Detect Advanced Persistent Threats

Joannès Murigneux, Lauriane Lelandais, Mathieu Boyer

École d'ingénieurs Télécom Physique Université de Strasbourg

UFR de mathématique et d'informatique Université de Strasbourg

Introduction

I. Présentation des données

II. Traitement des données

III. Application des modèles

Conclusion

Introduction au projet

Qu’est-ce qu’une APT ?

Advanced	Menace	→ Attaque informatique
Persistent	Persistante	→ Furtive et continue
Threat	Avancée	→ Haut niveau technique

Introduction au projet

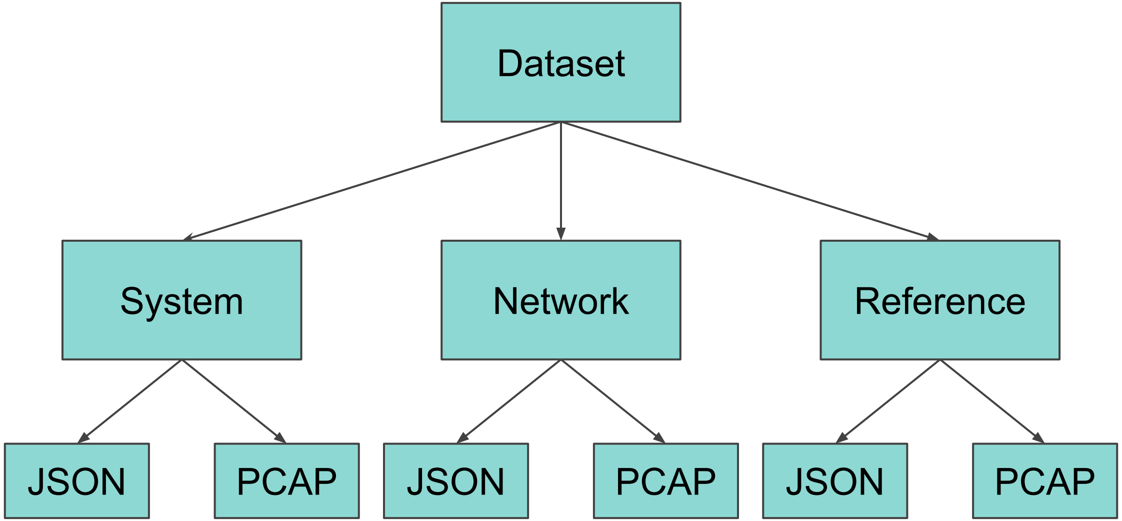
Comment les détecter ?

- Veille de surveillance sur le système
  - analyse des logs
  - analyse du traffic entre les appareils du système

Présentation des données

Présentation des données

PWNJUTSU



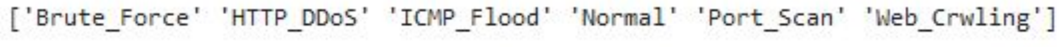
Présentation des données

PWNJUTSU

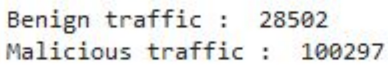
```
{
	"raw": "node=n11-vm3 type=PROCTITLE msg=audit(1620649063.784:20323): proctitle=\"/usr/local/sbin/sshd\"",
	"sourcetype": "linux_audit",
	"source": "/var/log/audit/audit.log",
	"time": "2021-05-10 12:17:43.784 UTC",
	"host": "n11-vm3"
}
```

Présentation des données

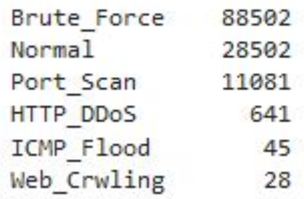
MSCAD



```
['Brute_Force' 'HTTP_DDoS' 'ICMP_Flood' 'Normal' 'Port_Scan' 'Web_Crwling']
```



```
Benign traffic :  28502
Malicious traffic :  100297
```



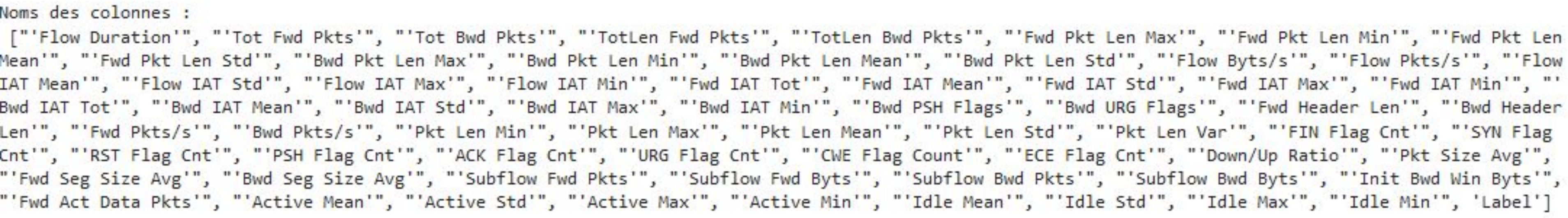
```
Brute_Force    88502
Normal         28502
Port_Scan      11081
HTTP_DDoS        641
ICMP_Flood        45
Web_Crwling       28
```

Jeu de données déséquilibré

Présentation des données

MSCAD

- 67 colonnes
- 128799 lignes



```
Noms des colonnes :
 ["'Flow Duration'", "'Tot Fwd Pkts'", "'Tot Bwd Pkts'", "'TotLen Fwd Pkts'", "'TotLen Bwd Pkts'", "'Fwd Pkt Len Max'", "'Fwd Pkt Len Min'", "'Fwd Pkt Len
Mean'", "'Fwd Pkt Len Std'", "'Bwd Pkt Len Max'", "'Bwd Pkt Len Min'", "'Bwd Pkt Len Mean'", "'Bwd Pkt Len Std'", "'Flow Byts/s'", "'Flow Pkts/s'", "'Flow
IAT Mean'", "'Flow IAT Std'", "'Flow IAT Max'", "'Flow IAT Min'", "'Fwd IAT Tot'", "'Fwd IAT Mean'", "'Fwd IAT Std'", "'Fwd IAT Max'", "'Fwd IAT Min'", "'
Bwd IAT Tot'", "'Bwd IAT Mean'", "'Bwd IAT Std'", "'Bwd IAT Max'", "'Bwd IAT Min'", "'Bwd PSH Flags'", "'Bwd URG Flags'", "'Fwd Header Len'", "'Bwd Header
Len'", "'Fwd Pkts/s'", "'Bwd Pkts/s'", "'Pkt Len Min'", "'Pkt Len Max'", "'Pkt Len Mean'", "'Pkt Len Std'", "'Pkt Len Var'", "'FIN Flag Cnt'", "'SYN Flag
Cnt'", "'RST Flag Cnt'", "'PSH Flag Cnt'", "'ACK Flag Cnt'", "'URG Flag Cnt'", "'CWE Flag Count'", "'ECE Flag Cnt'", "'Down/Up Ratio'", "'Pkt Size Avg'",
"'Fwd Seg Size Avg'", "'Bwd Seg Size Avg'", "'Subflow Fwd Pkts'", "'Subflow Fwd Byts'", "'Subflow Bwd Pkts'", "'Subflow Bwd Byts'", "'Init Bwd Win Byts'",
"'Fwd Act Data Pkts'", "'Active Mean'", "'Active Std'", "'Active Max'", "'Active Min'", "'Idle Mean'", "'Idle Std'", "'Idle Max'", "'Idle Min'", 'Label']
```

Présentation des données

APTGen

- Un outils de génération de dataset
- un dataset
- plus de 800 scénarios d’attaques

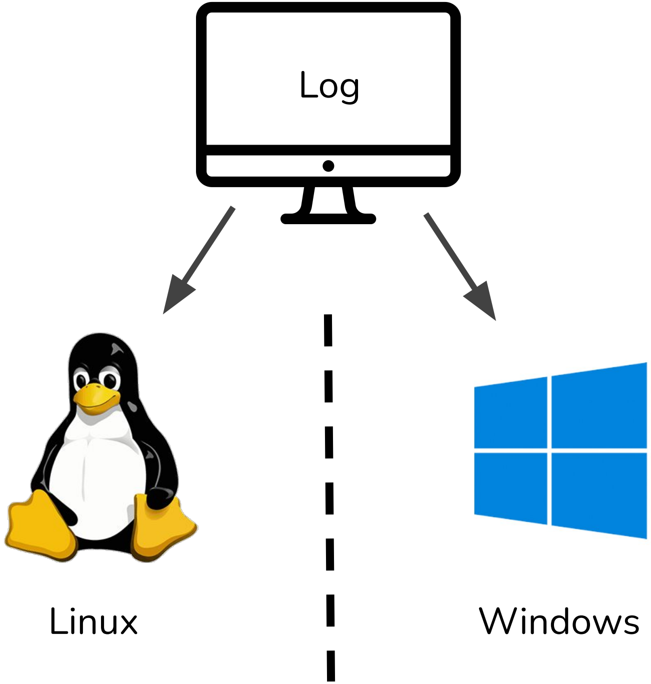
Traitement des données

Traitement des données

PWNJUTSU

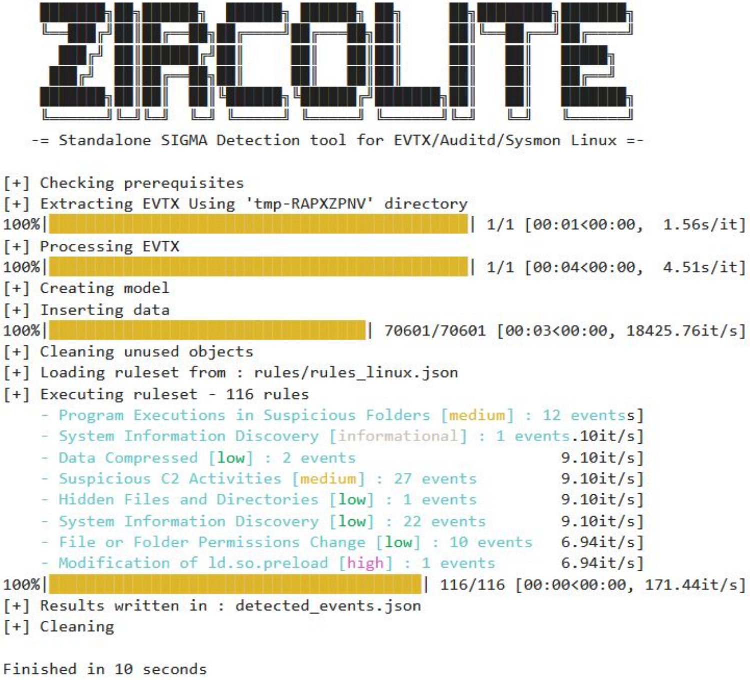
Traitement des données

PWNJUTSU



Traitement des données

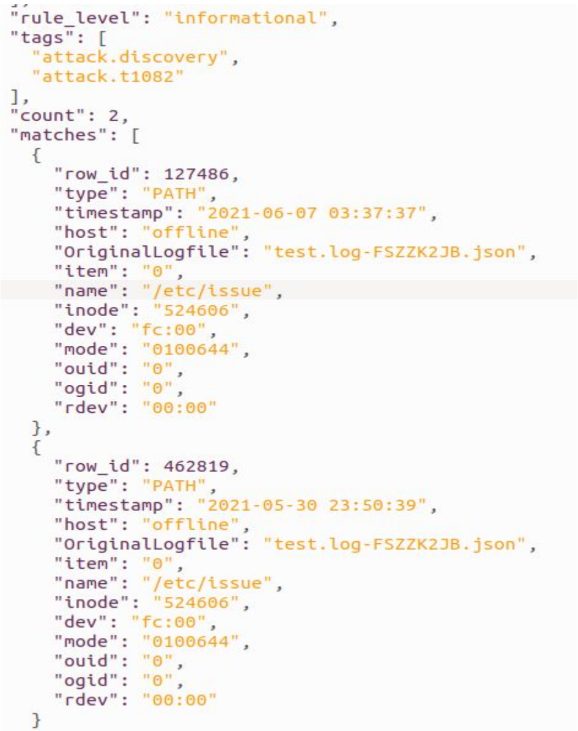
PWNJUTSU



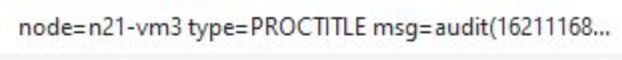
```
   -= Standalone SIGMA Detection tool for EVTX/Auditd/Sysmon Linux =-

[+] Checking prerequisites
[+] Extracting EVTX Using 'tmp-RAPXZPNV' directory
100%|████████████████████████████████████████| 1/1 [00:01<00:00,  1.56s/it]
[+] Processing EVTX
100%|████████████████████████████████████████| 1/1 [00:04<00:00,  4.51s/it]
[+] Creating model
[+] Inserting data
100%|██████████████████████████████| 70601/70601 [00:03<00:00, 18425.76it/s]
[+] Cleaning unused objects
[+] Loading ruleset from : rules/rules_linux.json
[+] Executing ruleset - 116 rules
    - Program Executions in Suspicious Folders [medium] : 12 eventss]
    - System Information Discovery [informational] : 1 events.10it/s]
    - Data Compressed [low] : 2 events                       9.10it/s]
    - Suspicious C2 Activities [medium] : 27 events          9.10it/s]
    - Hidden Files and Directories [low] : 1 events          9.10it/s]
    - System Information Discovery [low] : 22 events         9.10it/s]
    - File or Folder Permissions Change [low] : 10 events    6.94it/s]
    - Modification of ld.so.preload [high] : 1 events        6.94it/s]
100%|████████████████████████████████████| 116/116 [00:00<00:00, 171.44it/s]
[+] Results written in : detected_events.json
[+] Cleaning

Finished in 10 seconds
```

Traitement des données

PWNJUTSU



```
"rule_level": "informational",
"tags": [
  "attack.discovery",
  "attack.t1082"
],
"count": 2,
"matches": [
  {
    "row_id": 127486,
    "type": "PATH",
    "timestamp": "2021-06-07 03:37:37",
    "host": "offline",
    "OriginalLogfile": "test.log-FSZZK2JB.json",
    "item": "0",
    "name": "/etc/issue",
    "inode": "524606",
    "dev": "fc:00",
    "mode": "0100644",
    "ouid": "0",
    "ogid": "0",
    "rdev": "00:00"
  },
  {
    "row_id": 462819,
    "type": "PATH",
    "timestamp": "2021-05-30 23:50:39",
    "host": "offline",
    "OriginalLogfile": "test.log-FSZZK2JB.json",
    "item": "0",
    "name": "/etc/issue",
    "inode": "524606",
    "dev": "fc:00",
    "mode": "0100644",
    "ouid": "0",
    "ogid": "0",
    "rdev": "00:00"
  }
```



```
node=n21-vm3 type=PROCTITLE msg=audit(16211168...
```

Traitement des données

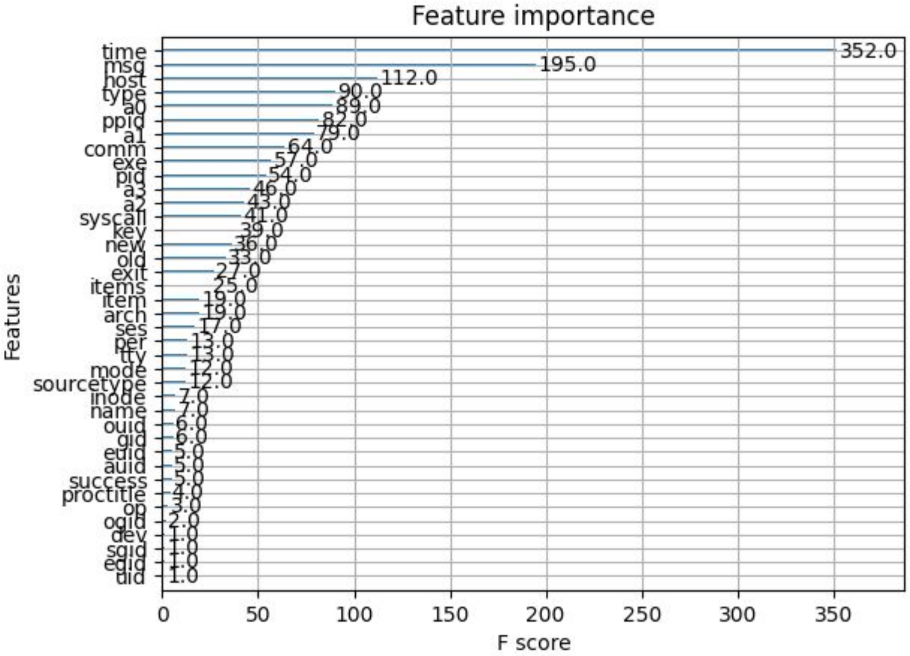
PWNJUTSU

2 défis:

- données manquantes
- encodage des données coûteux

Traitement des données

PWNJUTSU



Traitement des données

MSCAD

- Encodage des données catégorielles

Application des modèles

Application des modèles

Métriques

Données équilibrées

$$precision = \frac{\text{true positive}}{\text{true positive+false positive}}$$

$$tnr = \frac{\text{true negative}}{\text{true negative+false positive}}$$

$$recall = \frac{\text{true positive}}{\text{true positive+false negative}}$$

$$accuracy = \frac{\text{true positive+true negative}}{\text{true positive+false positive+true negative} + \text{false negative}}$$

Application des modèles

Métriques

Données déséquilibrées

$$f1\,score = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$$

$$balanced\,accuracy = \frac{\text{recall}+\text{tnr}}{2}$$

$$matthews\,correlation\,coefficient = \frac{tn \cdot tp - fn \cdot fp}{\sqrt{(tp+fp) \cdot (tp+fn) \cdot (tn+fp) \cdot (tn+fn)}}$$

Application des modèles

Modèles choisis

- XGBoost avec données catégorielles ou non
- KNN
- Cart
- Random Forest
- SVM
- MLP

Application des modèles

Résultats

métriques	XGBoost	KNN	CART	Random forest	SVM	MLP	XGBoost catégoriel
1 000 lignes d'entraînement							
précision	0.957	0.908	0.959	0.959		0.918	0.983
recall	0.969	0.978	0.968	0.968		0.978	0.920
TNR	0.997	0.993	0.997	0.997		0.994	0.998
accuracy	0.995	0.992	0.995	0.995		0.993	0.993
f1 score	0.963	0.942	0.964	0.964		0.947	0.950
balanced accuracy	0.983	0.986	0.983	0.983		0.986	0.959
matthews correlation coefficient	0.961	0.939	0.961	0.961		0.944	0.948
memory usage (MB)	5520	5937	5520	5520		5975	5092
time (sec)	28	195	29	50		30	121

Application des modèles

Résultats

métriques	XGBoost	KNN	CART	Random forest	SVM	MLP	XGBoost catégoriel
10 000 lignes d’entraînement							
précision	0.949	0.941	0.949	0.949	0.959	0.948	0.997
recall	0.976	0.969	0.976	0.976	0.968	0.976	0.779
TNR	0.996	0.995	0.996	0.996	0.997	0.996	0.999
accuracy	0.995	0.994	0.995	0.995	0.995	0.995	0.985
f1 score	0.962	0.955	0.962	0.962	0.964	0.962	0.875
balanced accuracy	0.986	0.982	0.986	0.986	0.983	0.986	0.889
matthews correlation coefficient	0.960	0.952	0.960	0.960	0.961	0.959	0.875
memory usage (MB)	5520	5938	5520	5520	5520	5968	5084
time (sec)	31	666	28	60	69	32	255

Application des modèles

Résultats

métriques	XGBoost	KNN	CART	Random forest	SVM	MLP	XGBoost catégoriel
100 000 lignes d’entraînement							
précision	0.959	0.958	0.959	0.959	0.959	0.959	0.645
recall	0.968	0.969	0.969	0.969	0.968	0.969	0.789
TNR	0.997	0.997	0.997	0.997	0.997	0.997	0.998
accuracy	0.995	0.995	0.995	0.995	0.995	0.995	0.934
f1 score	0.964	0.963	0.964	0.964	0.964	0.964	0.544
balanced accuracy	0.983	0.983	0.983	0.983	0.983	0.983	0.823
matthews correlation coefficient	0.961	0.961	0.962	0.962	0.961	0.962	0.634
memory usage (MB)	5520	5933	5520	5520	5520	5896	5005
time (sec)	43	5542	34	79	302	48	472

Application des modèles

Résultats

métriques	XGBoost	KNN	CART	Random forest	SVM	MLP	XGBoost catégoriel
1 000 000 000 lignes d’entraînement							
précision	0.960	0.960	0.960	0.960	0.960	0.960	0.378
recall	0.968	0.968	0.968	0.968	0.968	0.968	0.572
TNR	0.997	0.997	0.997	0.997	0.997	0.997	0.936
accuracy	0.995	0.995	0.995	0.995	0.995	0.995	0.913
f1 score	0.964	0.964	0.964	0.964	0.964	0.964	0.455
balanced accuracy	0.983	0.982	0.983	0.983	0.982	0.982	0.754
matthews correlation coefficient	0.962	0.961	0.962	0.962	0.962	0.962	0.421
memory usage (MB)	5520	5520	5520	5520	5520	5520	4213
time (sec)	181	4461	195	243	803	212	921

Conclusion

Merci pour votre attention