UNIVERSITÀ DEGLI STUDI FIRENZE

IWNC 2024

The intrinsic convenience of federated learning in malware IoT detection

Chiara Camerota - Tommaso Pecorella - Andrew D. Bagdanov

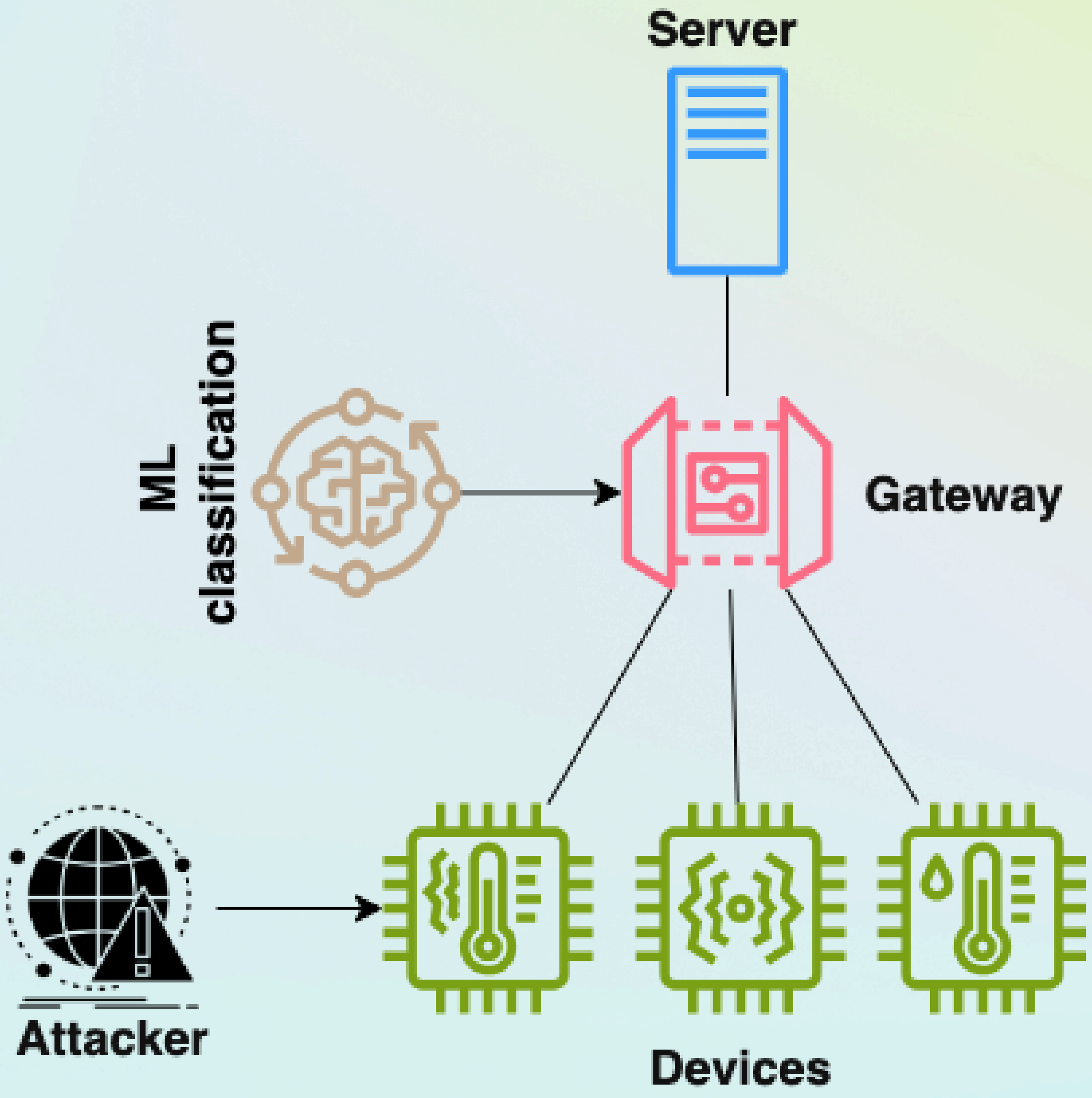
Agenda

- Internet of Things and Malware Detection
- Federated Learning in Internet of Things
- Proposed model and Methodology
- Results and Analysis
- Conclusions and Future work

Internet of Things

IoT is a network that interconnects billions of devices and objects that can collect, exchange, and analyze data

- Lightweight protocols and low power consumption
- High adaptability to various environments (smart cities, homes, etc.)
- Flexibility in wireless networks (e.g., LoRaWAN for cities, Z-Wave for homes)



Malware detection

Various techniques and tools designed to screen, alert, and block malware from gaining access to any device

Malware detection

Why and what is different in the ioT?

- Devices prioritize simplicity over robust security
- In the case of the IoT, resource constraints must be taken into account

Malware detection

Machine Learning (ML) enhances detection by learning malicious behavior patterns and detecting anomalies in IoT networks

- ML models can predict and mitigate emerging threats by analyzing large data sets and device communication in real-time
- It is essential to choose the best ML techniques for the task, given the constraints

Malware detection

Model consideration

High Overhead Traffic
Constant data exchange between devices and a central server leads to heavy traffic on the network

Privacy Concerns
Sending all device data to a central server may expose sensitive information, posing significant privacy risks

Malware detection

Centralized vs Federated models

Centralized Models
Data from all IoT devices is aggregated to a central server where machine learning models are trained

Federated Models
ML models are trained locally on IoT devices, with only model updates sent to a central server

Federated Learning

Models

Federated Averaging (FedAvg)

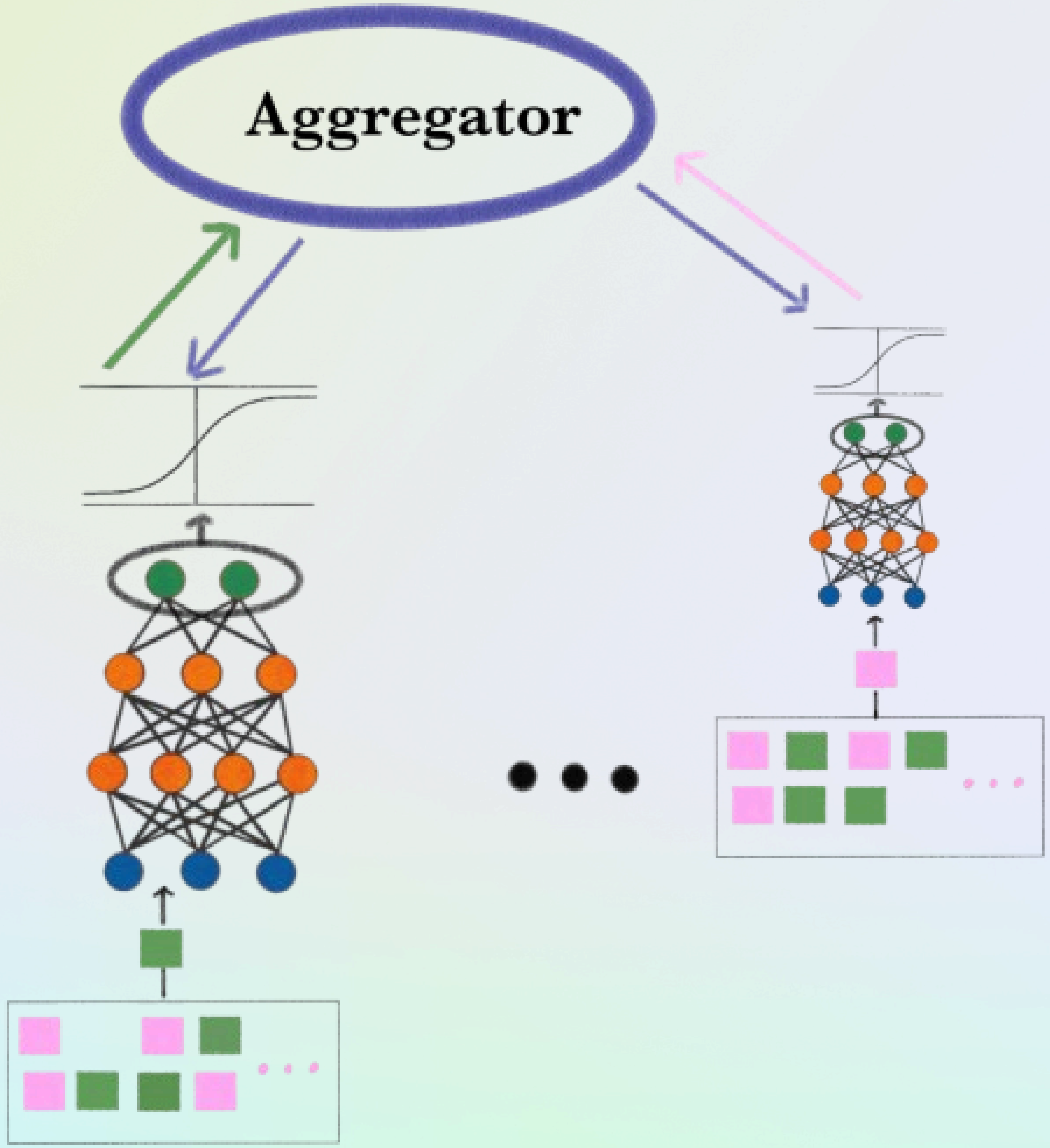
Local models are trained on distributed devices, and a central server averages their parameters to create a global model

Federated Knowledge Distillation

devices share knowledge through distilled model outputs (logits), raw data and model parameters

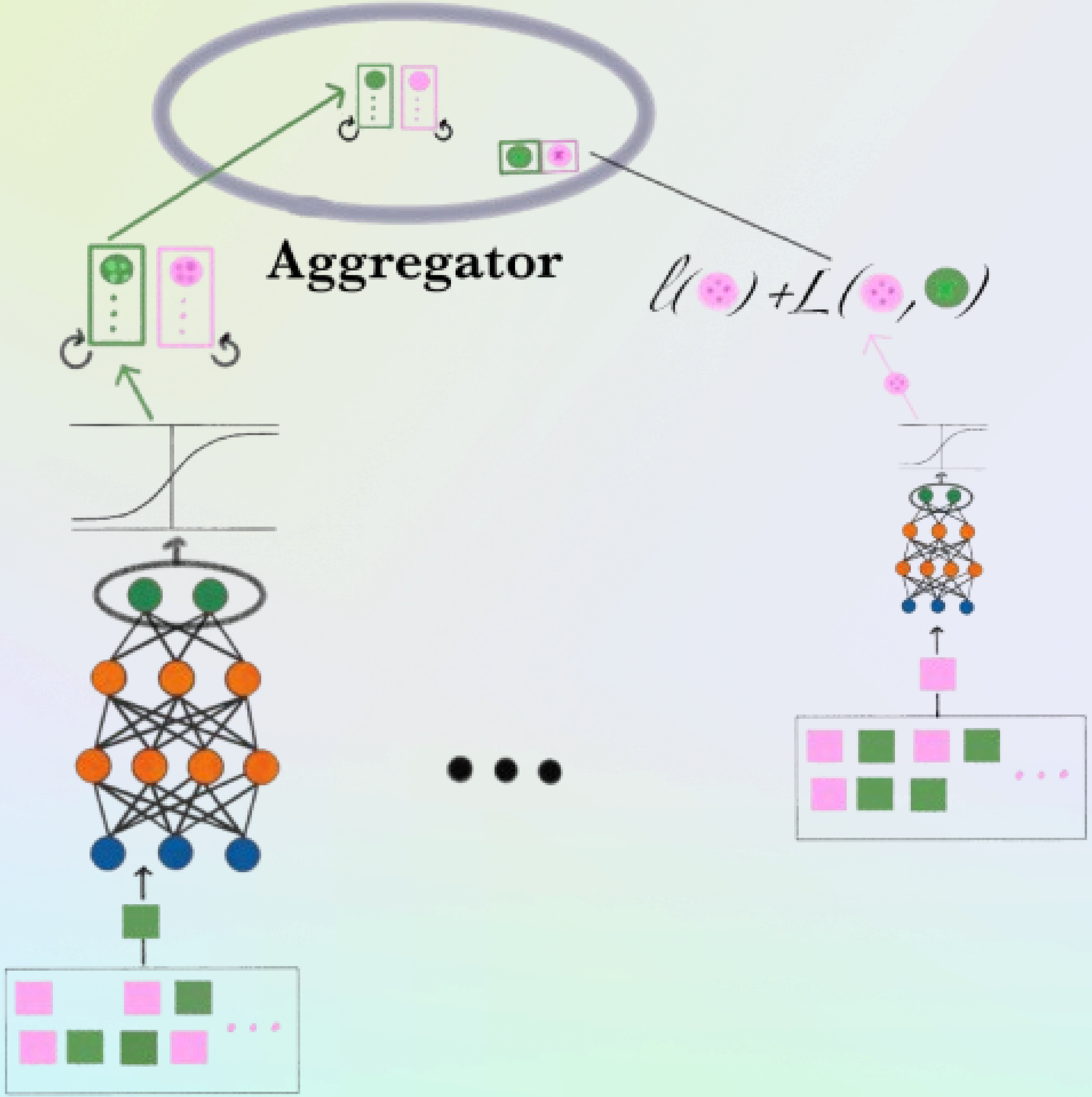
Federated Learning

FedAvg



Federated Learning

Federated Knowledge Distillation



Federated Learning

Advantages

- Suitability for Decentralized
- Robustness Against Non-IID Data
- Minimization of Data Exchange

Federated Learning

Our contributions

Weights updated
The updating weights are a weighted average of the previous and the new values

Non-Stationarity system
Weights are updated when the loss change exceeds a threshold

Minimization of Data Exchange

Evaluation set up

Dataset and features

- Using the public and available IoT-23 Dataset
- Features are all numeric (duration, origin bytes, missed bytes, original packets, origin IP bytes, response packets, response IP bytes)

Input data

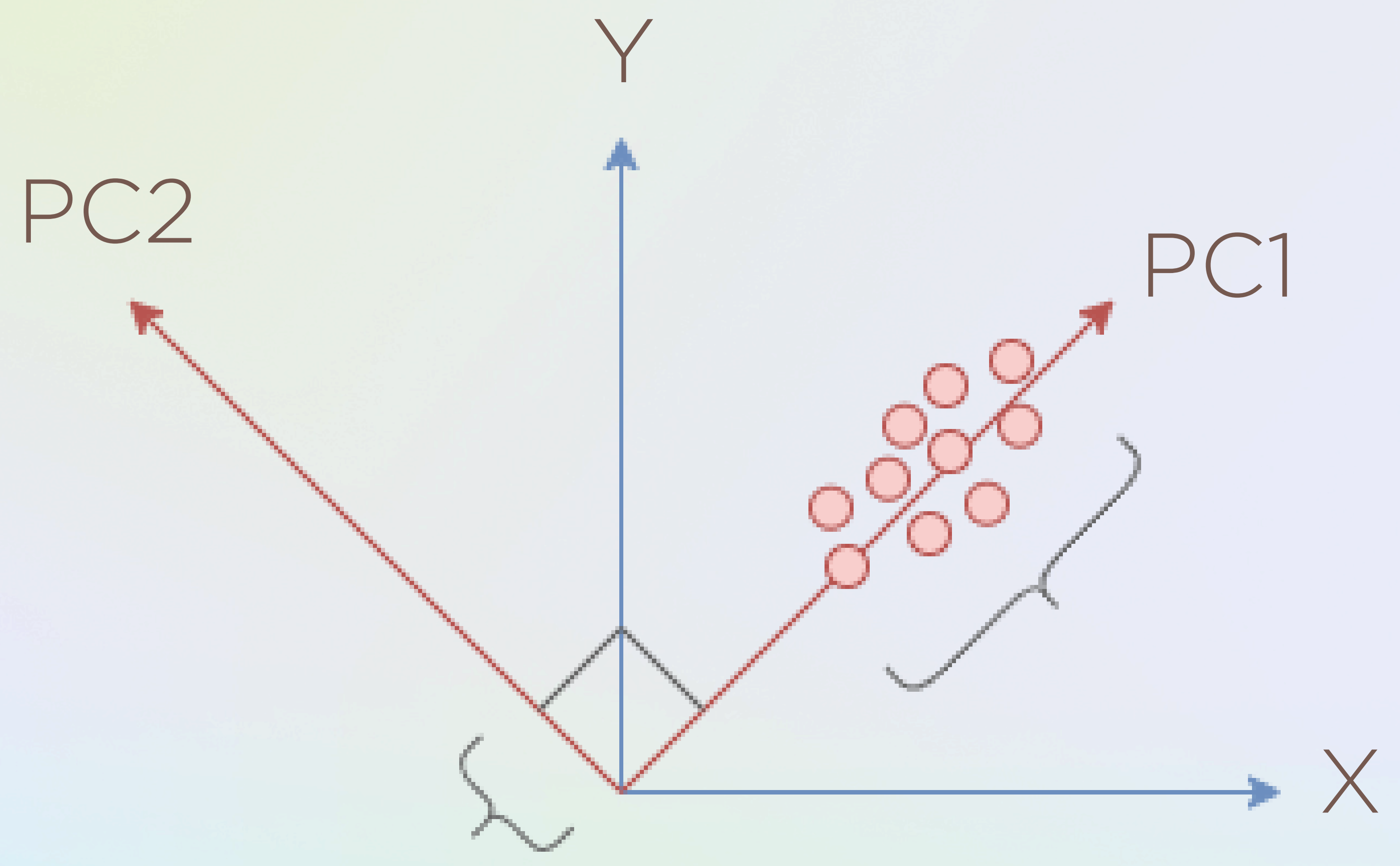
Standardization

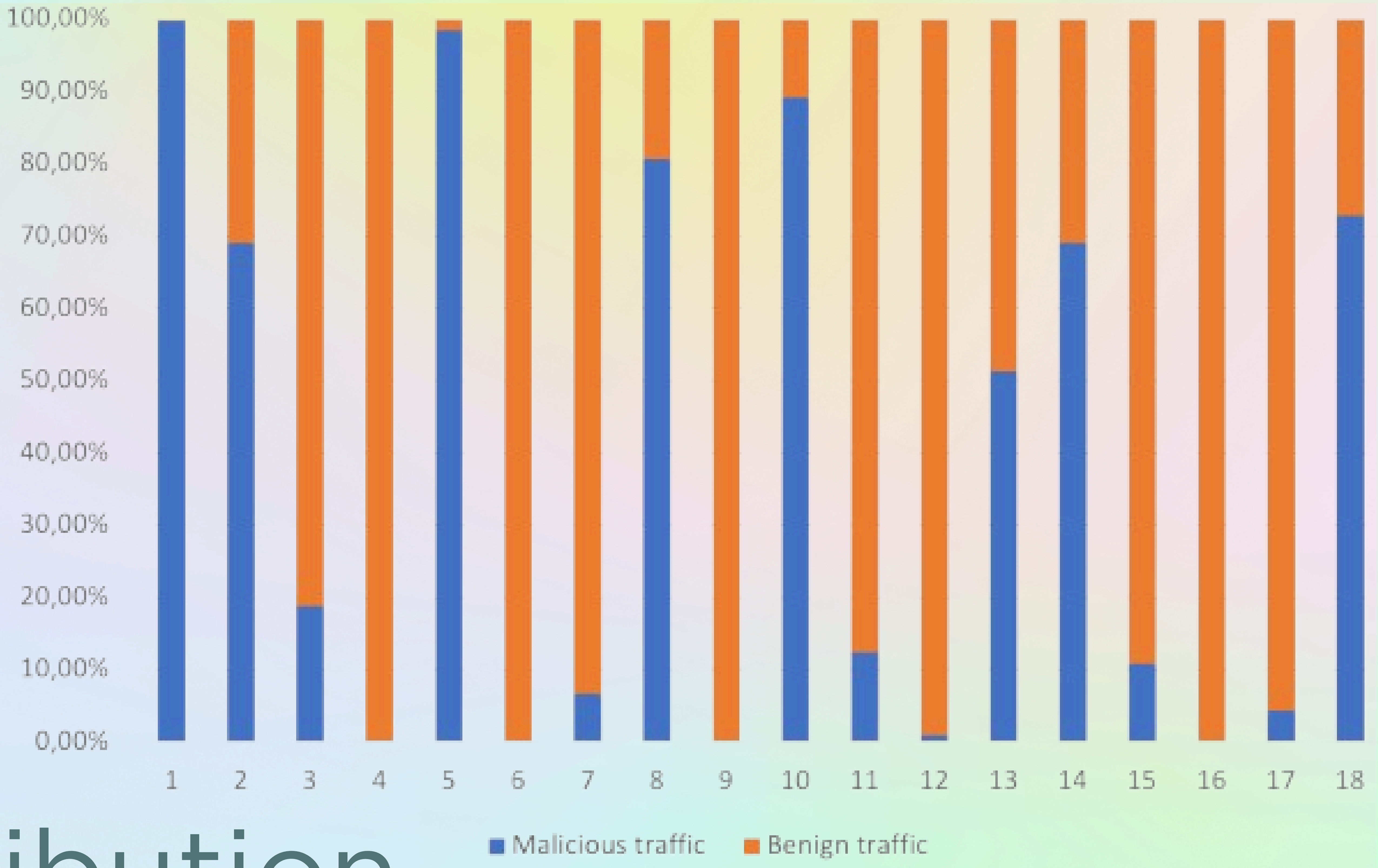
$$\text{INPUT} = \frac{\text{example - MEAN}}{\text{STANDARD DEVIATION}}$$

- Process of rescaling data so that it has a mean of zero and a standard deviation of one
- Can be **global** (indices based on all data clients) or **local** (indices based on client data)

Input data

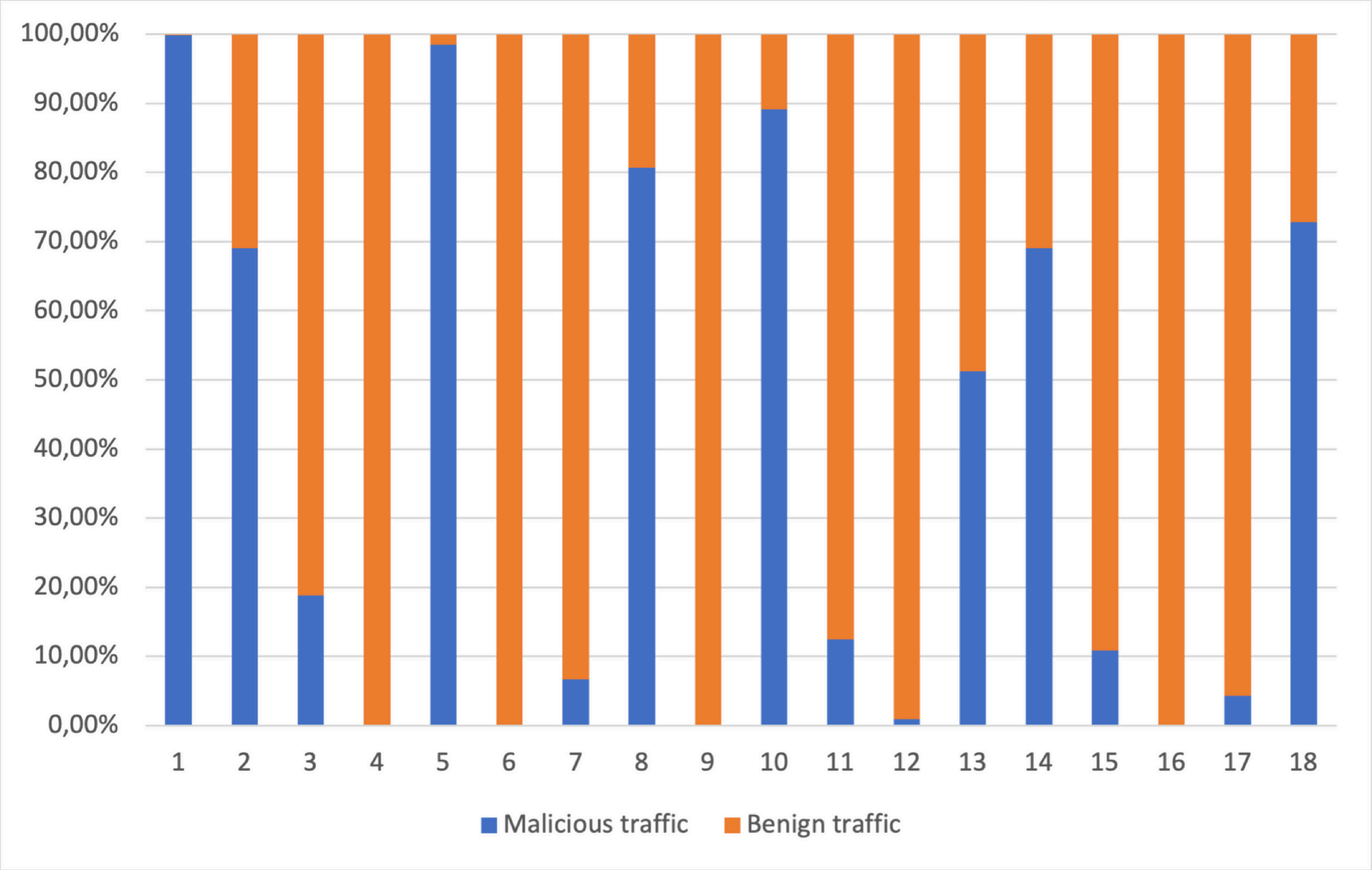
Principal Component Analisys
Reduce the dimensionality of a dataset while preserving as much variance (information) as possible





Label distribution

All data sets are re-balanced by ImbalancedDatasetSampler, which uses the resampling technique

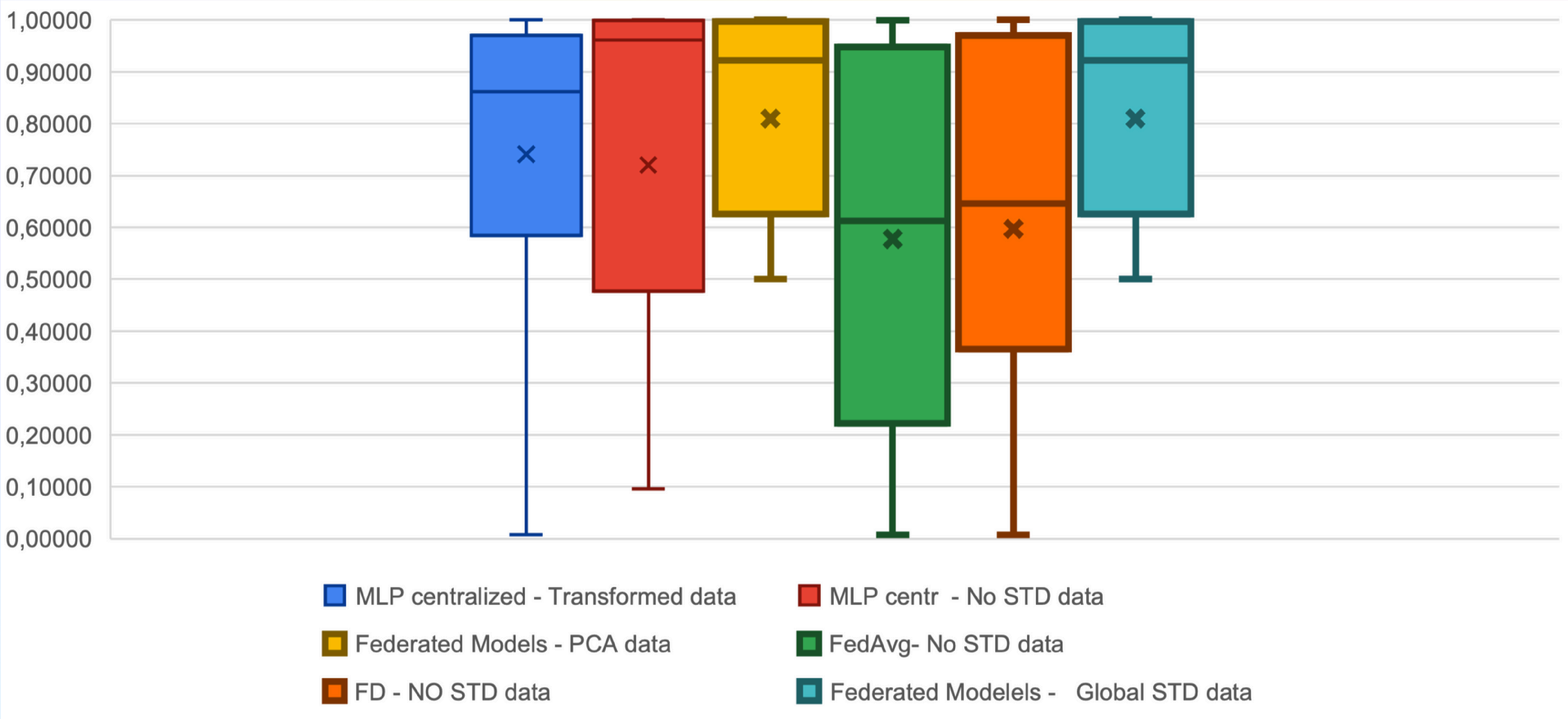


Label distribution

All data sets are re-balanced by ImbalancedDatasetSampler, which uses the resampling technique

Results

Box- plot of AUPRC across all clients by model



Results

Average AUPRC ratio between FD models and Centralized

$$\text{AUPRC-ratio} = \frac{\text{AUPRC(FD)}}{\text{AUPRC (Centr)}}$$

Model	Federated No STD	Federated Global STD	Federated PCA
Centralized No STD	0.94 (FedAvg) 1.07 (FD)	1.64	1.65
Centralized Data Transformation	0.82 (FedAvg) 0.97 (FD)	4.9	4.91

Results

Chi test on AUPRC index performed on the client AUPRC distribution

Chi test	p-value
PCA data	0.04
No STD data - FD	0.99
No STD data - FedAvg	0.99
Glob STD data	0.04

Results

GPU Usage and Time of execution

- **Average data size for each client:** 140 MB
- **GPU utilization per example:** 3.51 MB for centralized models and 2.15 MB for Federated approaches
- **Execution time for 1 MB:** 5 seconds on average for the centralized model and 4.83 seconds on average for the Federated models

Future Challenges

New security challenges require better ways to classify data and explain machine learning decisions

A novel approach using computer vision and explainable AI, such as saliency maps, helps to visualize raw data and highlight important features

Another area of research is to improve models' adaptability and resilience to address the forgetting problem

Conclusions

A federated approach for binary classification optimizes learning while ensuring data security. It leverages the decentralized nature of IoT devices.

Federated models outperform traditional centralised approaches in the global area under the precision-recall curve and have lower variance.

Q & A

Session

The intrinsic convenience of federated learning in malware IoT detection

Thank you!

Chiara Camerota
chiara.camerota@unifi.it

Results

GPU usage based on the test set size (number of examples).
The bars indicate the confidence interval at level 95%.