Use Cases of Applying Machine Learning Mechanism with Network Traffic

(draft-jiang-nmlrg-traffic-machine-learning-00)

Sheng Jiang (Huawei)
Bing Liu (speaker) (Huawei)
Jerome Francois (Inria)
Giovane C. M. Moura (SIDN Labs)
Pere Barlet (Network Polygraph)

June 27， EuCNC2016

Background

- NMLRG meeting in IETF95
  - Focused on ML (Machine Learning) applied in network traffic handling
  - Some distinct use cases from different organizations were presented in the meeting
  - This presentation is the collection of the use cases

Why Focused on Network Traffic?

- The user contents within traffic is becoming more diverse due to the development of various network services, and increasing use of encryption.
- It is more and more challenging for administrators to get aware of the network's running status (such as performance, failures, and security etc.) and efficiently manage the network traffic flows.
- It is natural to utilize machine learning technology to analyze the large mount of data regarding network traffic , to understand the network's status. The analyzed objectives could be:
  - Measurable properties: latency, packet number, duration etc.
  - Protocol metadata: headers, source/dest IP addresses, port number etc.
  - User content: webs, audio, video etc.
  - Network signaling, routing signaling, MPLS-TE, P2P etc.

#1: HTTPS Traffic Classification

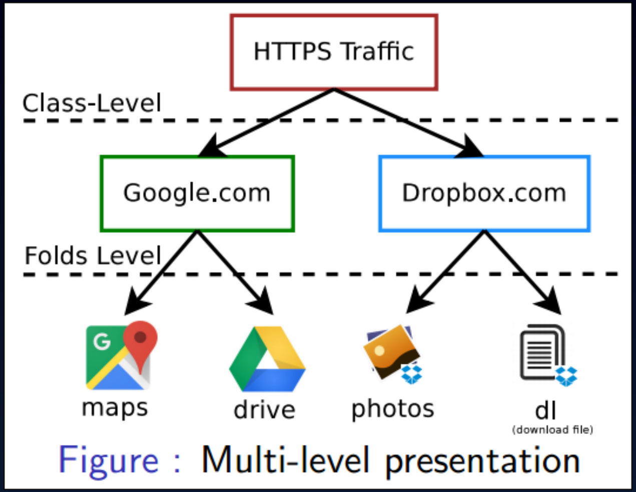
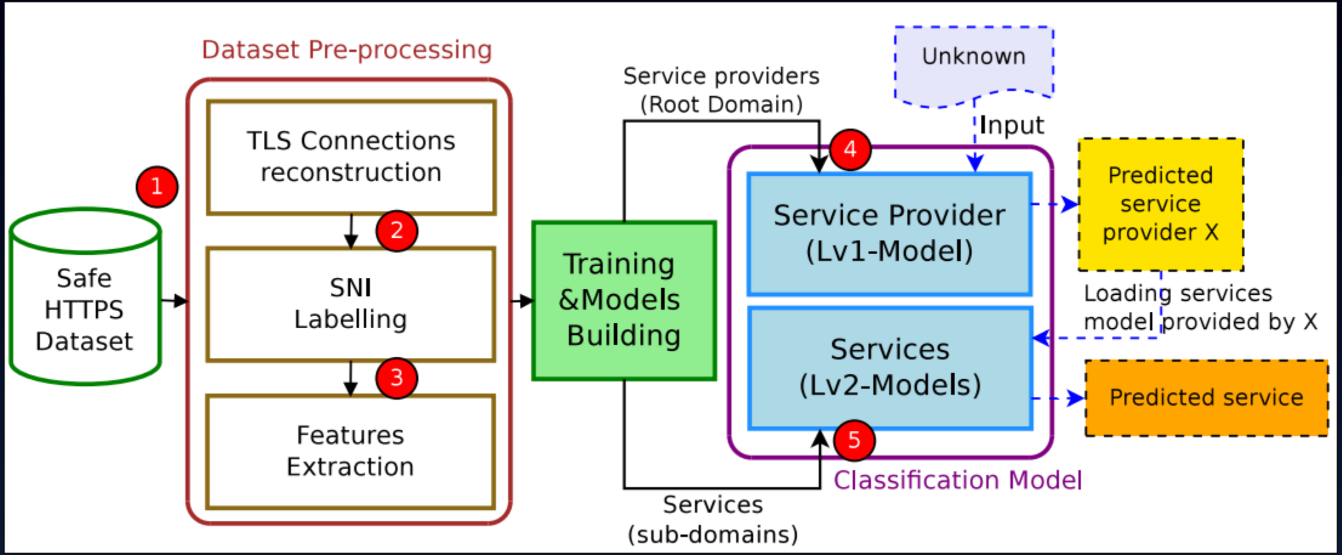


Figure : Multi-level presentation


Dataset Pre-processing
Safe HTTPS Dataset
1
TLS Connections reconstruction
2
SNI Labelling
3
Features Extraction
Training &Models Building
Service providers (Root Domain)
Services (sub-domains)
Unknown
Input
4
Service Provider (Lv1-Model)
Services (Lv2-Models)
5
Classification Model
Predicted service provider X
Loading services model provided by X
Predicted service

- The objective is to automatically label an HTTPS connection by the service and service provider associated with.

- A multi-level ML approach has been proposed:
  - a first level model (L1 model) whose the goal is to identify the service provider
  - a set of second level models (L2 models), one for each service provider to identify specific service of a service provider

Results

Evaluate the framework as black-box (Level1&2)

Results show that we achieve 93.10% of Perfect identification and 2.9% of Partial identification.

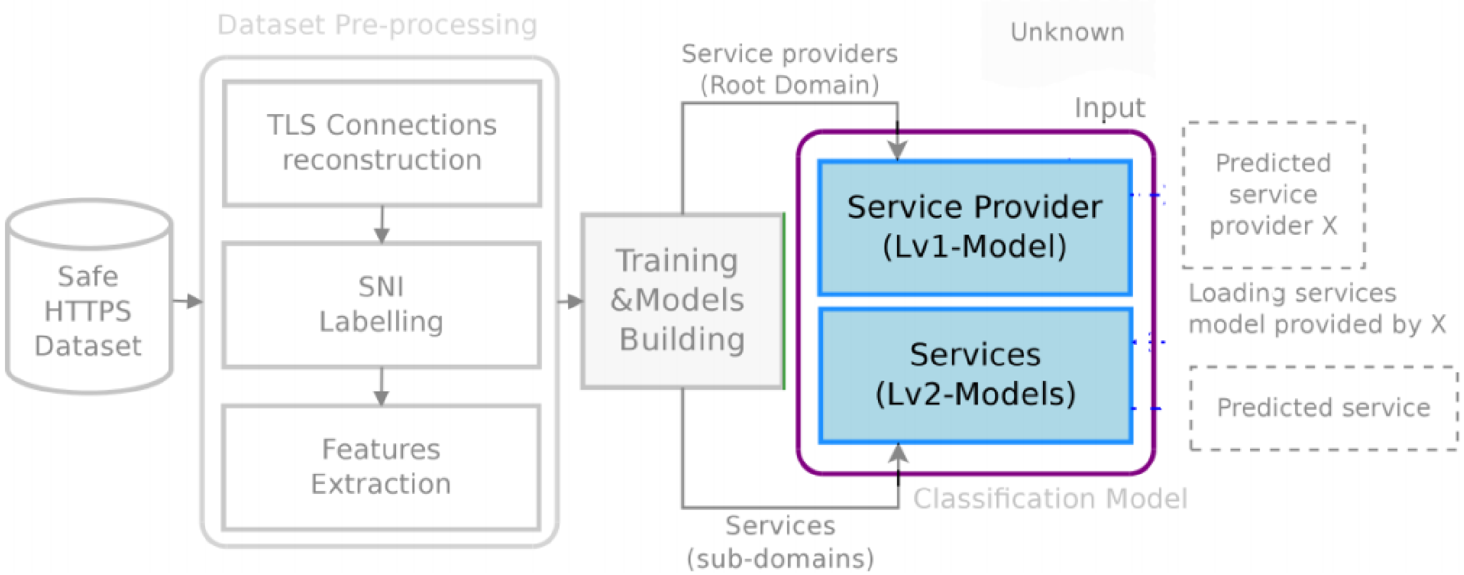

Dataset Pre-processing
Safe HTTPS Dataset
TLS Connections reconstruction
SNI Labelling
Features Extraction
Training &Models Building
Service providers (Root Domain)
Services (sub-domains)
Unknown
Input
Service Provider (Lv1-Model)
Services (Lv2-Models)
Classification Model
Predicted service provider X
Loading services model provided by X
Predicted service

Figure : The complete classification model

#2: Automatic Malicious Domain Detection with DNS Traffic Analysis

- Data to be analyzed:

  - a domains registration database and

  - authoritative DNS server traffic data, which is typically the case for Top-Level Domains (TLD) registries.

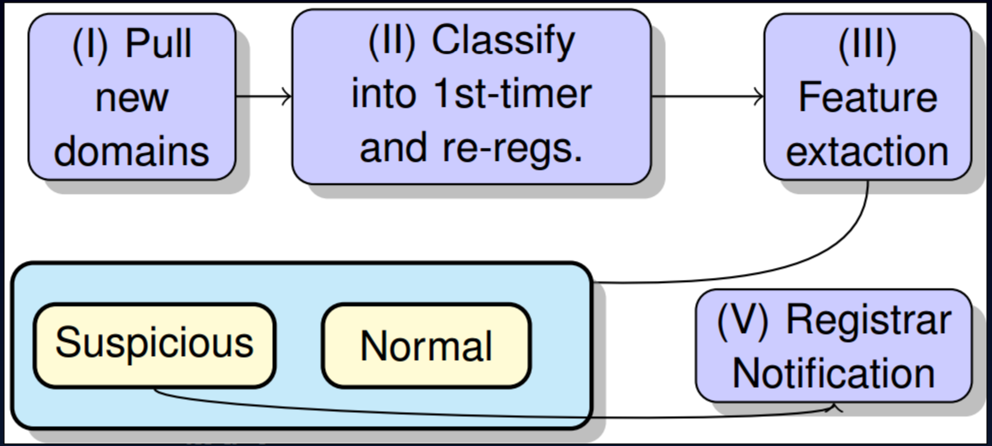
- These domains are classified using k-means as a clustering method into two clusters using four features extracted from the analyzed DNS traffic:

  - DNS queries

  - IP addresses

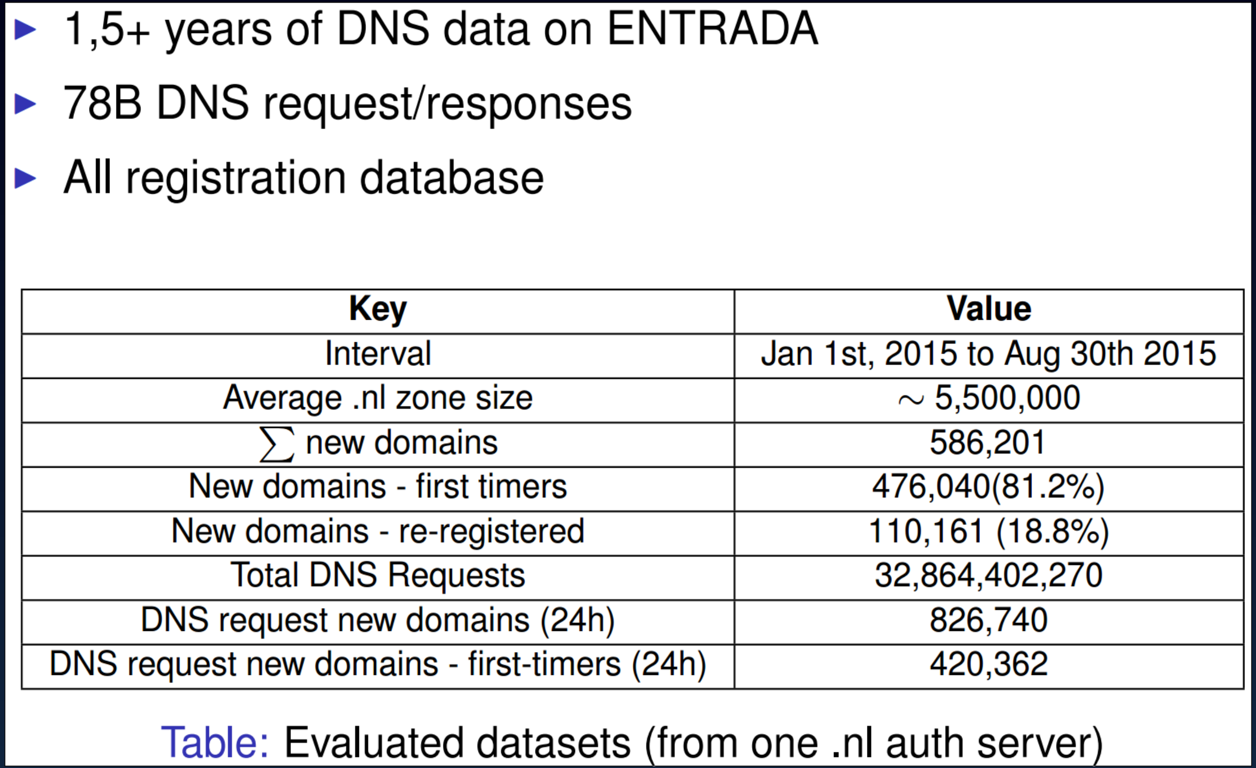
  - Autonomous Systems (Ases)

  - Countries, which were chosen empirically


(I) Pull new domains
(II) Classify into 1st-timer and re-regs.
(III) Feature extaction
Suspicious
Normal
(V) Registrar Notification

- nDEWS (New Domains Early Warning System), a tool that classifies the newly registered domains based on their initial lookup pattern.

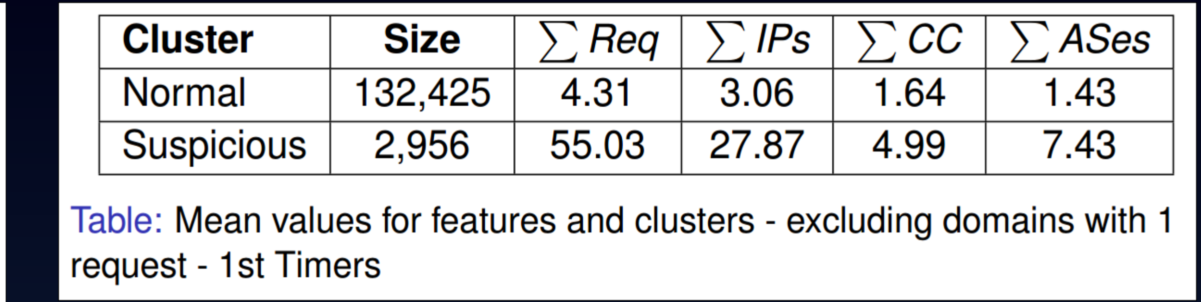
Results



- 1,5+ years of DNS data on ENTRADA
- 78B DNS request/responses
- All registration database

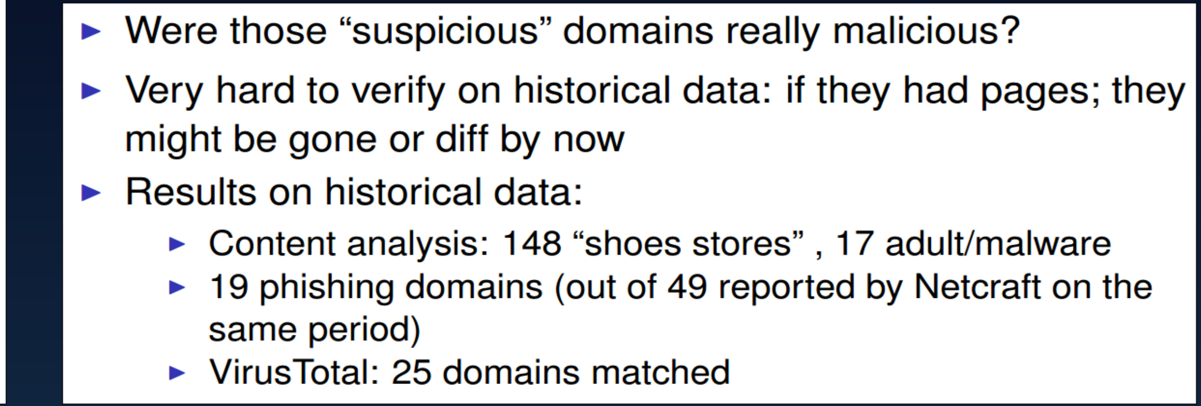
Key	Value
Interval	Jan 1st, 2015 to Aug 30th 2015
Average .nl zone size	$\sim$ 5,500,000
$\sum$ new domains	586,201
New domains - first timers	476,040(81.2%)
New domains - re-registered	110,161 (18.8%)
Total DNS Requests	32,864,402,270
DNS request new domains (24h)	826,740
DNS request new domains - first-timers (24h)	420,362

Table: Evaluated datasets (from one .nl auth server)



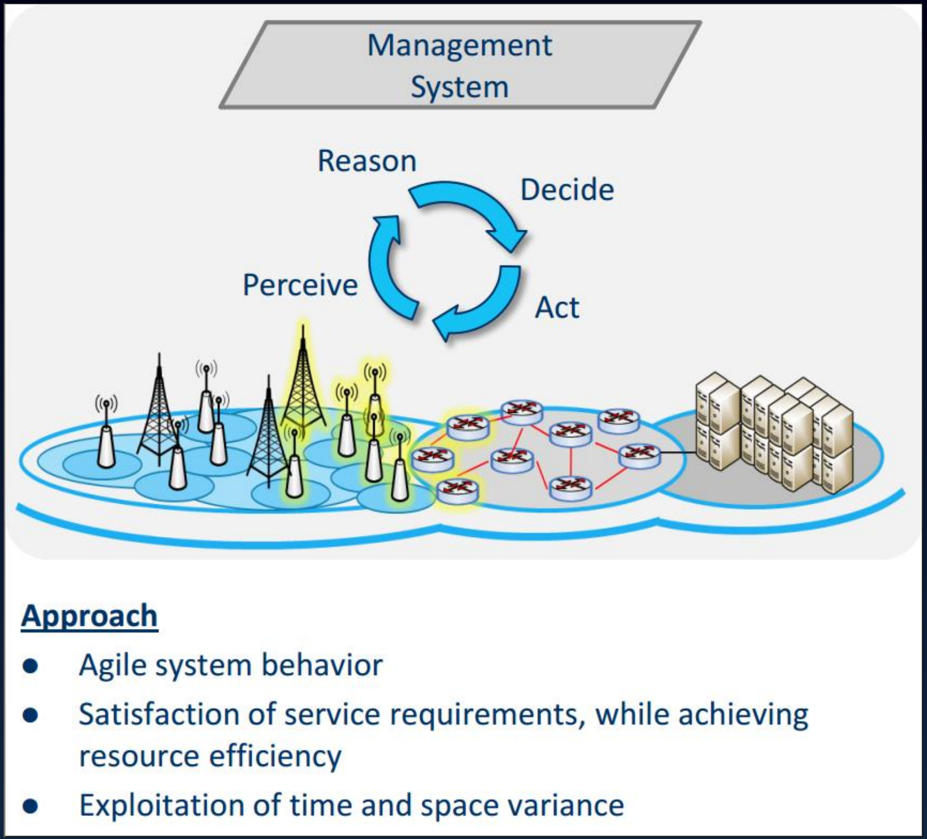
Cluster	Size	$\sum$ *Req*	$\sum$ *IPs*	$\sum$ *CC*	$\sum$ *ASes*
Normal	132,425	4.31	3.06	1.64	1.43
Suspicious	2,956	55.03	27.87	4.99	7.43

Table: Mean values for features and clusters - excluding domains with 1 request - 1st Timers



- Were those “suspicious” domains really malicious?
- Very hard to verify on historical data: if they had pages; they might be gone or diff by now
- Results on historical data:
  - Content analysis: 148 “shoes stores” , 17 adult/malware
  - 19 phishing domains (out of 49 reported by Netcraft on the same period)
  - VirusTotal: 25 domains matched

#3: ML based Policy Derivation and Evaluation in Broadband Networks

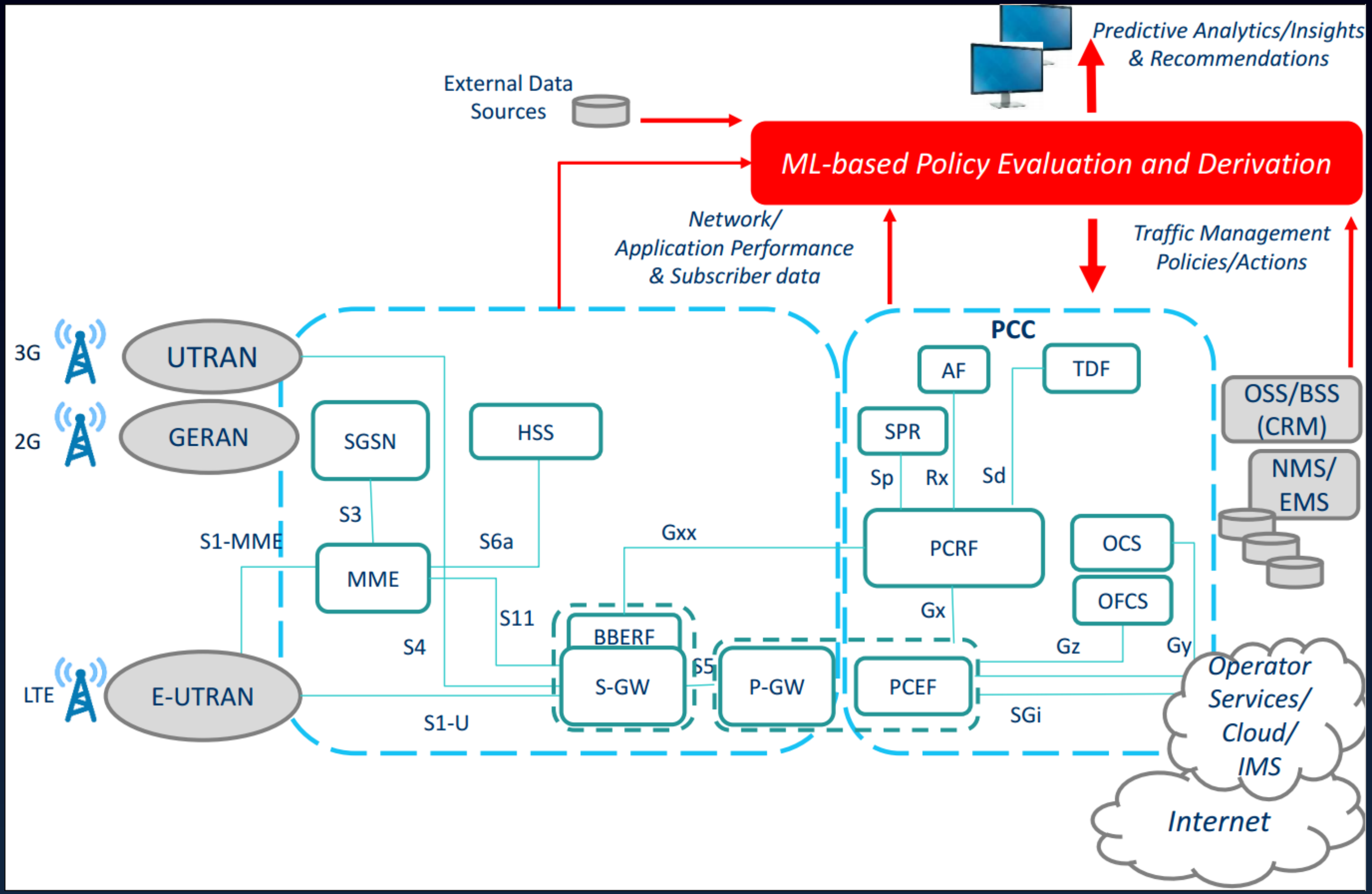


Approach

- Agile system behavior
- Satisfaction of service requirements, while achieving resource efficiency
- Exploitation of time and space variance

- Objective function
  - – QoS, resource consumption
- Input
  - – User information
  - – Services
  - – Traffic, mobility in time and space
  - – System capabilities
  - – …
- Output
  - – Network configuration
  - – Traffic allocation
- Difficulty
  - – Computationally hard problems

#3: ML based Policy Derivation and Evaluation in Broadband Networks



Predictive Analytics/Insights & Recommendations
External Data Sources
ML-based Policy Evaluation and Derivation
Network/ Application Performance & Subscriber data
Traffic Management Policies/Actions
PCC
3G
UTRAN
2G
GERAN
LTE
E-UTRAN
SGSN
HSS
S3
S1-MME
S6a
MME
Gxx
S11
S4
BBERF
S-GW
S5
P-GW
PCEF
S1-U
AF
TDF
SPR
Sp
Rx
Sd
PCRF
OCS
OFCS
Gx
Gz
Gy
SGi
OSS/BSS (CRM)
NMS/ EMS
Operator Services/ Cloud/ IMS
Internet

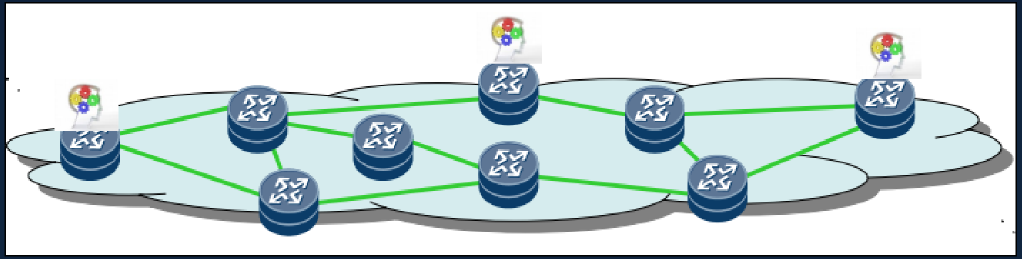
Through machine learning it will be possible to provide faster and targeted solutions to specific network problems.

Moreover, it is possible cluster various usage profiles and prioritize the traffic according to the criticality level.

#4: Traffic Anomaly Detection in the Router

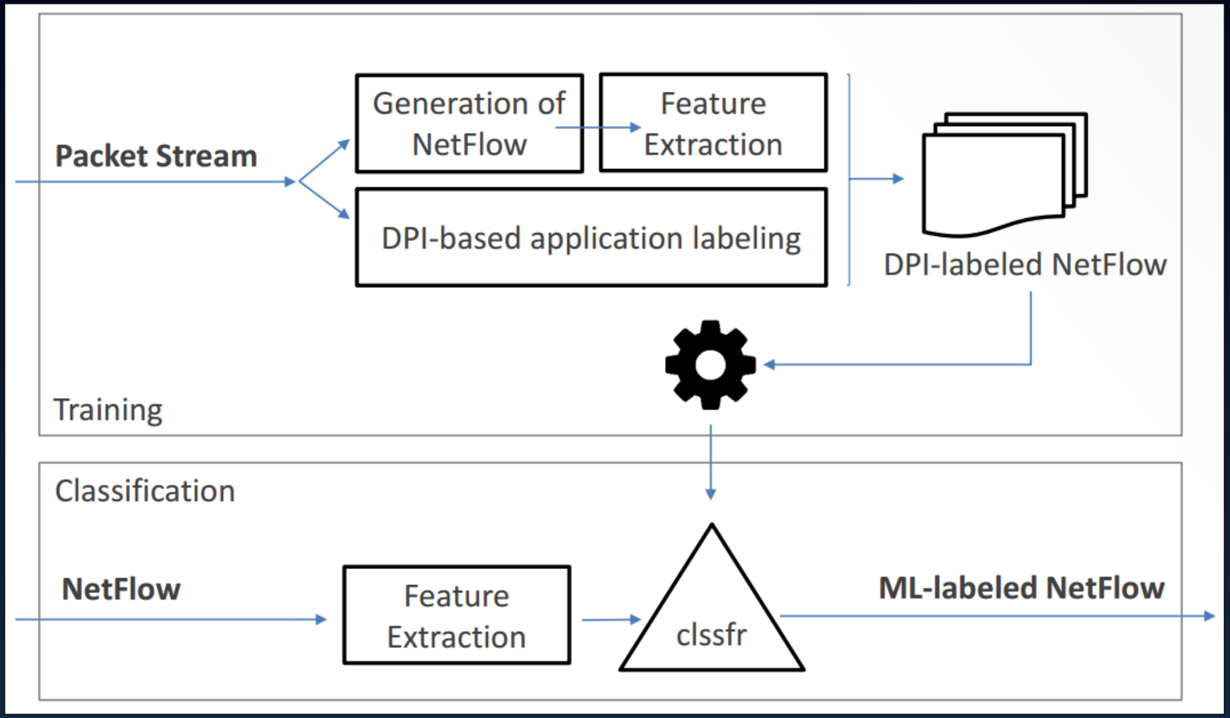
- When interface traffic exceeds a certain threshold, the router will consider it as an anomaly event and report it to the NMS
  - *E.g. trap-threshold { input-rate | output-rate }*
- However, network traffic is usually changing. static configuration could not effectively identify the traffic anomaly events.

- Wavelets are employed to analyze time-series network traffic for anomaly detection. In some certain interval, the routers measure, record, and analyze the input and output traffic rates respectively, or in the form of rate sums.
- Running for some time, the router would get a set of "time-rate” data, collected as time-series waves for further wavelet analysis. Besides wavelets, this use case proposes other machine learning techniques such as outlier detection. For this way, features are to be extracted from wavelets for supervised or unsupervised learning.



#5: Applications of ML to Flow Monitoring

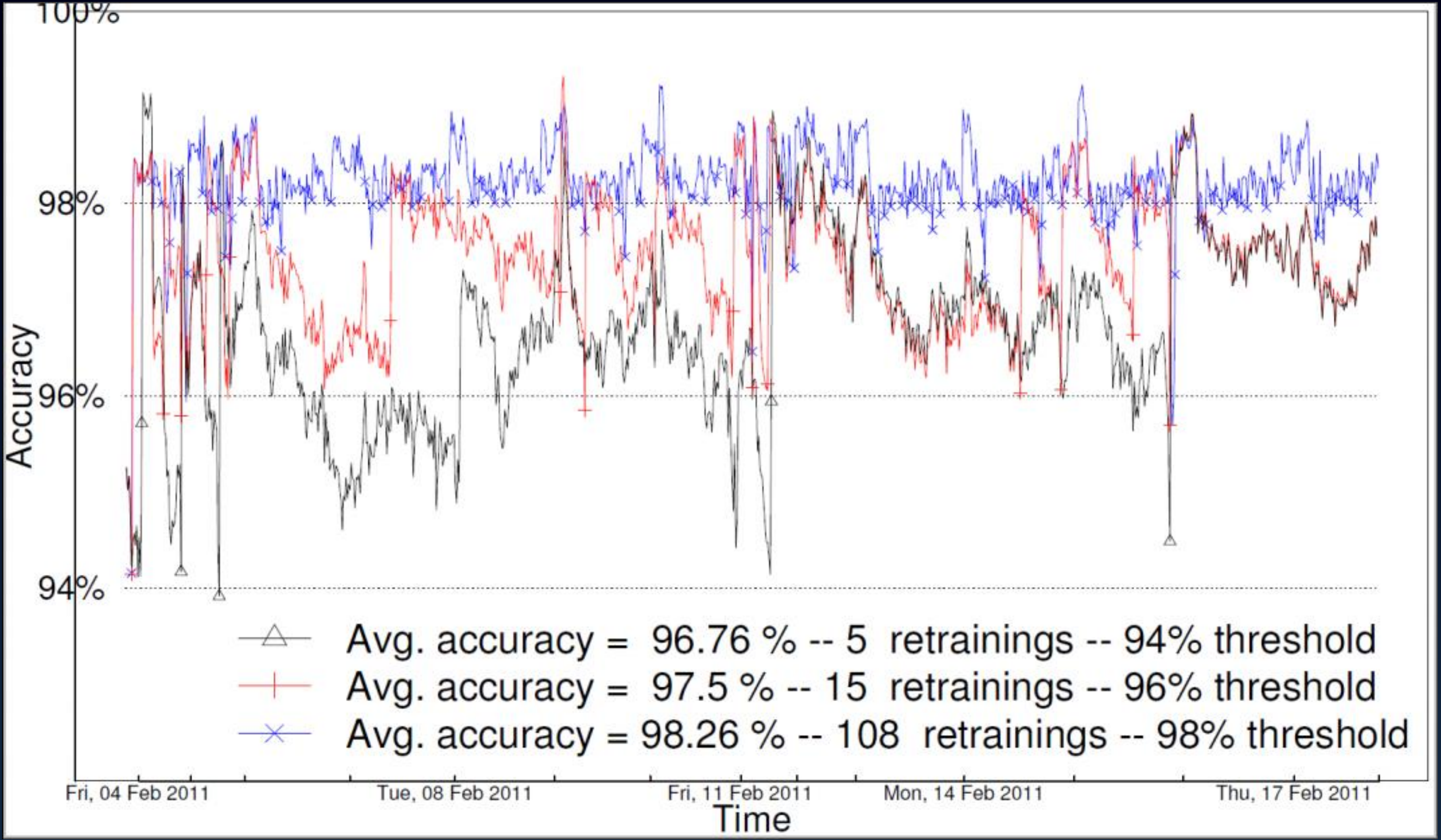
- How to identify applications w/o payloads? e.g., identify Netflix, BitTorrent, Skype..



1. Continuous training process:
   - Collect traffic (with payload), run through DPI
   - Build “NetFlow-derived features -> app” dataset
   - Machine learning to build a classifier

2. Classification process:
   - Collect NetFlow and extract features
   - Run through classifier

Results


100%
98%
96%
94%
Accuracy
Avg. accuracy = 96.76 % -- 5 retrainings -- 94% threshold
Avg. accuracy = 97.5 % -- 15 retrainings -- 96% threshold
Avg. accuracy = 98.26 % -- 108 retrainings -- 98% threshold
Fri, 04 Feb 2011
Tue, 08 Feb 2011
Fri, 11 Feb 2011
Mon, 14 Feb 2011
Thu, 17 Feb 2011
Time

Reference

- NMLRG IETF95 meeting materials:
  - https://www.ietf.org/proceedings/95/nmlrg.html
- IETF draft of these slides:
  - https://tools.ietf.org/html/draft-jiang-nmlrg-traffic-machine-learning

Thank you