POLITECNICO
MILANO 1863

Machine Learning Methods for Communication Networks and Systems

Francesco Musumeci

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)

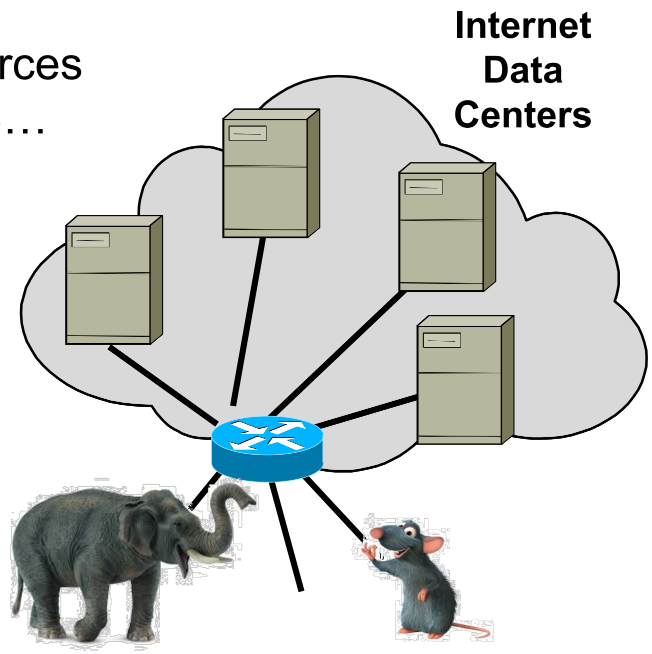
Politecnico di Milano, Milano, Italy

Part II – 9: Flow classification

Network layer domain

Traffic classification

- Communication networks usually serve heterogeneous traffic flows in terms of:
  - protocols (http, ftp, smtp…)
  - services (fixed vs mobile, VoD, data transfer, text messages…)
  - requirements (latency, bandwidth, jitter…)
  - network "customers" (human end-users, companies, sensors, macines, servers…)
    - E.g., "mice" vs "elephant" flows in Data Centers
- Distinguish between different flows is crucial for resources (i.e., capacity) allocation, scheduling, security/privacy, QoS…
- Traditional classification uses partial information (source/dest IP address, protocol, port number etc.)
  - often unavailable (e.g., due to tunneling or cryptography)
  - sometimes insufficient (e.g., same protocols can carry flows with highly different characteristics)
  - maybe misleading: different protocols can carry flows with similar characteristics
- ML
  - enables traffic features extraction from direct observation of traffic flows
  - allows simultaneous use of heterogeneous features

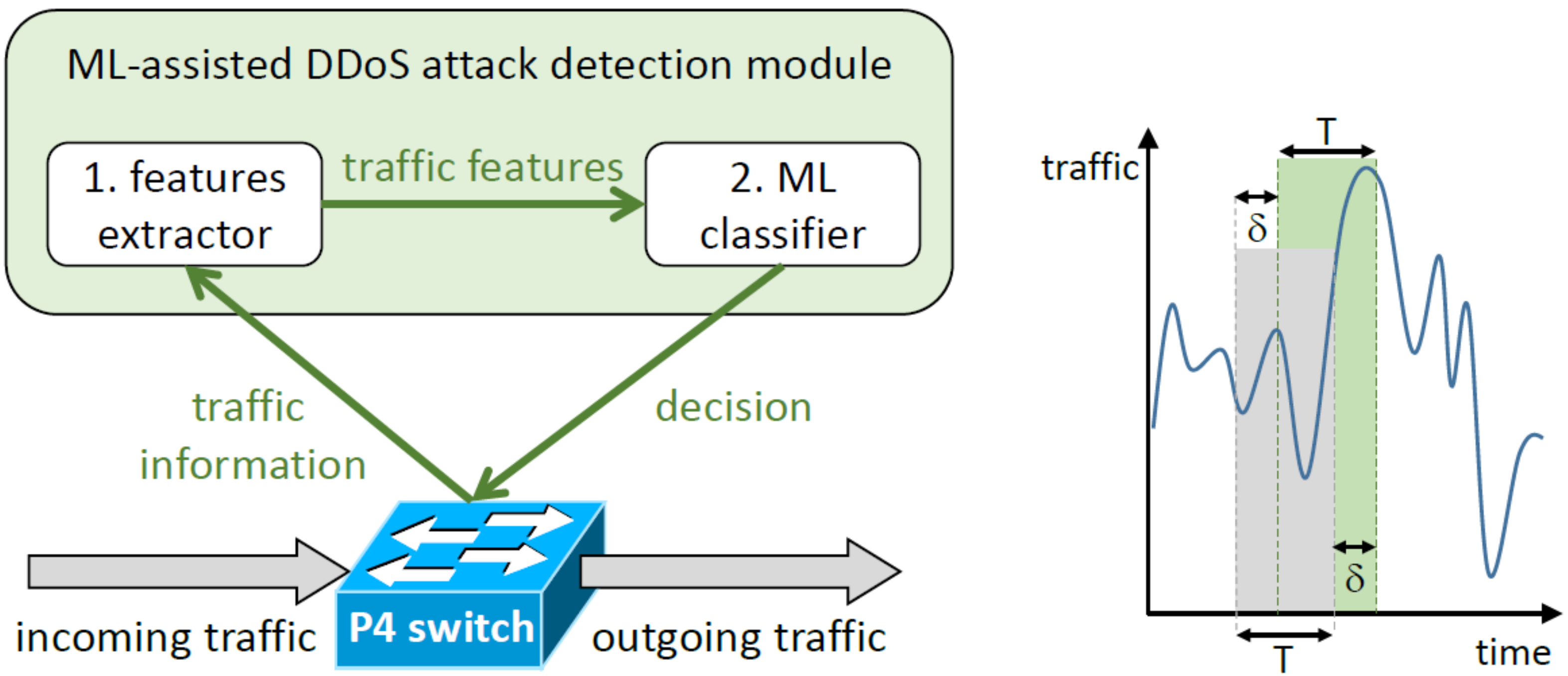


Traffic classification

Source 1

- F. Musumeci *et al.*, “Machine-Learning-enabled DDoS attacks detectionin P4 programmable networks”, *Springer Journal of Network and Systems Management, 30 (21) Nov. 2021*
- Paper objective: detect Distributed Denial of Service (DDoS) attacks
  - input
    - features extracted from headers of IP packets
  - output
    - labeled "windows" (set of packets within a time frame) indicating if at least one attack packet is present
  - ML algorithms: KNN, SVM, RF, ANN

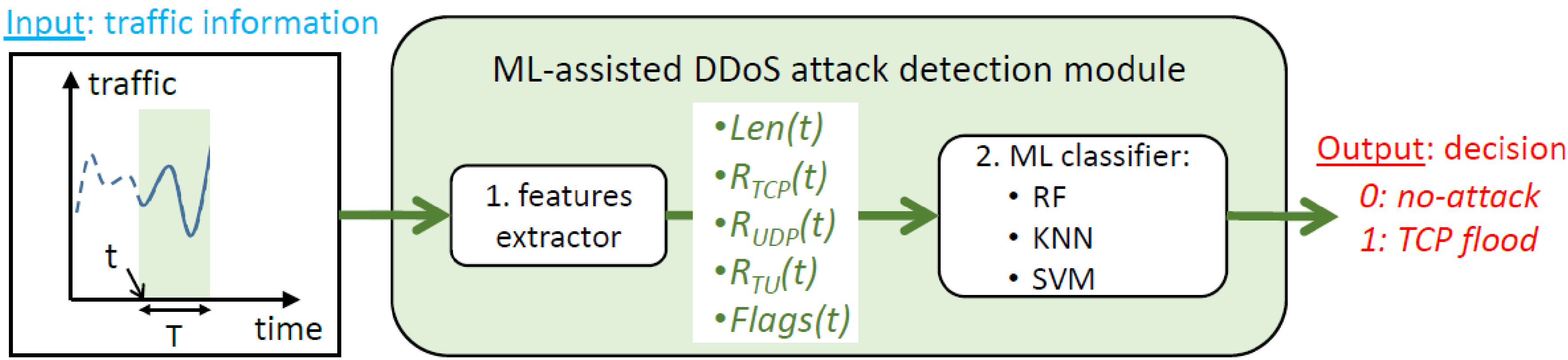
Traffic classification
Source 1



(a) Detection framework and functional blocks (b) Traffic window

Traffic classification
Source 1

- Features
  - Len(t): average size in bytes of packets in window (t, t + T)
  - R_{TCP}(t): percentage of TCP packets in window (t, t + T)
  - R_{UDP}(t): percentage of UDP packets
  - R_{TU}(t): ratio between TCP and UDP packets in window (t, t+T)
  - Flags(t): percentage of TCP packets with an active SYN flag out of the total in window (t, t + T)



Traffic classification

Source 1

- Data set

Parameter	Value
Traces duration	15 minutes
TCP traffic bit rate	13.5 Mbit/s
UDP traffic bit rate	11.4 Mbit/s
IP traffic bit rate	5.1 Mbit/s
Attack traffic bit rate	26.5 kbit/s
Attack Type	SYN flood
Window duration	$T \in \{0.5; 1; 2; 10\}$ seconds
Windows distance	$\delta \in \{0.01; 0.05; 0.1; 0.2; 0.5; 1\}$ seconds

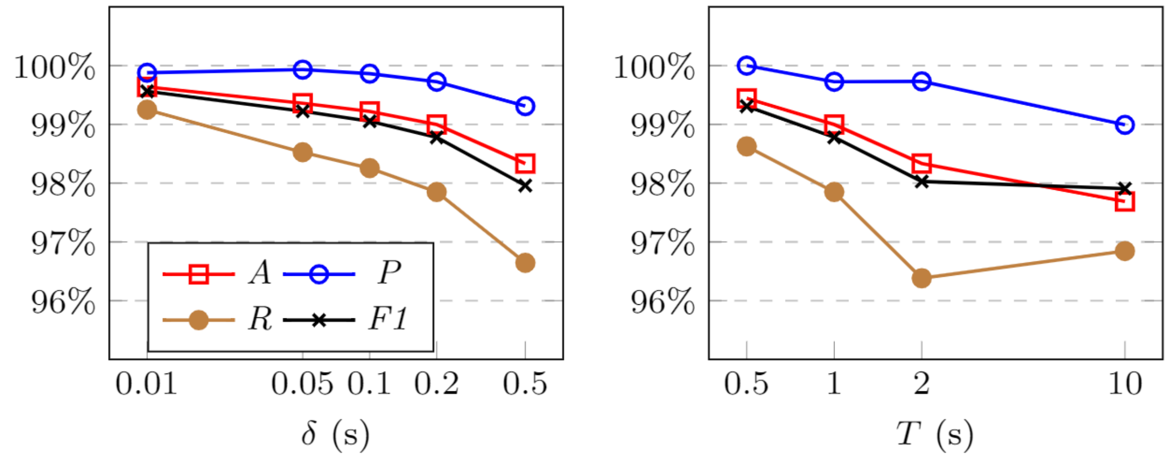
- Hyperparameters

Algorithm	Parameter	Tested values	Selected value
KNN	no. of neighbors K	{3, 4, 5, 6, 7, 8, 9, 10}	3
	neighbors weight	{uniform, distance-based}	uniform
RF	splitting criterium	{Gini, Entropy}	Gini
	no. of trees	{10, 20, 30, 40, 50, 60, 70, 80, 90, 100}	10
SVM	kernel	{sigmoid, rbf, polynomial}	rbf
	regul. param. C	$\{1, 10, 10^2, 10^3, 10^4\}$	10^3
	kernel coefficient γ	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$	10^{-2}
ANN	no. of hidden layers	{1,2}	2
	no. of neurons per layer	{5,10,11}	10
	activation function	{sigmoid, relu, elu}	elu

Traffic classification

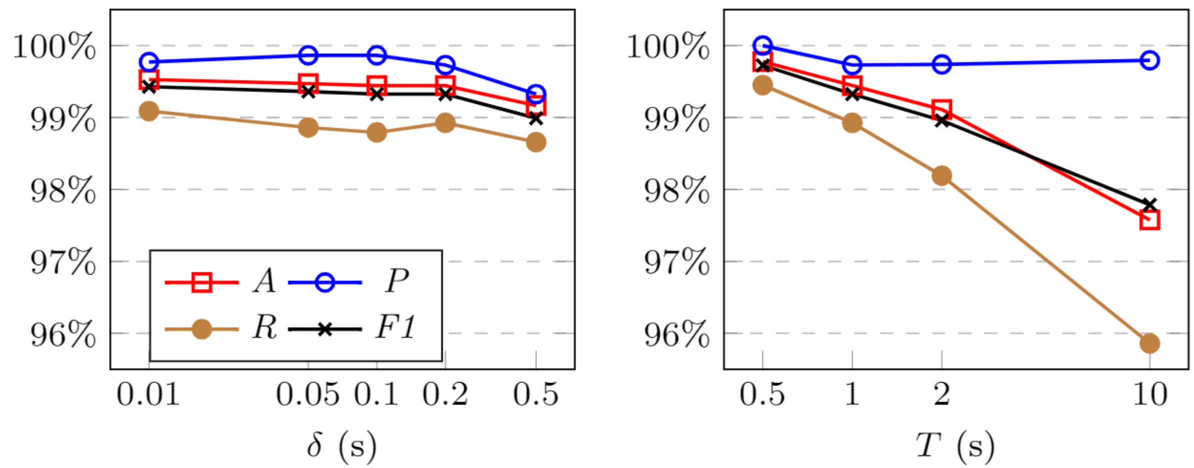
Source 1

- KNN



(a) Comparison over δ ($T = 1$ s) (b) Comparison over T ($\delta = 0.2$ s)

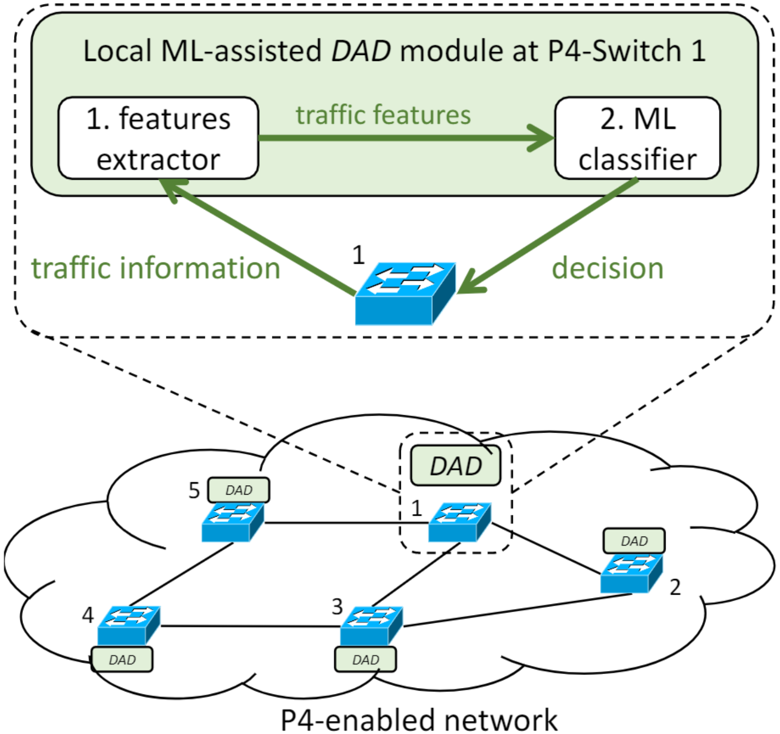
- ANN



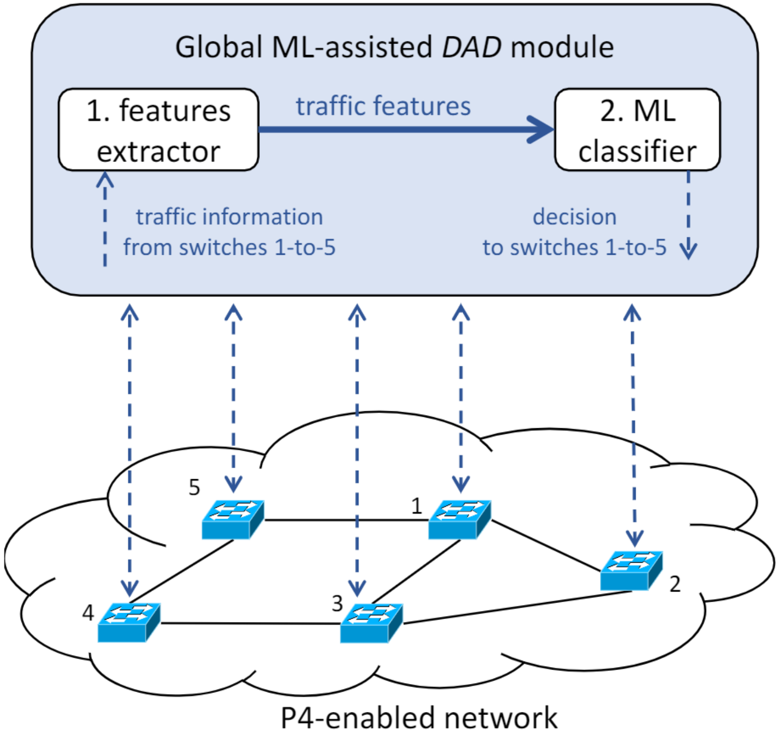
(a) Comparison over δ ($T = 1$ s) (b) Comparison over T ($\delta = 0.2$ s)

Traffic classification
Source 1

- Standalone vs correlated DDoS attack detection (DAD)



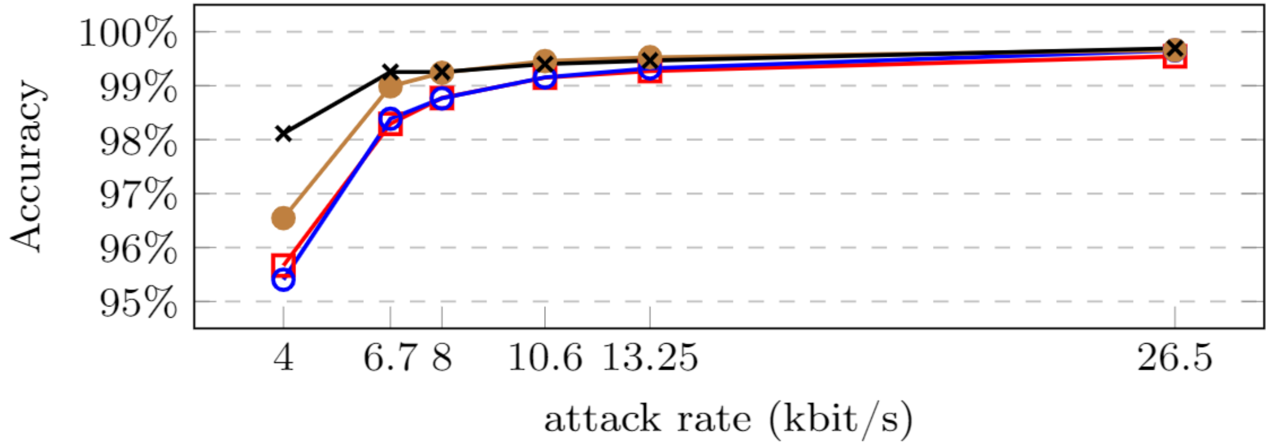
(a) *Standalone DAD*



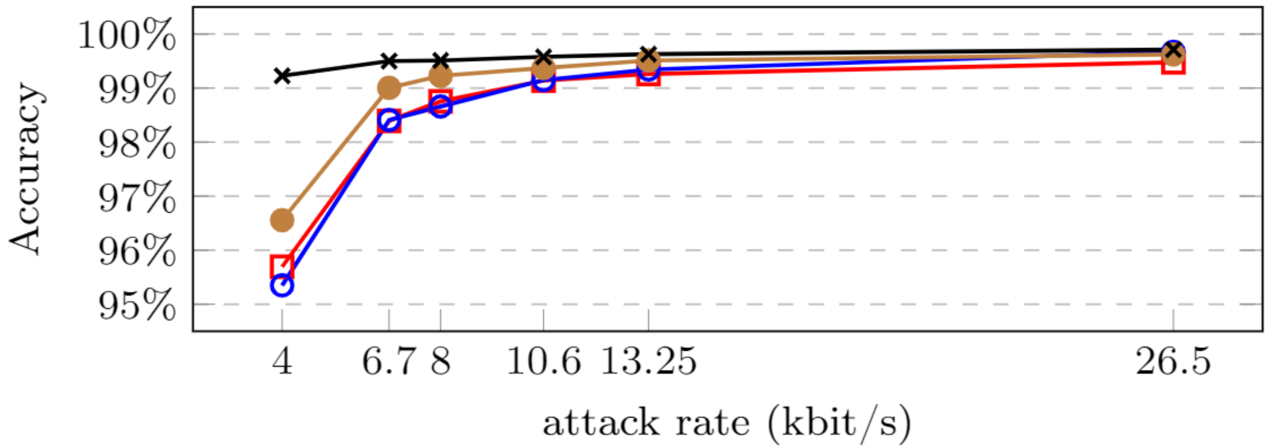
(b) *Correlated DAD*

Traffic classification

Source 1



(b) RF



(c) SVM

Traffic classification

Source 2

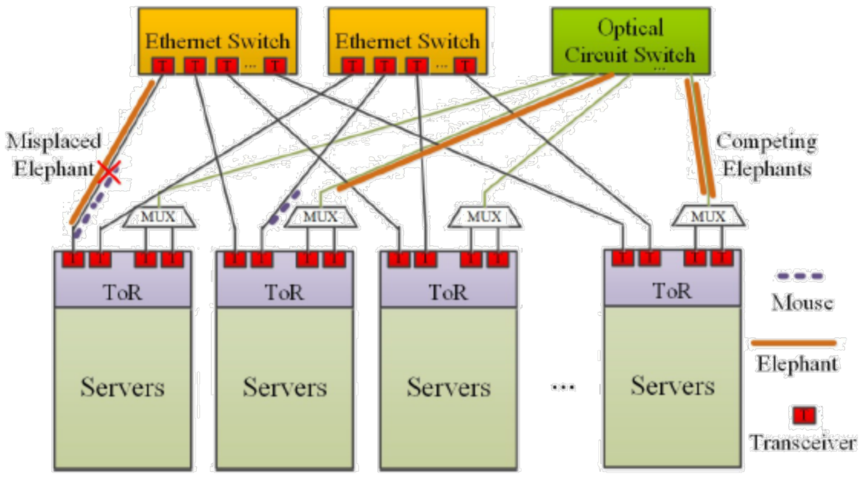
- Viljoen *et al.*, “Machine Learning Based Adaptive Flow Classification for Optically Interconnected Data Centers”, in *ICTON 2016*, July 2016
- Paper objective: optimal flow allocation in multi-tenant DC networks
  - input
    - information retrieved from incoming packets headers (40 packets per flow)
  - output
    - Labeled flows (mice or elephant)
  - ML algorithm: Neural Network

Traffic classification

Source 2

*elephant flows: > 100MB

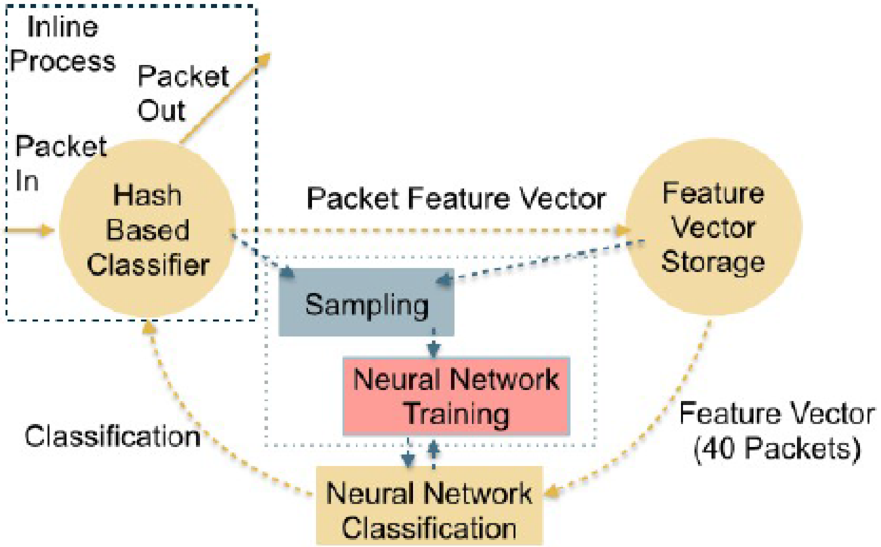
- All-electrical DCs: elephant flows* are typically distributed uniformly across the DC links
- Hybrid electrical-optical DCs: elephant flows tend to be assigned to optical links and switches
  - larger bandwidth & lower latency
- Proper classification of mice and elephant flows can be useful to allocate flows to proper resources within a DC (i.e., servers, switches, tx/rx equipment…)



- Misclassification can lead to
  - resource underutilization (mice flows assigned to optical links/switches)
  - lack of resources (bulk data transfer, i.e., elephant flows, assigned to electrical links/switches)

Traffic classification

Source 2



- NN characteristics
  - 4 hidden layers
  - Features of flows
    - src/dest IP address
    - src/dest port
    - protocol
    - packet size
    - intra-flow timings (within first 40 packets of a flow)
- NN is compared to existing heuristic approach:
  - Flows increasing their bandwidth more than 10% in 1 second are tagged as "elephant"
- Data set of 24h sampled every 20 minutes
  - 4% traffic flows as elephant (summing to 94% of data transferred)
  - different traffic types dominant at every hour of the day

Traffic classification

Source 2

- Results: mice vs elephant NN-classifier <u>accuracy</u>

Flow classification: +22% wrt heuristic

Total Flows 78674

Neural Net

	P_e	P_m	
T_e	2332	739	TPR 0.76
T_m	2060	73542	TNR 0.97
	Sn. 0.53	Sp. 0.99	

Heuristic

	P_e	P_m	
T_e	1644	1426	TPR 0.54
T_m	1122	74482	TNR 0.98
	Sn. 0.59	Sp. 0.99	

Per-byte classification: +22% wrt heuristic

Total Bytes: 2.26TB

Neural Net

	P_e	P_m	
T_e	0.70	0.24	TPR 0.75
T_m	0.017	0.043	TNR 0.72
	Sn. 0.98	Sp. 0.15	

Heuristic

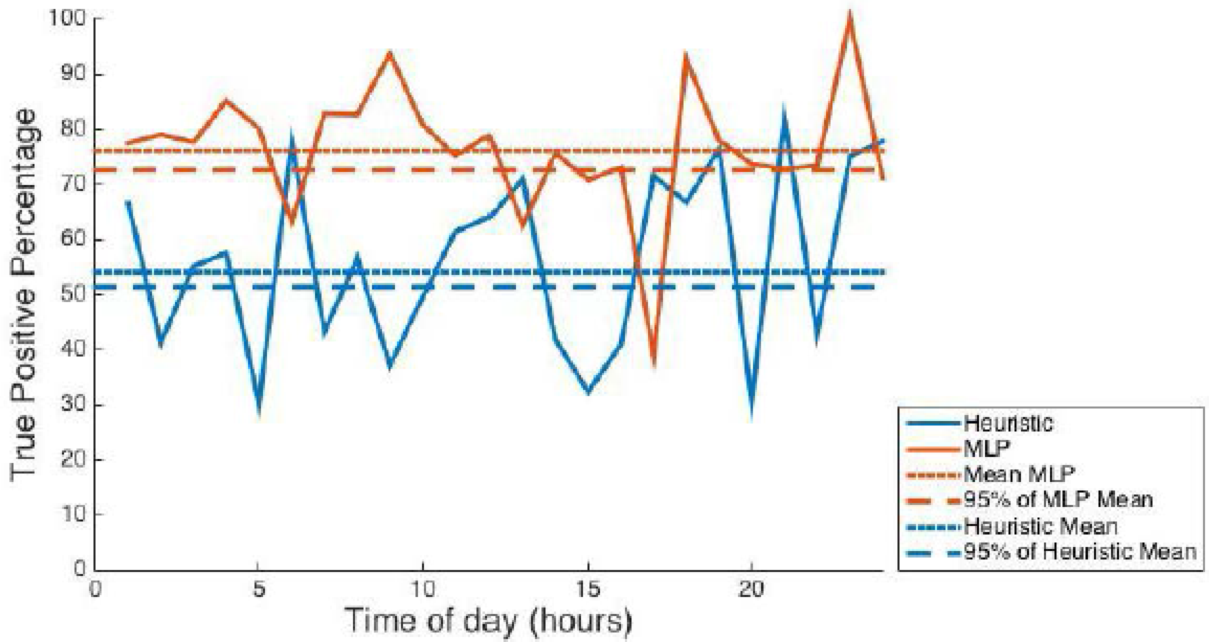
	P_e	P_m	
T_e	0.46	0.48	TPR 0.49
T_m	0.002	0.058	TNR 0.96
	Sn. 1.00	Sp. 0.11	

TPR = True Positive Ratio
TNR = True Negative Ratio
Sn = Sensitivity
Sp = Specificity

Traffic classification

Source 2

- Results: prediction consistency
  - Lower variance with time evolving situations in true positives obtained w/ NN (aka Multilayer Perceptron, MLP)
  - At most 1h period of performance 5% below the mean (see h16-h17)



Traffic classification

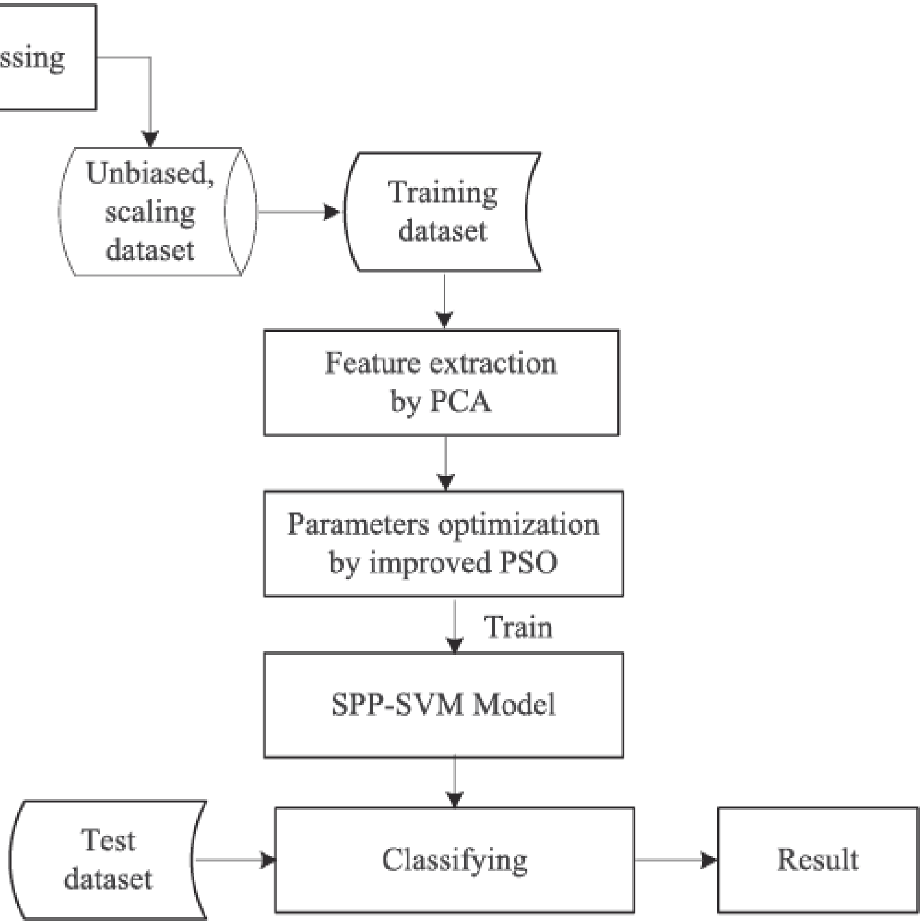
Source 3

- Cao *et al.*, "An accurate traffic classification model based on support vector machines", *International Journal on Network Management*, 27:e1962, 2017.
- Paper objective: classify internet traffic from/into a research facility hosting about 1k users
  - input
    - features extracted from IP packets headers
  - output
    - labeled flows
  - ML algorithm: SVM

Traffic classification

Source 3

- Original dataset[1]
  - 240+ features
    - Clt/server port number, IAT, and various statistics…
  - Preprocessed to have normalized features
- PCA to reduce dimensionality
- SVM characteristics
  - 2-class (1 vs all) & multiclass (any vs any)
  - 4 different kernels: linear, polynomial, radial basis function (RBF) and sigmoid
    - Hyperparameters optimized via particle swarm optimization[2]



[1] Moore AW, *et al.* "Discriminators for use in flow-based classification". *Tech. Rep. RR-05-13, Department of Computer Science, Queen Mary, University of London*, 2005;1–16.

[2] Tayal VK, Lather JS. Reduced order H∞ TCSC controller & PSO optimized fuzzy PSS design in mitigating small signal oscillations in a wide range. *Int. J. of Electrical Power and Energy Systems*. 2015;68:123–131.

Traffic classification
Source 3

- Dataset: 10 different classes (flow-types)

Traffic class	WWW	Mail	FTP-control	FTP-pasv	Attack
Representative applications	HTTP and HTTPS	Pop2/3, smtp, and imap	FTP	FTP	worm and virus
Samples of flows	2999	2999	2990	2989	1793

Traffic class	P2P	Database	FTP-data	Multimedia	Services
Representative applications	Kazaa, BitTorrent, nd Gnutella	Postgres, sqlnet, oracle, and ingres	FTP	Voice and video streaming	X11, dns, ident, and ntp
Samples of flows	2391	2943	2997	576	2220

Traffic classification

Source 3

- Results: impact of SVM-kernel on accuracy
  - Highest Avg accuracy for 2-class is w/ RBF(85%)

	Two-class SVM				
Classifier	WWW	Mail	FTP-control	FTP-pasv	Attack
Linear	99.9036	87.4839	12.01	30.1816	88.6166
Polynomial	99.7269	95.5174	75.0321	12.0019	86.8011
RBF	87.9579	88.1748	88.0141	87.9981	92.794
Sigmoid	87.9499	82.222	87.982	87.9981	92.794
Classifier	P2P	Database	FTP-data	Multimedia	Services
Linear	52.0566	44.2561	67.0389	64.8538	77.3618
Polynomial	84.8249	11.8172	99.4457	95.9672	99.0681
RBF	90.4804	88.3194	88.3355	97.6944	91.5087
Sigmoid	87.5884	88.1668	87.9579	97.6864	91.0749

Traffic classification

Source 3

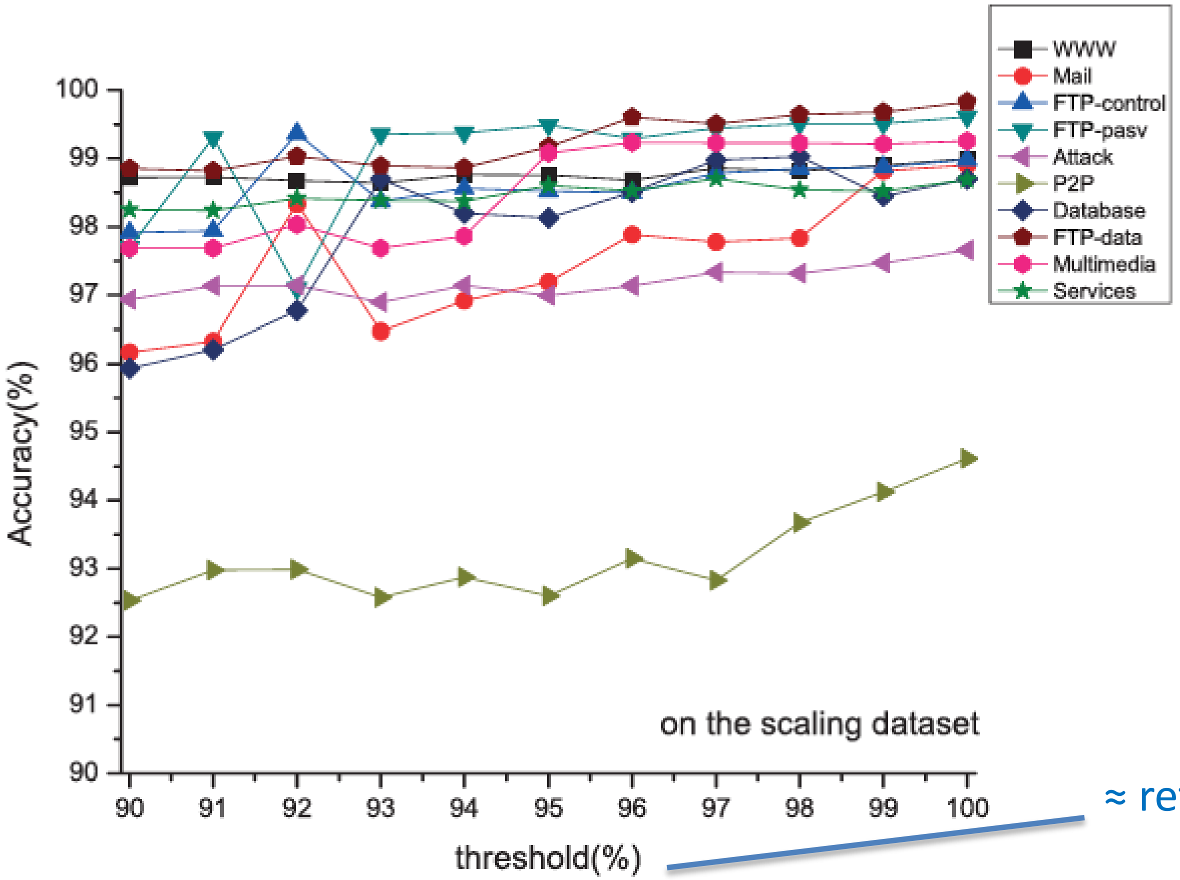
- Results: impact of features scaling on accuracy
  - 2-class RBF only: accuracy >94% (max =99.8%)

	Two-class SVM				
Classifier	**WWW**	**Mail**	**FTP-control**	**FTP-pasv**	**Attack**
Accuracy (%) on original data	87.9579	88.1748	88.0141	87.9981	92.794
Accuracy (%) on scaling data	98.7789	98.9878	98.9798	99.6546	97.5016
Classifier	**P2P**	**Database**	**FTP-data**	**Multimedia**	**Services**
Accuracy (%) on original data	90.4804	88.3194	88.3355	97.6944	91.5087
Accuracy (%) on scaling data	94.0312	99.036	99.8233	99.2449	98.6102

Traffic classification

Source 3

- Results: impact of dimension reduction (PCA) on accuracy
  - Original dataset



≈ retained variance

Traffic classification
Source 3

- Results: impact of dimension reduction (PCA) on accuracy
  - Dataset after features scaling (more stable accuracy)