THALES

Novel anomaly detection and classification algorithms for IP and mobile networks

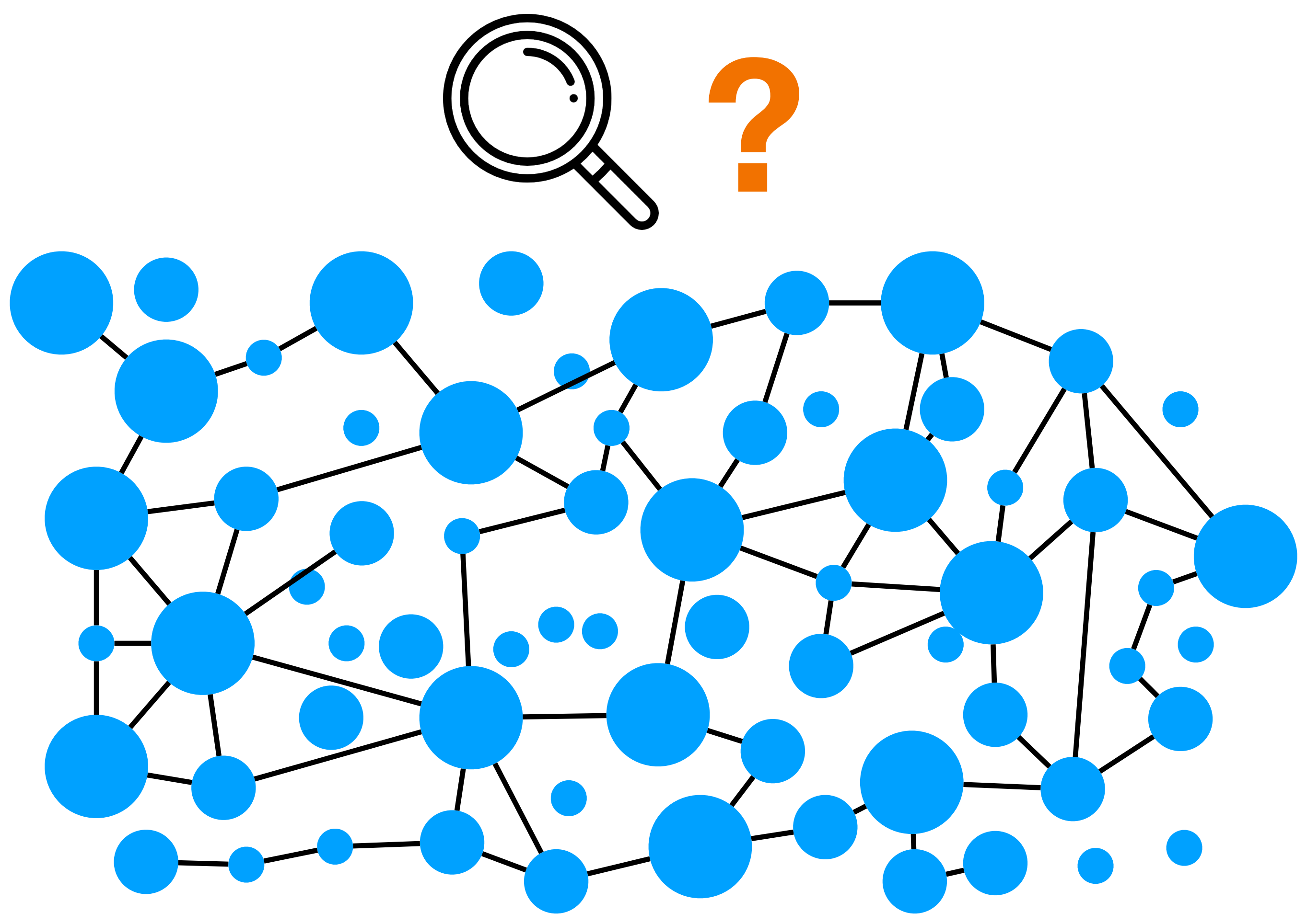
Agathe Blaise

Journée thématique du GT SSLR 2021 sur la sécurité des réseaux - 11 mai 2021

Data analysis

Data: logs of communications, list of transactions, actions of the users, etc.

Potentially **thousands of logs** to handle each day

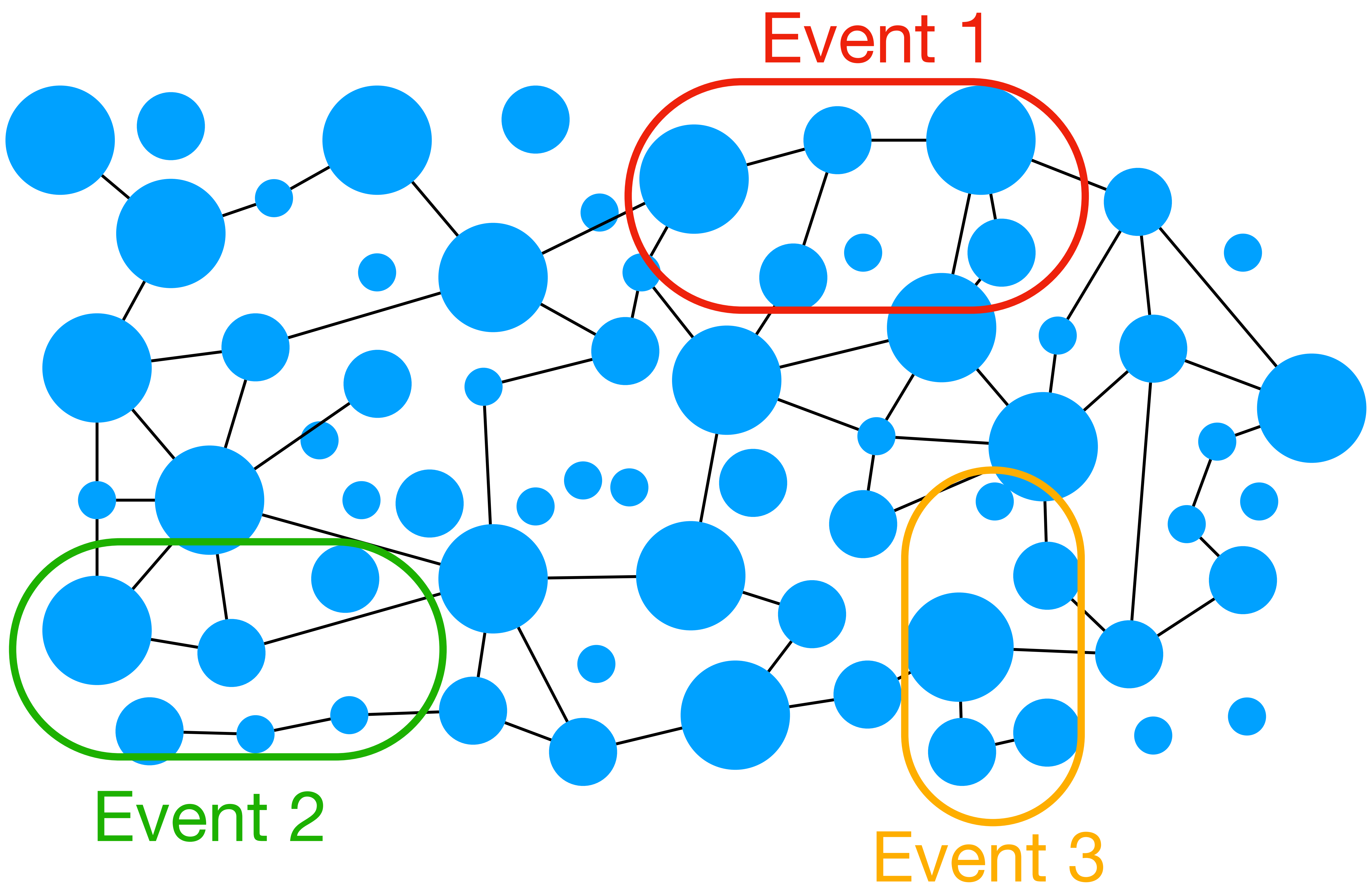


At first sight: **indecipherable**
and no obvious patterns

Knowledge discovery:

- Find underlying patterns
- Define generic model for learning

Data analysis techniques



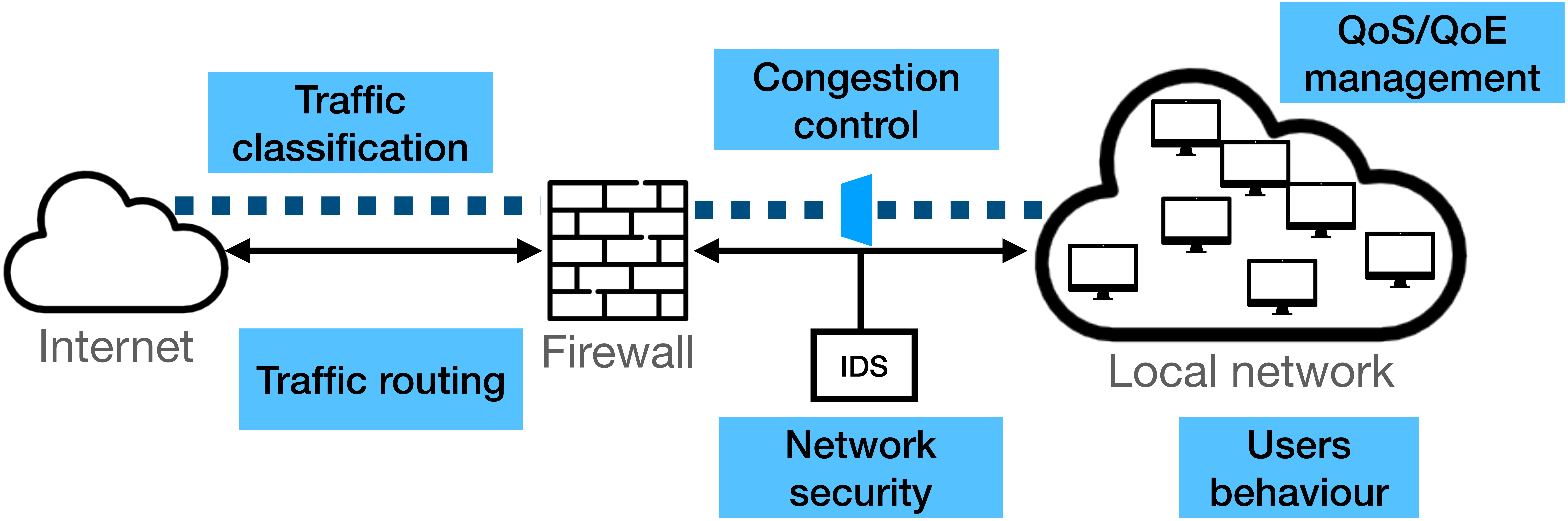
Numerous anomalies

- Correlate them to find **events**
- Investigate **root causes**, **identity** of attackers, modus operandi…

<u>**Supervised** learning</u>: learning based on example input-output pairs.
→ classification and regression techniques

<u>**Unsupervised** learning</u>: learns patterns from unlabelled data.
→ clustering and rule extraction techniques

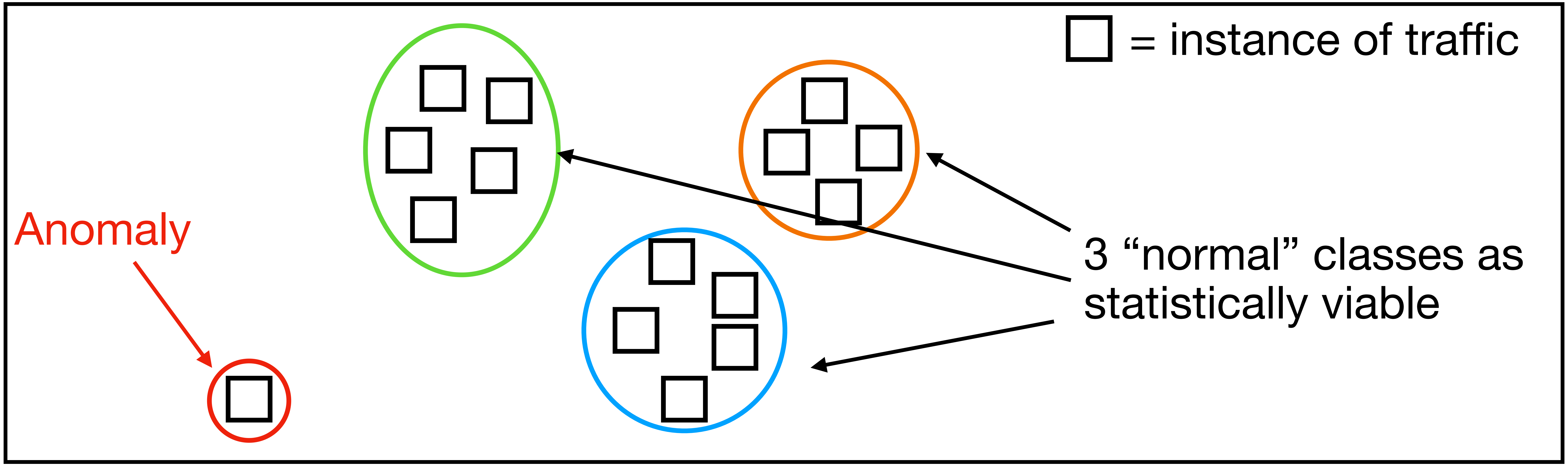
Network behaviour analysis



Targets of data analysis

- **Malicious behaviour** from users
  - Denial-of-service attacks, network scanning, click fraud, man-in-the-middle
- **Unusual behaviour** from users
  - Bursts of traffic, special events, point-to-multipoint communications
- **Operational events**
  - Outages from the network or cloud operator, hardware failures, bad configurations

Data analysis



1. Aggregation level

□ = host, flow?

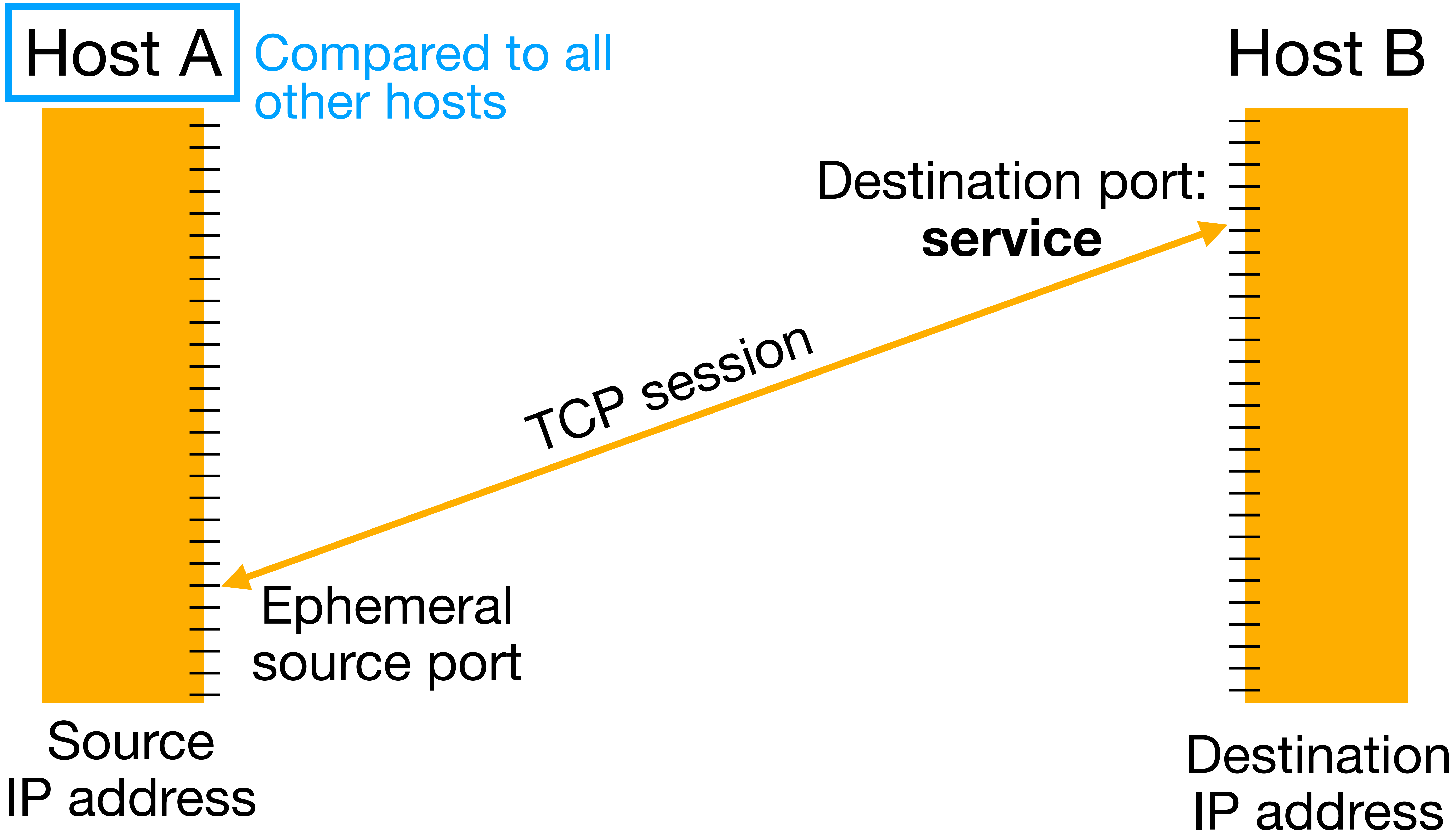
What to characterise?

2. Features choice

→ Attributes of the element

How to characterise it?

Aggregation levels



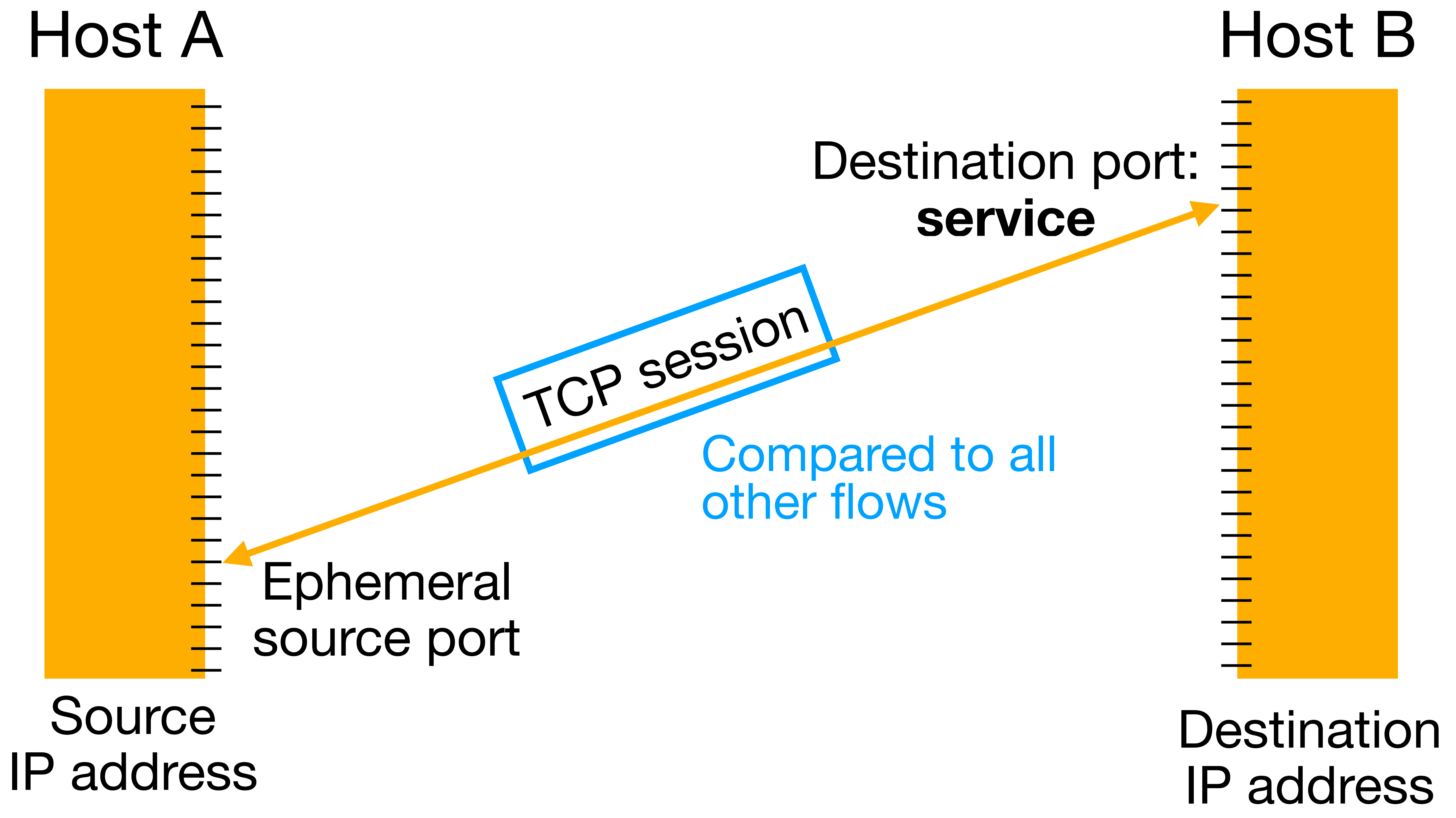
1. **Aggregation level**

Host behaviour

2. **Features**

Packet counts, frequency of communications, protocols

Aggregation levels



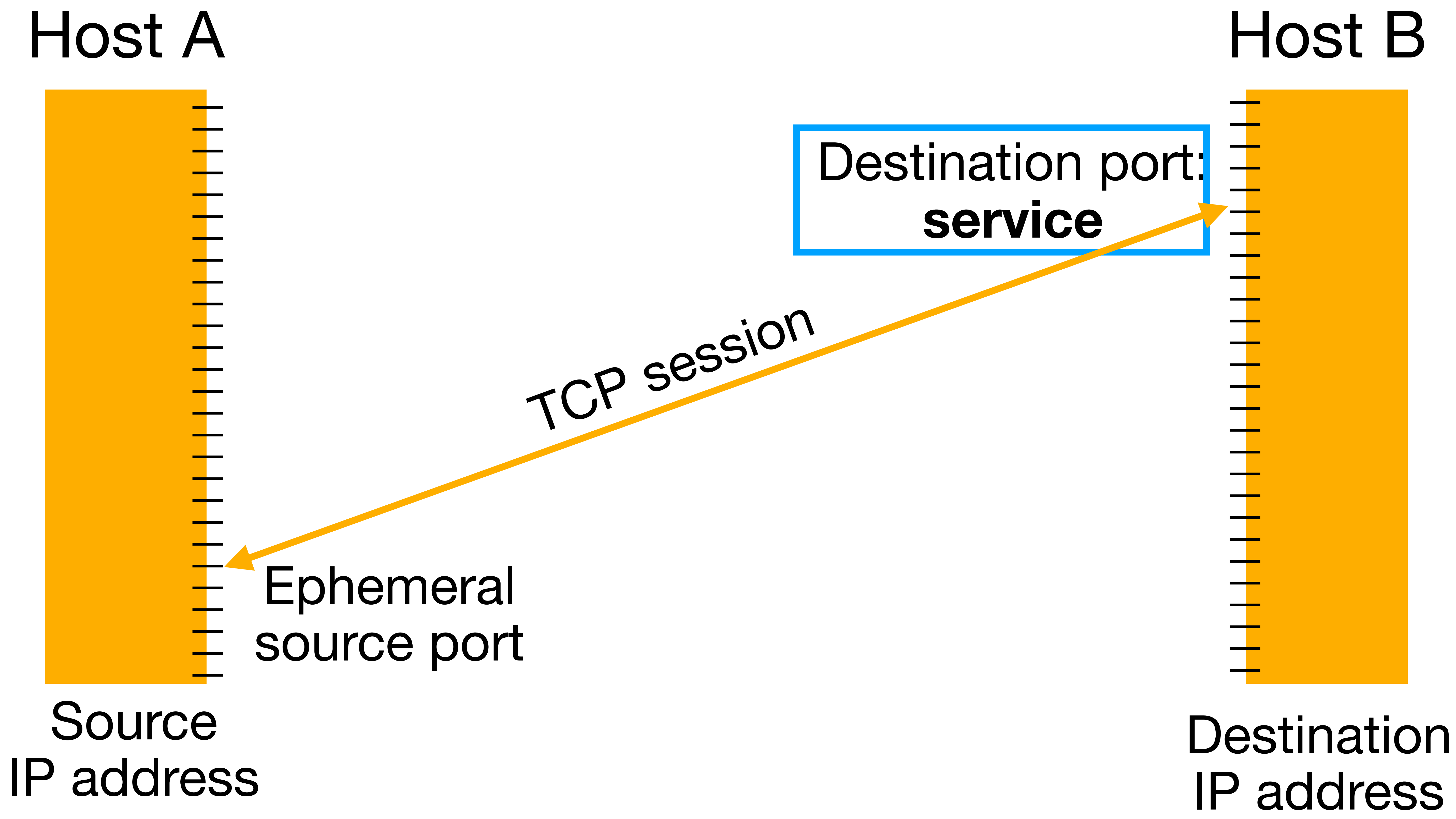
1. **Aggregation level**

Flow features

2. **Features**

Flow duration, flow volume, mean packet length, packet inter-arrival time, entropy

Aggregation levels



1. **Aggregation level**

Port number / service id

2. **Features**

Packet counts, diversity indices, protocols

→ Port or service-level **rarely analyzed**

Applications

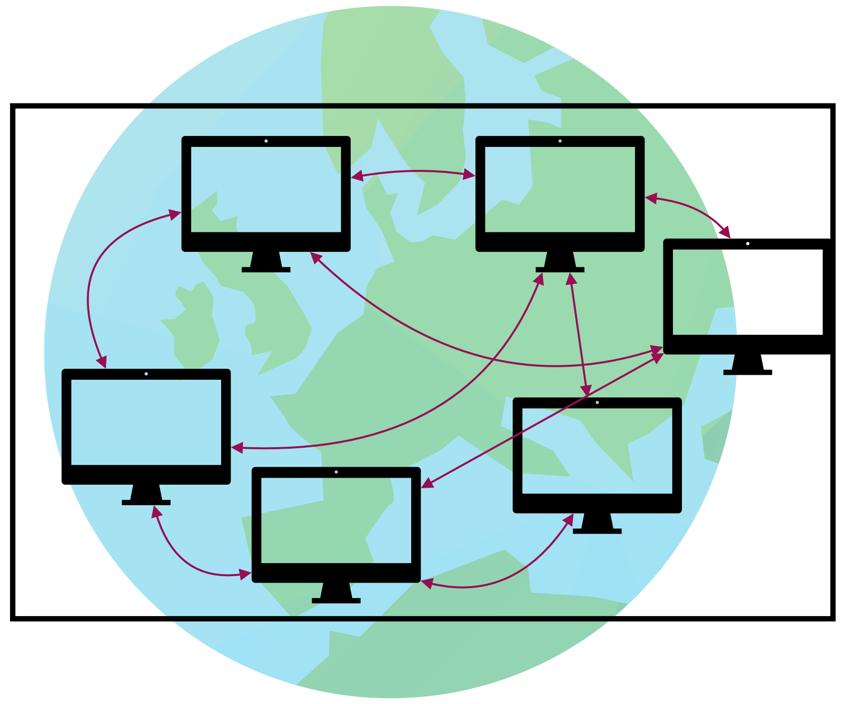
Analysis of the usage of **services, applications and port numbers**

- **State-of the art:** reasons why unused technique
- **Objective:** assessing its **benefits** through **lightweight** techniques
- Examples in 3 different contexts:

Split-and-Merge

Internet-carrier level

BotFingerPrinting



Local (corporate) network

ASTECH

Cellular networks

Security aspects

Behavioural analysis

Per-service detection

Rather **underused** method:

- Numerous elements to analyse
  - In IP networks: 65,536 ports
  - In cellular networks: all services or mobile apps

→ Requires an algorithm of low-complexity

- Traffic obfuscation to avoid firewalls / encrypted traffic

→ Deep Packet Inspection to induce used applications

Per-service detection

Ports and applications **universally and permanently** used

Able to identify uncommon behaviours **not seen with flows and IP adresses**:

- **Long-term** detection as ports subsist over time

→ Detection of attackers **slowly spreading**

- **Several vantage points** as ports universally used

→ **Cross-validation**

- Application **failover** or **update**, **vulnerability scan** on a given port

→ Not visible by analysing IP addresses and flows

Our objectives

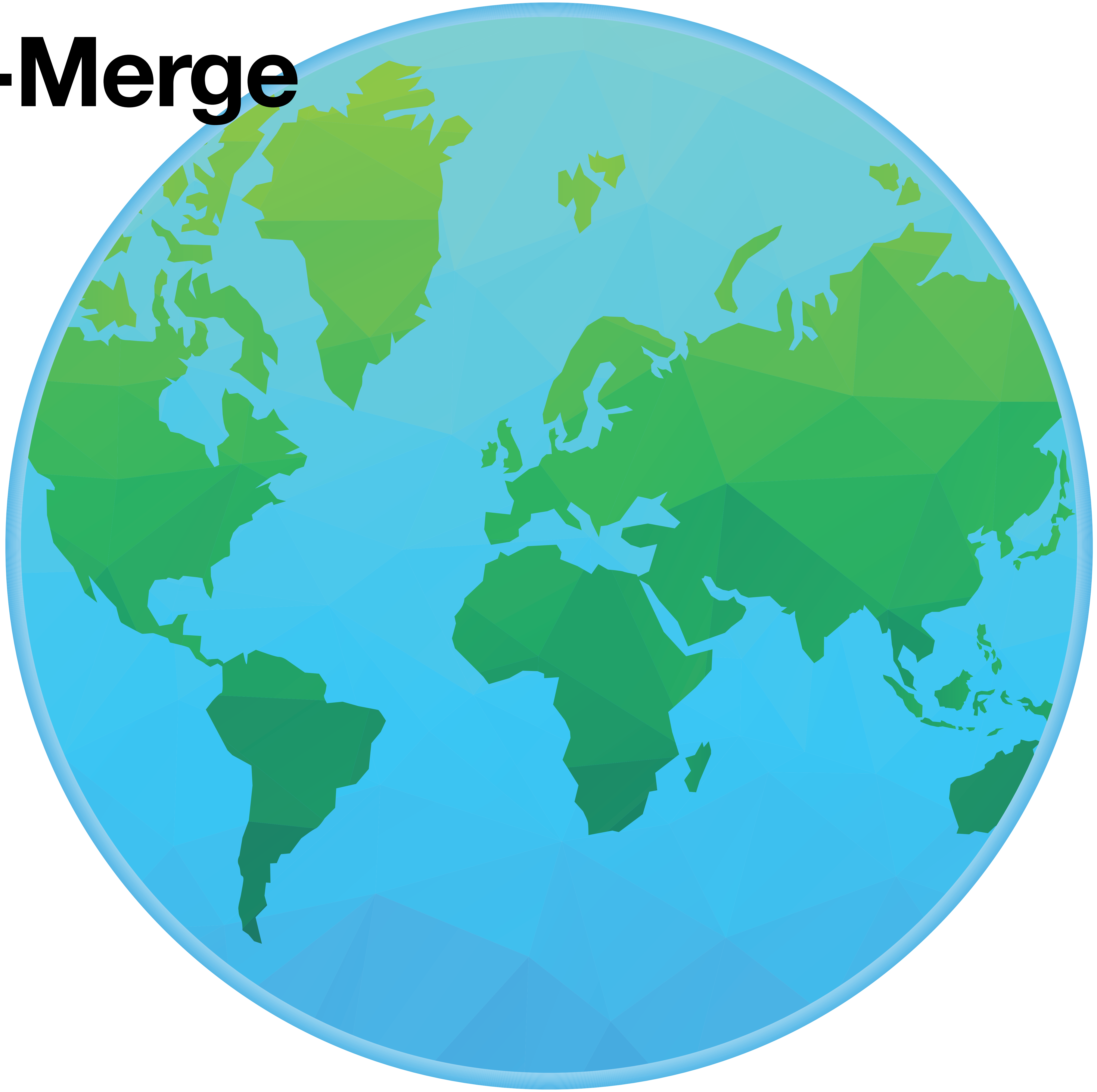
- **State-of-the art:** complex approaches, not fit for real networks

Objective: provide a **pragmatic approach**, lightweight, efficient and scalable

- Through the analysis of **ports, services and applications** usages
- Using **statistical and machine learning** techniques: classification, clustering, anomaly detection
- In various contexts: at IP-level, in local networks, in cellular networks

Split-and-Merge

Split-and-Merge



Split-and-Merge

At Internet carrier-level

- Detection of large-scale **attacks**: vulnerability scans
- Trend of major **botnets** spreading

Split-and-Merge

Challenge: major botnets spreading **not detected** by traditional Intrusion Detection Systems

Our approach:

- **Long-term** analysis of ports usage
- **Cross-validation** in several subnetworks

Our contribution: detection of large-scale **vulnerability scans** and **botnets** spreading

Server vulnerabilities

Exposed to the Internet, open ports, no authentication

Common Vulnerabilities and Exposures:

- CVE-2018-1000115 (memcached) port 11211
- CVE-2017-17215 (Huawei HG532 routers) port 37215

IoT devices vulnerabilities

Low computational power to run **security functions**

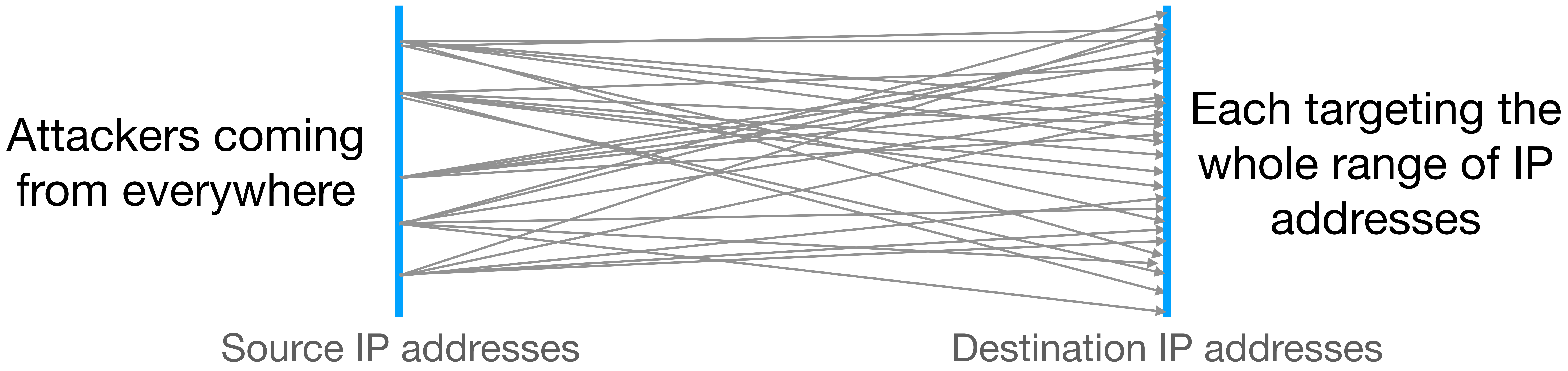
- CVE-2018-7445 (MikroTik devices) port 8291
- CVE-2018-11653 & CVE-2018-11654 (Netwave IP cameras) port 8000

→ Identification of these services or devices by port number.

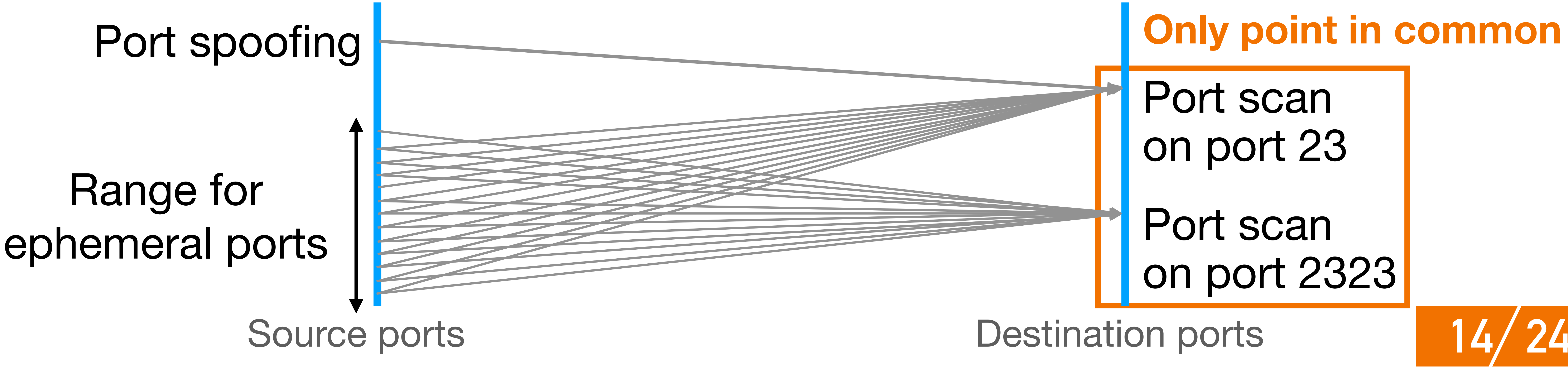
Vulnerability scan

Port scan to identify devices hosting **vulnerable services**

- IP addresses



- Port numbers



Split-and-Merge

Overview

- **Long-term analysis of the usage of ports:**

  1 - Features computation

  2 - Local anomaly detection

  3 - Central correlation

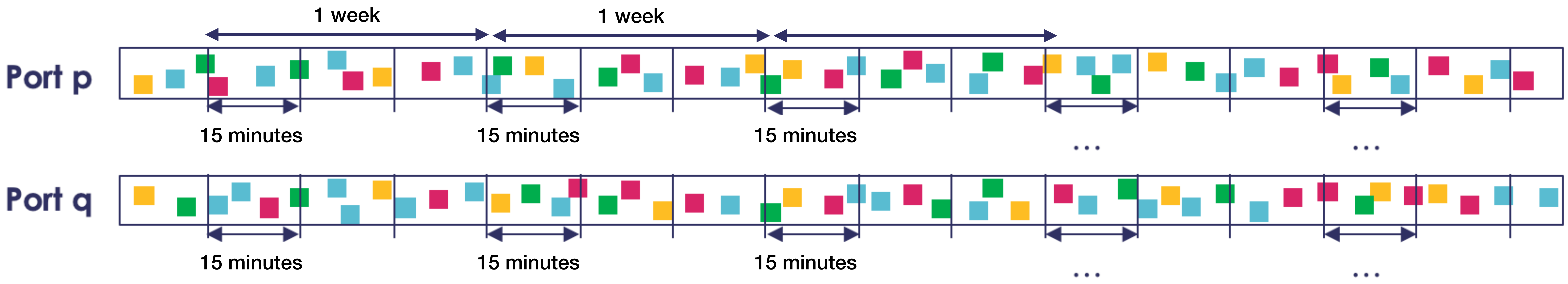
  4 - Fine-grained anomaly characterisation

Split-and-Merge

1 - Features computation

For each port *p:*

- Source diversity index
- Destination diversity index
- Port diversity index
- Mean packet size
- Standard deviation of packet size
- Percentage of SYN packets

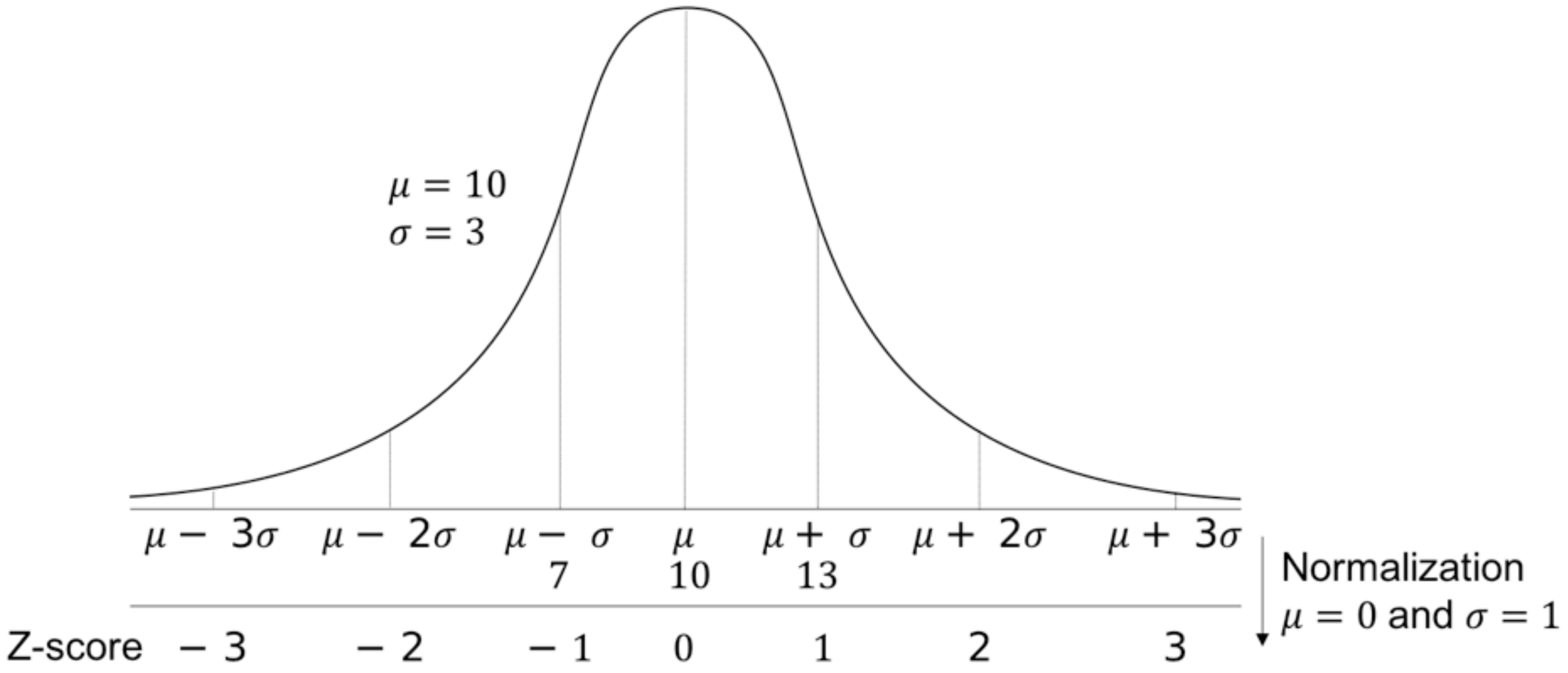


Split-and-Merge

2 - Local anomaly detection

Time series x → normal distribution $\mathcal{N}(\mu, \sigma^2)$ of mean μ and std σ

port *p*	x_1	x_2	x_3
Feature	7	13	30
Feature	54	50	53



- Z-score of x_i : $Z = \dfrac{x_i - \mu}{\sigma}$

  → **not resistant to outliers**

- Modified Z-score using median and median std

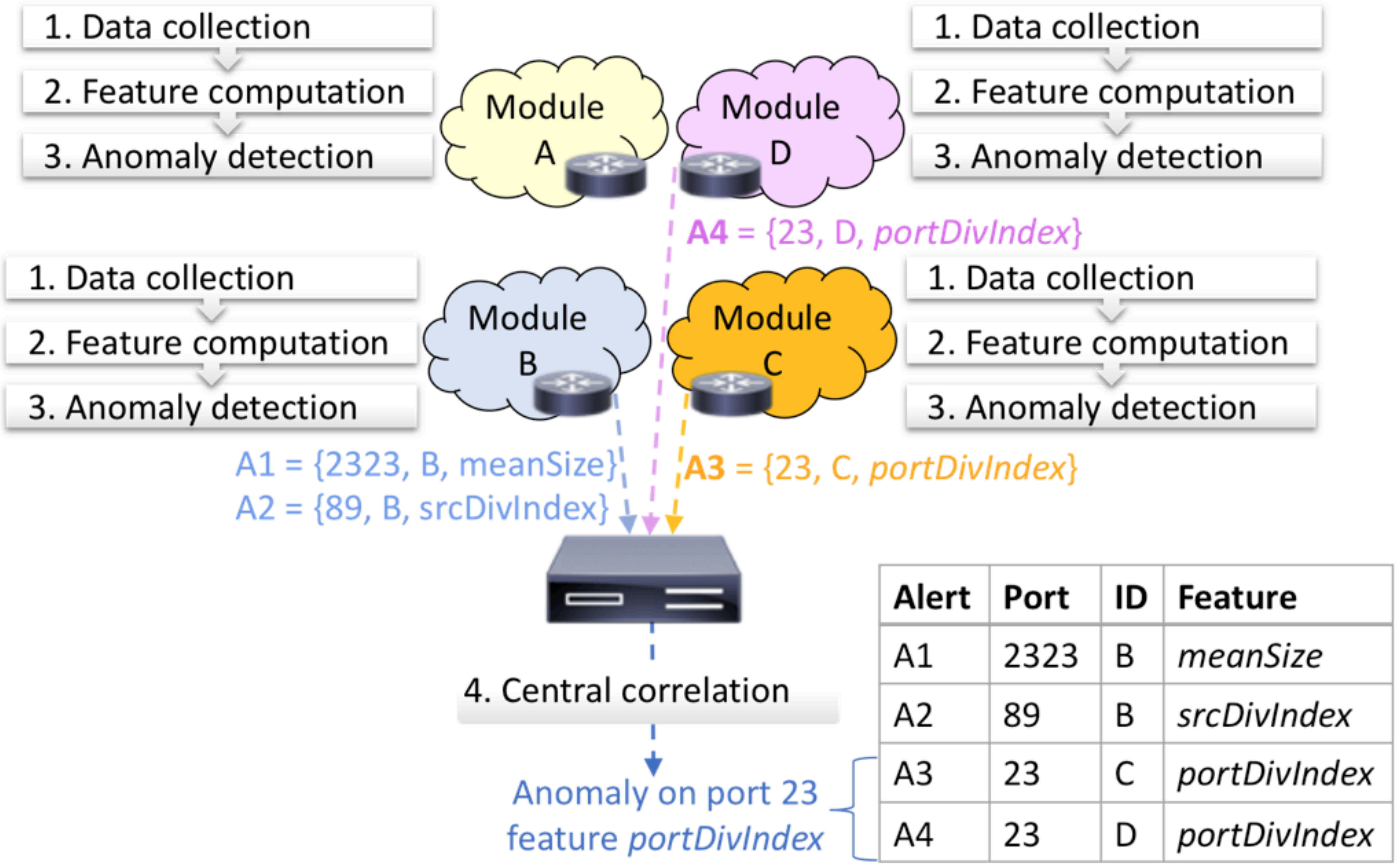
If M > threshold T = 3.5 → **anomaly**

Split-and-Merge

3 - Central correlation

To reduce false positives: Split-and-Merge architecture

Central controller: keep only **distributed** anomalies



Alert	Port	ID	Feature
A1	2323	B	*meanSize*
A2	89	B	*srcDivIndex*
A3	23	C	*portDivIndex*
A4	23	D	*portDivIndex*

Split-and-Merge

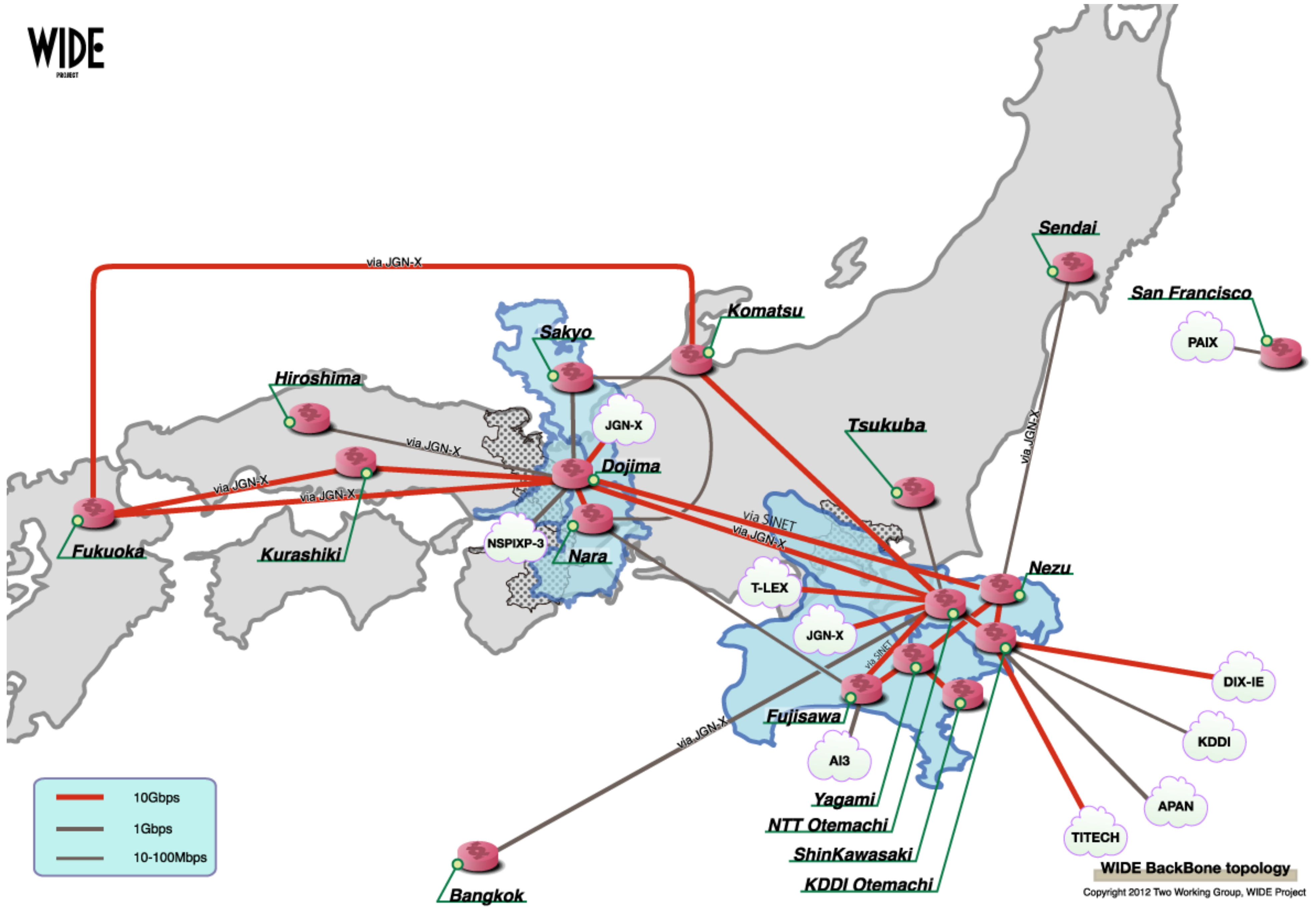
4 - Fine-grained characterisation through expert rules

Classes	Characteristics
More normal packets	`+meanSize, +stdSize`
More forged packets	`-meanSize, -stdSize`
Large scan	`-srcDivIndex, +destDivIndex, -meanSize`
DDoS	`+srcDivIndex, -destDivIndex`
Botnet scan	`+srcDivIndex, +destDivIndex, -meanSize`
Botnet expansion	`+srcDivIndex, +destDivIndex, -stdSize`
Targeted scan	`-srcDivIndex, -destDivIndex`
Less botnet scan	`-srcDivIndex, -destDivIndex, +meanSize, -stdSize`

Evaluation on real-world traces

MAWI dataset (WIDE Project):

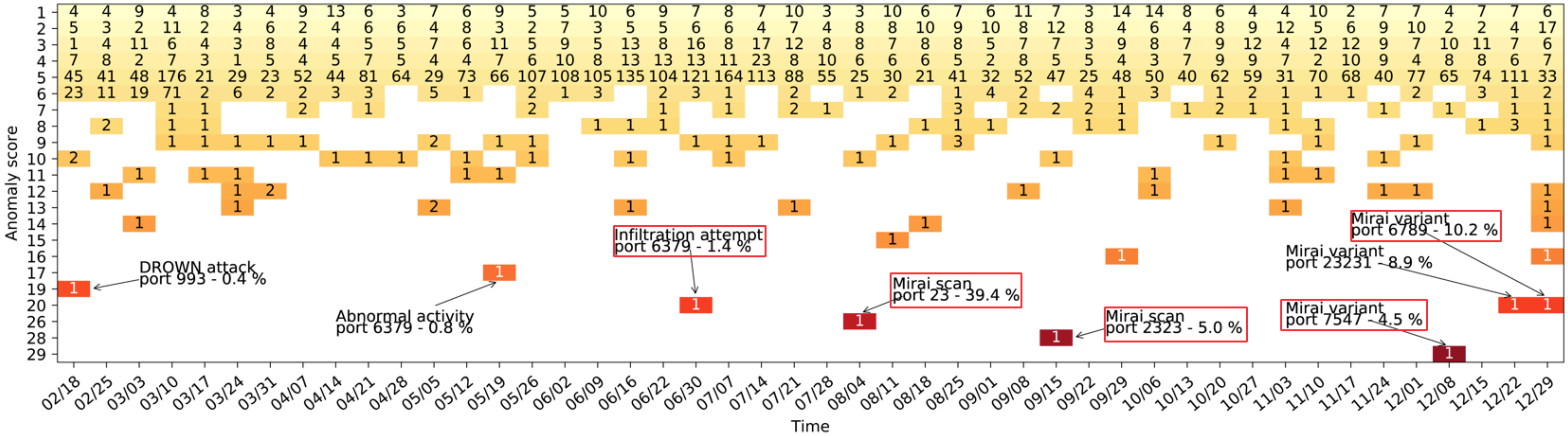
- **Daily files** of 15 minutes of traffic from a transpacific link
- Captured between the **MAWI network and the upstream ISP**



Evaluation (2016)

Anomaly score: number of anomalies for one port

→ Considering **all subnetworks** and **all features**



- Very low number of anomalies
- **Not detected** by traditional IDSs (MAWILab, ORUNADA)

MAWILab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking, Co-NEXT, 2010.

Online and Scalable Unsupervised Network Anomaly Detection Method, IEEE Transactions on Networks and Service Management, 2016.

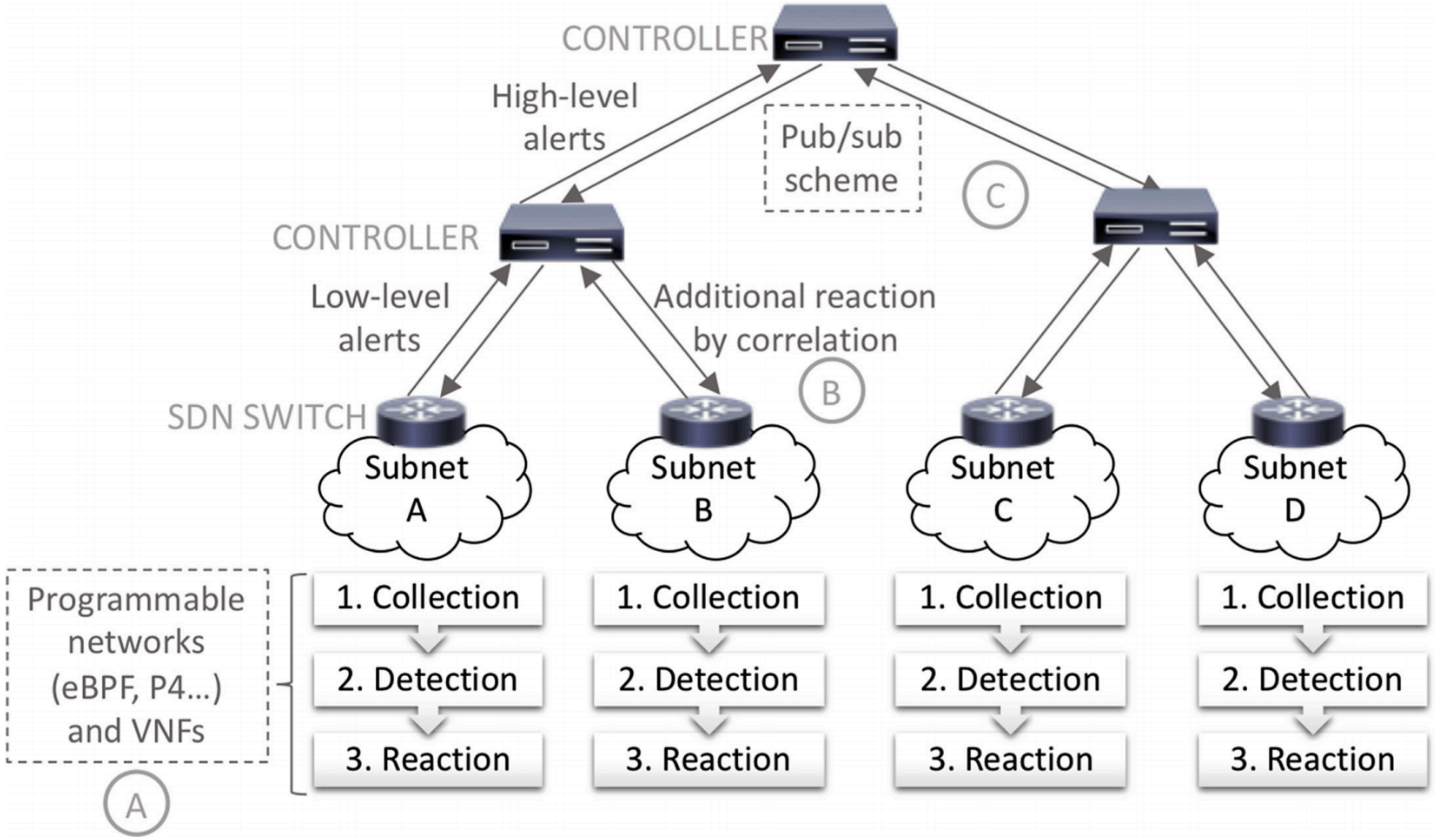
Retrospective of major botnets

- Mirai (ports 23, 2323, 7547, 6789, 2222, 23231)
- Hajime (port 5358)
- Reaper (port 20480)
- Satori (ports 37215, 52869, 8000)
- ADB.Miner (port 5555)
- Memcached (port 11211)
- Wannacry (port 445)

Implementation

Local detection at the data plane enhanced by collaboration between ISPs

- **A**: data plane programming greatly easing the detection and prediction tasks
- **B**: controller aggregating high-level alerts to detect distributed attacks
- **C**: various controllers communicating using a pub/sub communication scheme



Split-and-Merge conclusion

Benefits of **per-port detection:**

- Focus on **port numbers:** detection of **world-wide attacks**, not seen by traditional IDS

- **Long-term** analysis: possible only when using **port numbers**

- **Cross-validation** in different subnetworks: very few **false positives**

Lightweight algorithm: ideally running at the switch-level

THALES

Novel anomaly detection and classification algorithms for IP and mobile networks

Agathe Blaise

Journée thématique du GT SSLR 2021 sur la sécurité des réseaux - 11 mai 2021