

# Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems

Creach Lab - October 10<sup>th</sup>, 2023 - Inria Rennes

---

Maxime Lanvin<sup>2</sup>   Pierre-François Gimenez<sup>2</sup>   Yufei Han<sup>1</sup>   Frédéric Majorczyk<sup>3</sup>   Ludovic Mé<sup>1</sup>   Éric Totel<sup>4</sup>

<sup>1</sup>Inria, Univ. Rennes, IRISA, {firstname.lastname}@inria.com

<sup>2</sup>CentraleSupélec, Univ. Rennes, IRISA, {firstname.lastname}@centralesupelec.fr

<sup>3</sup>DGA-MI, Univ. Rennes, IRISA, frederic.majorczyk@intradef.gouv.fr

<sup>4</sup>Samovar, Télécom SudParis, Institut Polytechnique de Paris, eric.totel@telecom-sudparis.eu



## **Introduction on NIDS & motivation**

---

# Two paradigms for NIDS

## Intrusion Detection

- Intrusion Detection Systems (**IDS**) offer a way to detect attacks and let operators react according to the alerts
- We focus in this work on Network IDS (**NIDS**)

## Paradigms

- **Signature**-based : detection of signature associated with known attacks
- **Anomaly**-based : detection of deviation from a normal behavior

# Comparison of the two paradigms

## Signature based

### Supervision

- ts: 2023-01-19T14:02:46.143Z
- dst\_address: "192.168.101.3"
- dst\_port: 47426
- src\_address: "192.168.101.26"
- src\_port: 1389
- signature: "ET ATTACK\_RESPONSE Possible  
CVE-2021-44228 Payload via LDAPv3  
Response"
- category: "Attempted Administrator  
Privilege Gain"
- severity: 1
- CVE: **CVE\_2021\_4422**



## Anomaly based

### Supervision

- ts: 2023-01-19T14:02:46.143Z
- dst\_address: "192.168.101.3"
- dst\_port: 47426
- src\_address: "192.168.101.26"
- src\_port: 1389

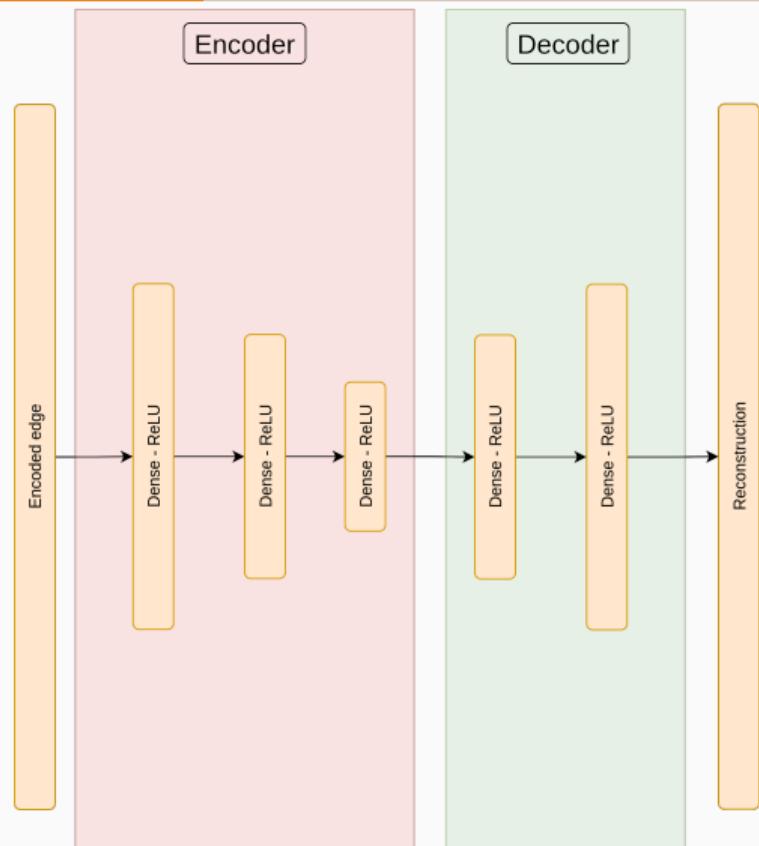


Good Luck ! Enjoy !



1. Introduction on NIDS & motivation
2. AE-pvalues
3. Benchmark XAI techniques
4. Using explanations on CICIDS2017 dataset
5. Conclusion

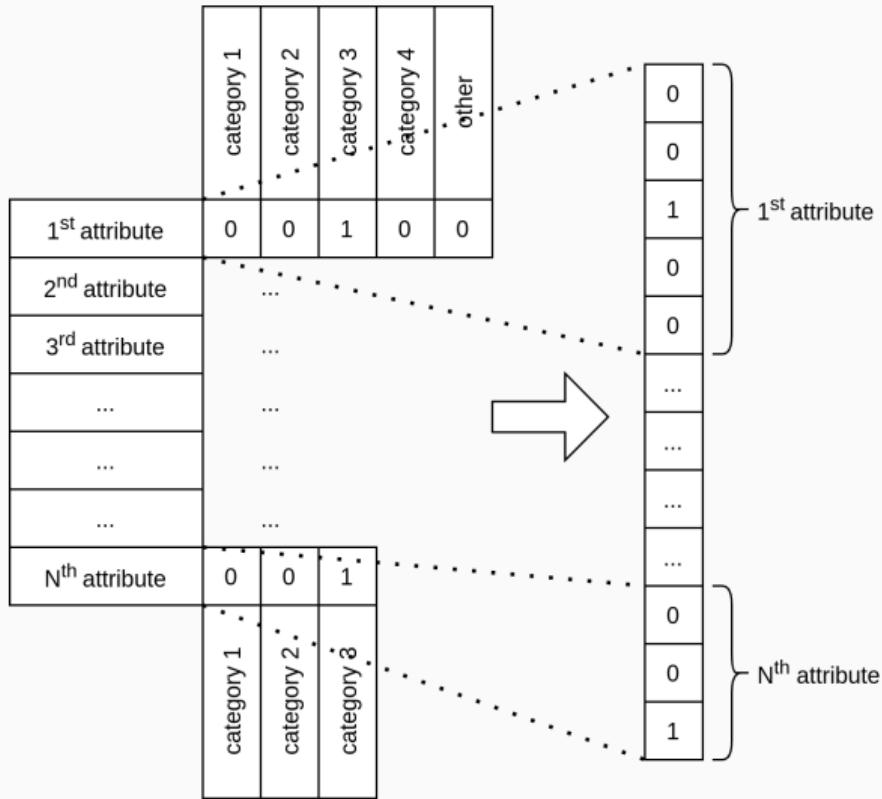
# Unsupervised anomaly detection : Autoencoder



## Learning

Minimisation of the reconstruction error between the input vector and its reconstructed version.

## One Hot Encoding - Meaning of the vectors



# Providing explanations

---

## Goal

In our context providing explanations consists in providing the list of the **network attributes ranked** from the most abnormal to the least abnormal.

## Example

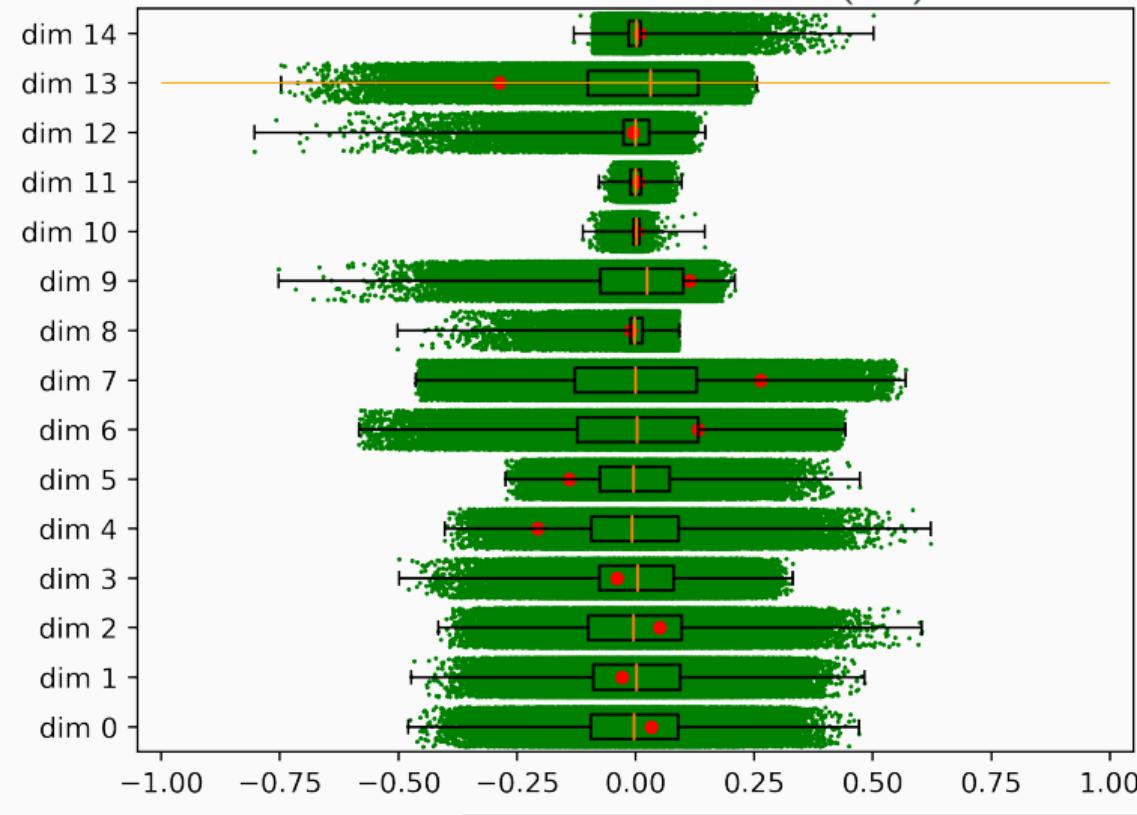
[connection\_duration, user\_agent, ..., http\_method, ..., destination\_port]

## **AE-pvalues**

---

# XAI techniques for Autoencoders

Reconstruction error distribution (AE)



## Possible methods

- Ranking by **absolute** values
- Ranking by **shapley** values
- Ranking by **p-values**

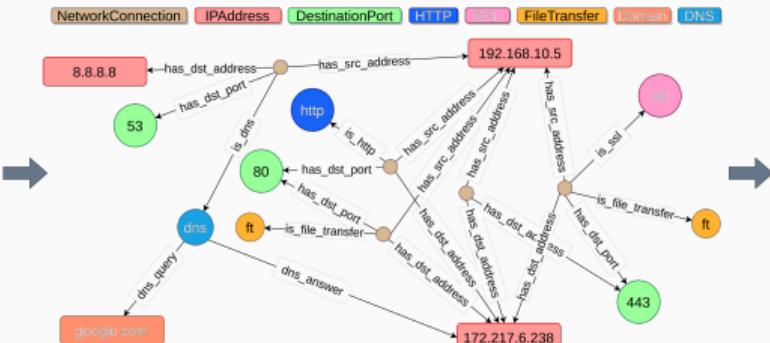
## Observation

The highest reconstruction error is not always an indication of the most abnormal dimension.

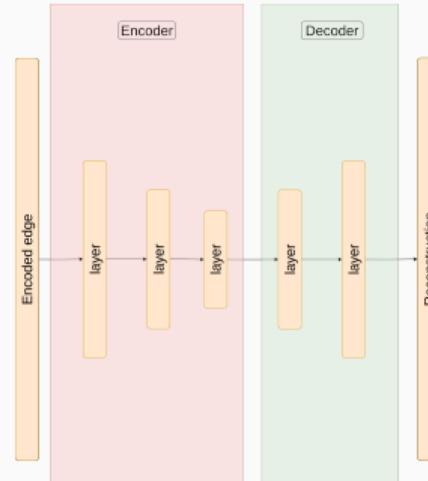
## Benchmark XAI techniques

---

# Sec2graph : An anomaly detection NIDS



0	0	1
0	0	0
1	1	0
0	0	1
0	1	0
...	...	...
...	...	...
...	...	...
0	0	0
0	0	1
1	1	1

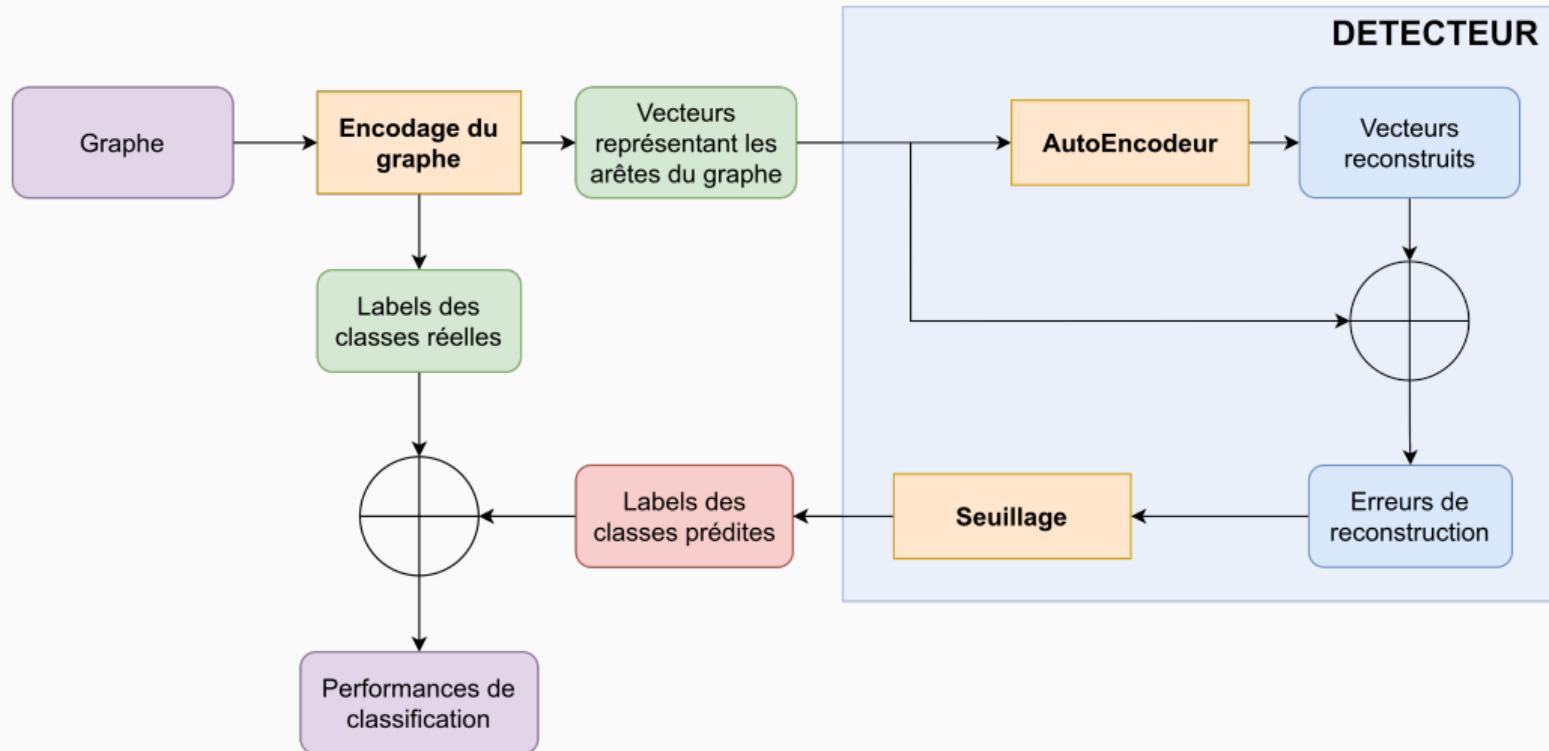


Autoencoder

## Assumptions

- **Unsupervised** : no attacks used for the training
- **Anomaly-based NIDS** : detect drift from normal behaviours
- **The graph** helps to create a meaningful neighborhood

# Anomaly Detection - Autoencoder

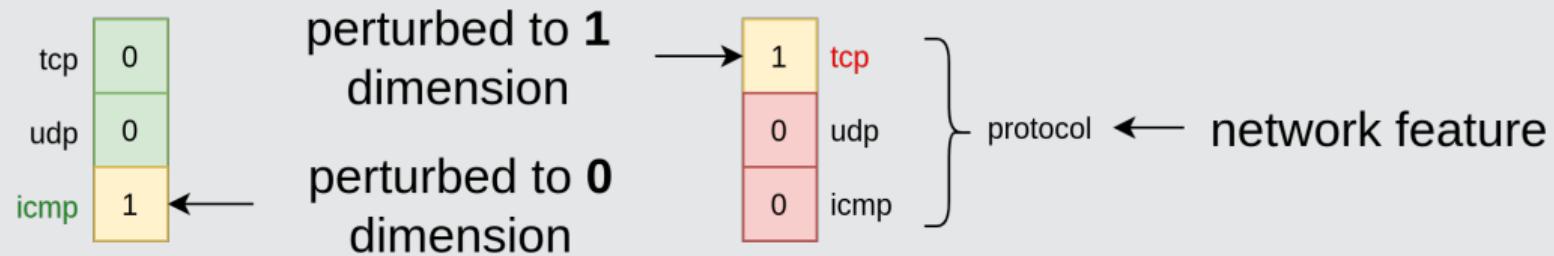


# Methodology for the comparison

## Methods

- Inject noise in a known network characteristic of vectors
- Assess ability of XAI methods to find the noisy network characteristic

## Exemple of noise insertion in the protocol characteristic

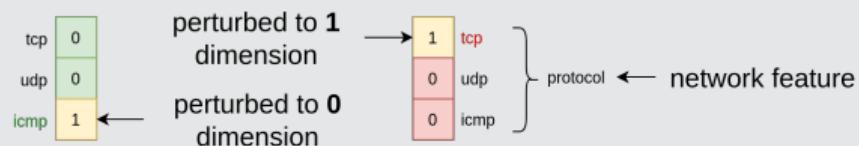


## Evaluation : correlated attributes

http\_status\_code = 200 is equivalent to http\_status\_msg = OK

# Benchmark results

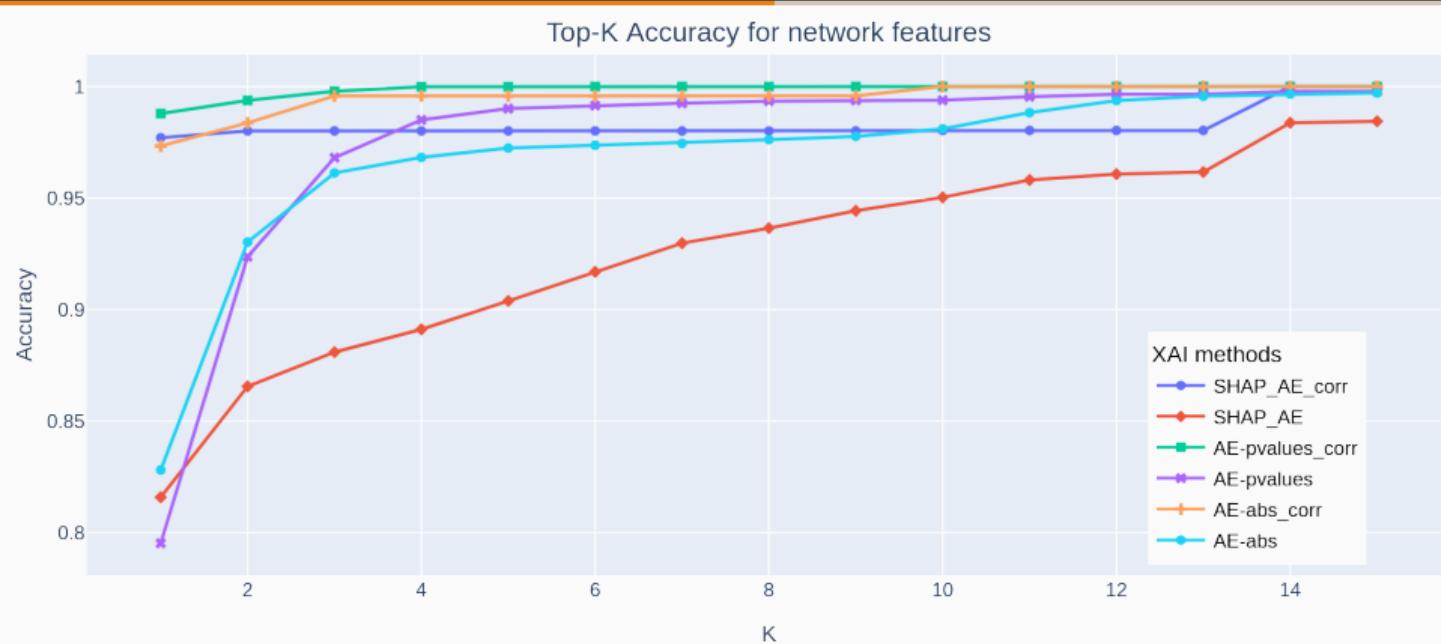
## Vocabulary reminder



explaining method	Mean rank of the perturbed to 0 dimension	Mean rank of the perturbed to 1 dimension	Mean rank of the network feature ↓
<b>AE-pvalues_corr</b>	<b>2.96</b>	1.63	<b>1.02</b>
AE-abs_corr	3.89	<b>1.61</b>	1.07
SHAP_AE_corr	4.71	4.44	1.26
Random_corr	5.68	16.3	1.85
<b>AE-pvalues</b>	<b>4.61</b>	<b>3.07</b>	<b>1.39</b>
AE-abs	5.78	4.78	1.49
SHAP_AE	18.96	7.18	2.15
Random	26.93	27.13	7.8

Table of mean ranks of the perturbed to 0 or 1 dimensions, and the network feature where the noise is inserted.

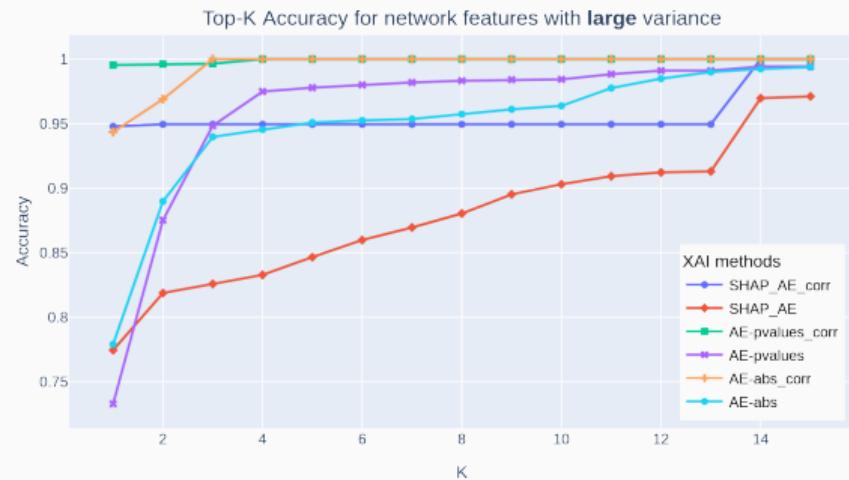
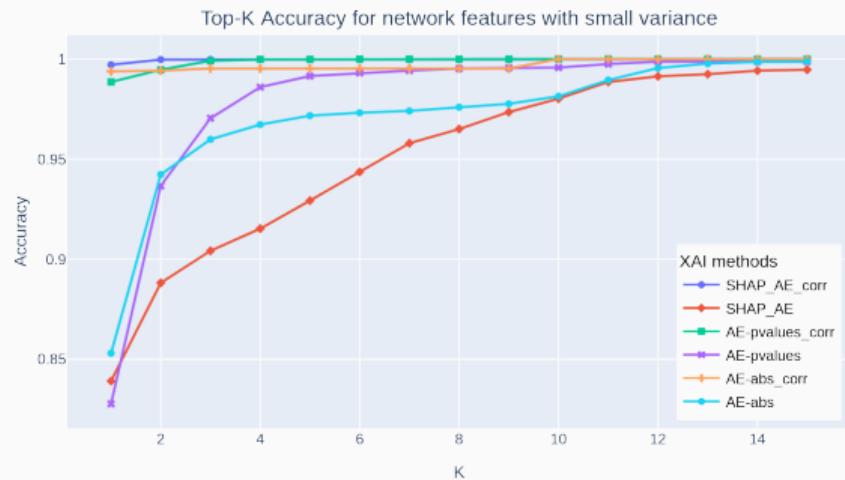
# Benchmark results



## Top-K accuracy

Proportion of samples for which the right explanation is among the Top-K explanations.

# Benchmark results



## Benchmark results

Method	Processing time per sample
SHAP_AE	28 s
AE-pvalues	1.9 ms
AE-abs	1.0 ms

Processing time for one sample for each explaining method

## Conclusion

AE-pvalues is approximately 10,000 faster than the SHAP\_AE method.

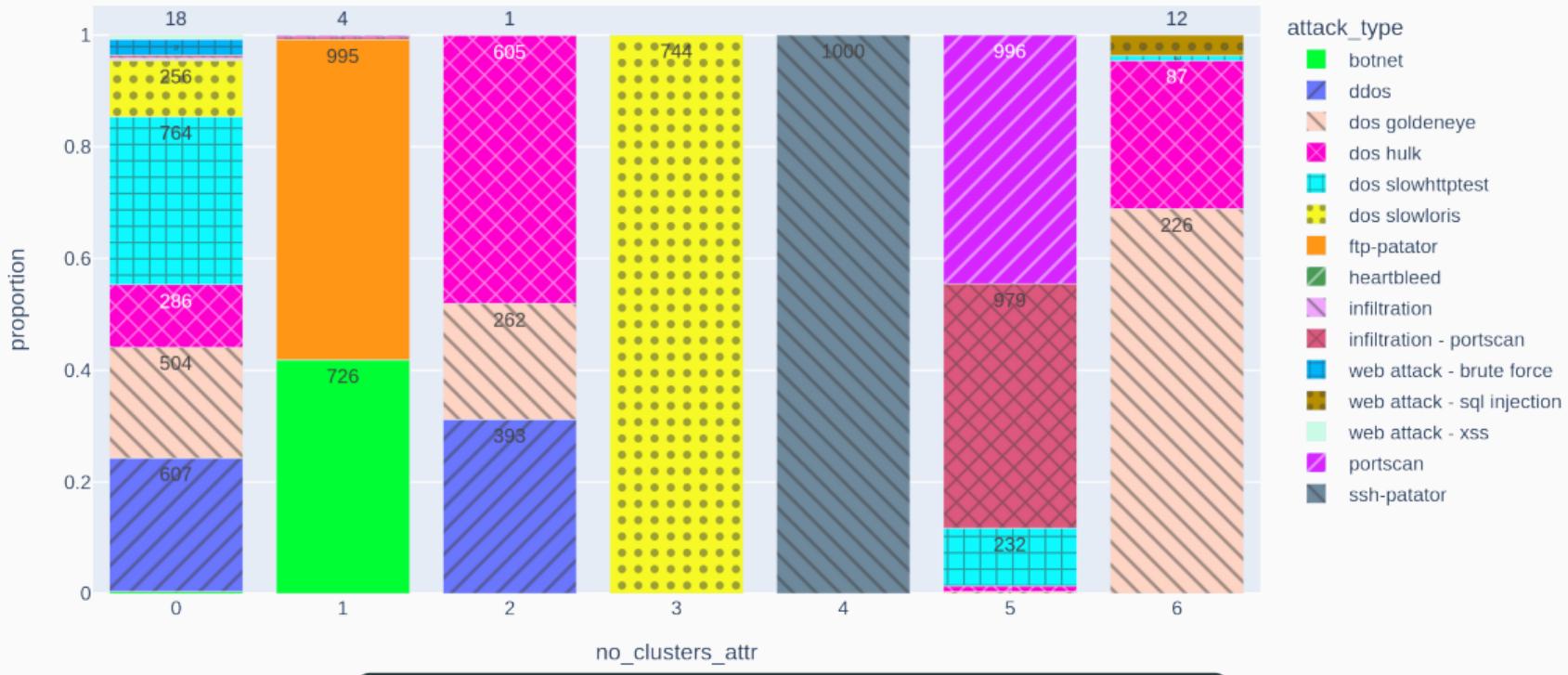
## **Using explanations on CICIDS2017 dataset**

---

# Applications - Clustering

## Principle

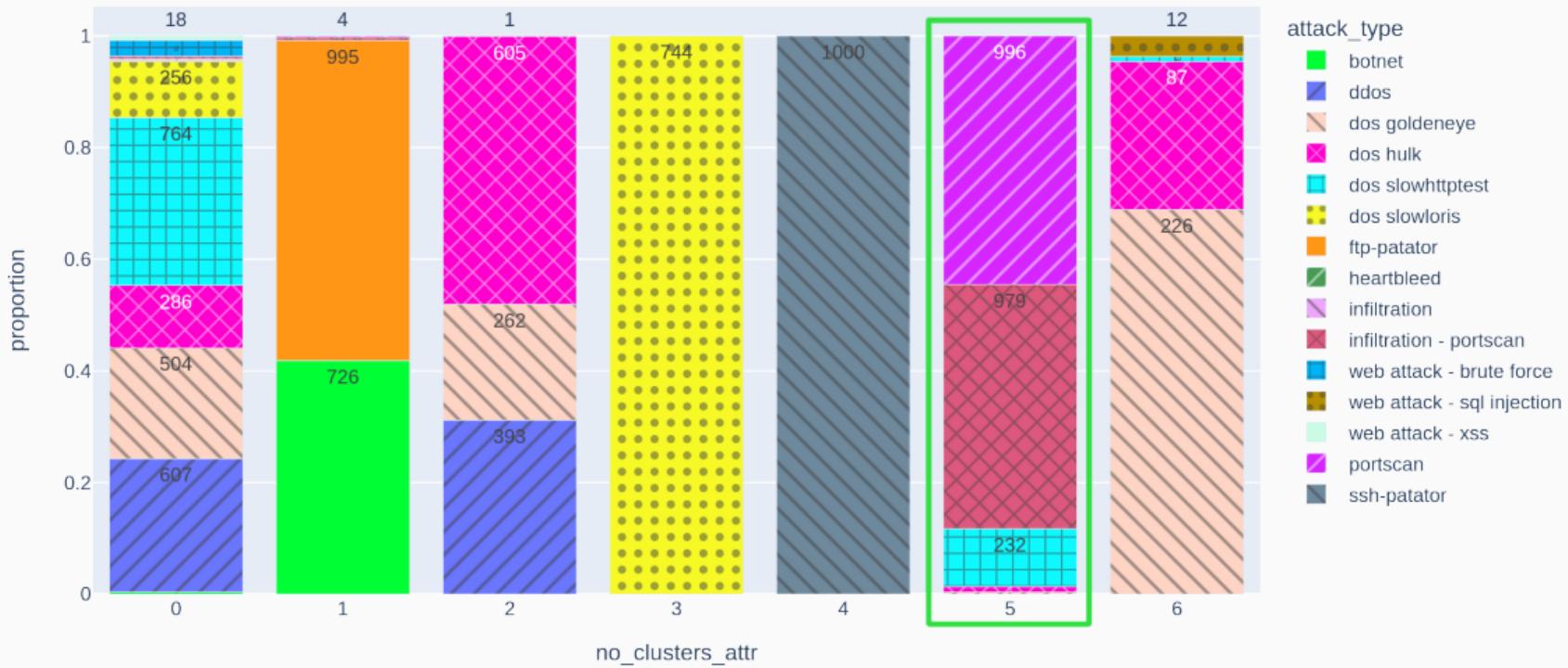
Clustering of the alerts based on the explanations



# Applications - Clustering

## Principle

Clustering of the alerts based on the explanations



# Applications - Feature contribution to attack types

	http_trans_depth	http_status_code	address_value	port_value	service	http_version	http_method	ua_uaos	ua_browser	filetransfer_conn_state	filetransfer_mime_type	duration	weird_name	weird_peer	http_info_code	http_info_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_alg	ssh_client
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	1.6	1.6	0.0	0.0
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	15.7	15.7	0.0	0.0
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0

# Applications - Feature contribution to attack types

	http_trans_depth	http_status_code	address_value	port_value	service	http_version	http_method	ua_uaos	ua_browser	filetransfer_conn_state	filetransfer_mime_type	duration	weird_name	weird_peer	http_info_code	http_info_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client	
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	
	ddos	-20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	
	dos goldeneye	-18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	
	dos hulk	-13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	1.6	1.6	0.0	0.0
	dos slowloris	-4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	15.7	15.7	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	20.0
	web attack - brute force	-20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0
	web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0

# Applications - Feature contribution to attack types

	http_trans_depth	http_status_code	address_value	port_value	service	http_version	http_method	ua_ua_os	ua_browser	filetransfer_conn_state	filetransfer_mime_type	duration	weird_name	weird_peer	http_info_code	http_info_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client	
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	1.6	1.6	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	15.7	15.7	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0	0.0
	web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

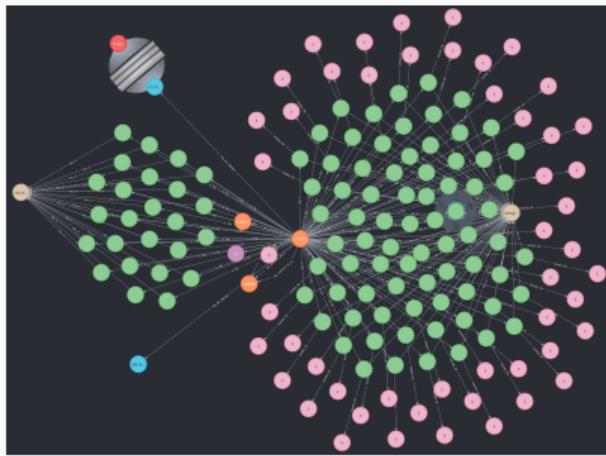
# Applications - Feature contribution to attack types

	http_trans_depth	http_status_code	address_value	port_value	service	http_version	http_method	ua_ua_os	ua_browser	filetransfer_conn_state	filetransfer_mime_type	duration	weird_name	weird_peer	http_info_code	http_info_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client	
	botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	
	heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	
	infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	
	portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	
	dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	
	dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	
	dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	1.6	1.6	0.0	0.0
	dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	15.7	15.7	0.0	0.0
	ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	20.0
	web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0
	web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0
	web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0

# Applications - Feature contribution to attack types

	http_trans_depth	http_status_code	address_value	port_value	service	http_version	http_method	ua_ua_os	ua_browser	filetransfer_conn_state	filetransfer_mime_type	duration	weird_name	weird_peer	http_info_code	http_info_msg	ssh_host_key_alg	ssh_host_key	ssh_cipher_algo	ssh_client
botnet	0.1	0.0	1.8	0.0	20.0	2.4	17.4	0.0	17.5	19.2	18.0	1.8	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0
heartbleed	0.0	0.0	20.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
infiltration	0.0	0.0	20.0	2.9	17.1	20.0	0.0	0.0	0.0	0.0	14.3	17.1	0.0	2.9	2.9	0.0	0.0	0.0	0.0	0.0
infiltration - portscan	0.0	0.0	19.9	19.8	19.8	0.2	0.0	0.0	0.0	0.0	19.8	19.9	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0
portscan	0.0	0.0	20.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ddos	20.0	20.0	7.0	0.0	0.0	0.0	0.3	20.0	20.0	0.0	0.0	0.0	0.0	12.7	0.0	0.0	0.0	0.0	0.0	0.0
dos goldeneye	18.4	14.8	11.2	0.2	0.8	0.3	1.0	18.3	15.3	7.6	8.9	1.1	1.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0
dos hulk	13.5	14.0	1.9	0.5	3.1	0.0	10.9	13.5	15.9	6.2	6.2	1.0	0.5	5.5	3.2	2.9	1.0	0.0	0.0	0.0
dos slowhttptest	0.4	7.2	5.0	5.1	2.5	0.0	1.7	4.1	3.5	0.1	0.1	12.4	4.6	8.2	13.2	13.2	1.6	1.6	0.0	0.0
dos slowloris	4.3	16.1	0.0	0.8	0.0	0.0	16.9	20.0	3.0	3.1	3.1	0.0	0.0	1.3	0.0	0.0	15.7	15.7	0.0	0.0
ftp-patator	0.0	0.0	20.0	0.1	19.9	20.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ssh-patator	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	19.9	19.9	20.0
web attack - brute force	20.0	19.7	0.0	0.0	0.0	0.0	0.3	0.3	0.5	19.7	19.7	0.0	0.0	0.0	0.0	0.0	19.7	0.0	0.0	0.0
web attack - sql injection	0.0	0.0	3.1	20.0	0.0	20.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	16.9	0.0	0.0	0.0	0.0	0.0	0.0
web attack - xss	0.0	18.9	0.0	20.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0

# Applications - Forensic analysis - Web attack : Brute Force



single connection graph

network_feature	value
http_method	POST
http_referrer	http://205.174.165.68/dv/login.php
http_request_body_len	130
http_status_code	302
http_status_msg	Found
http_trans_depth	84
user_agent_browser	Mozilla/5.0
user_agent_os	Linux x86_64

## Top 5 explanations

user\_agent\_browser - user\_agent\_os - http\_status\_msg

http\_status\_code - http\_trans\_depth

## Conclusion

---

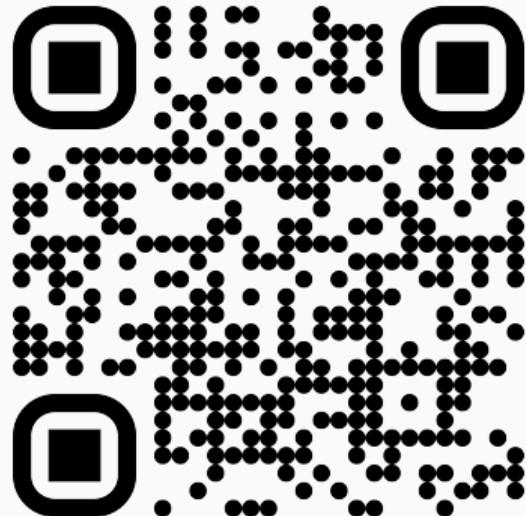
# Conclusion<sup>1</sup>

## Summary :

- Explanation technique for alerts raised by AutoEncoder-based NIDS
- Clustering alerts based on explanations
- Help manual analysis

## Future works

Leverage explanation techniques for the detection and alert triage



gitlab code for *AE-pvalues*  
[gitlab.inria.fr/mlanvin/ae-pvalues](https://gitlab.inria.fr/mlanvin/ae-pvalues)

---

1. Work accepted at RAID 2023 conference <https://doi.org/10.1145/3607199.3607247>