Application of Machine Learning to Flow-based Network Monitoring

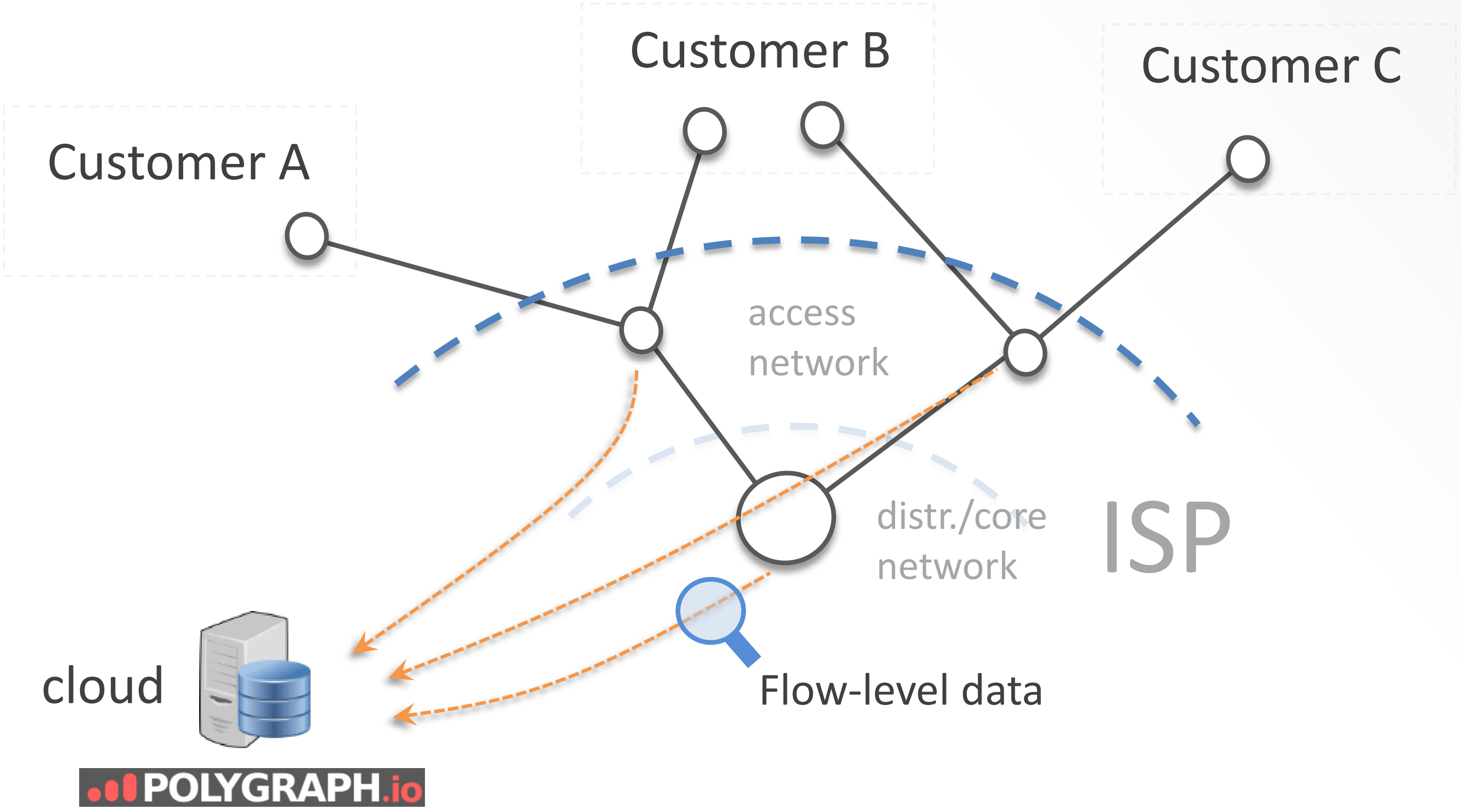
Josep Sanjuas – jsanjuas@polygraph.io

POLYGRAPH.io

Network Monitoring: Approaches

- Deep Packet Inspection
  - Dedicated hardware to intercept & scan packets
  - High cost, high visibility
- Flow-based monitoring
  - Data collection performed by routers
  - Lower cost, but less information available

Cloud-based Flow Monitoring



Flow-based Monitoring Protocols

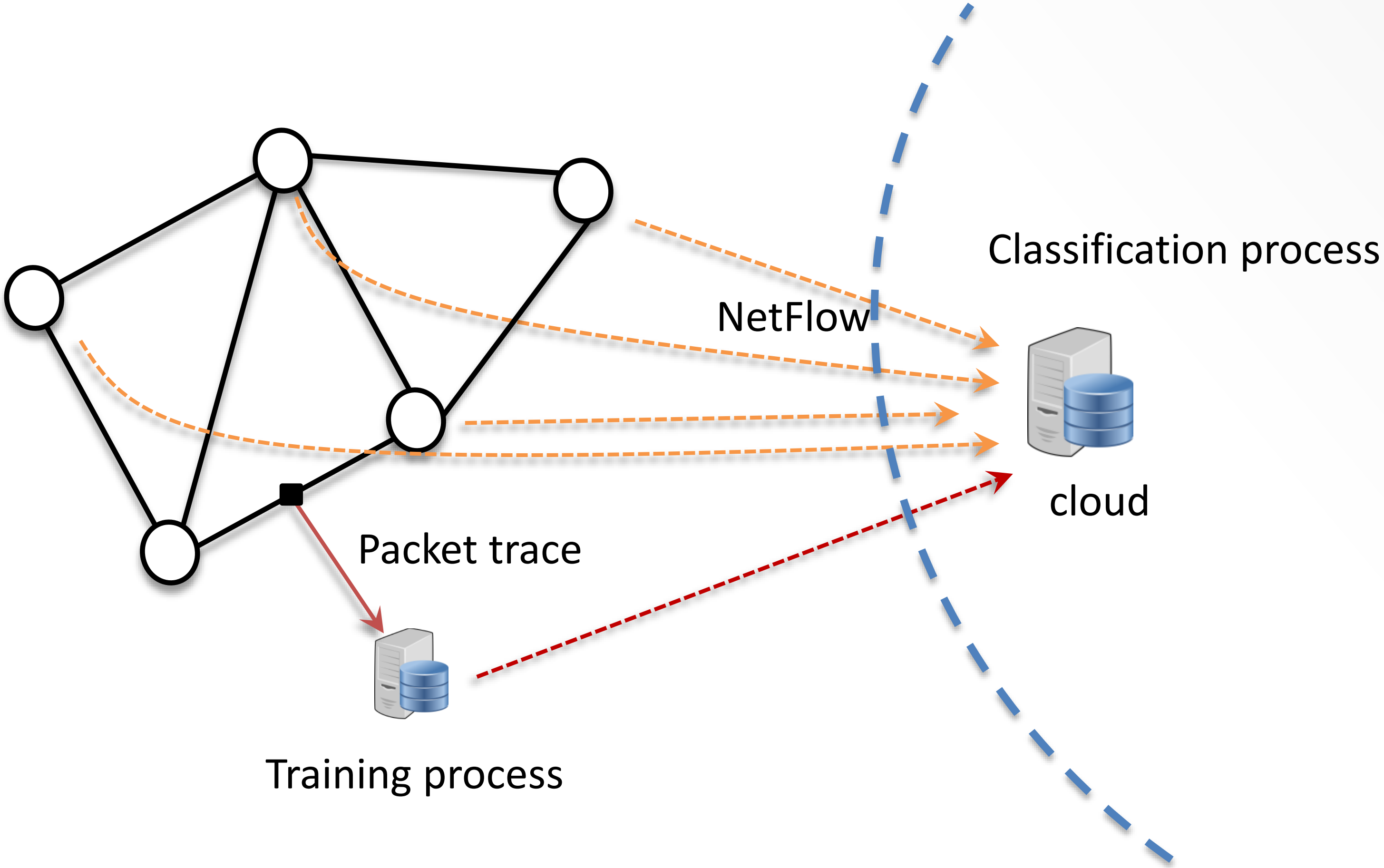
- sFlow
  - Samples individual packets, sends them to a monitor
- NetFlow (Cisco), IPFIX (IETF standard)
  - Send flow aggregates to software collector
  - Support for packet sampling to reduce overhead

<src_ip, dst_ip, sport, dport, proto: ts$_0$, ts$_f$, #bytes, #pkts>

Requirement: Application Identification

- Packet payloads are not available
- How to identify applications w/o payloads?
  - – e.g., identify Netflix, BitTorrent, Skype..
- Naïve approach: port-based classification
  - – misses apps using dynamic ports
  - – port 80 and 443 carry wildly different apps
- Solution: machine learning!

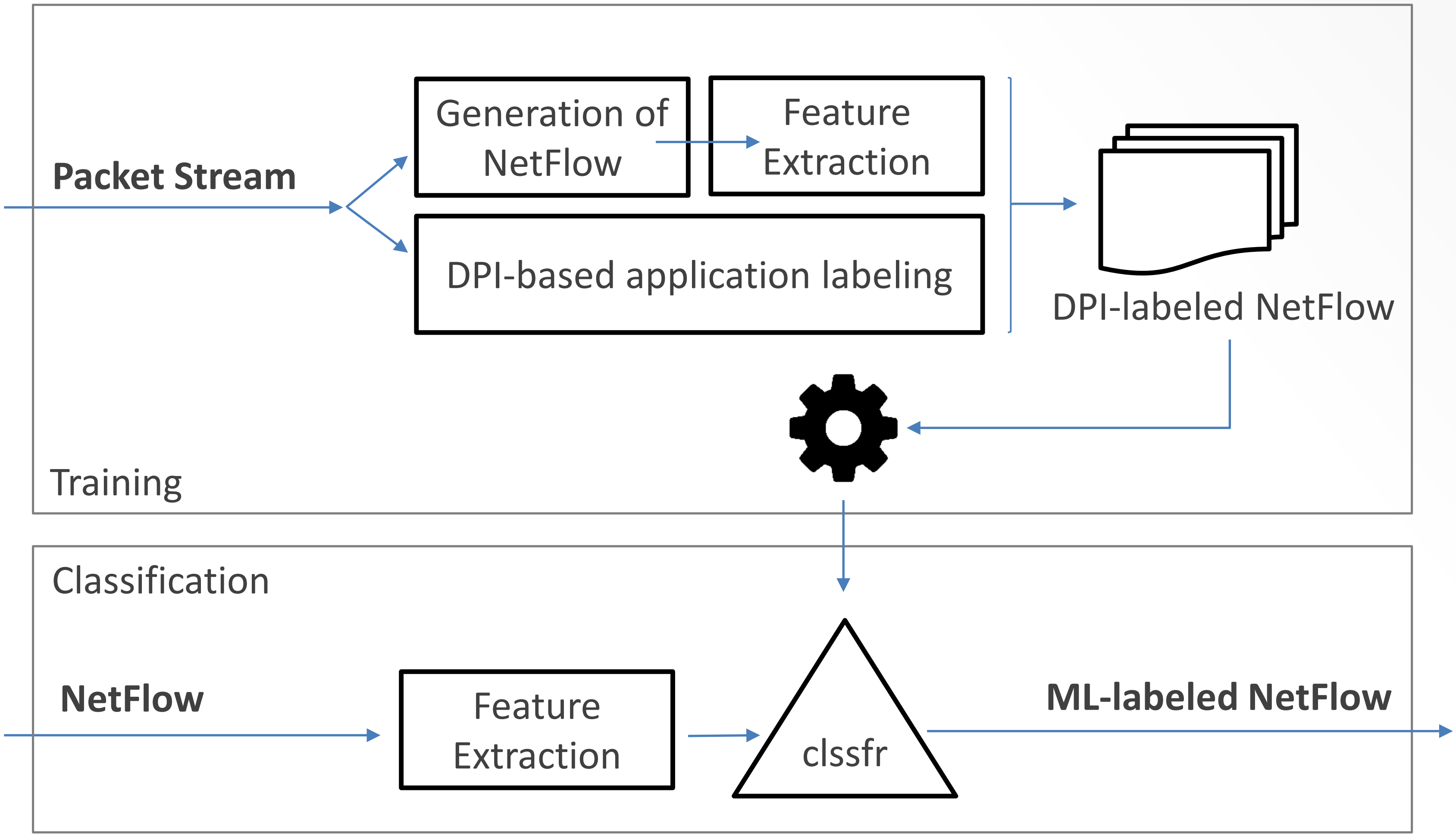
Deployment Architecture



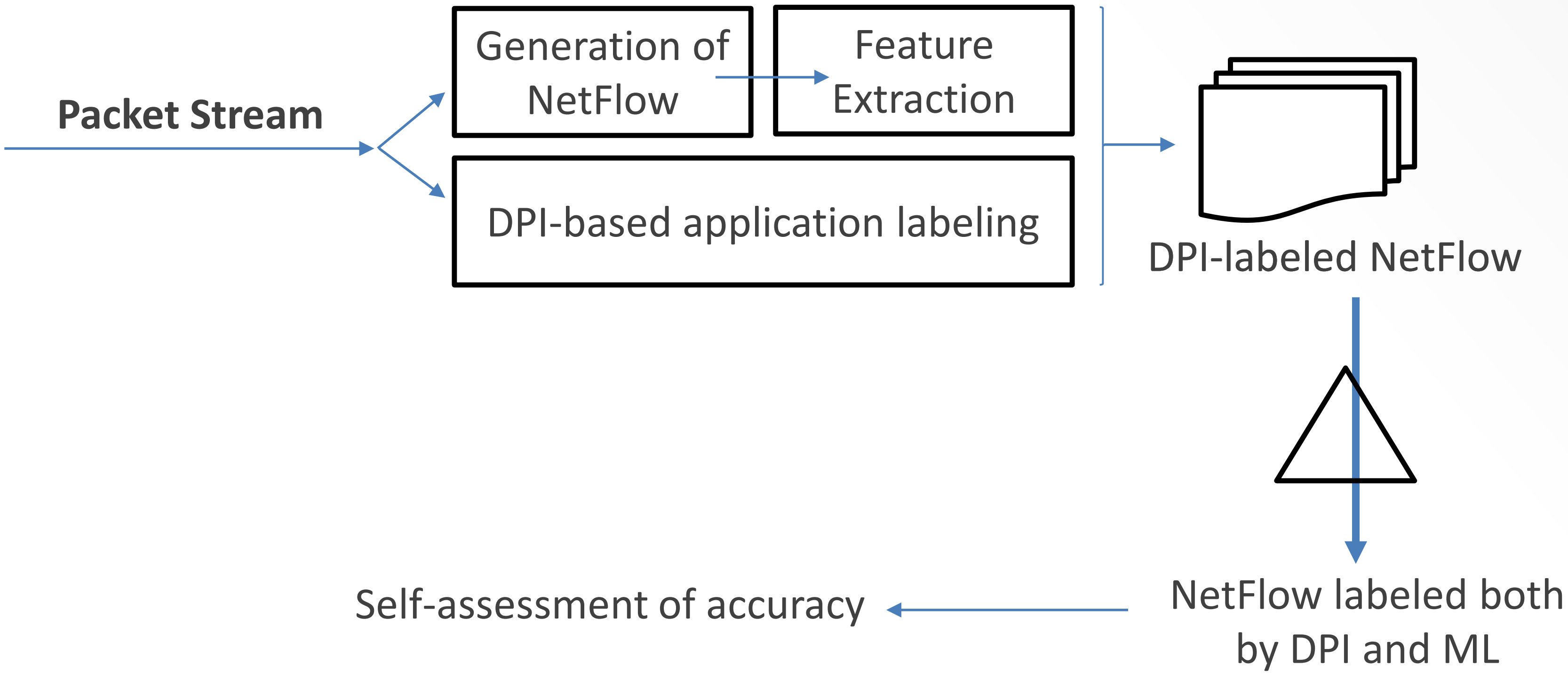
High-level Approach

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

ML-based Traffic Classification



Self-Assessment of Accuracy



Packet Stream

Generation of NetFlow

Feature Extraction

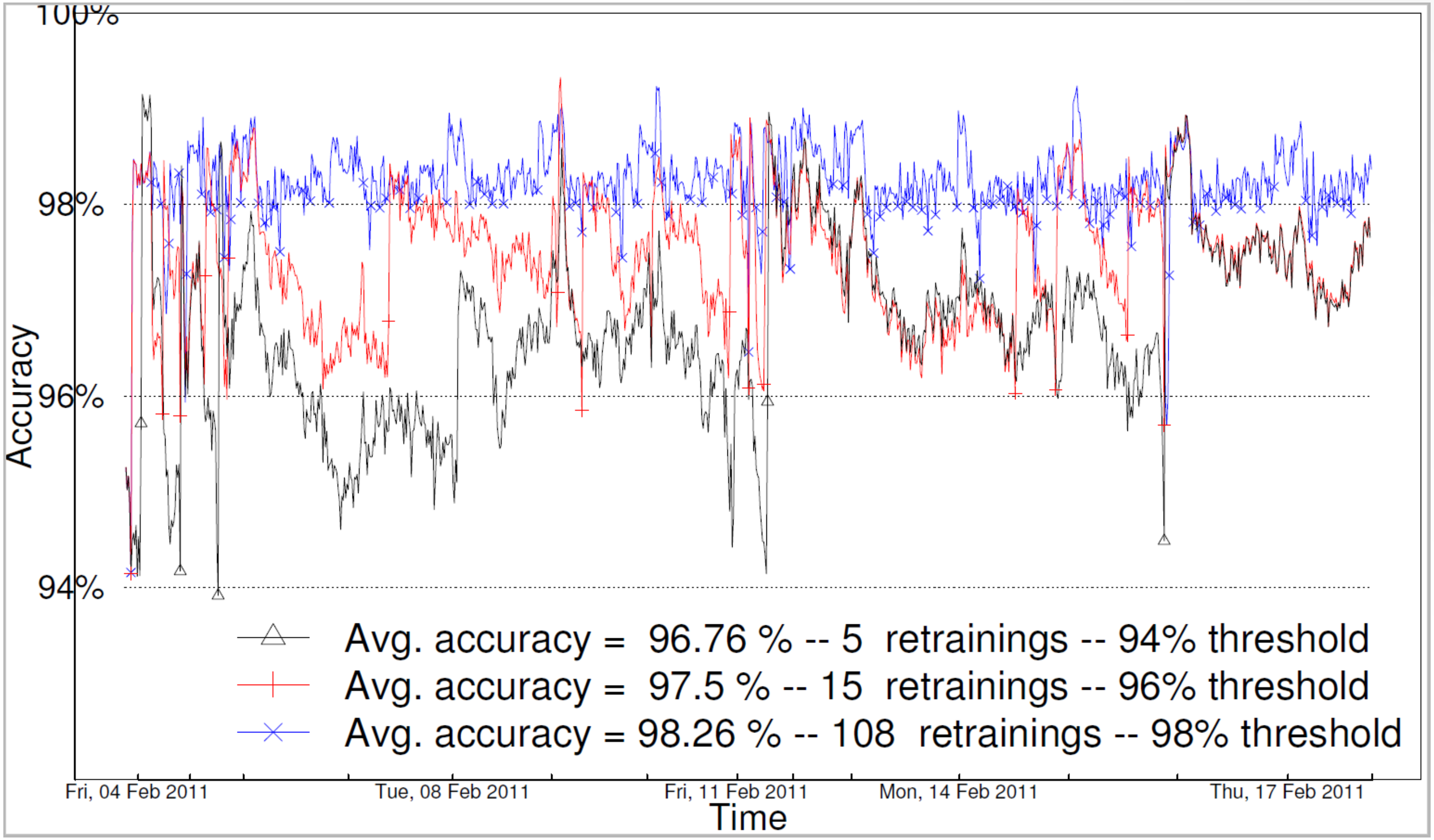
DPI-based application labeling

DPI-labeled NetFlow

NetFlow labeled both by DPI and ML

Self-assessment of accuracy

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

Summary

- Environment: flow-based network monitoring in the cloud
- Objective: per-application traffic classification
- Challenge: packet contents not available
- Solution:
  - collect packet payloads, use ML algorithms to generate a classifier based on NetFlow info
  - Use the model to classify NetFlow traffic

Future Work

- Enhance accuracy for web apps (& CDN traffic)
- Automated generation of traffic datasets for popular applications
- Combining ground truths / classification models from several vantage points

Network Polygraph

Talaia Networks, S.L.
K2M – Parc UPC Campus Nord
Jordi Girona, 1-3
Barcelona (08034)
Spain

Telephone: +34 93 405 45 87
contact@polygraph.io
https://polygraph.io

POLYGRAPH.io