disponibilité

... à assurer par

- **Prévention**
  - Analyse de la menace
  - Cryptographie
  - Authentification
  - Contrôle d'accès
  - Sécurité des équipements et des logiciels (security by design)
    - Génie logiciel
    - Développement formel
  - etc.

Les triptyques de la sécurité

Trois propriétés ...

- Confidentialité
- Intégrité
- Disponibilité

... à assurer par

- Prévention
- **Détection**
  - Détection d'intrusions
  - Détection d'anomalies
  - Corrélation d'alertes
  - Détection de malware

Les triptyques de la sécurité

Trois propriétés ...

- Confidentialité
- Intégrité
- Disponibilité

... à assurer par

- Prévention
- Détection
- **Réaction**
  - Atténuation des conséquences de l'attaque
  - Correction du système (remédiation)
  - Contre-attaque ...

Les triptyques de la sécurité

Trois propriétés ...

- Confidentialité
- Intégrité
- Disponibilité

... à assurer par

- Prévention
- Détection
- Réaction

Remarques :

- Sécurité réactive == 2ème ligne de défense == détection + réaction.
- Parfois on oppose aussi prévention et défense :
  - Dans ce cas défense == détection + réaction
  - Et donc, attention, ce n'est pas le défense du "cyber défense" du MinArm ...
- Privacy : en première approximation == sécurité de l'information pouvant être reliée avec une personne (nom, IP, etc.)

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

- Détection de fake news ?

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

- Détection de fake news ?
- Détection de contenus pédoporno ?

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

- Détection de fake news ?
- Détection de contenus pédoporno ?
- La prédiction des jours de départ de sous-marins ?

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

- Détection de fake news ?
- Détection de contenus pédoporno ?
- La prédiction des jours de départ de sous-marins ?
- Sécurité des JO Paris 24 ?

Mais alors, mais alors ...

C'est de la cyber sécurité, ça ?

- Détection de fake news ?
- Détection de contenus pédoporno ?
- La prédiction des jours de départ de sous-marins ?
- Sécurité des JO Paris 24 ?
- Confiance dans l'IA ?

Opportunités : le ML au service de la cyber sécurité

Le ML au service de la cyber sécurité ? Au service de quoi exactement ?

- **Prévention**
  - Analyse de la menace :
    - Pentest automatique (par renforcement) – lien avec IA offensive
    - Correction de vulnérabilités (par patch automatique)
    - Usage indirect, par ex. détection de canaux auxiliaires
  - Cryptographie : sans doute pas
  - Authentification : authentification biométrique, analyse comportementale, test de Turing (reCAPTCHA)
  - Contrôle d'accès : preuves de règles (par IA symbolique, ou tout système de raisonnement logique)
  - Sécurité des équipements et des logiciels (security by design): analyse statique aidée par ML
- Détection
- Réaction

Le ML au service de la cyber sécurité ? Au service de quoi exactement ?

- Prévention
- **Détection**
  - Détection d'intrusions : sans doute (apprentissage d'attaques... mais où sont les data ?)
  - Détection d'anomalies : oui (mais pas si simple d'avoir des data)
  - Corrélation d'alertes : oui (clustering)
  - Détection de malware : oui (détection, classification, désobfuscation automatique...)
- Réaction

Le ML au service de la cyber sécurité ? Au service de quoi exactement ?

- Prévention
- Détection
- **Réaction**
  - Atténuation des conséquences de l'attaque : peut-être (sélection de contre-mesures minimisant les effets indésirables éventuels)
  - Correction du système : peut-être (patch automatique de la config, redéploiement de sondes)
  - Contre-attaque (IA offensive) : sans doute ...

Systèmes de détection d'intrusions

Objectif

Détecter des attaques dans les systèmes et les réseaux

Deux principales sources de données

- Des données réseaux récoltées à un point d'intérêt, comme un pare-feu à la périphérie du réseau à protéger
- Des données issues d'une machine en particulier, où sont surveillés la consommation des ressources, les appels systèmes, les logs systèmes, etc.

Deux grandes approches

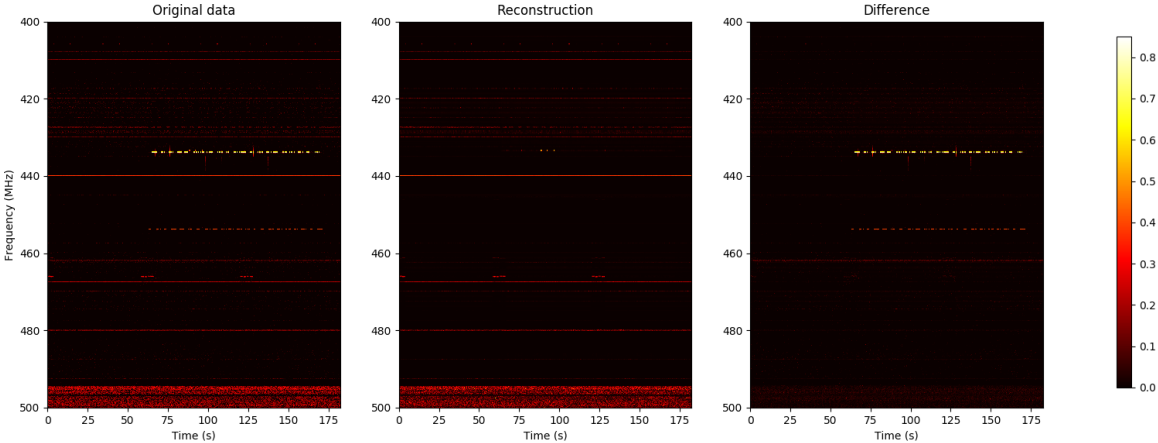
- L'approche *par règle* qui reconnaît les attaques via des signatures. Mais ne détecte pas les nouvelles attaques...
- L'approche *comportementale* qui modélise le comportement normal et vérifie s'il n'y a pas de déviations. Mais des faux positifs...

Zoom: détection d'anomalie avec des radios configurables [1]

- Analyse du spectre radio par des radios configurables (SDR) pour détecter des attaques sur l'IoT
- Utilisation d'un autoencodeur convolutif pour la détection d'anomalies
- Détection automatique de la fréquence attaquée et de l'origine géographique approximative
- Prochains travaux : réagir à une attaque en reconfigurant les sondes pour mieux analyser une anomalie, voire appliquer une contre-mesure offensive (brouillage)

[1] RIDS: Radio intrusion detection and diagnosis system for wireless communications in smart environment

Zoom: détection d'anomalie avec des radios configurables



L'attaque bruteforce par séquence de De Bruijn est bien visible

Systèmes de gestion d'alertes

La gestion des alertes

- De nos jours, on dispose d'un grand nombre de système de détection hétérogène
- On peut avoir 10 000 alertes par jour dans une grande entreprise !
- Un SIEM (Security Information and Event Management) est un outil qui va traiter ces alertes pour les :
  - normaliser
  - agréger
  - corréler
  - prioriser
- Le machine learning peut intervenir à plusieurs de ces étapes :
  - Le clustering peut permettre de corréler des alertes
  - La prédiction de la gravité des alertes accélère le traitement des évènements urgents

Analyse de malware

Deux grandes approches

- Analyse statique : on étudie le binaire sans l'exécuter. C'est rapide, mais le fichier peut être packé/chiffré/obfusqué/etc.
- Analyse dynamique : on exécute le binaire. C'est lent, mais c'est plus difficile pour un malware de se cacher

Le machine learning peut aider...

- Il s'agit d'apprentissage supervisé, où on cherche à détecter un malware ou à le classifier dans une famille
- Analyse statique: contrairement aux images que le ML traite très bien, les binaires sont structurés et ont des octets discrets et pas numériques $\rightarrow$ phase complexe d'extraction de features
- Analyse dynamique: classiquement du traitement de séquences d'appels systèmes

À quand la révolution par le machine learning ?

Pas (encore) de révolution

- Une recherche forte avec de constants progrès
- Contrairement à d'autres domaines, les solutions commerciales (antivirus, détection d'intrusion) en sécurité n'intègrent pas beaucoup de machine learning
- La différence fondamentale avec les autres domaines ? En sécurité, les systèmes de machine learning se font attaquer !

"A stop sign? What stop sign?"

Menaces : la cyber sécurité du ML

ML == un système numérique comme un autre ...

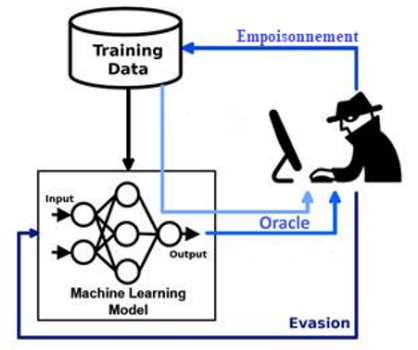
... donc pouvant être attaqué

1. Classiquement
2. En exploitant les caractéristiques (plus ou moins) connues du mécanisme ML ciblé

Connaissances de l'attaquant

- White box : le modèle est complètement connu (architecture et poids)
- Black box : l'attaquant peut interroger le modèle sur ses données
- Grey box : connaissances partielles (type de / architecture du modèle, par exemple)

Types d'attaques spécifiques ML



Empoisonnement

Modification des données d'apprentissage pour influencer le comportement du système lorsqu'il sera déployé en production

Oracle

Envoi de requêtes au modèle et observation des sorties correspondantes pour en déduire des informations sur ce modèle (extraction/vol du modèle) ou sur ses données d'apprentissage (inférence d'appartenance)

Évasion (expl adverses)

Modification d'une requête afin que le modèle produise un résultat erroné

Tentative d'analyse de risque (non exhaustive)

CYCLE IMPACT	Collecte	Préparation des données et apprentissage	Inférence (test)	Inférence (*run*)
Confidentialité	*Écoute : vol des données en transit sur le canal de collecte*	*Vol de données/modèle via attaques classiques*	• *Vol de données de test / modèle via attaques classiques* • **Oracle dans les phases finales de test**	**Oracle sur :** • **Données d'apprentissage : inversion, appartenance** • **Modèle : extraction**
Intégrité	• *Modification des données en transit sur le canal de collecte* • **Empoisonnement des données par perturbation de la source de ces données**	• **Empoisonnement des données d'entrainement (modification, injection)** • ***Empoisonnement du modèle sur disque ou en mémoire***	*Empoisonnement du modèle sur disque (remplacement / substitution du modèle)*	• **Évasion : Entrainant une mauvaise prédiction en sortie**
Disponibilité	*Perturbation du canal de collecte*	**Empoisonnement** • **Préparation à indisponibilité future : atteinte à l'intégrité des données d'apprentissage afin d'avoir une indisponibilité durant l'inférence**	N/A (possible mais peu d'intérêt pour l'attaquant, à moins de vouloir augmenter le délai de livraison du modèle)	**Évasion suite à empoisonnement** • **Déclenchement de l'indisponibilité : présentation du *trigger* (input déclencheur)**

Travail fait en commun avec Alexandre Gakic (BNP Paribas), dans le cadre du GT IA du campus cyber

Zoom : évasion de détecteur de malware [2]

Exemple d'éléments que les attaquants peuvent modifier dans un binaire pour déjouer l'analyse statique et/ou dynamique :

- ajout d'appels API inutiles, inversions d'appels systèmes
- modification des métadonnées du binaire (PE header)
- ajout d'octets de padding à la fin du fichier
- ajout de NOP sémantiques, de fonctions inutiles
- modification du graphe de flot de contrôle

Attaques :

- en boîte blanche: facile ! C'est une descente de gradient
- en boîte grise/noire: modification aléatoire. Faisable mais plus coûteux

[2] A Survey on Adversarial Attacks for Malware Analysis

Zoom : désanonymisation de Tor

- Tor est un réseau d'anonymisation (on ne sait pas quel utilisateur consulte quel site)
- On peut attaquer cette anonymat avec du ML en observant les communications (chiffrées)
- Fingerprinting de site avec du ML (SVM)
- Jusqu'à 55% des sites identifiées !
- Mais cela nécessite de fingerprinter les sites à l'avance

[3] Website fingerprinting in onion routing based anonymization networks

ML == un système numérique comme un autre ...

... donc devant être protégé et défendu

1. Classiquement
2. En exploitant les caractéristiques du mécanisme ML à protéger/défendre

Protection/défense tout le long du pipe-line

- Collecte des données
- Préparation des données
- Apprentissage
- Inférence en test et en run

Tentative de typologie des contre-mesures (non exhaustives)

CYCLE IMPACT	Collecte	Préparation des données et apprentissage	Inférence (test)	Inférence (*run*)
Confidentialité	*Chiffrement du canal de communication*	• *Techniques de sécurité classiques (authentification, contrôle d'accès, chiffrement des données, détection des fuites d'information)* • *Technique de tatouage pour détecter après-coup un vol de données ou de modèle.* • **Differential pricacy, Cryptographie homomorphe (techniques de sécurité classiques appliquées aux données, mais demandant une adaptation spécifique sur les modèles de ML)**	• *Techniques de sécurité classiques (idem colonne de gauche)* • **Differential pricacy, Cryptographie homomorphe** • **Limitation du nombre des interrogations**	**Limitation du nombre des interrogations : pas d'accès illimité lorsque cela n'est pas nécessaire (sauf un chatbot grand public par exemple) et peu verbeux sur les réponses (pas de précision sur la confiance en cas de classification par exemple)**
Intégrité	• *Signature des données (HMAC) à émettre* • *Signature de code*	• *Techniques de sécurité classiques : authentification, contrôle d'accès, signature numérique.* • **« Désinfection » des données** • **Statistiques robustes** • **Augmentation de la robustesse : apprentissage contradictoire, masquage par gradient, méthodes d'ensemble, feature Squeezing, reformers**	*Techniques de sécurité classiques (idem colonne de gauche)*	**Bannissement de l'apprentissage continu**
Disponibilité	• *Redondance du canal* • *Évasion de fréquence*	• *Techniques de sécurité classiques : authentification, contrôle d'accès, signature des données.*	N/A	**Détection (à postériori) du trigger (par monitoring et explicabilité)**

Zoom : protection contre l'inférence de contenu par confidentialité différentielle

- Les sites de streaming (Youtube, Netflix...) envoient du flux vidéo. Ils optimisent l'utilisation du réseau avec des heuristiques complexes qui dépendent de la qualité du réseau et de la vidéo compressée
- Le machine learning peut analyser ces schémas d'envoi pour identifier quelle vidéo est visionnée (ou quel service de VOD est utilisé), même si le flux est chiffré !
- On pourrait s'en protéger avec des exemples adversariaux...
- Une autre piste : la confidentialité différentielle
- En altérant les envois de paquets de manière aléatoire, il est possible d'empêcher l'analyse
- Il ne faut ni trop réduire ces paquets (sinon l'utilisateur doit attendre plus que nécessaire), ni trop les agrandir artificiellement (gâchis de bande passante)

[4] Statistical Privacy for Streaming Traffic

Des solutions ?

Des pistes actuellement recherchées

- Améliorer l'explicabilité des prédictions pour mieux détecter ces attaques
- Bruiter les données pour rendre le modèle plus robuste (confidentialité différentielle)
- Modifier le modèle avec des contraintes mathématiques pour s'assurer que de petites perturbations en entrée ne produisent pas de grosses perturbations en sortie
- Utiliser du chiffrement homomorphique pour que les données d'apprentissage ne soient pas inférables du modèle

$\Rightarrow$ ces méthodes ont tendance à faire baisser l'efficacité des modèles ou à beaucoup augmenter le temps de calcul.

La recherche est très active dans ce domaine !

En résumé

En résumé

Le machine learning apporte à la fois

- Des opportunités
  - Des résultats intéressants mais pas (encore ?) de rupture majeure
  - Besoins en explicabilité forts
  - Encore du travail, donc !
- Des risques
  - Augmentation de la surface d'attaque
  - Des solutions classiques, des solutions spécifiques
    - Attention : risque de dégradation des performances spécifiques ML (classification, prédiction, etc.)
    - Là aussi, encore du travail !

Merci de votre attention !