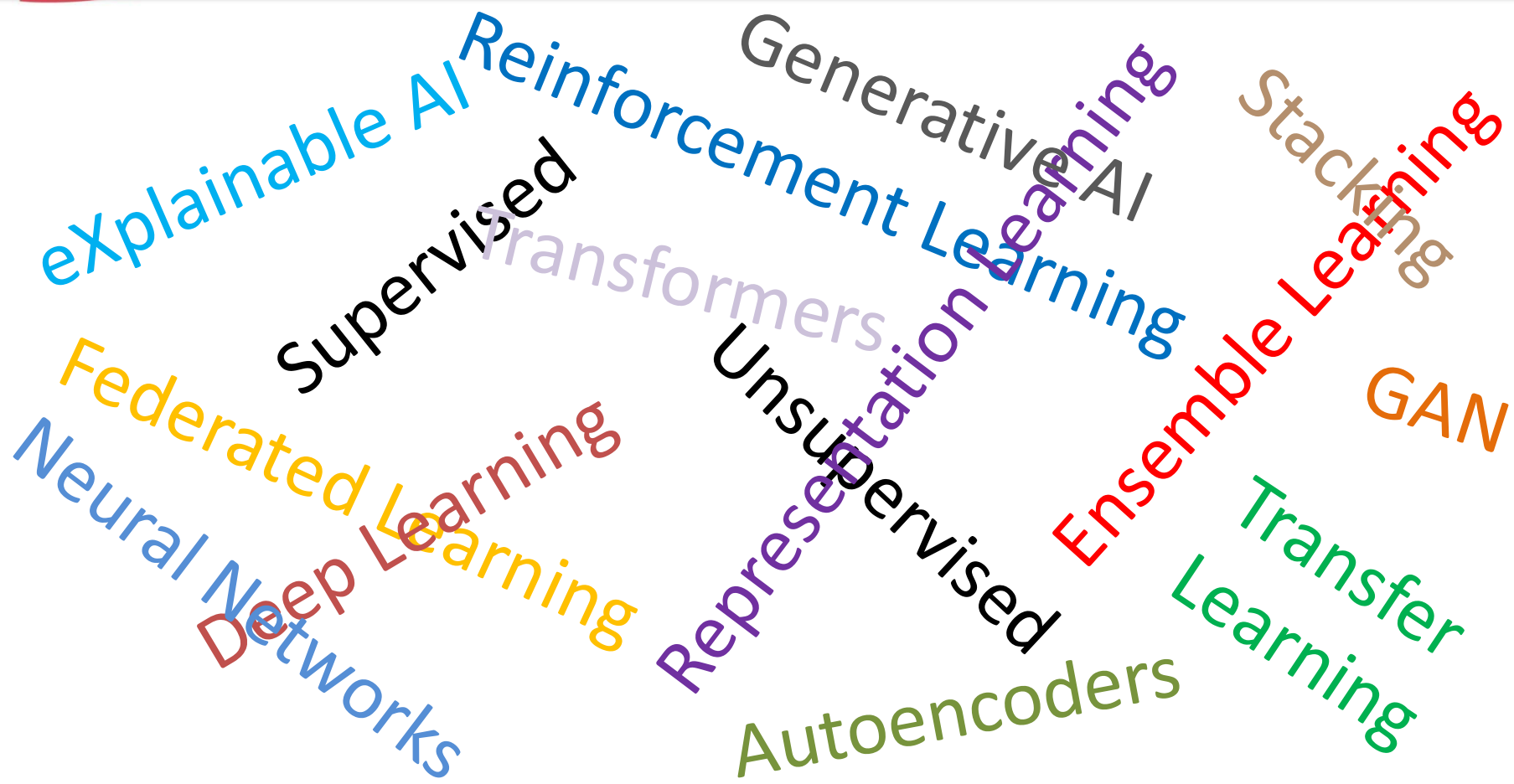
A Short History of Machine Learning for Network Anomaly Detection

Philippe Owezarski

LAAS-CNRS

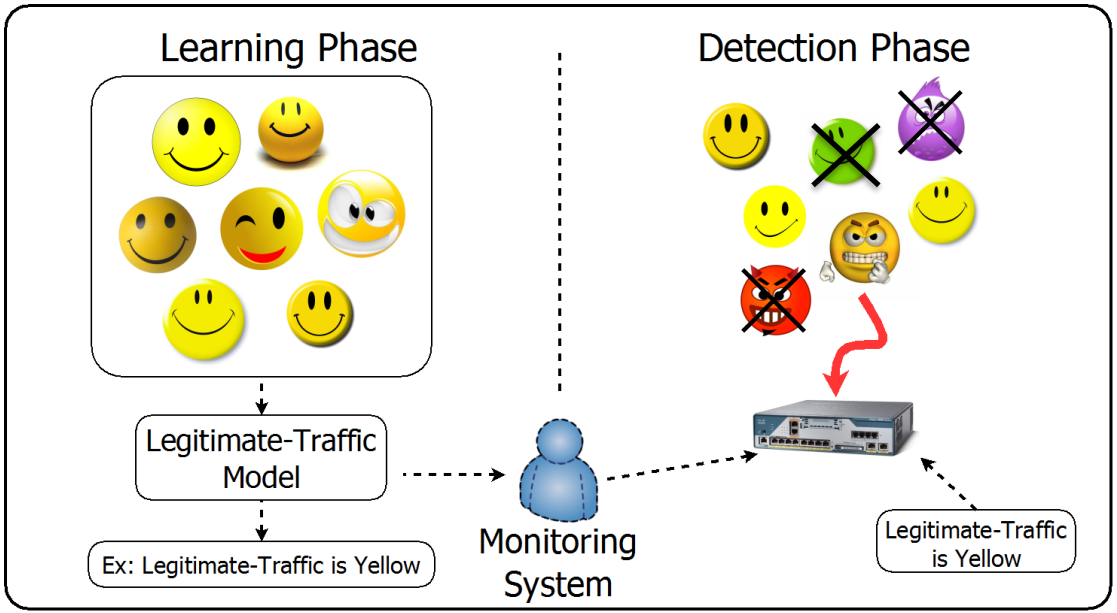
Workshop Superviz, Rennes, 16-17 décembre 2024

Many keywords related to ML…



Supervised learning-based AD

- Detect what is different from what I know



(+) It can detect new anomalies out-of the baseline

(-) Requires training on anomaly-free traffic

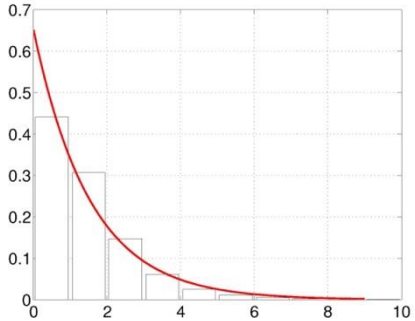
(-) Robust and adaptive models are difficult to conceive

Traffic models

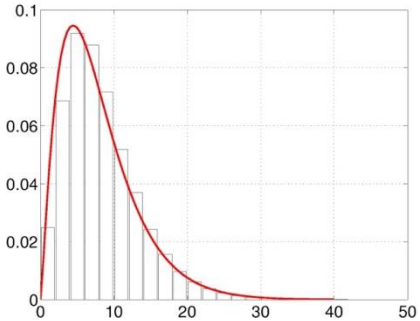
- Borrowed to the performance evaluation domain
  - Poisson
  - Exponentials
  - Markov Chains
  - GMM : Gaussian Mixture models
  - Convex approximations
  - …

Marginal laws and multiscale issue

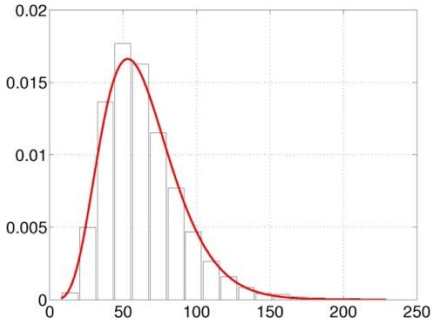
- Distributions of empirical probabilities LBL-TCP-3



Δ=4ms



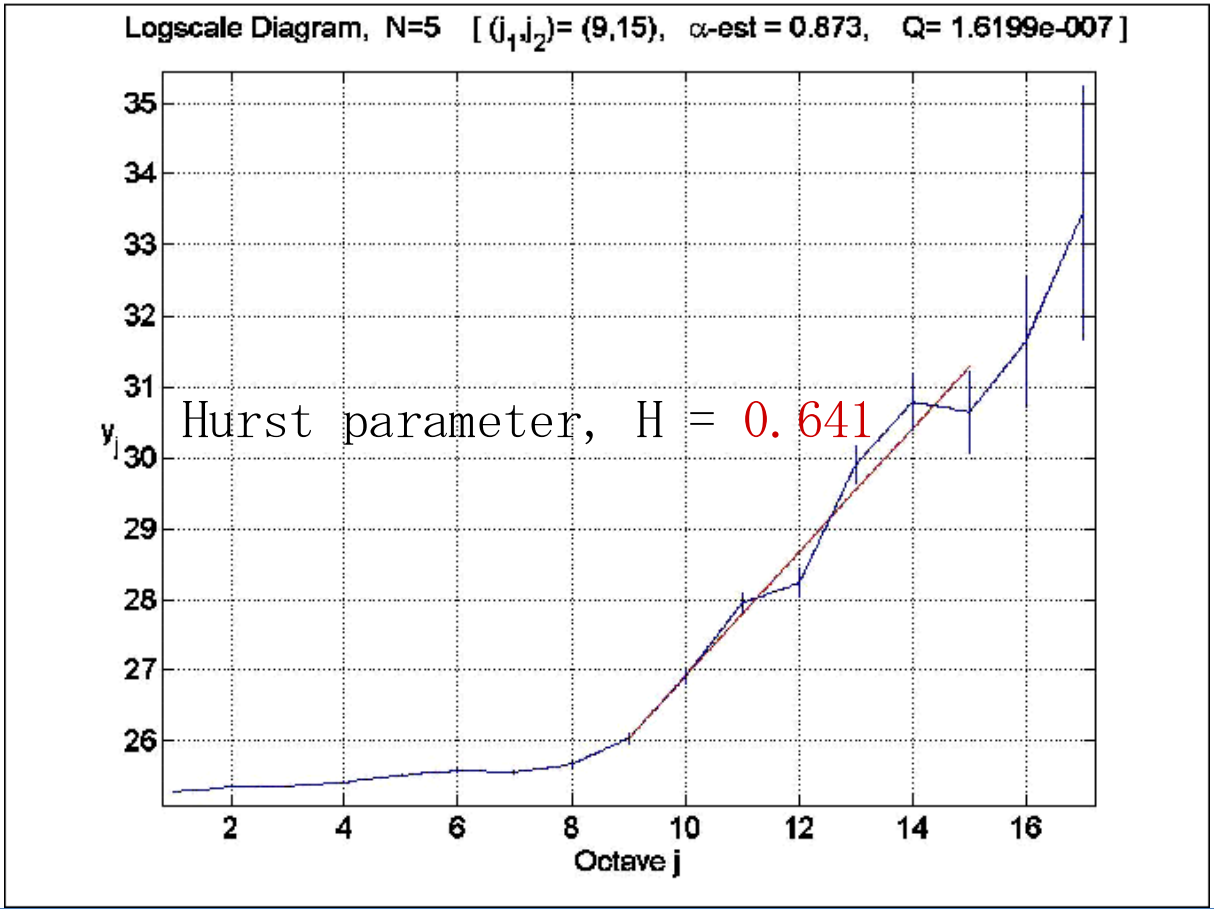
Δ=32ms



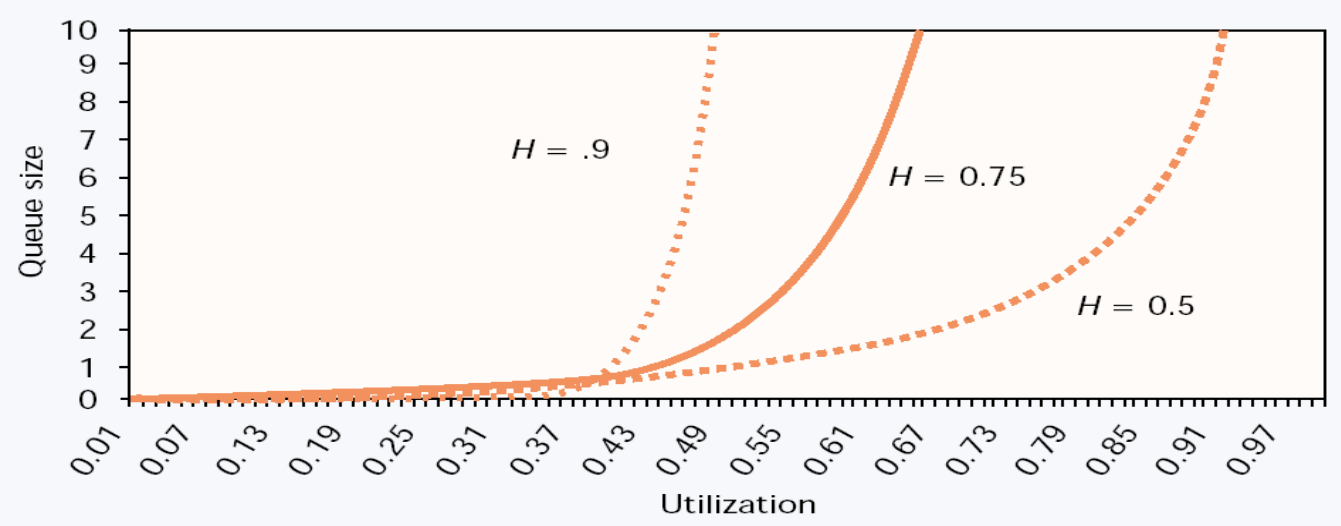
Δ=256ms

- Poisson model? Exponential law? Gaussian?
- What aggregation level to select?

Traffic correlation (SRD and LRD)



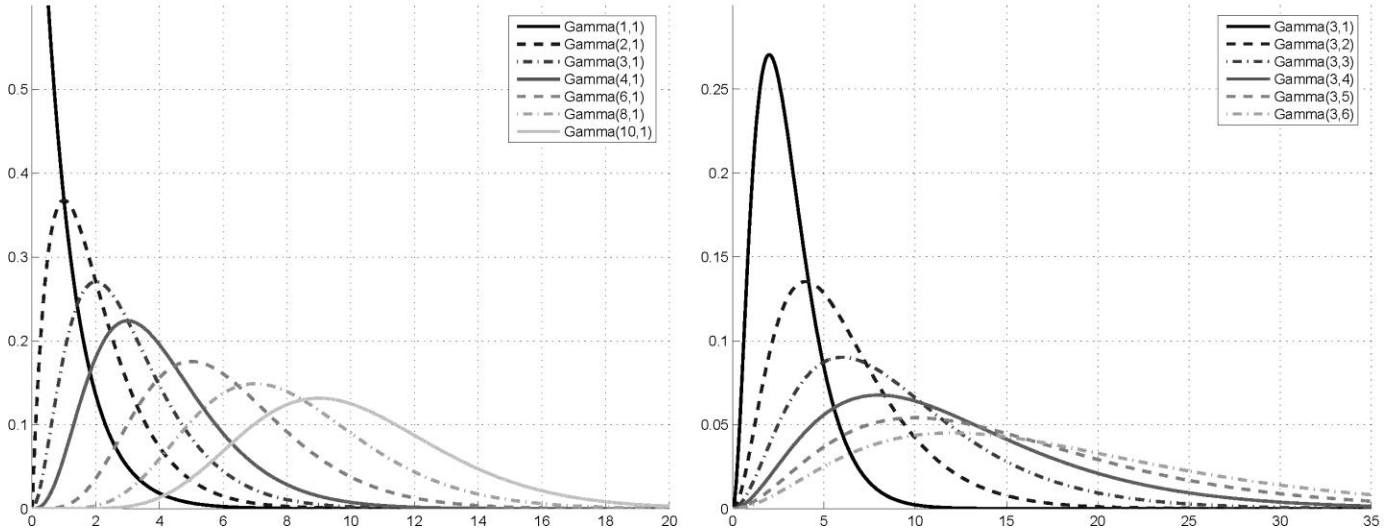
Impact on LDRD on network performances



Relation between LRD , network usage and queue sizes in routers

Gamma distributions

$$\Gamma\alpha,\beta(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)$$



Shape parameter α : can model from Gaussian to exponential ;

$1/\alpha \approx$ distance to Gaussian

Scale parameter β : multiplicative factor

Long memory from a farima model

- Long range dependence

covariance is a non-summable power-law → spectrum $f_{X\Delta}(\nu)$:

$f_{X\Delta}(\nu) \sim C|\nu|^{-\gamma}$, $|\nu| \to 0$, with $0<\gamma<1$

- Farima = fractionnaly integrated ARMA
  1. Fractional integration with parameter d→LRD with $\gamma=2d$
  2. Short range correlation of an ARMA(1, 1)

     →parameters θ, ϕ

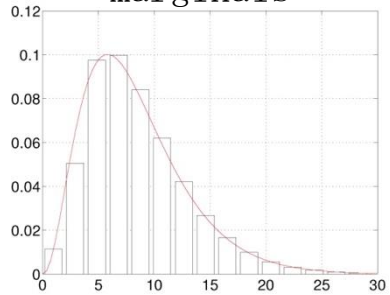
$$f_{X_\Delta}(\nu) = \sigma_\varepsilon^2 \left|1-e^{-i2\pi\nu}\right|^{-2d} \frac{\left|1-\theta e^{-i2\pi\nu}\right|^2}{\left|1-\phi e^{-i2\pi\nu}\right|^2}$$

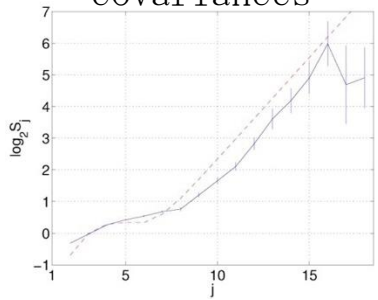
$\Gamma_{\alpha,\beta}$ – farima (ϕ, d, θ) fits

marginals

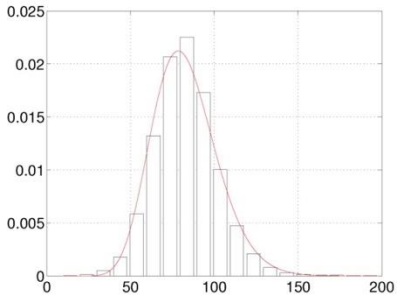
covariances

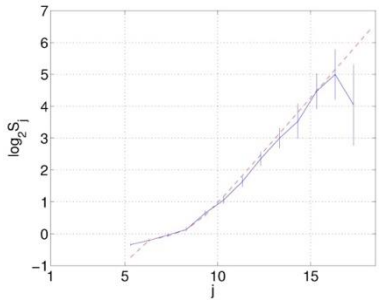
Δ=10ms



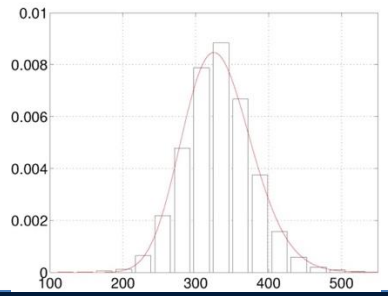


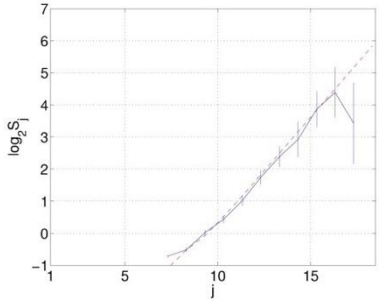
Δ=100ms





Δ=400ms

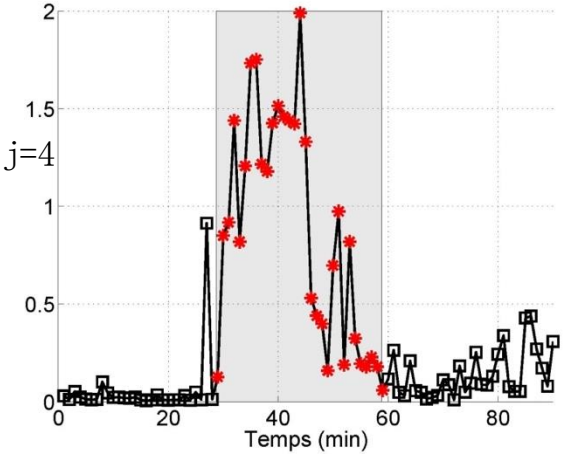




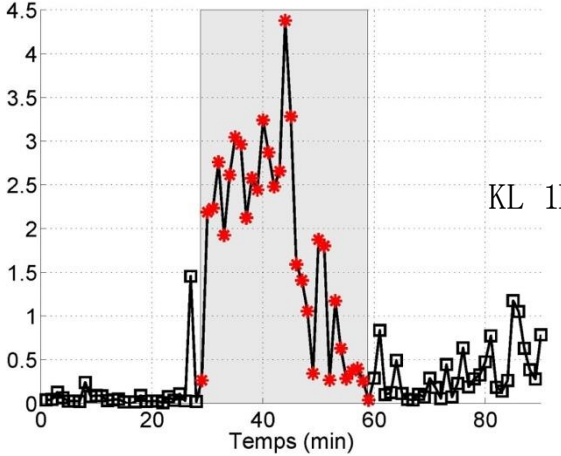
j=1 corresponds to 10 ms

Distance-based detection

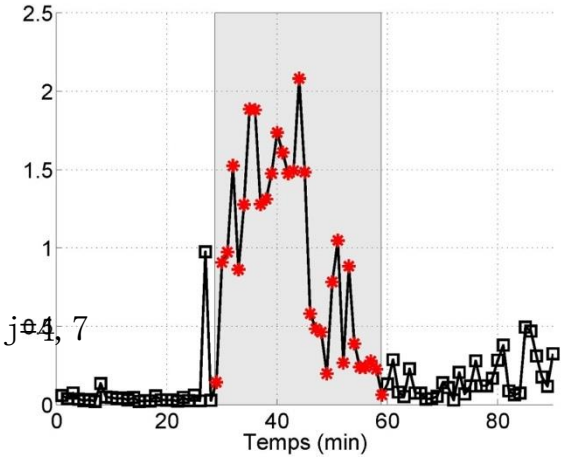
KL 1D, j=4



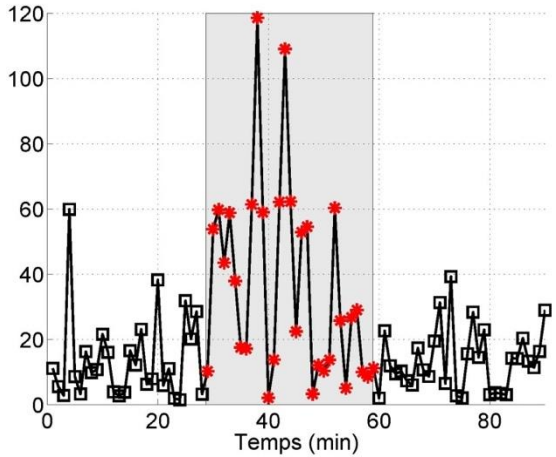
KL 1D, j=7



KL 2D, j=4, 7



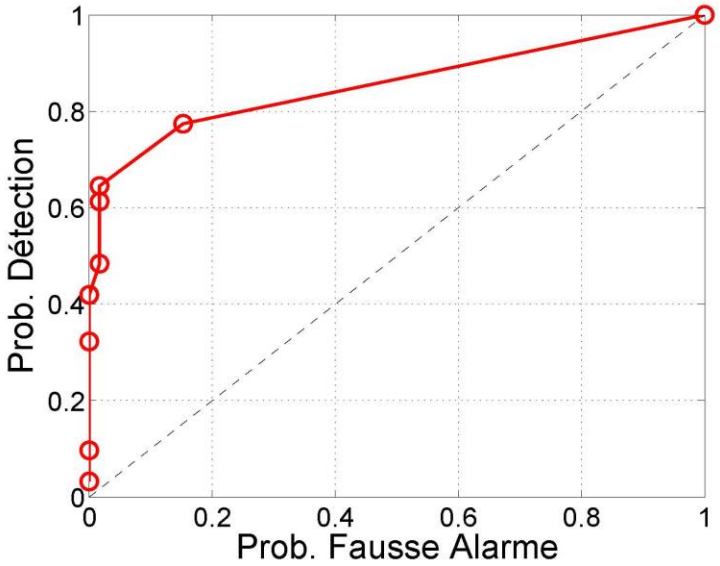
D_α

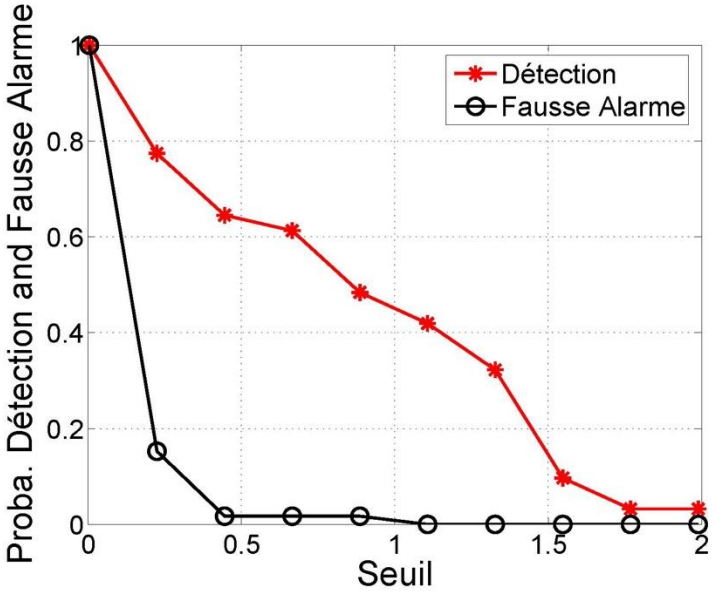


KL: Kullback-Leibler
D: Quadratic

Statistical performance: ROC curves

- ROC curves: detection probability according to the fixed probability of false alarms
- $P_D=f(P_{FA})$ or $P_D=f(\lambda)$, $P_{FA}=f(\lambda)$





From supervised to unsupervised AD

- Current Anomaly Detection (AD) approaches are based on an "acquired knowledge" perspective
  - →signature based
  - →Supervised approaches
- But
  - Network anomalies are a moving target
  - New attacks as well as new variants to already known attacks arise
  - New services and applications are constantly emerging
- And
  - Defense is reactive, often hand made, slow, costly
  - Network and system remain unprotected for long periods

From supervised to unsupervised AD

- Can we detect what we don't know in an evolving Internet ?
- Is current anomaly-detection perspective rich-enough to handle the problem ?
- Is it possible to manage the network security in a self-aware basis to improve performance and reduce operation costs ?

➔unsupervised learning is the idea

- For proactive security (e.g. 0d anomaly detection)
- For autonomous defense system (cost reduction)

A detailed ex. of unsupervised AD

> Approach based on Clustering

> Benefits

(+) no previous knowledge: neither labeled data nor traffic signatures

(+) no need for traffic modeling or training (labeling traffic flows is difficult, time-consuming, and costly)

(+) can detect unknown traffic anomalies

(+) a major step towards self-aware monitoring

> Challenges with clustering

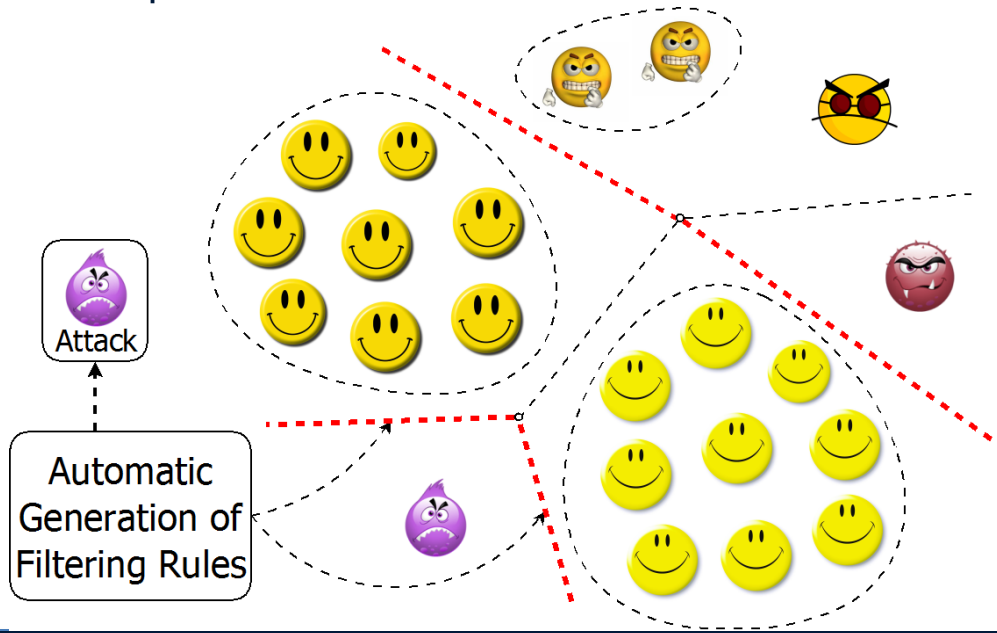
(-) lack of robustness: general clustering algorithms are sensitive to initialization, specification of number of clusters, etc.

(-) difficult to cluster high-dimensional data: structure-masking by irrelevant features, sparse spaces ("the curve of dimensionality")

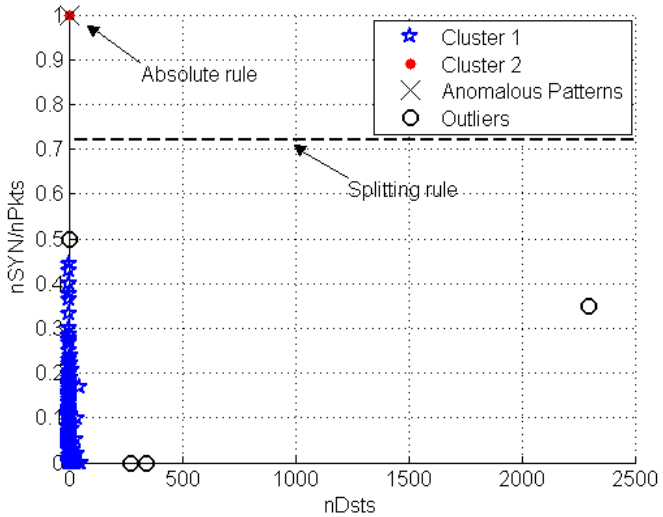
(-) clustering is used only for outliers detection

Filtering rules for anomaly characterization

- Automatically produce a set of filtering rules to correctly isolate and characterize detected anomalous flows
- Select the “best” features to construct a signature of the anomaly, combining the top-K filtering rules
- In a nutshell, select those sub-spaces where anomalous traffic is isolated the best



Detection of a SYN Distributed Denial of Service (DDoS) attack in MAWI



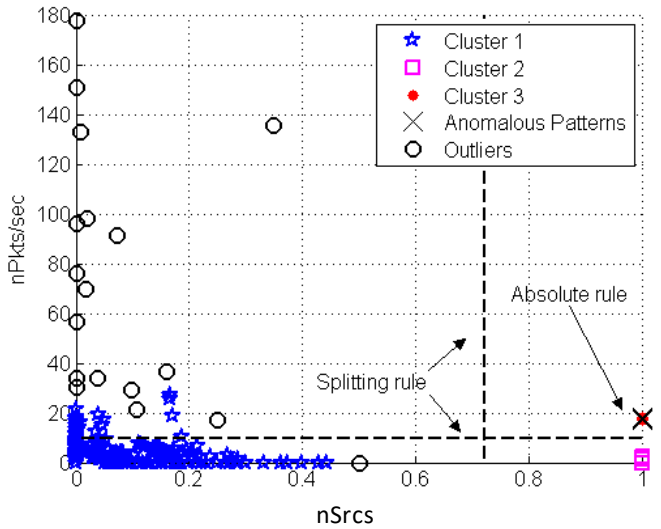


Illustration of clustering graphical results

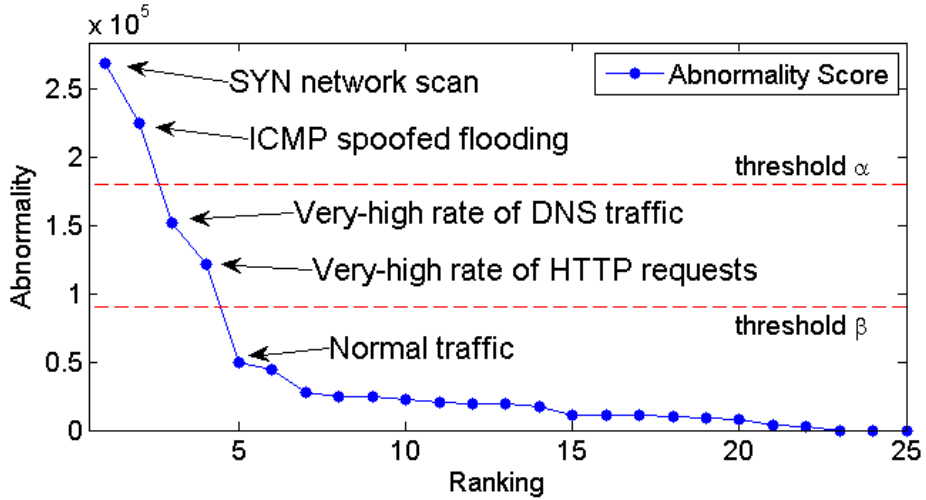
(a) SYN DDoS (1/2) (b) SYN DDoS (2/2)

Generated signature

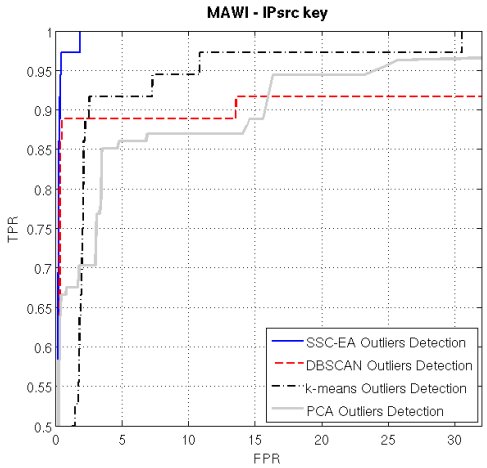
$(nDsts == 1) \wedge (nSYN/nPkts > \lambda_3) \wedge (nPkts/sec > \lambda_4) \wedge (nSrcs > \lambda_5)$

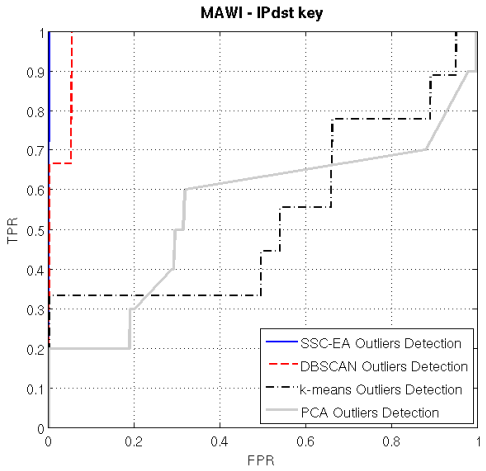
Attacks detection & characterization in MAWI traffic

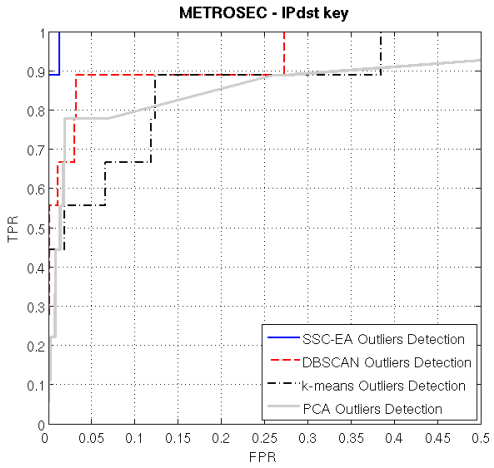
> Detect network attacks that are not the biggest elephant flows



Comparison between ≠ unsupervised techniques







Comparison of detection performance of several detection algorithms

ROC (receiver Operating Characteristic) curves presenting True Positive Rate (TPR) vs. False positive rate (TPR)

Then ML became fashionable !

- Starting with unsupervised ML algos
  - Isolation Forest (IF): algo based on trees
  - Local Outlier Factor (LOF): based on density
  - Random Forest (RF): Based on many Decision Trees (DT)
  - One-Class SVM (OCSVM): using a hyperplane
  - Unsupervised KNN: algo based on distance
  - COPOD: probabilistic algo
  - …

Accuracy of base learners

	Source		Destination	
	TPR	**FPR**	**TPR**	**FPR**
IF	0,6630	0,1462	0,38630	0,1636
LOF	0,8923	0,3754	0,9041	0,3366
DBSCAN	0,2007	0,1032	0,7808	0,0696
OCSVM	0,8394	0,4634	0,8676	0,2375
KNN	0,2579	0,1568	0,8904	0,1862
COPOD	0,7652	0,1467	0,8744	0,1795
RF	**0,9591**	**0,0423**	**0,9690**	**0,0285**

Example of detection decisions

	LOF	KNN	COPOD	Expected decision
Accuracy	90 %	89 %	87 %	
Flow 1	Benign	Attack	Attack	Benign
Flow 2	Attack	Attack	Benign	Attack
Flow 3	Attack	Attack	Benign	Attack
Flow 4	Benign	Benign	Benign	Benign
Flow 5	Benign	Benign	Attack	Attack

Ensemble Learning + Stacking

- Combines several base learners
- Stacking for decision making
  - Majority voting
  - Weighted voting w.r.t. model performance
  - Meta-model

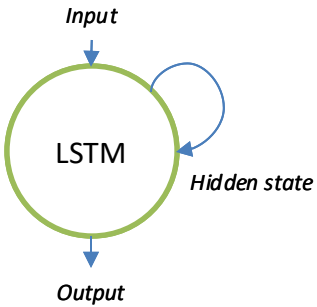
Combination of base learners	TPR	FPR
IF – LOF – OCSVM – COPOD	0,9914	0,6796
LOF – OCSVM – COPOD	0,9878	0,6764
LOF – OCSVM	0,9726	0,6587

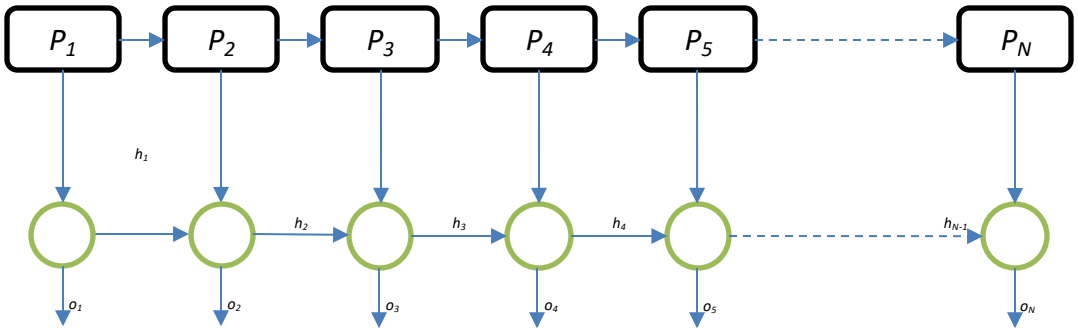
Ens. Learning + stacking and XAI

- State of the Art
  - LIME , SHAP: Explainability based on most impacting features on the decision
- Stacking with a Meta-learner (CNN based) → eXplainability by design
  - Attend PhD defense of Céline Minh on December 20th, 10 am
  - At LAAS-CNRS, Toulouse
    - or
  - Streamed on live.laas.fr

… And Neural Networks invaded our world!

- Deep Learning is the new buzz word
- Especially **Recurrent NN** for intrusion detection
  - Able to make decision based on past events

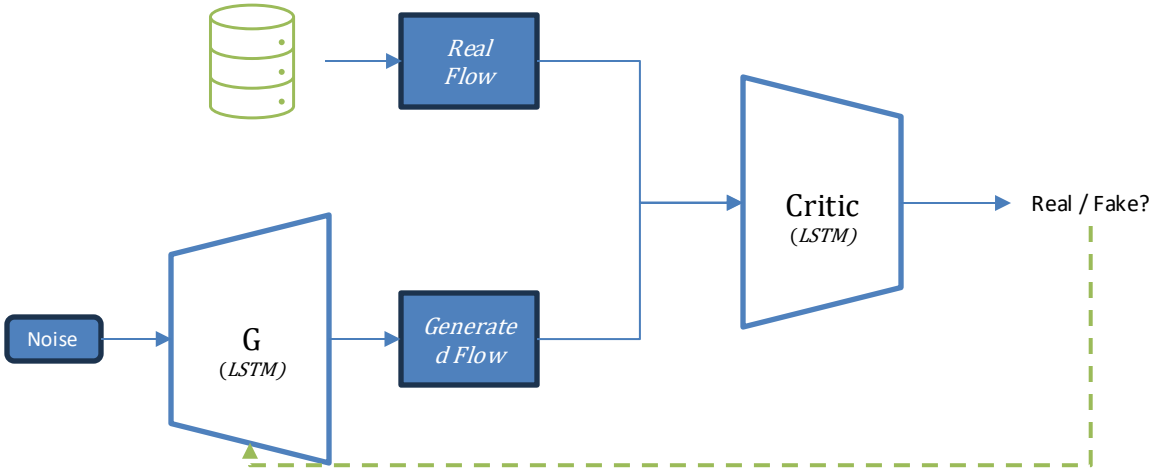




Then Comes the Generative AI Tsunami!

> GAN: Generative Adversarial learning

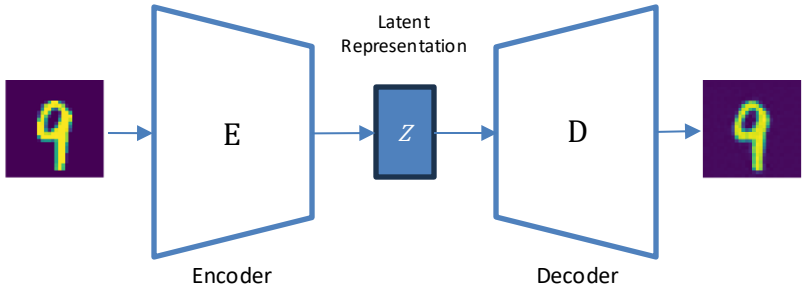
- Detection model
- Generation model

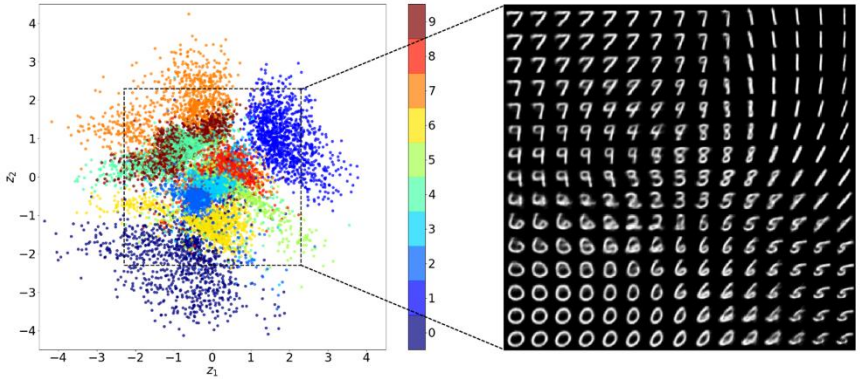


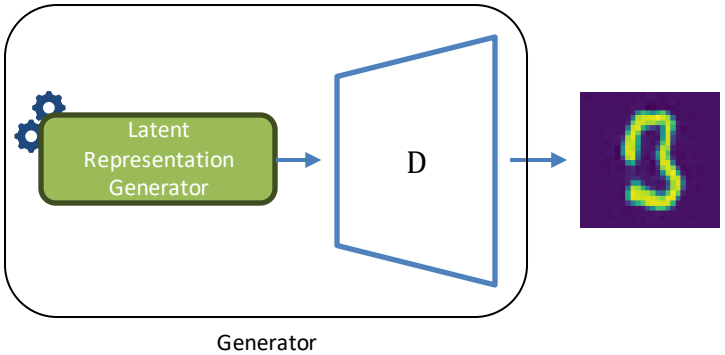
Difficult to generate sequence of discrete data

Representation Learning

> Autoencoders







Source : https://argma.github.io/blog/deep-rae/#rezende2014

Present and future?

- Nowadays challenges
  - Stacking
  - Generative AI
- XAI is the fundament of Trustable AI

- Future trends
  - Introduction of semantic features
    - May NLP techniques help (BERT, … ?)