Doctoral study report:

TIME SERIES ANALYSIS OF NETWORK TRAFFIC AND DETECTION OF SECURITY THREATS

Ing. Josef Koumar

Supervisor
doc. Ing. Tomáš Čejka, Ph.D.

MOTIVATION

- Maintaining network security has become increasingly challenging in recent years due to mass traffic encryption and consequent reduced visibility, for example:
    - Encryption of TLS certificates by TLS1.3
    - Deployment of encrypted DNS
    - Encrypted Client Hello proposal
- New generation detection systems must use packet length and time
    - It is challenging to build an accurate model that will universally work
    - Detection system must adapt to dynamic and evolving threats
    - Detection system must operate efficiently without slowing down network

WHY TO USE TIME SERIES ANALYSIS?

- Time series are a natural representations of network traffic
- Long term view on the data can bring better precision of models
- Classical approaches can be combined with novel time-series-based approaches to obtain better precision

- The statistical view at time series can be used for multiple purposes in network monitoring, for example:
    - Network traffic classification
    - Anomaly detection

RESEARCH GOALS

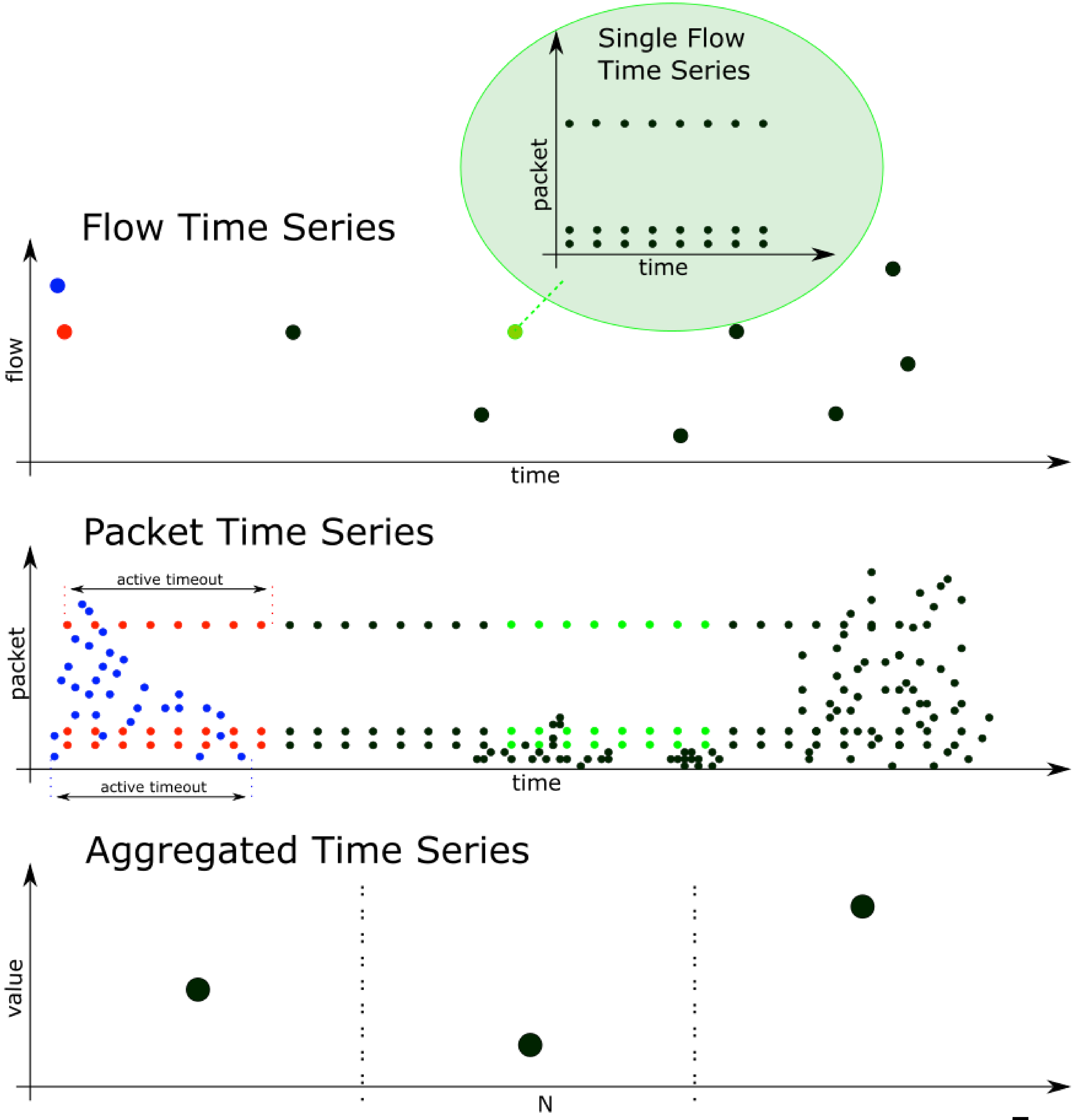
- Supervised Classification
  - Classification of network traffic and detection of security threats using Time Series Analysis (TSA)
  - Deployability of TSA-based detection into real-world environments

- Unsupervised Classification
  - Anomaly detection in ISP network
  - Deployability of anomaly detection methods in real-world environments

- Time Series Analysis of network traffic
  - Statistical properties of time series from network traffic
  - Data drift detection in network traffic

TIME SERIES FROM NETWORK TRAFFIC

- My definition of a time series from network traffic
- Describe the pros and cons of different types of time series
- Find out the statistical properties of unevenly spaced time series



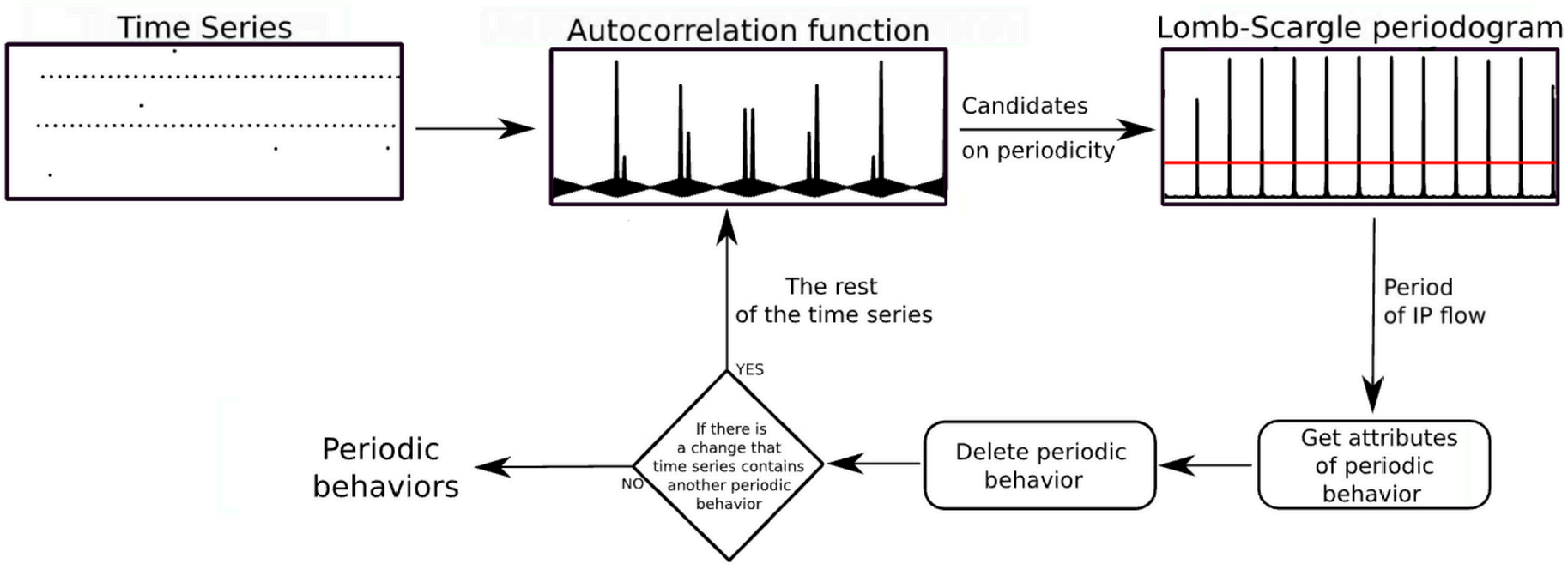
TMA 2023, CONFERENCE

Koumar J., Čejka T."*Unevenly spaced time series from network traffic.*" 2023 7th Network Traffic Measurement and Analysis Conference (TMA). IEEE,2023.

PERIODICITY DETECTION

The network traffic of many processes tends to be periodic in time. Therefore, we use periodicity for detection using the following method:

1. Network traffic is divided into Network dependencies and FTS are created
2. Apply Autocorrelation function and Lomb-Scargle periodogram
3. Feature mining on periodic time series
4. Machine Learning detection



PERIODICITY DETECTION

Case study 1: Applications, Services, and Operating systems

Dataset from CESNET2 network was created with the following categories:

- social networks (Facebook, MS teams, Slack, ...)
- remote storage (Google Drive, OneDrive, Github, ...)
- updates of operating systems (Windows, Android, Fedora, ...)
- antivirus programs (Eset, Avast, Kaspersky, ...)
- network services and protocols (Keep-alive, HTTP2 ping, DNS, ...)
- email browser viewers and clients (Gmail, Outlook, ...)
- multimedia streaming (youtube, itunes, spotify)

The best-performing classification algorithm was XGBoost with F1-score 90%.

CNSM 2022, CORE B CONFERENCE

Koumar J., Čejka T. "*Network traffic classification based on periodic behavior detection.*" 2022 18th International Conference on Network and Service Management (CNSM). IEEE, 2022.

PERIODICITY DETECTION

Case study 2: Cryptomining

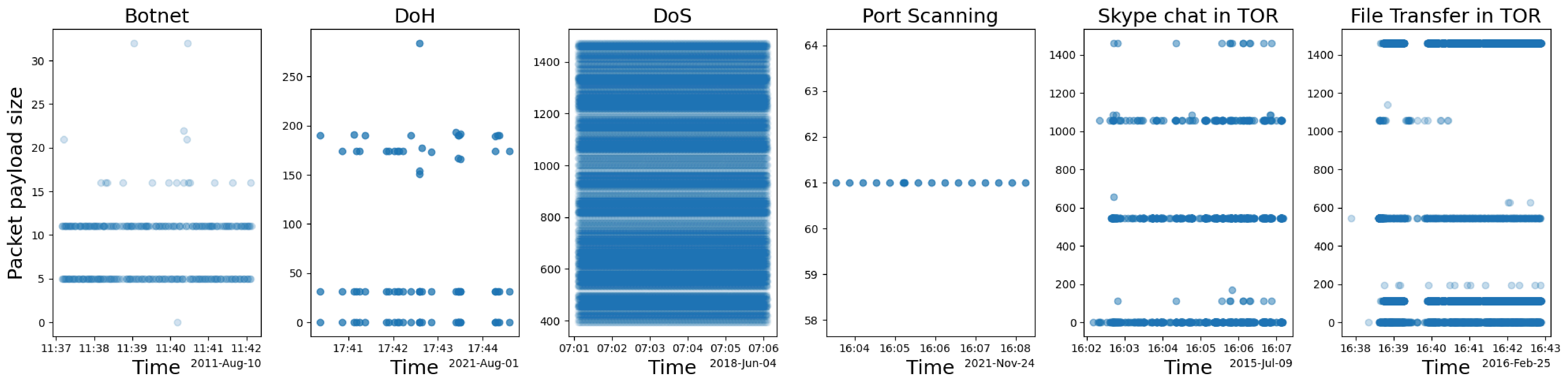
- Dataset from ISP network CESNET2
- ML model achieves more than 92% F1-score.
- Moreover, the DeCrypto system enhanced by periodicity detection achieved:
    - 97.25% Accuracy -- improvement by 2.95%
    - 99.99% Precision -- improvement by 0.001%
    - 92.47% Recall -- improvement by 7.74%
    - 96.08% F1-score -- improvement by 4.37%
- Deployable by DP3 system

CNSM 2023, CORE B CONFERENCE

Koumar J., Plný R., Čejka T. "*Enhancing DeCrypto: Finding Cryptocurrency Miners Based on Periodic Behavior.*" 2023 19th International Conference on Network and Service Management (CNSM). IEEE, 2023

CLASSIFICATION OF IP FLOWS BASED ON TSA

1. In the IP flow exporter are packets for each IP flow put into time series which is called Single Flow Time Series (SFTS).
2. The Single Flow Time Series is analysed by Time Series Analysis
3. Set of time-series based features is exported as novel extended IP flow
4. Novel IP flow is used for classification using Machine Learning



CLASSIFICATION OF IP FLOWS BASED ON TSA

Case study 1: Classification based on Single Flow Time Series

- The 69 features are exported as novel extended IP flow
- The features can be organized into five categories:
  - statistical, time, frequency, distribution, and behavior
- Evaluated on 15 well-known network datasets -- both binary and multiclass task.
- Results compared to best-performing classifiers from related works:
  - Outperforming on 12/23 tasks
  - Achieved similar results on 7/23 task
  - Cannot achieve similar results only 3/23

CNSM 2023, CORE B CONFERENCE

Koumar J., Hynek K., Čejka T. "*Network traffic classification based on single flow time series analysis.*" 2023 19th International Conference on Network and Service Management (CNSM). IEEE, 2023.

- Cited in: K. Dietz et al., "The Missing Link in Network Intrusion Detection: Taking AI/ML Research Efforts to Users," in IEEE Access

CLASSIFICATION OF IP FLOWS BASED ON TSA

Case study 2: NetTiSA flow

- The 20 features are exported as novel extended IP flow NetTiSA
- Deployable into high-speed ISP networks (deployed on CESNET3 network)

  - previous approach can be deployed only in small or medium networks
- Evaluated on 15 well-known network datasets -- both binary and multiclass task.
- Results compared to best-performing classifiers from related works:
  - Outperforming on 12/23 tasks
  - Achieved similar results on 7/23 task
  - Cannot achieve similar results only 3/23
- Lowest network telemetry from all approaches

COMPUTER NETWORKS, Q1 JOURNAL

Koumar J., Hynek K., Pešek J., Čejka T. "*NetTiSA: Extended IP flow with time-series features for universal bandwidth-constrained high-speed network traffic classification.*" Computer Networks 240 (2024): 110147.

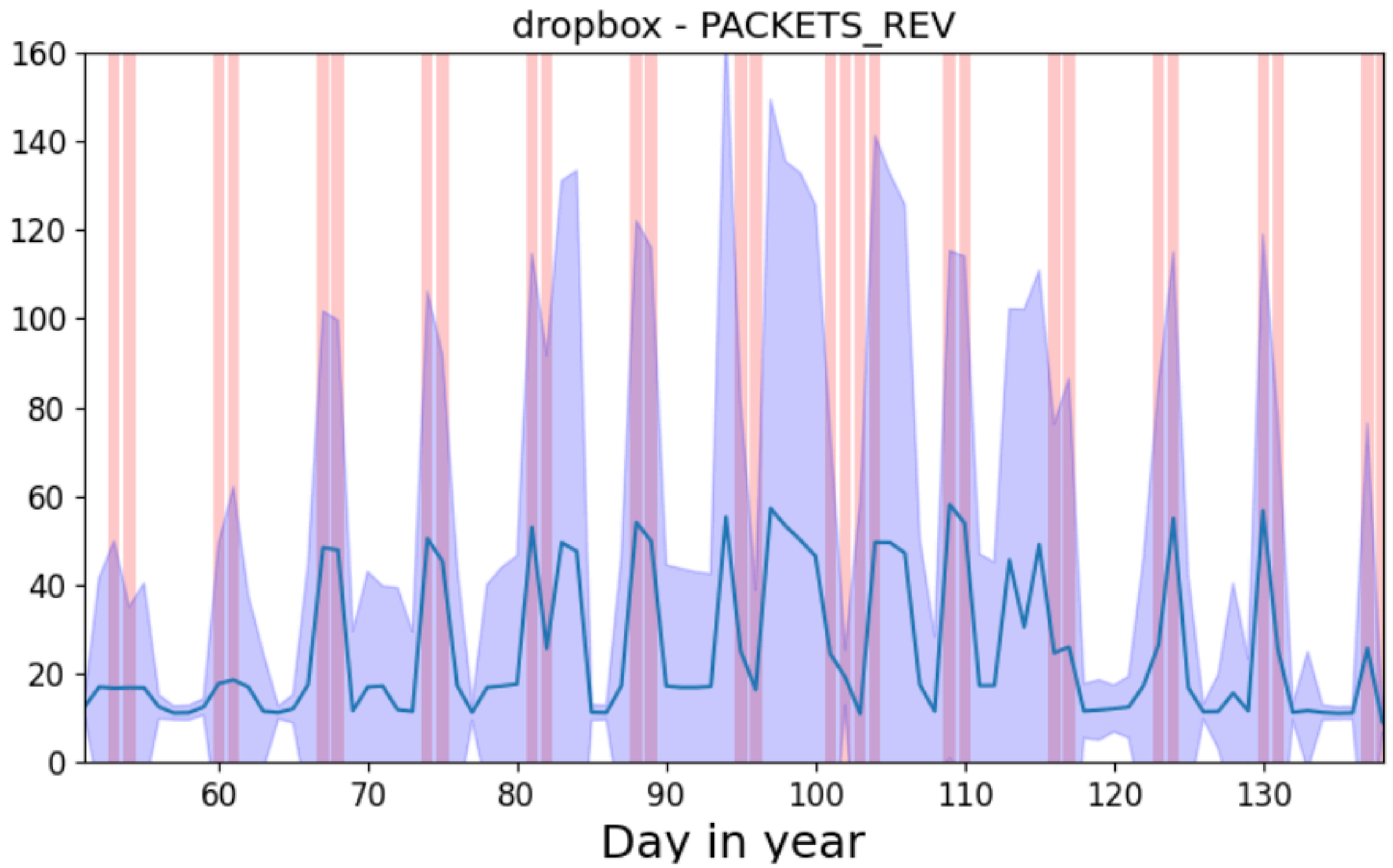
DATA DRIFT EVALUATION

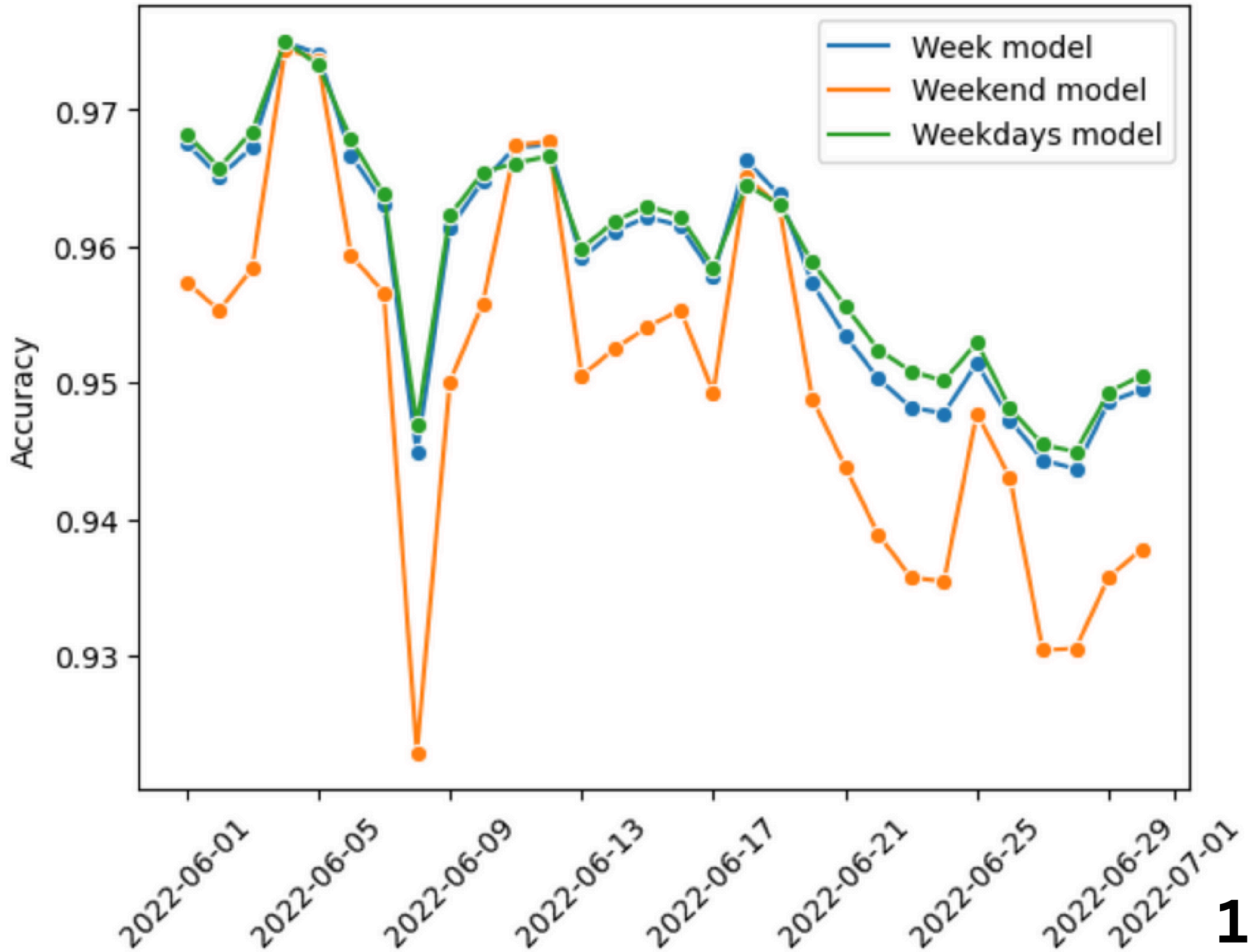
- Accuracy of ML models deployed into real-world environment occurs with fast decrease of accuracy

  => data / concept drift
- Time series analysis of data distribution is used for drift detection
- Description of the Weekend phenomenon

NOMS 2024, CORE B CONFERENCE

Jančička L., Koumar J., Soukup D, Čejka T. "*Analysis of Statistical Distribution Changes of Input Features in Network Traffic Classification Domain*". NOMS, 2024

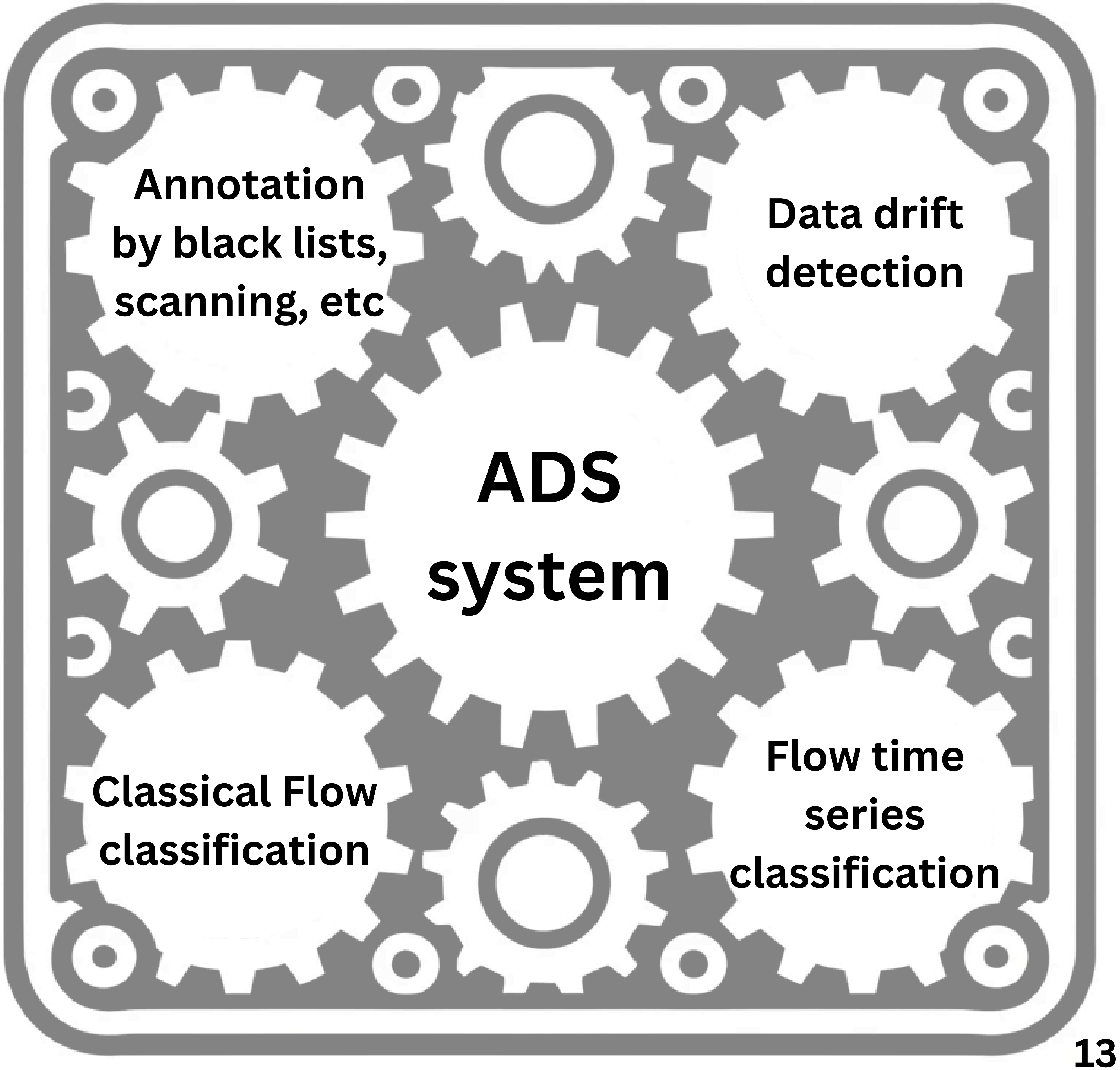




VISION

- Interconnect all deployable approaches for detection threats in one deployable IDS system
- Cooperation supervised and unsupervised methods
- Retraining in real-time
- Proper annotation of real-time datasets using black lists, scanning, etc

IDS system in Flow Collector


Annotation by black lists, scanning, etc
Data drift detection
ADS system
Classical Flow classification
Flow time series classification

PARTICIPATION IN PROJECTS

- Smart ADS (TH04010073, cooperation with Flowmon Networks supported by Technology Agency (TAČR))
- CYBERTHRETS (OYCESNET20221, Use of artificial intelligence for defence against cyber security attacks)
- Flow based Encrypted Traffic Analysis (VJ02010024, security research challenge IMPAKT 1, Ministry of Interior Czech Republic)
- SGS (internal grant of CTU in Prague)

VOLUNTEERING FOR THE COMMUNITY

- Senator in the Academic Senate of Faculty of Information Technology CTU
- Vice-Chairman of the Commission for Evaluation of Education Quality of the Academic Senate of Faculty of Information Technology CTU
- President of IEEE Student Branch for Czech Technical University
- Poster Co-Chair at International Conference on Network and Service Management (CNSM) 2024
- Student poster session Co-Chair at 12th Prague Embedded Systems Workshop (PESW) 2024
- Supervision of 3 theses and opposition of 3 theses
- Preparation of Workshop for Ph.D. students on NECS Winter School 2024

MY PUBLICATIONS

Jančička L., Koumar J., Soukup D, Čejka T. "*Analysis of Statistical Distribution Changes of Input Features in Network Traffic Classification Domain*". NOMS, 2024

Koumar J., Hynek K., Pešek J., Čejka T. "*NetTiSA: Extended IP flow with time-series features for universal bandwidth-constrained high-speed network traffic classification.*" Computer Networks 240 (2024): 110147.

Koumar J., Hynek K., Čejka T. "*Network traffic classification based on single flow time series analysis.*" 2023 19th International Conference on Network and Service Management (CNSM). IEEE, 2023.

Koumar J., Plný R., Čejka T. "*Enhancing DeCrypto: Finding Cryptocurrency Miners Based on Periodic Behavior.*" 2023 19th International Conference on Network and Service Management (CNSM). IEEE, 2023

Pešek J., Plny R., Koumar J., Jeřábek K., Čejka T."*Augmenting Monitoring Infrastructure For Dynamic Software-Defined Networks.*" SpliTech. IEEE, 2023.

Koumar J., Čejka T."*Unevenly spaced time series from network traffic.*" 2023 7th Network Traffic Measurement and Analysis Conference (TMA). IEEE,2023.

Koumar J., Čejka T. "*Network traffic classification based on periodic behavior detection.*" 2022 18th International Conference on Network and Service Management (CNSM). IEEE, 2022.

PUBLISHED DATASETS

NETWORK TRAFFIC DATASETS WITH NOVEL EXTENDED IP FLOW CALLED NETTISA FLOW

NETWORK TRAFFIC DATASETS CREATED BY SINGLE FLOW TIME SERIES ANALYSIS

CESNET-MINER22-TS: PERIODIC BEHAVIOR FEATURES OF CRYPTOMINING COMMUNICATION

CESNET-USTS23: A BENCHMARK DATASET OF UNEVENLY SPACED TIME SERIES FROM NETWORK TRAFFIC

PLANNED PUBLICATIONS FOR NEXT 2 YEARS

Koumar J., Šiška P., Čejka T. "*CESNET-TimeSeries-2023/2024: Real-world time series dataset from ISP network traffic.*" -- article writing in process (Nature Scientific Data)

Mudruňka K., Koumar J., Čejka T. "*Device type classification based on time series clustering of ISP network traffic*" -- final experiments in process

Koumar J., Čejka T. "*Real-world Time Series Benchmarks for Anomaly Detection and Network Traffic Forecasting from ISP Network*" -- preparations (Computer Networks)

Koumar J., Čejka T. "*Evaluation of statistical anomaly detection algorithms on high-speed ISP network traffic*"

Koumar J., Čejka T. "*ADS-CORE: A highly configurable system for deploying time-series-based anomaly detection applicable to any area*"

Němec F., Koumar J., Čejka T. "Time Series Analysis in Application Specific Networks"

Jančička L., Soukup D., Koumar J., Čejka T. "*Evaluation of Data Drifts in Network Traffic Classification Domain*"

Smolnej T., Koumar J., Čejka T. "*Evaluation of existing neural network-based anomaly detection methods on high-speed network traffic*"

Oškera D., Koumar J., Čejka T. "*Botnet detection using periodic behavior of network traffic*"

FUTURE DISSERTATION THESIS

Title of the future dissertation thesis:

Threat Detection in Network Traffic using Time Series Analysis

Builded on these three pillars:

- *Supervised classification based on time series analysis*
  - Periodicity, NetTiSA flow
- *Unsupervised classification based on time series analysis*
  - Evaluation of AD on ISP networks, ADS system
- *Time series based datasets*
  - *Datasets for anomaly detection and network traffic prediction*
  - *Datasets for supervised classification*