



WAVESTONE

Benign EXE.mple
un générateur de virus adverses
PFEE EPITA SCIA

Ilyass ELOMRI – Numa RESPLANDY – Paul KONG

Mars 2022

WAVESTONE

AGENDA

/ **01** Contexte et Enjeux

/ **02** Objectifs et démarche du PFEE



/ **01**

Contexte et Enjeux

Apprentissage automatique

L'apprentissage automatique **est un sous-domaine de l'intelligence artificielle** qui permet de donner aux ordinateurs la capacité d'« apprendre » à partir de données d'entraînement. Un algorithme d'apprentissage automatique prendra donc en entrée une base de données d'entraînement et retournera en sortie un modèle qui contiendra toutes les propriétés permettant de faire le lien entre les données d'entraînement.

1

Entraînement

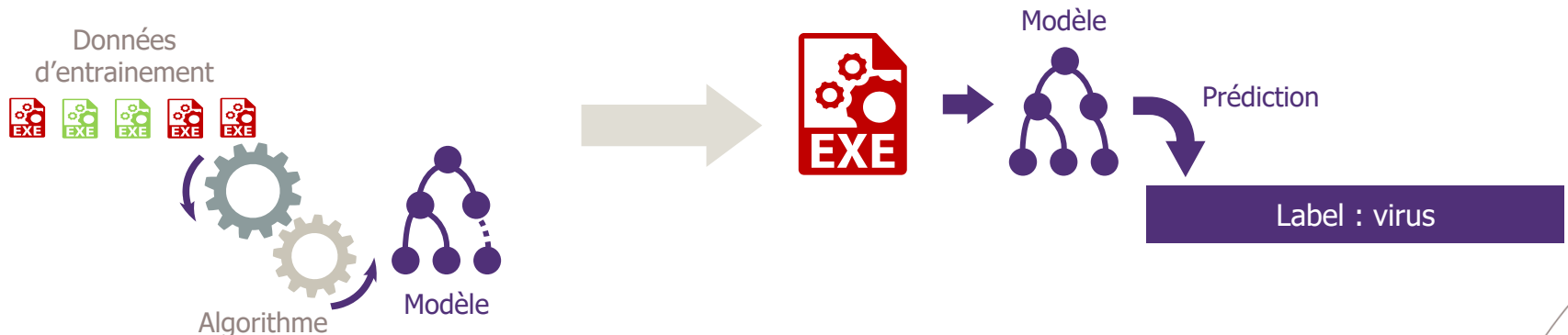
Entraînement d'un algorithme sur les données traitées en vue de réaliser des prédictions

2

Prédiction

Réalisation de prédictions sur de nouvelles données à l'aide du modèle entraîné

Exemple d'algorithme d'identification de virus



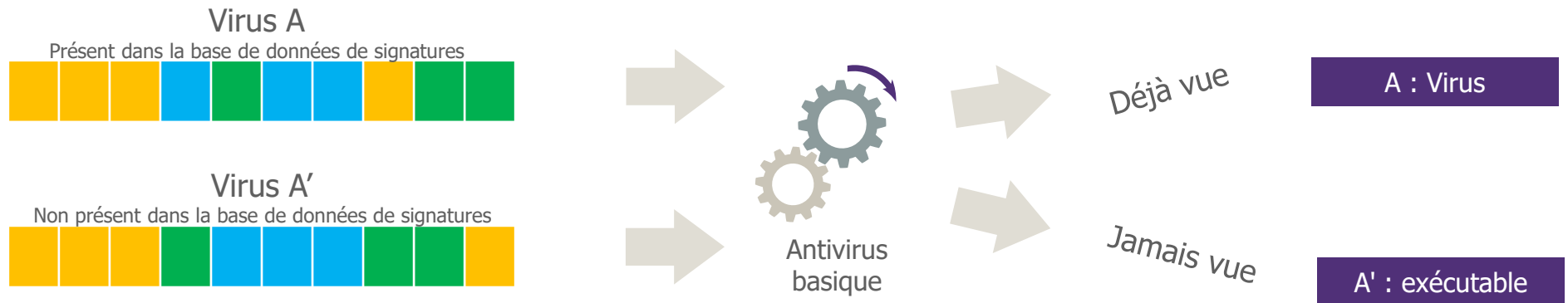
La détection de virus par apprentissage automatique

Lorsque les systèmes de détection de logiciels malveillants sur les ordinateurs ont été créés, ils étaient basés sur des calculs simples effectués sur des exécutable, tels que la comparaison de :

- fragments de code
- hashes de fragments de code ou du fichier entier
- propriétés du fichier
- de combinaisons de ces fonctionnalités.

Cependant, il s'est avéré très tôt que ces méthodes de comparaison étaient inefficaces face à de nouvelles méthodes telles que le métamorphisme ou le polymorphisme.

La détection de virus par apprentissage automatique utilise des méthodes plus poussées qui permettent lors de la phase d'apprentissage d'extraire des informations insensibles à de petites variations utilisées par les attaquants.



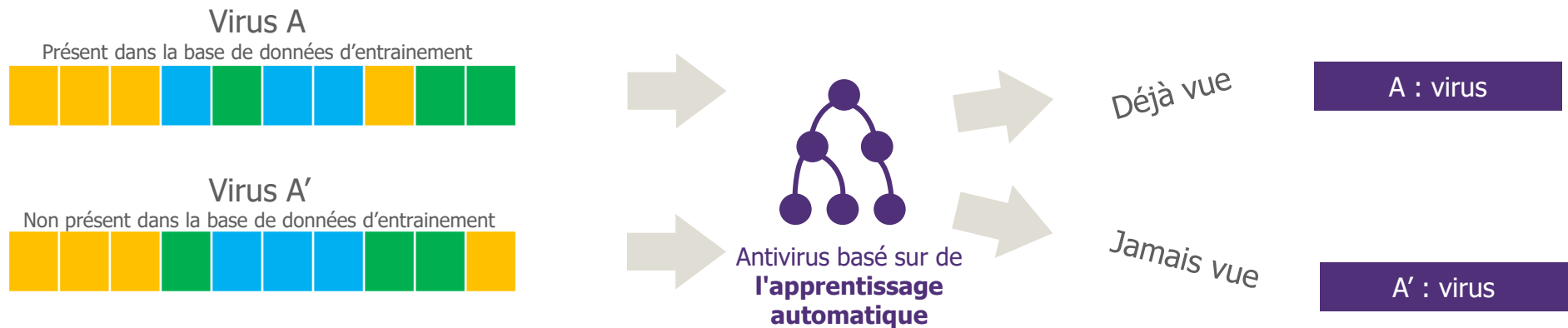
La détection de virus par apprentissage automatique

Lorsque les systèmes de détection de logiciels malveillants sur les ordinateurs ont été créés, ils étaient basés sur des calculs simples effectués sur des exécutables, tels que la comparaison de :

- fragments de code
- hashes de fragments de code ou du fichier entier
- propriétés du fichier
- de combinaisons de ces fonctionnalités.

Cependant, il s'est avéré très tôt que ces méthodes de comparaison étaient inefficaces face à de nouvelles méthodes telles que le métamorphisme ou le polymorphisme.

La détection de virus par apprentissage automatique utilise des méthodes plus poussées qui permettent lors de la phase d'apprentissage d'extraire des informations insensibles à de petites variations utilisées par les attaquants.



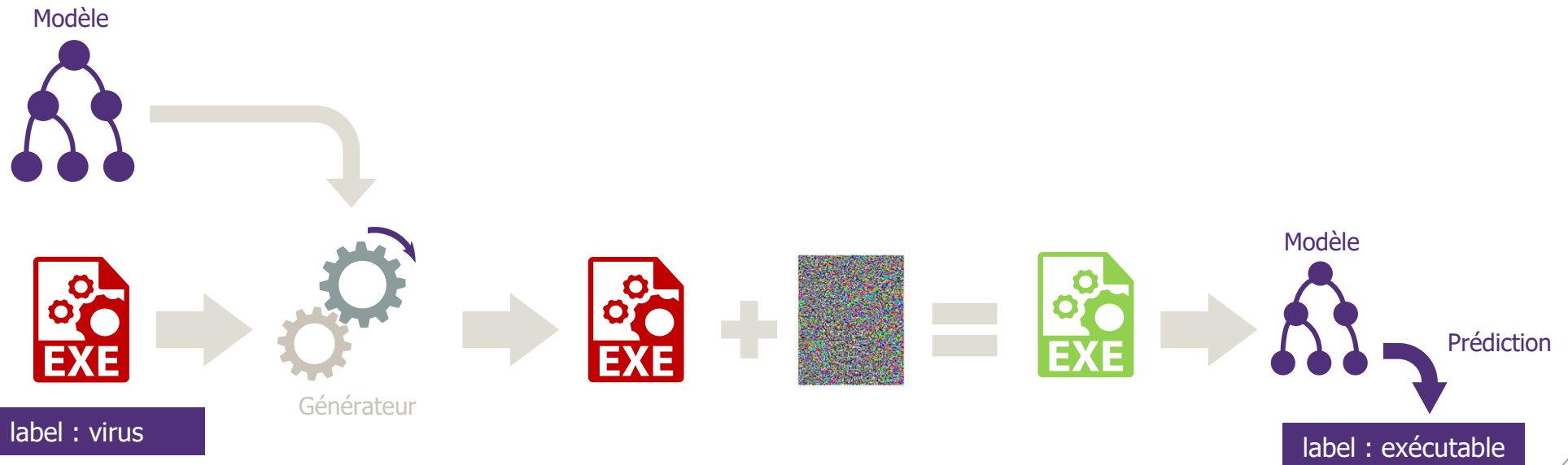
Générateur de virus adverses

Définition

L'**adversarial machine learning**, est une technique employée dans le domaine de l'apprentissage automatique **qui tente de tromper des modèles** d'apprentissage par le biais d'une saisie malveillante appelés, **exemples adverses**.

Dans le cas d'un algorithme de reconnaissance d'objet sur une image, **un exemple adverse** est une image contenant des perturbations de pixels difficilement visibles à l'œil nu et qui cause une mauvaise classification de l'image.

La distance entre l'image original et l'image adverse, appelée **epsilon**, doit être la plus faible possible.



Adversarial Machine Learning : Rendre les algorithmes plus robustes

Adversarial training

Méthode qui a pour but de créer puis d'incorporer des exemples adverses dans le processus d'entraînement du modèle afin qu'il soit moins sensible à ce type d'attaque.

Defensive distillation

Technique consiste à utiliser deux modèles successifs afin de minimiser les erreurs de décision



Gradient masking

Stratégies de défense qui permet d'obtenir un modèle plus lisse, rendant plus difficile la génération d'exemples adverses.

Randomization

Consiste à ajouter un bruit aléatoire à chaque donnée. Permet de rendre plus difficile pour un attaquant de prédire la perturbation à ajouter à une entrée.



/ **02**

Objectif du et démarche du PFEE

Un projet en 4 étapes clés



Sprint 1 :
Cadrage du sujet

- Définition de l'**équipe de travail**
- Partage de la **démarche** à suivre
- **Réajustement** du sujet avec la prise en compte des idées des étudiants



Sprint 2 :
PoC de l'attaque adverse

- Investigation des différentes approches d'attaques adverses sur les systèmes de reconnaissance de virus
- Sélection de la **méthode d'attaque adverse** pour notre générateur de virus adverses
- Développement d'une **application d'attaques adverses** sur des systèmes de reconnaissance de virus



Sprint 3 :
Système de génération d'exécutable

- Développement d'un **générateur d'exécutables** reconnus comme bénins à partir de virus
- **Optimisation** du générateur de virus adverses



Sprint 4 (**Bonus**):
PoC des méthodes de défense

- Formalisation de **bonnes pratiques** contre les attaques adverses
- Etudes des performances de ces **méthodes de défense**
- Développement d'une **application de défense contre les virus adverses**



PoC d'une application d'une attaque



Générateur de virus adverses



Application d'une défense contre les attaques adverses



Ilyass ELOMRI

Consultant Cybersécurité & Confiance Numérique
ilyass.elomri@wavestone.com

Paul KONG

Consultant Cybersécurité & Confiance Numérique
paul.kong@wavestone.com

Numa RESPLANDY

Consultant Cybersécurité & Confiance Numérique
numa.resplandy@wavestone.com

PARIS

LONDON

NEW YORK

HONG KONG

SINGAPORE *

DUBAI *

BRUSSELS

LUXEMBOURG

GENEVA

CASABLANCA

LYON

MARSEILLE

NANTES

* Partenaires stratégiques

WAVESTONE

