POLITECNICO
MILANO 1863

Passive Classification of Wi-Fi enabled devices

Alessandro E. C. Redondi, **Davide Sanvito**, Matteo Cesana

POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

Italian Networking Workshop (INW) - Falcade (BL), Italy - January 11-13, 2017

Outline

- Motivations
- Data collection, features extraction and classification
- Use case: traffic analysis
- Conclusions

Motivations

- Network traffic from wireless devices will soon exceed traffic from wired devices
- Increasing attention towards analyzing and profiling Wi-Fi traffic, especially for personal devices (BYOD).
- Two classes of Wi-Fi enabled devices:
  - Mobile handheld devices (MHD)
  - Non handheld devices (NHD)

Wi-Fi device classification

- Two main groups of classification methods:
  - Medium Access Control (MAC) informations
  - Packet inspection (DHCP log/HTTP User-Agent)

- We propose an effective method to perform device classification
  - Entirely passive
  - No traffic probes on network edge devices
  - No DPI
  - Based on capturing and processing Wi-Fi probe requests

- Main idea: extract features from captured probe request frames and train a Machine Learning classifier to recognize MHD and NHD devices.

Data collection

- Network data traces collected during hands-on university classes
  - Students have their own laptop and smartphone

- Students are asked to turn on Wi-Fi and fill out an anonymous form

MAC address	Device type

- Linux laptop + Wi-Fi card in monitor mode (802.11 ch 1)
  - *tshark* collects only probe requests

timestamp	MAC source	OUI	RSS	SSID

Data collection (2)

- The first database contains 279 labelled devices

- The second database is filtered to keep just:
  - Probes from known devices (survey)
  - Probes from devices with known label (OUI)

- We collected 200000 probe req spanning 10 hours over 5 days

Features extraction

- Classification is based on four main informations about
  - Temporal process
    - avg and std dev of Inter-Probe Period (IPP) μ_p, σ_p [s]
    - coefficient of variation $c_p = \frac{\mu_p}{\sigma_p}$
  - Power levels
    - avg and std dev of Received Signal Strength (RSS) μ_r, σ_r [dBm]
    - coefficient of variation $c_r = \frac{\mu_r}{\sigma_r}$
  - SSID data
  - Device manufacturer

Features extraction (2)

- Classification is based on four main informations about
  - Temporal process
  - Power levels
  - SSID data
    - Number of probe req with known/*Broadcast* SSID N_k, N_b
    - Proportion of known/*Broadcast* SSID $\frac{N_k}{N_k+N_b}, \frac{N_b}{N_k+N_b}$
    - Number of unique SSID N_u
  - Device manufacturer
    - V dummy binary variables $d_i = \begin{cases} 1 & if\ device\ is\ from\ i-th\ vendor \\ 0 & otherwise \end{cases}$
    - V = number of different vendors in the database

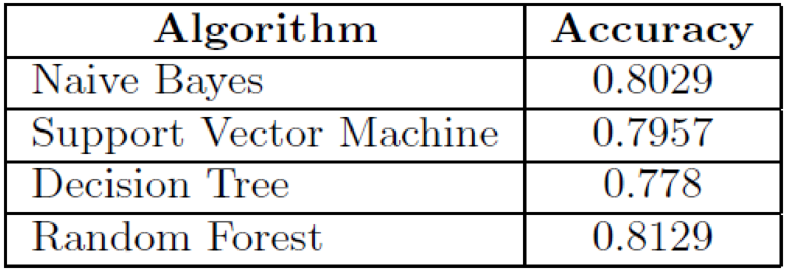
Classification algorithms

- The dataset has been fed to four supervised learning algorithms:
  - Naïve Bayes (NB)
  - Support Vector Machine (SVM)
  - Decision Tree (DT)
  - Random Forest (RF)
- Performance tested in three different scenarios:
  - Dummy features only (DF)
  - Quantitative features only (QF)
  - All features (AF)

Classification performances

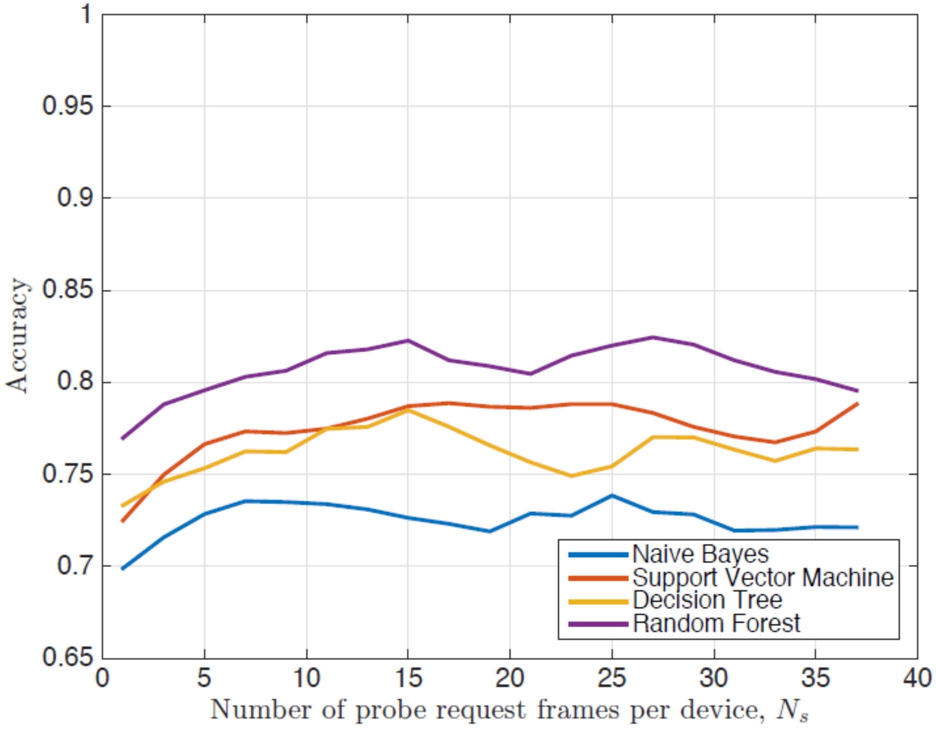
- K-fold cross validation (k=5)

Dummy features only (DF)

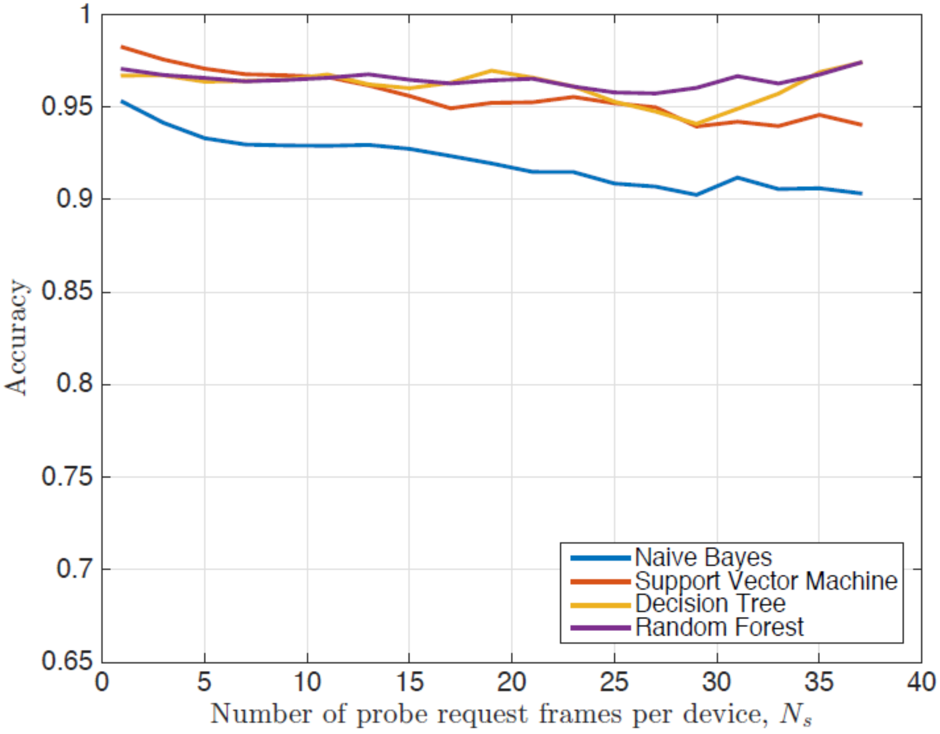


Algorithm	Accuracy
Naive Bayes	0.8029
Support Vector Machine	0.7957
Decision Tree	0.778
Random Forest	0.8129

Quantitative features only (QF)



All features (AF)



Use case: traffic analysis

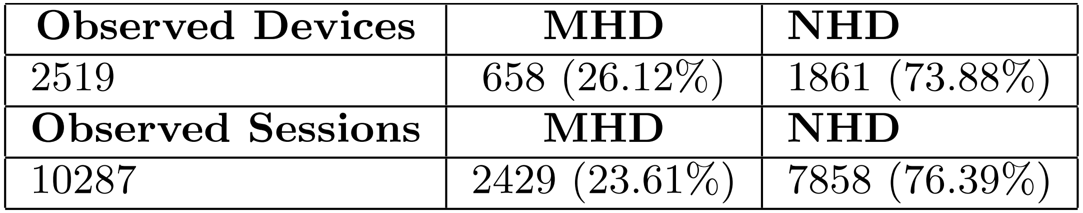
- Classification as pre-processing stage for network traffic analysis

- Network traffic is collected via AirWave Management Platform
  - MAC address of associated devices
  - Timestamp of the association with the AP
  - Duration of the session
  - Avg and variance of the bandwidth in the session
  - Avg and variance of the signal quality in the session

Use case: traffic analysis (2)

- We analyzed a period of two weeks of Wi-Fi sessions from MHD and NHD devices in a building of our university

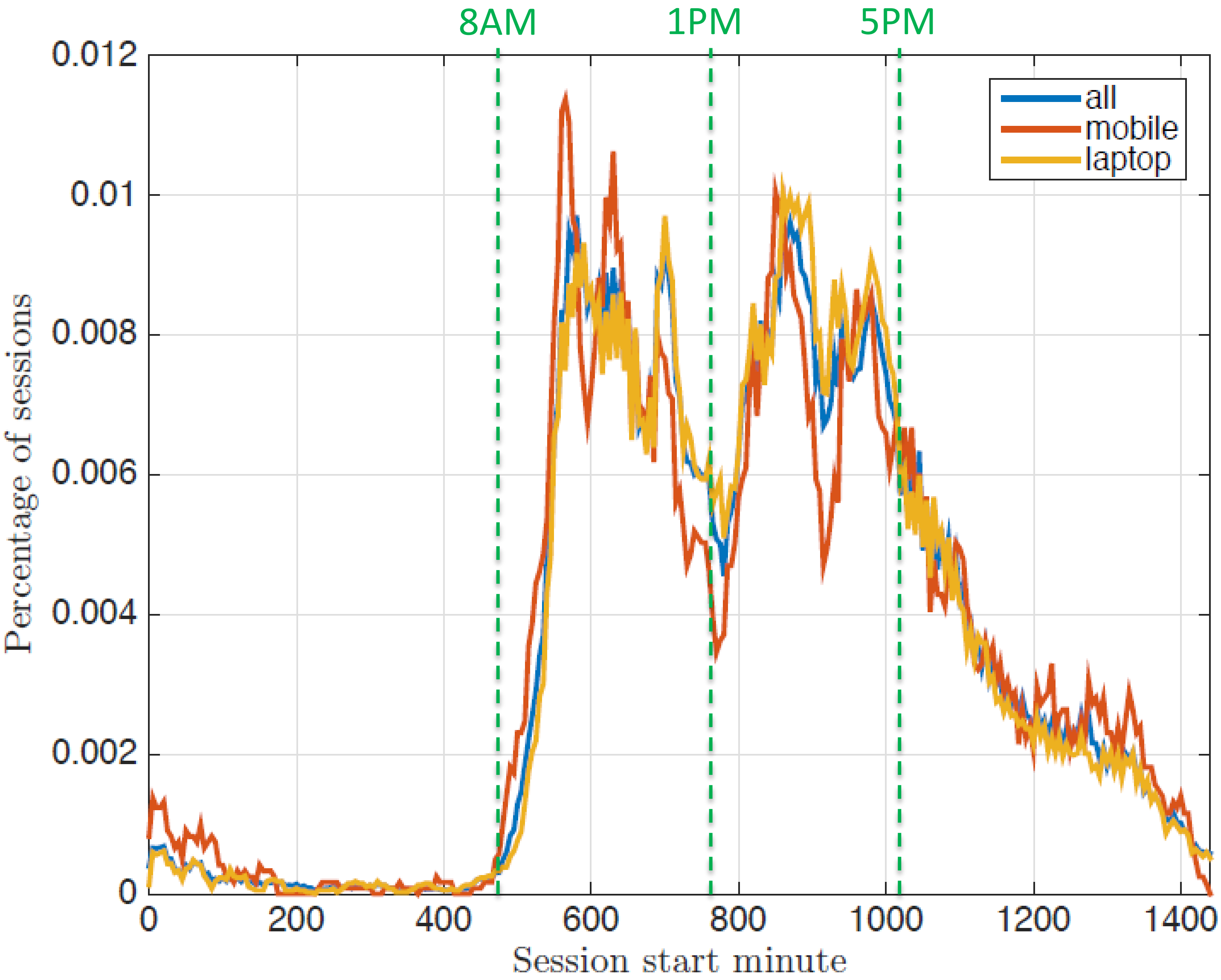
- A single Raspberry PI 3 captured probe request in an open space and we run a RF classifier to label each device as «Laptop» or «Smartphone»



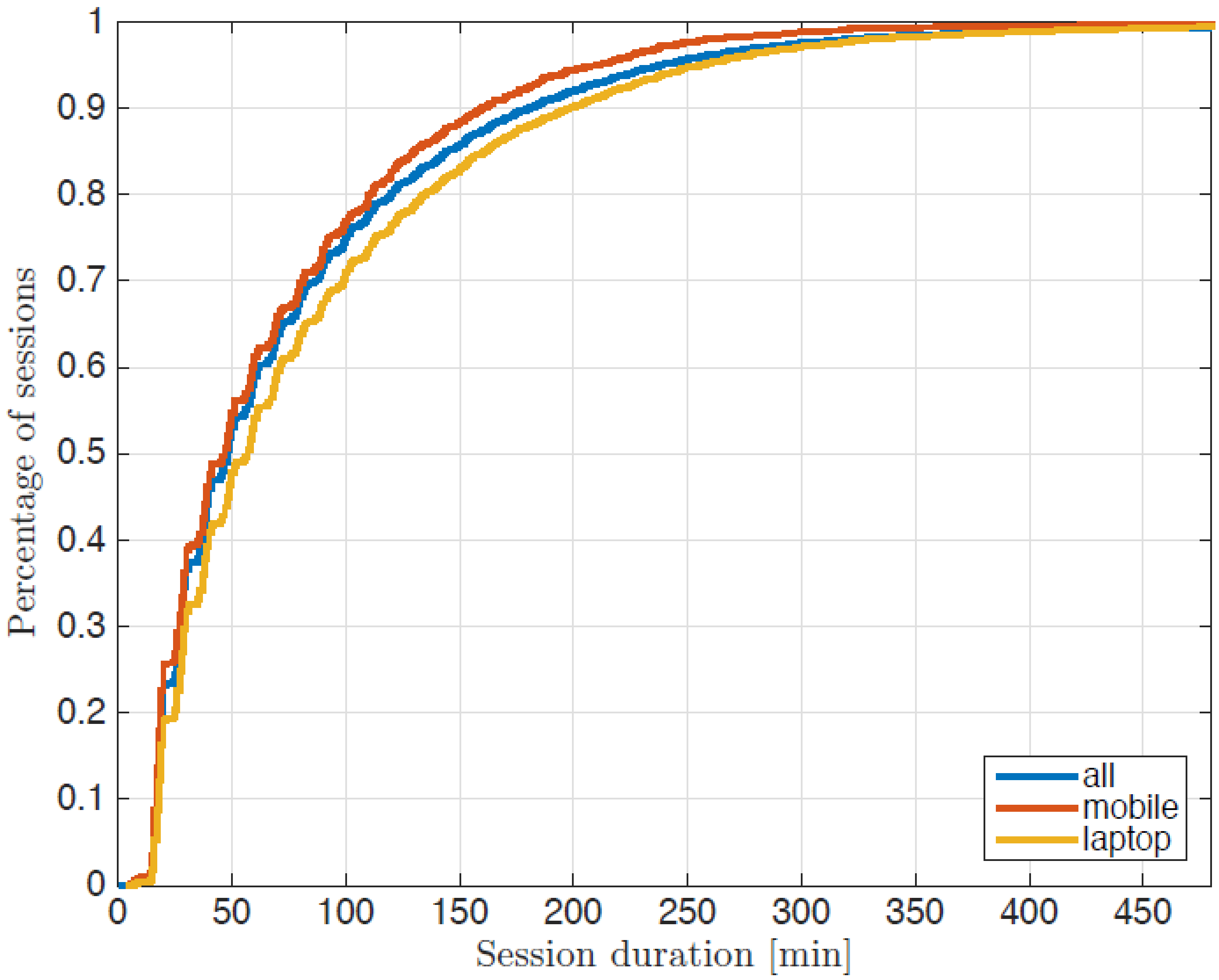
Observed Devices	MHD	NHD
2519	658 (26.12%)	1861 (73.88%)
Observed Sessions	**MHD**	**NHD**
10287	2429 (23.61%)	7858 (76.39%)

- We analyzed session start time, duration and average bandwidth usage only for those devices seen and classified by our method

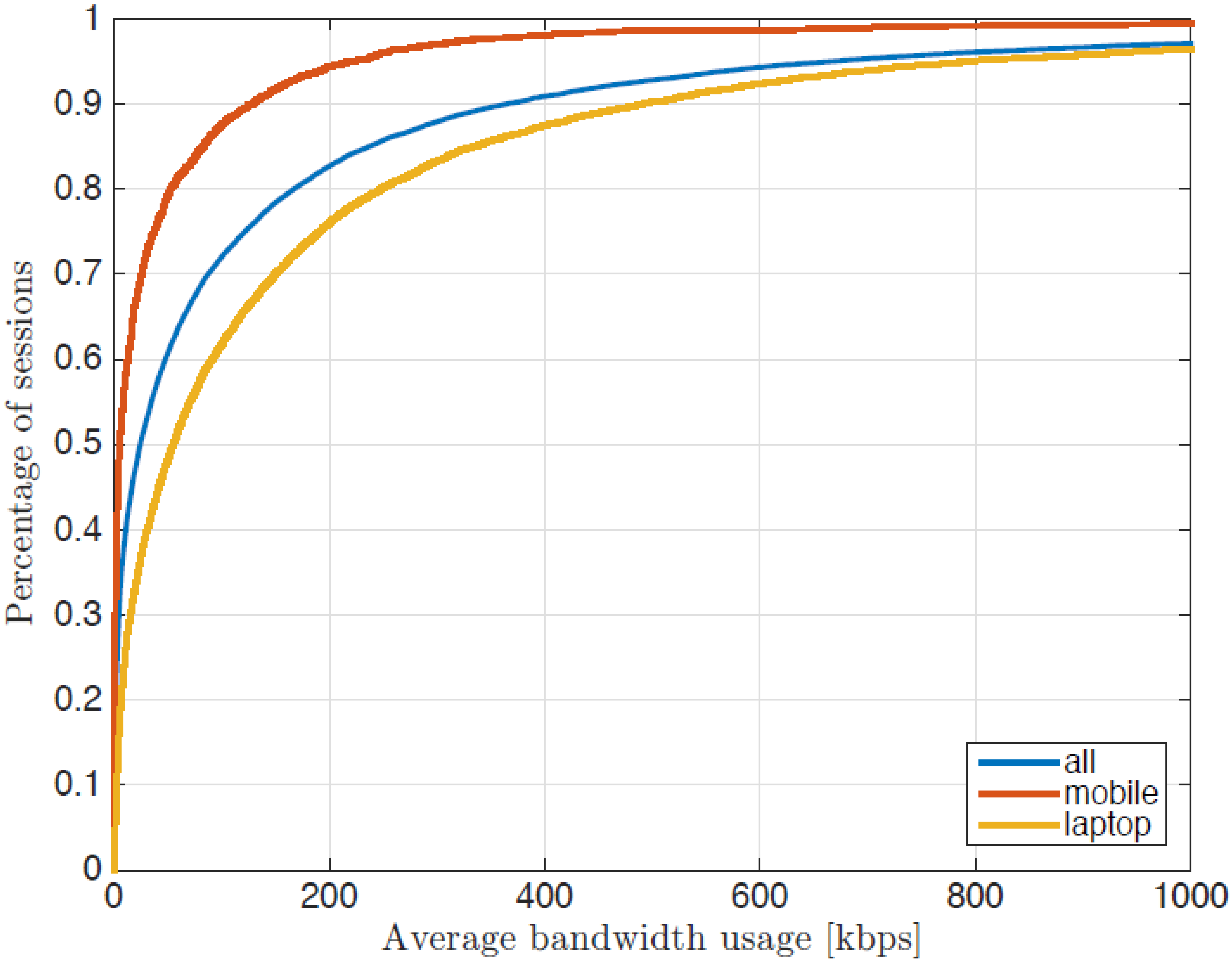
Session start time



Session duration CDF



Average bandwidth usage CDF



Conclusions

- Method for classifying wireless devices as MHD or NHD
- Our solution correctly classifies more than 95% of the devices
- Applications
  - Pre-processing stages of network data analysis
  - Improve performance of indoor localization systems

Thank you!

davide.sanvito@polimi.it