You Are What You Broadcast: Identification of Mobile and IoT Devices from (Public) WiFi

Lingjing Yu, Bo Luo, Jun Ma, Zhaoyu Zhou, and Qingyun Liu

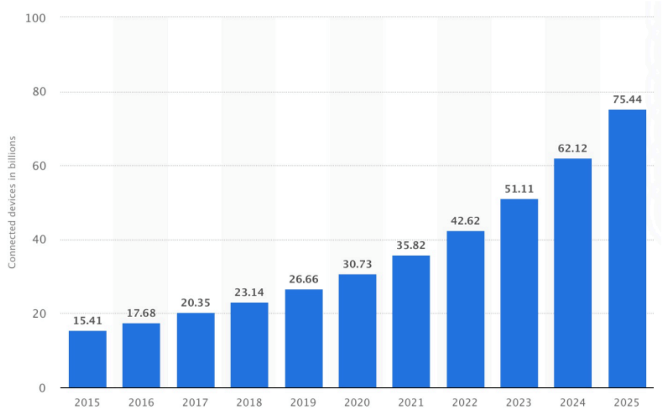
USENIX Security Symposium, Boston, MA. August 2020

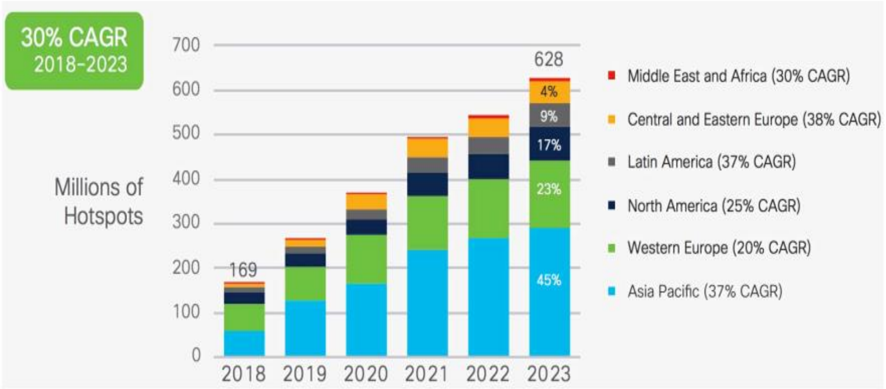
中國科學院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

THE UNIVERSITY OF KANSAS

TSINGHUA UNIVERSITY

Background





Ubiquitous Connectivity

Public WiFi Hotspots

Security & Privacy?

Device Identification: Why?

Device identification in communication and computing	Network measurement Cyber situational awareness
Device management	Malicious device discovery

Task 1: Device Identification

When a mobile/IoT device connects to a WiFi network, we want to know that it is.

To identify the **manufacturer, type, model** of mobile devices using public information.

Task 2: Malicious Device Detection

When a **malicious** device connects to a WiFi network, we want an alert.

To detect the **abnormal** devices, whose BC/MC traffic deviates from benign patterns.

OWL: Overhearing on WiFi for Device Identification

Core Idea: Use Features from Broadcast/Multicast Packets

When a device is connected to a wireless network, it sends out broadcast or multicast packets: DHCP, mDNS, SSDP, etc

Use these packets to fingerprint devices.

Device Identification: Data Collection

Open Public WiFi

Public WiFi with Captive Portals

Encrypted WiFi

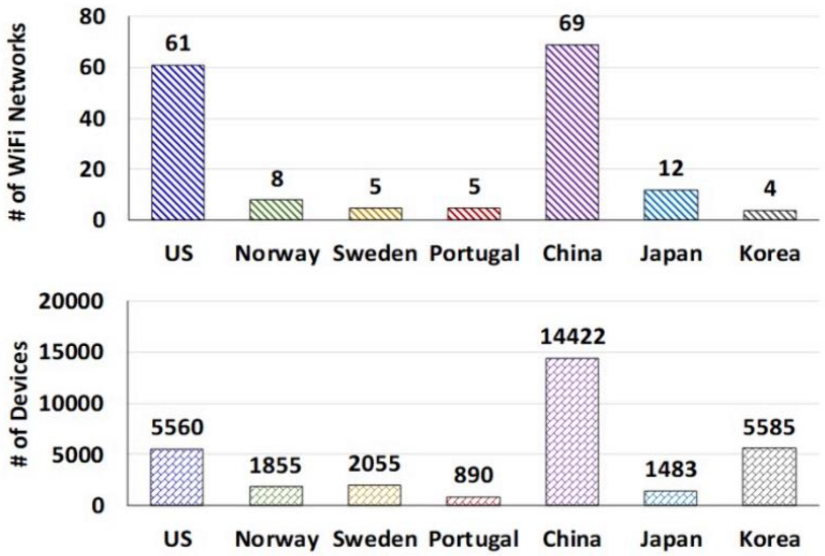
Device Identification: Data Collection

7 Countries: US, Norway, Sweden, Portugal, China, Japan, Korea

Collected Broadcast/Multicast traffic from 176 WiFi Networks, 12 networks disabled BC/MC traffic

Locations: coffee shops, restaurants, retail stores, airports, hotels, corporate guest networks, universities, authors' own homes

BC/MC packets from 31,850 unique devices



Device Identification: Data Collection

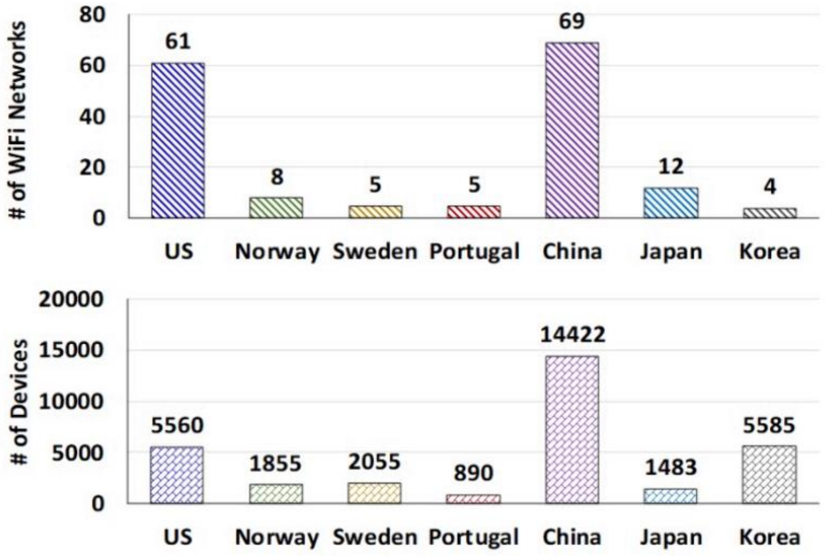
We collected data through a completely passive approach.

We did not turn on promiscuous mode: we were the legitimate and intended receivers of these BC/MC packets. NO unicast traffic.

No violation of Terms and Conditions to our best knowledge.

Post-processing to remove any potential personal identifier.

Discussed with two Institutional Review Boards.



Device Identification: Data Collection

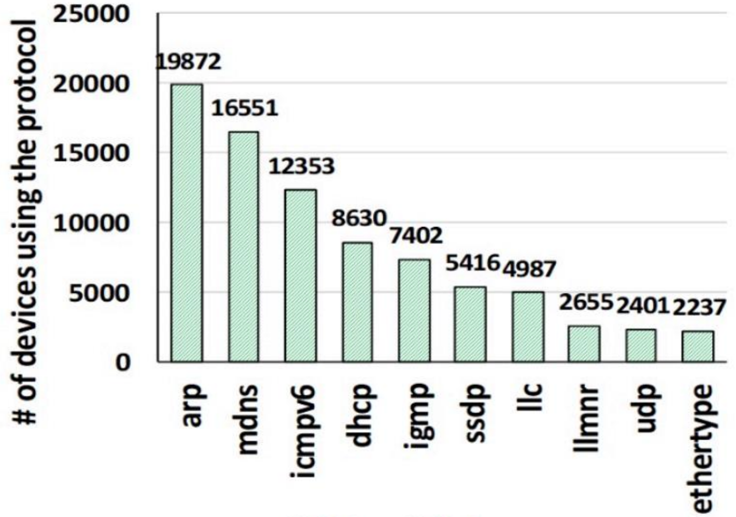
275 different BC/MC protocols were identified in our data

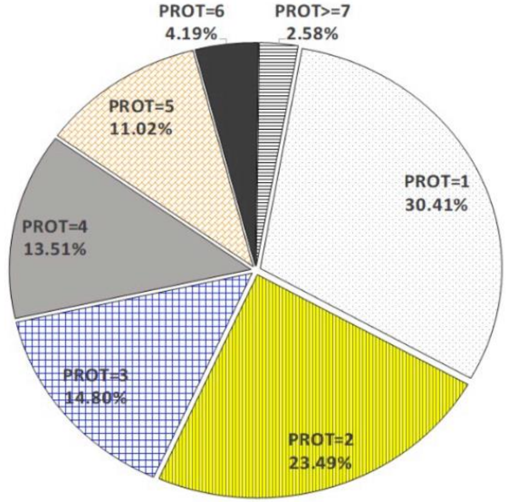
51.9% of the devices use mDNS. Several other application-layer protocols are also widely used.

69% of the devices use more than two protocols

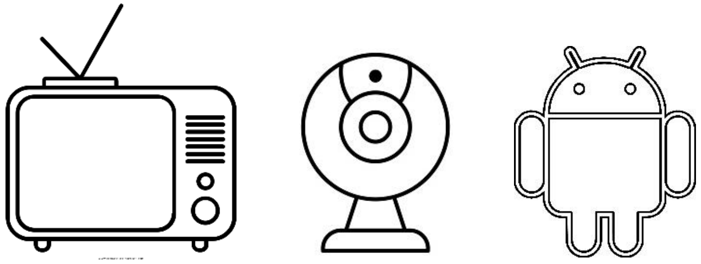
Popularity of protocols appear to be relatively consistent across countries, with a few small exceptions

Some (proprietary) protocols were only discovered from one manufacturer/type/model of devices, e.g., KINK in Samsung TVs



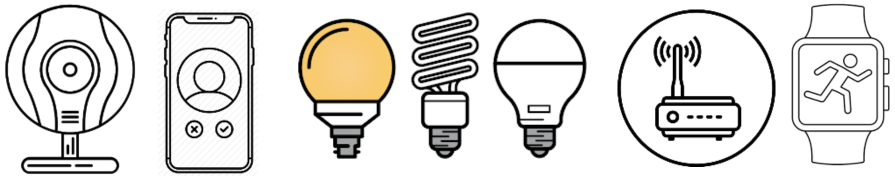


Device Identification: Data Collection



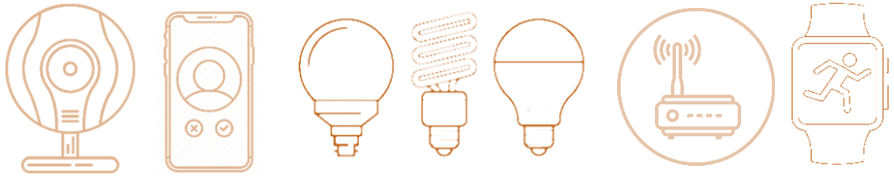
The ground truth dataset

- The identity of each device is physically verified
- **423 devices** with {manufacturer, type, model} labels
- E.g., {D-link, camera, dcs-930lb}



The annotated dataset

- Labeled by annotators based on human-readable content
- **4064 devices** with {manufacturer, type, model} labels
- **6519 devices** with {manufacturer, type} labels
- **15895 devices** with only {manufacturer} label



The sanitized dataset

- Removed all human interpretable textual features from the annotated dataset.
- MAC prefixes are also removed.

Device Identification: Feature Extraction

Identifiers. MAC prefix, HostName in DHCP, etc.

- ✓ Informative, unique
- × not always available, may be tampered

Main Features. Key-value pairs, pseudo natural language features

- × Not unique identifiers
- ✓ Robust, available, provide good discriminatory power

Auxiliary Features. SSDP notify → URL → device description file

- ✓ Include identifiers (device names)
- × Need to actively retrieve the file, not always available

Table 3.2 Examples of data fields that may contain identifiers

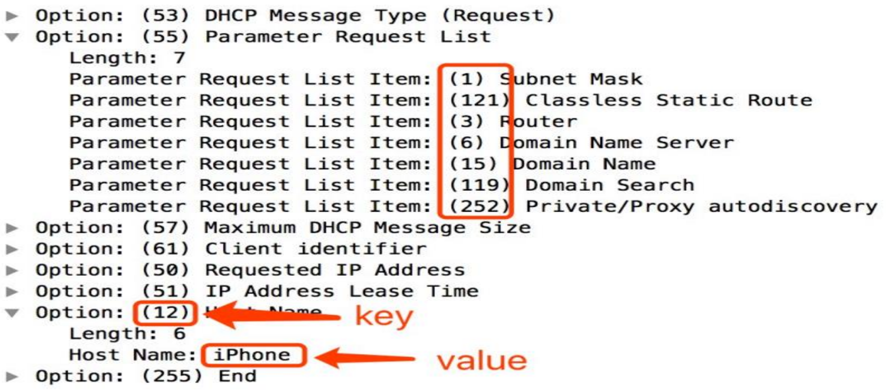
priority	Protocol	Fields
1	–	MAC prefix
2	DHCP	Option12 (HostName)
3	DHCP	Option60 (VendorClass)
4	DHCP	Option77 (ModuleName)
5	DHCPv6	Option39 (ClientFQDN)
6	MDNS	answer names in response messages
7	SSDP.MSEARCH	user-agent
8	SSDP.MSEARCH	X-AV-Client-Info
9	LLMNR	query name
10	BROWSER	query name
11	NBNS	query name
12	UDP	device name

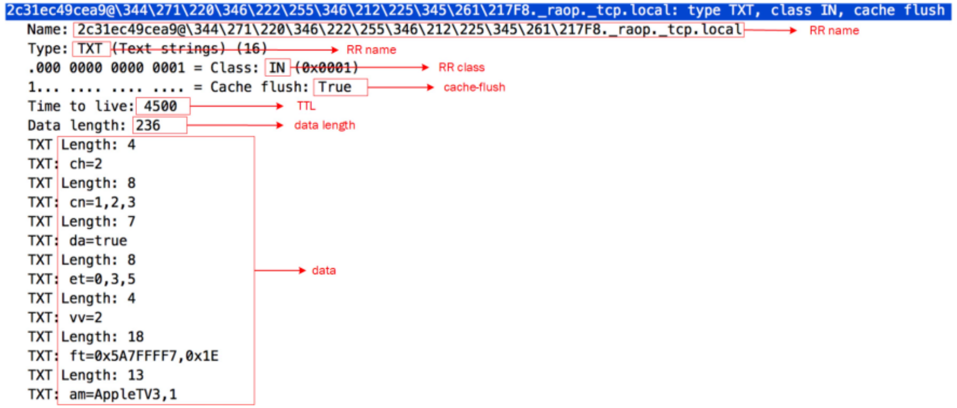
Device Identification: Feature Extraction

Identifiers. MAC prefix, HostName in DHCP, etc.
- ✓ Informative, unique
- × not always available, may be tampered

Main Features. Key-value pairs, pseudo natural language features
- × Not unique identifiers
- ✓ Robust, available, provide good discriminatory power

Auxiliary Features. SSDP notify → URL → device description file
- ✓ Include identifiers (device names)
- × Need to actively retrieve the file, not always available





Device Identification: Feature Extraction

Identifiers. MAC prefix, HostName in DHCP, etc.

- ✓ Informative, unique
- × not always available, may be tampered

Main Features. Key-value pairs, pseudo natural language features

- × Not unique identifiers
- ✓ Robust, available, provide good discriminatory power

Auxiliary Features. SSDP notify → URL → device description file

- ✓ Include identifiers (device names)
- × Need to actively retrieve the file, only used in evaluation


```
▼<root xmlns="urn:schemas-upnp-org:device-1-0">
   <script/>
  ▼<specVersion>
     <major>1</major>
     <minor>0</minor>
   </specVersion>
  ▼<device>
     <UDN>uuid:3f2c04b9-6d19-4d3f-b4a2-fd1f0cf1eec8</UDN>
     <friendlyName>DELL-PC: dell:</friendlyName>
     <deviceType>urn:schemas-upnp-org:device:MediaServer:1</deviceType>
     <manufacturer>Microsoft Corporation</manufacturer>
     <manufacturerURL>http://www.microsoft.com</manufacturerURL>
```

Device Identification: Feature Extraction

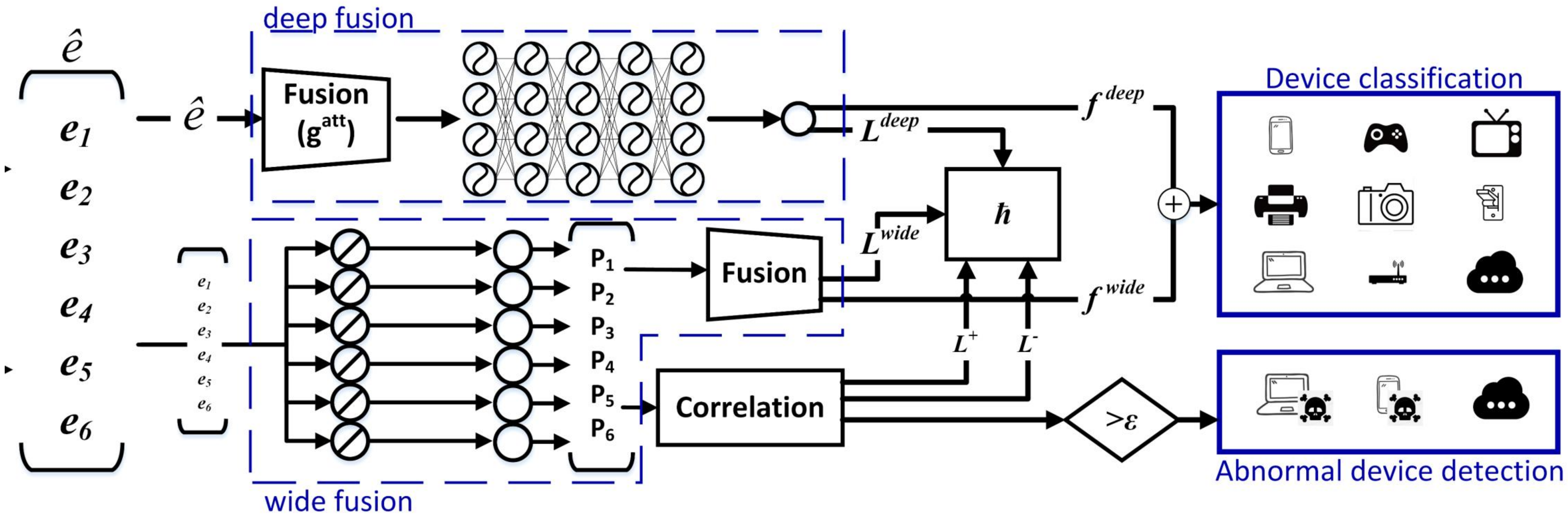
Each protocol generates an independent set of features

Features from different protocols complement each other

Not all protocols are available in all devices

Multi-view Classification

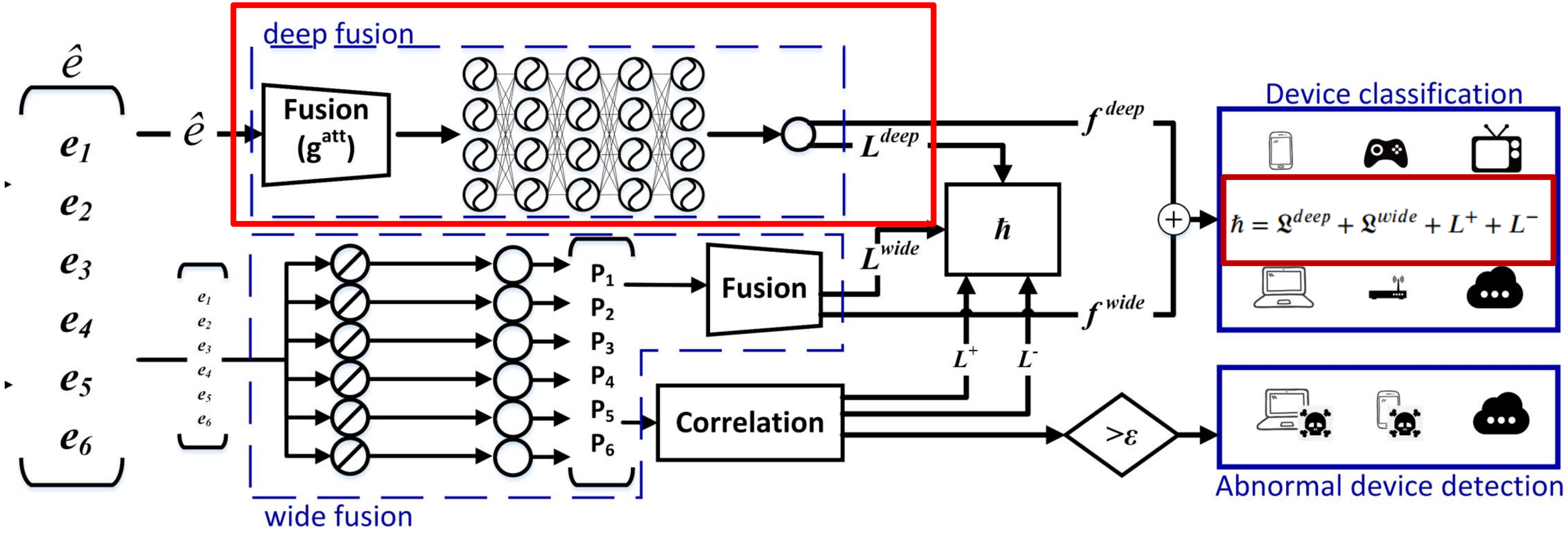
Device Identification: Multi-view Wide and Deep Learning



The deep component: early fusion; maximize the generalization performance

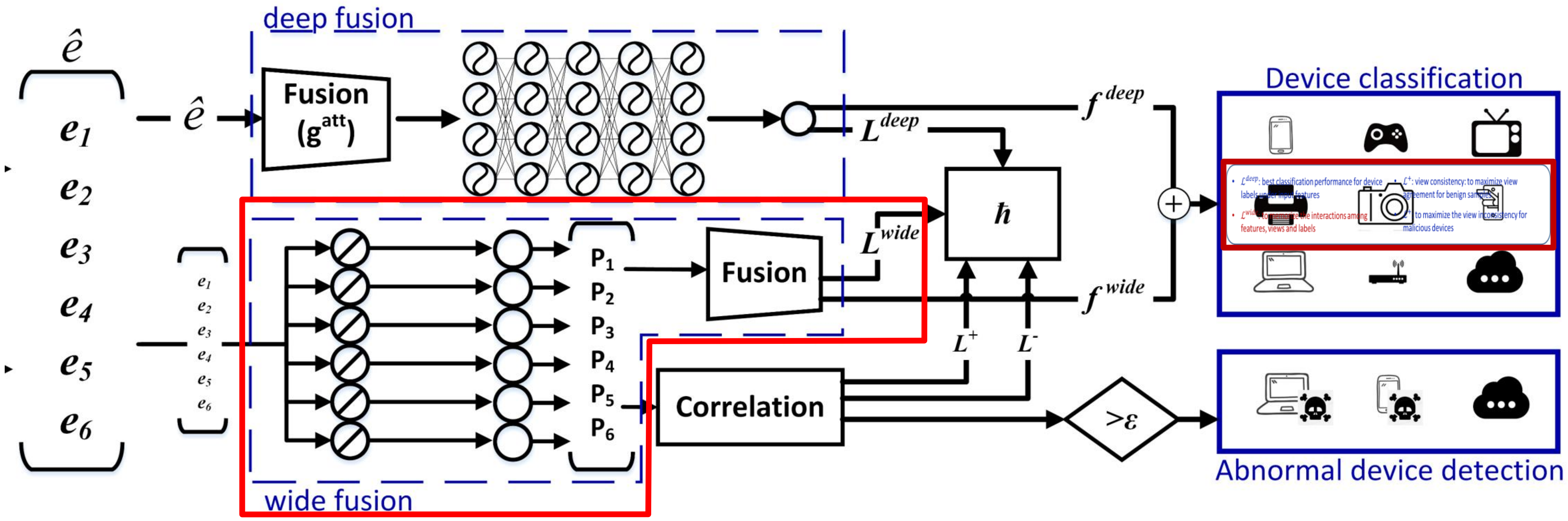
The wide component: late fusion; improve the memorization of label-view interaction

Device Identification: Multi-view Wide and Deep Learning



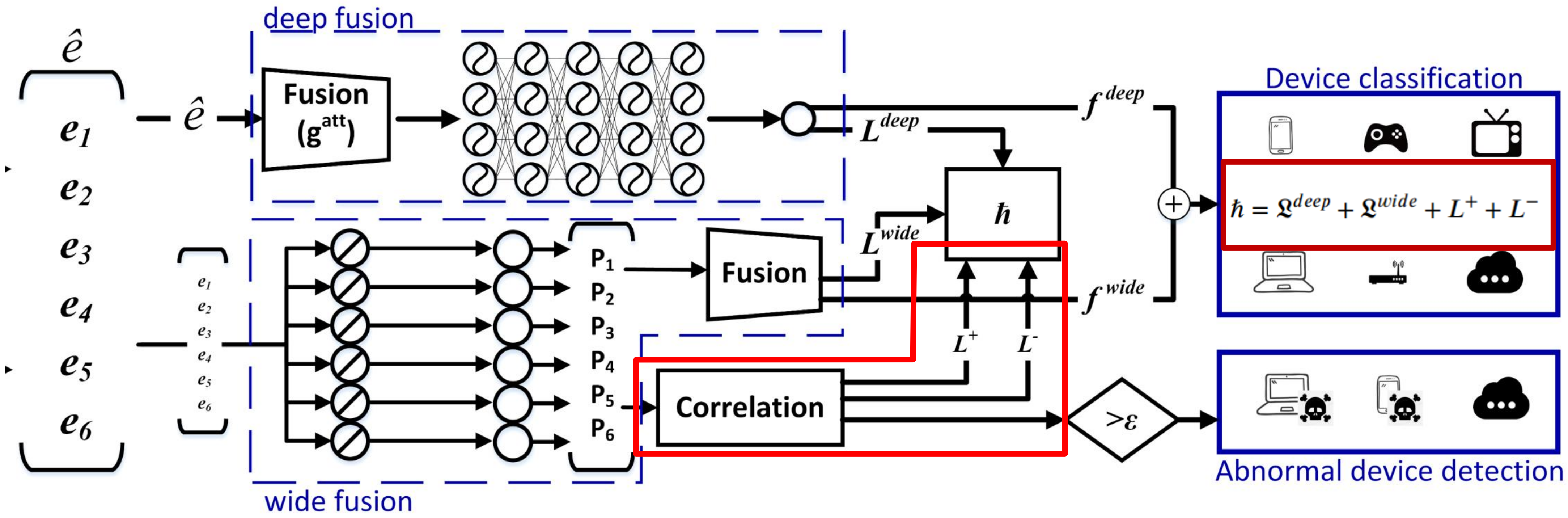
- $\mathcal{L}^{deep}$: best classification performance for device labels under input features
- $\mathcal{L}^{wide}$: to optimize classification performance on each view
- $\mathcal{L}^{+}$: view consistency: to maximize view agreement for benign samples
- $\mathcal{L}^{+}$: to maximize the view inconsistency for malicious devices

Device Identification: Multi-view Wide and Deep Learning



- $\mathcal{L}^{deep}$: best classification performance for device labels under input features
- $\mathcal{L}^{wide}$: to optimize classification performance on each view
- $\mathcal{L}^{+}$: view consistency: to maximize view agreement for benign samples
- $\mathcal{L}^{+}$: to maximize the view inconsistency for malicious devices

Device Identification: Multi-view Wide and Deep Learning



- $\mathcal{L}^{deep}$: best classification performance for device labels under input features
- $\mathcal{L}^{wide}$: to optimize classification performance on each view
- $\mathcal{L}^{+}$: view consistency: to maximize view agreement for benign samples
- $\mathcal{L}^{-}$: to maximize the view inconsistency for known malicious devices

Device Identification: Experiments

Coverage: the fraction of all devices that OWL could generate a label for.

$$C = |\{\text{labeled devices}\}| / |\{\text{all devices}\}|$$

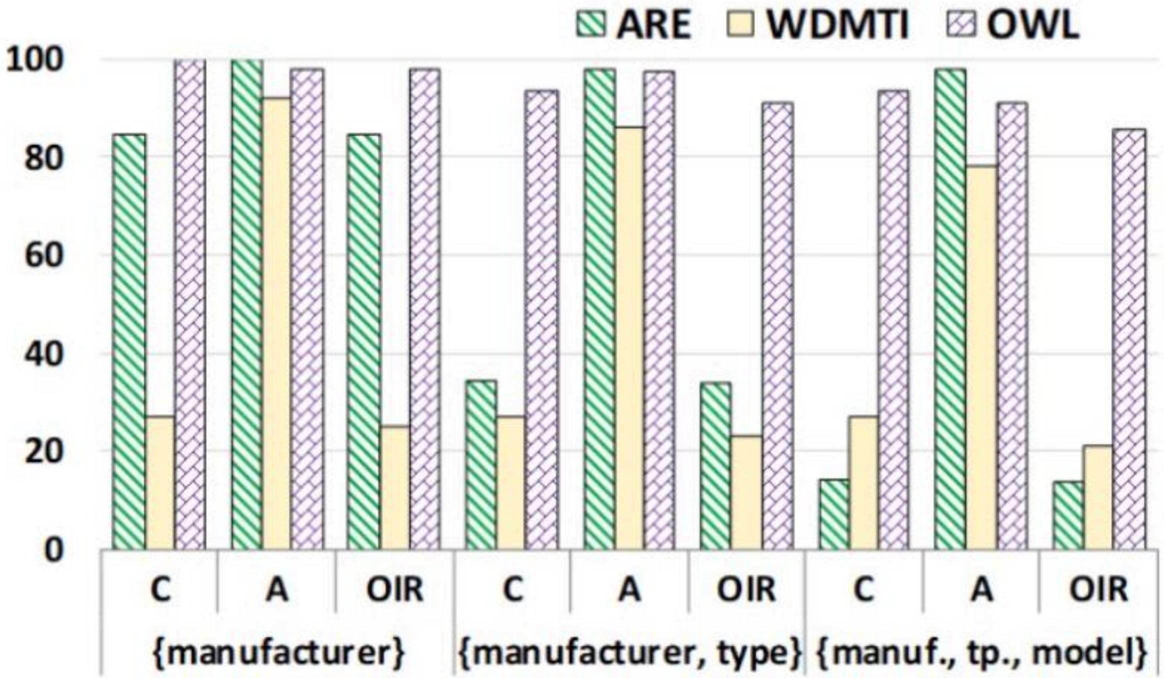
Accuracy: the fraction of labeled devices that are correctly labeled

$$A = \frac{|\{\text{correctly labeled devices}\}|}{|\{\text{labeled devices}\}|}$$

Overall Identification Rate: the faction of all devices that are correctly labeled.

$$OIR = \frac{|\{\text{correctly labeled devices}\}|}{|\{\text{all devices}\}|} = C \times A$$

Device Identification: Experiments



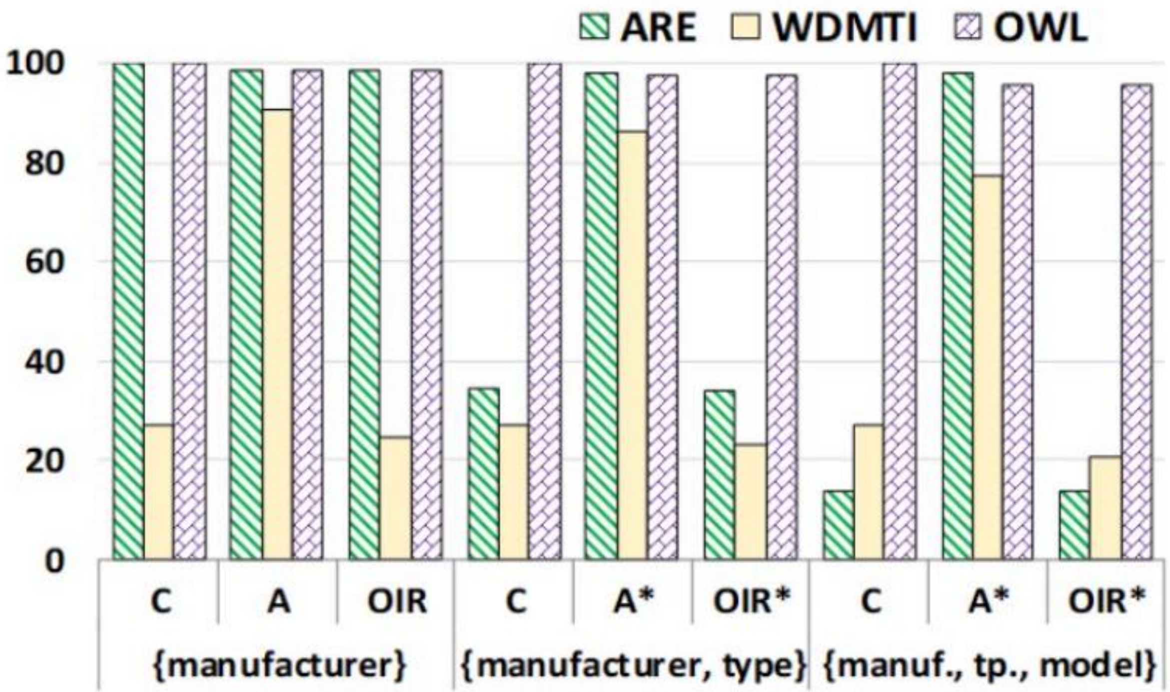
Performance on Ground truth Data

- OWL provides the best overall performance (OIR) at all granularity levels.
- OWL's coverage is consistently the highest.
- At finer granularity, OWL significantly outperforms both ARE and WDMTI in OIR.
- ARE has the best accuracy but limited coverage, especially at fine granularity levels.
- WDMTI's coverage is always limited.

[ARE] Xuan Feng, Qiang Li, Haining Wang, and Limin Sun. Acquisitional rule-based engine for discovering internet-of-things devices. In *USENIX Security*, 2018.

[WDMTI] Lingjing Yu, Tao Liu, Zhaoyu Zhou, Yujia Zhu, Qingyun Liu, and Jianlong Tan. WDMTI: Wireless Device Manufacturer and Type Identification using Hierarchical Dirichlet Process. In *IEEE MASS*, 2018.

Device Identification: Experiments



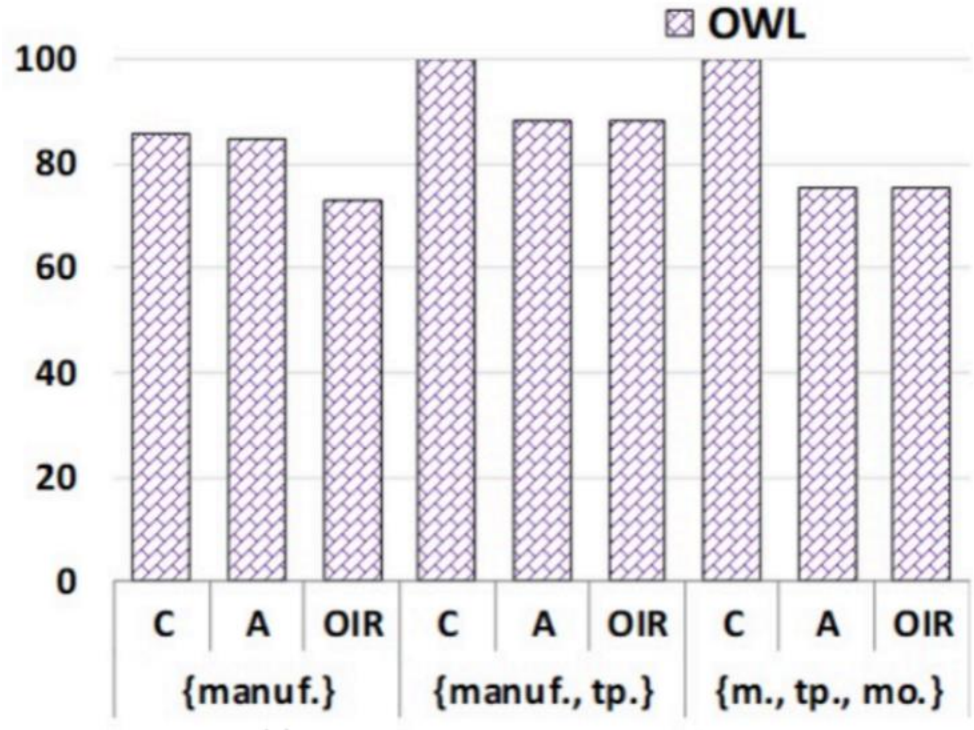
Performance on Annotated Data

- Again, OWL provides the best overall performance (OIR) at all granularity levels.
- Accuracy (A*) and OIR (OIR*) was only evaluated on partial data in {M, T} and {M, T, M} categories.
- A* and OIR* represent the upper-bound of the actual A and OIR.
- OIR* in the range of [0.95, 0.98]

[ARE] Xuan Feng, Qiang Li, Haining Wang, and Limin Sun. Acquisitional rule-based engine for discovering internet-of-things devices. In *USENIX Security*, 2018.

[WDMTI] Lingjing Yu, Tao Liu, Zhaoyu Zhou, Yujia Zhu, Qingyun Liu, and Jianlong Tan. WDMTI: Wireless Device Manufacturer and Type Identification using Hierarchical Dirichlet Process. In *IEEE MASS*, 2018.

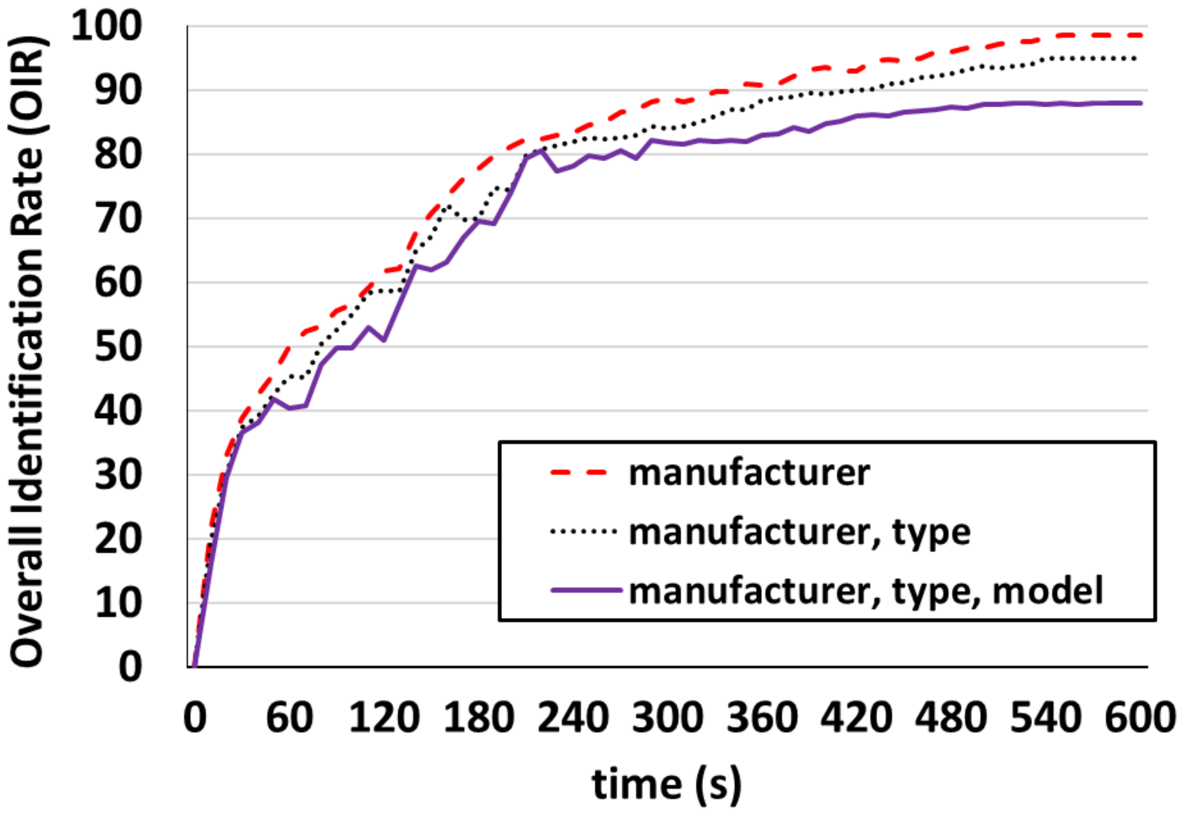
Device Identification: Experiments



Performance on Sanitized Data

- All human-interpretable contents are removed from annotated dataset.
- This is to evaluate OWL's performance in extreme conditions.
- A and OIR represent the lower-bound of the actual A and OIR.
- OWL's OIR is still high, in the range of [0.75, 0.88].

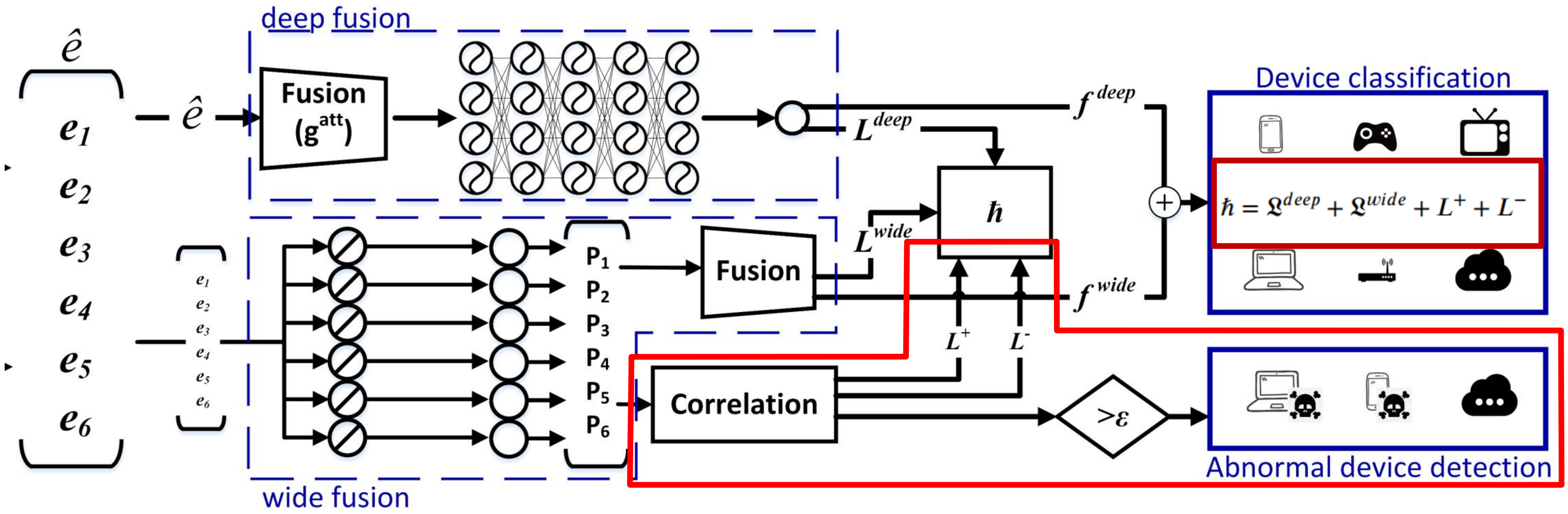
Device Identification: Experiments



Detection Speed

- When all the features are available, It only takes mini-seconds for a trained MvWDL model to classify a new device
- However, packets/features come to OWL slowly in real world settings.
- OWL was connected to the network at t_0, and started to see packets on the network.
- OIR increased rapidly for approximately 240 seconds, when OIR reached 80%.
- OIR peaked in about 500 seconds.

Malicious Device Detection: Approach



- $\mathcal{L}^{deep}$: best classification performance for device labels under input features
- $\mathcal{L}^{wide}$: to memorize the interactions among features, views and labels
- $\mathcal{L}^{+}$: view consistency: to maximize view agreement for benign samples
- $\mathcal{L}^{-}$: to maximize the view inconsistency for known malicious devices

Malicious Device Detection: Case Study

mDNS view

other views

Xiaomi,TV,4	Leshi,TV,x55	Leshi,TV,x65s
Gaoshengda,TV	Funshion,TV	Chuangwei,TV
Hisense,TV,vidaa	PPTV,TV	Changhong,TV,43s1
whaley,TV,w50j	MTN,TV	Changhong,TV,LED50
Rflink,TV	Nebula,TV	Tianmao,Magiccast,m18

Spoofed Apple TV (31 devices)

- **mDNS view: AppleTV, high confidence**
- **Other views: not AppleTV, high confidence**
- Labels: not AppleTV, some in ground truth dataset
- AirPlay: Apple's proprietary protocol suite for multimedia streaming over WiFi
- mDNS packets of these devices were similar to AppleTV, so that others may AirPlay on them
- They were all connected to a corporation named Lebo (or HappyCast)

Malicious Device Detection: Case Study

DHCP view

other views

Fake DHCP Server and Gateway (1 device)

- **DHCP view: router, high confidence**
- **Other views (mDNS, SSDP): Microsoft Surface**
- The device sent DHCP Offer and DHCP ACK messages to inform other devices the gateway of the network is itself.
- MAC prefix: Microsoft
- Explanation: Microsoft surface book spoofed a gateway to lure others to connect through it.
- Some devices were tricked (DHCP request)
- Simulated this attack in the lab

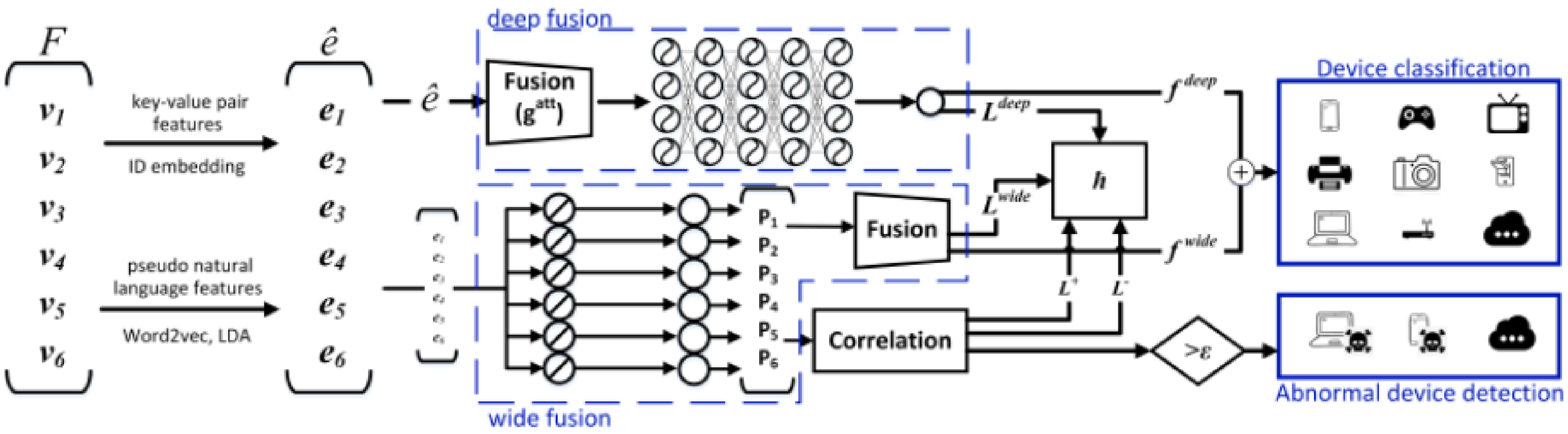
Malicious Device Detection: Case Study

(Hidden) Camera Detection

- Hidden cameras are often considered as sensitive or malicious devices that infringe users’ privacy.
- Existing solutions: traffic analysis
- Cannot detect cameras that are not actively transmitting (in stand-by)
- Attackers and record/store now and send later
- Cameras are still online, they send out BC/MC packets and they can be detected by OWL
- OWL achieved 100% accuracy in detecting cameras at {manufacturer, type} granularity

OWL: Overhearing on WiFi for Device Identification

OWL



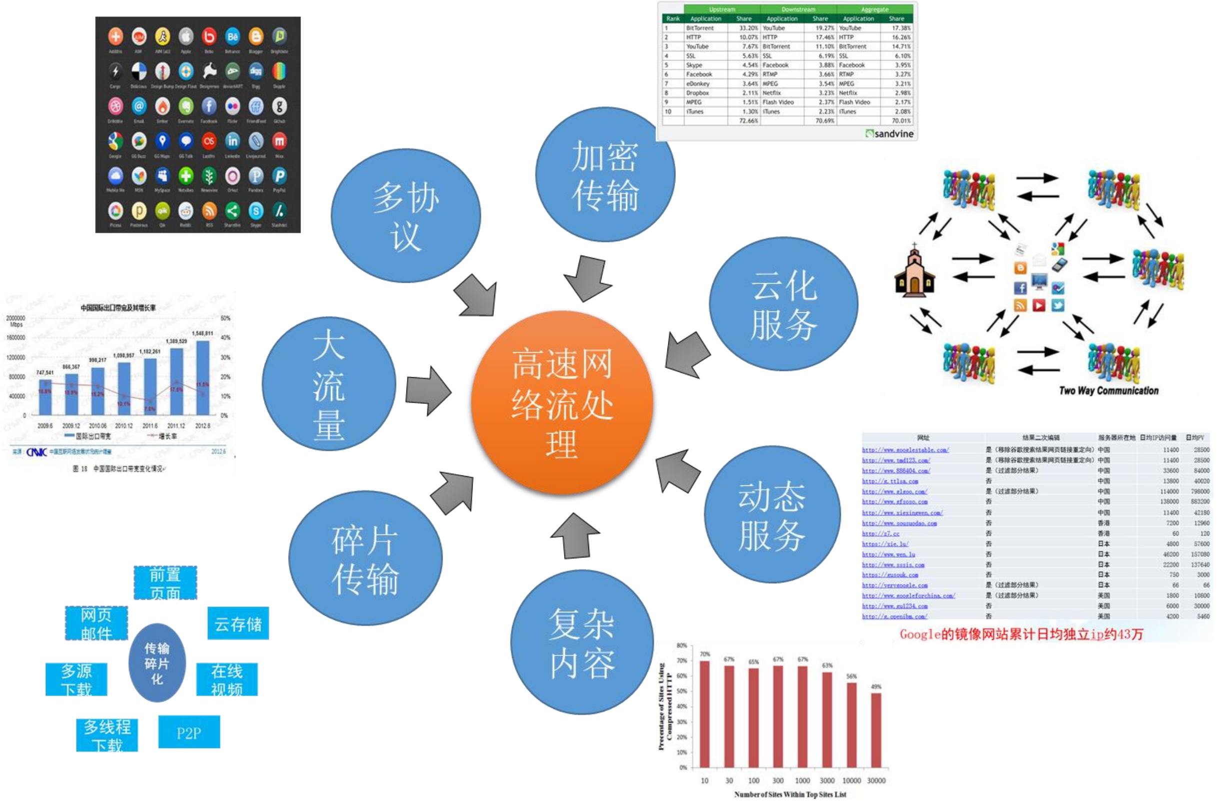
MvWDL

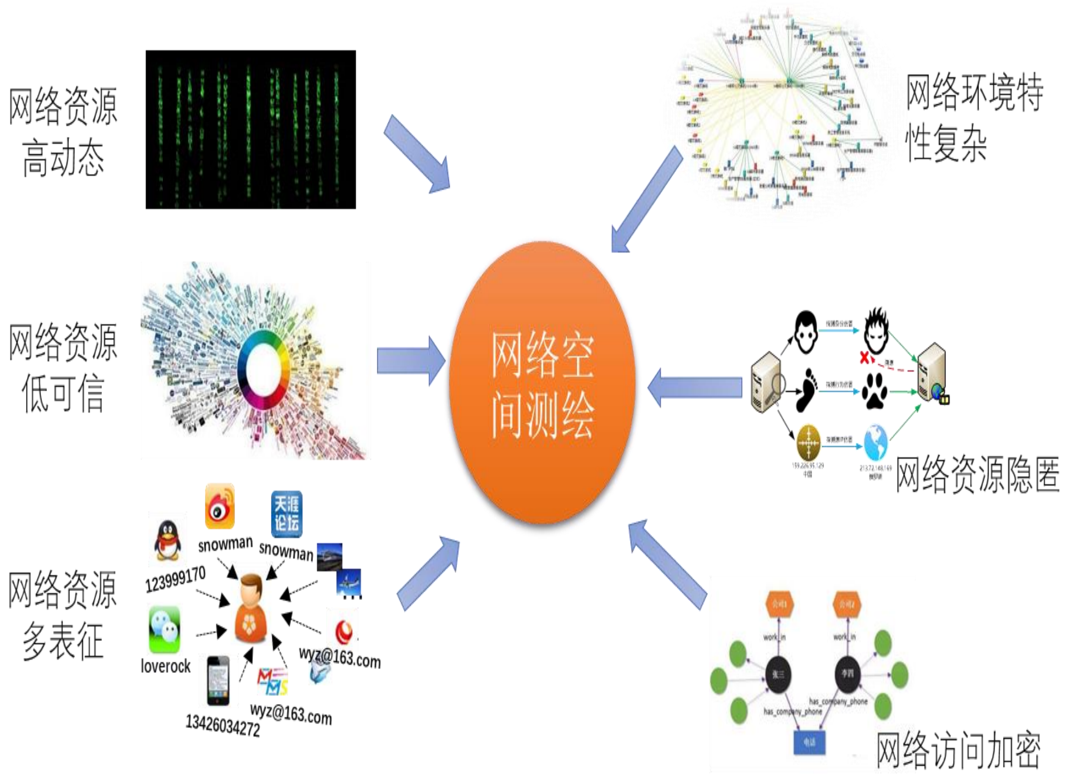
第二研究室：信息智能处理研究室

面向网络空间安全，研究大规模网络智能信息处理的基本理论、模型、算法和关键技术，研制可扩展、高可用、易使用的网络数据处理与分析系统。

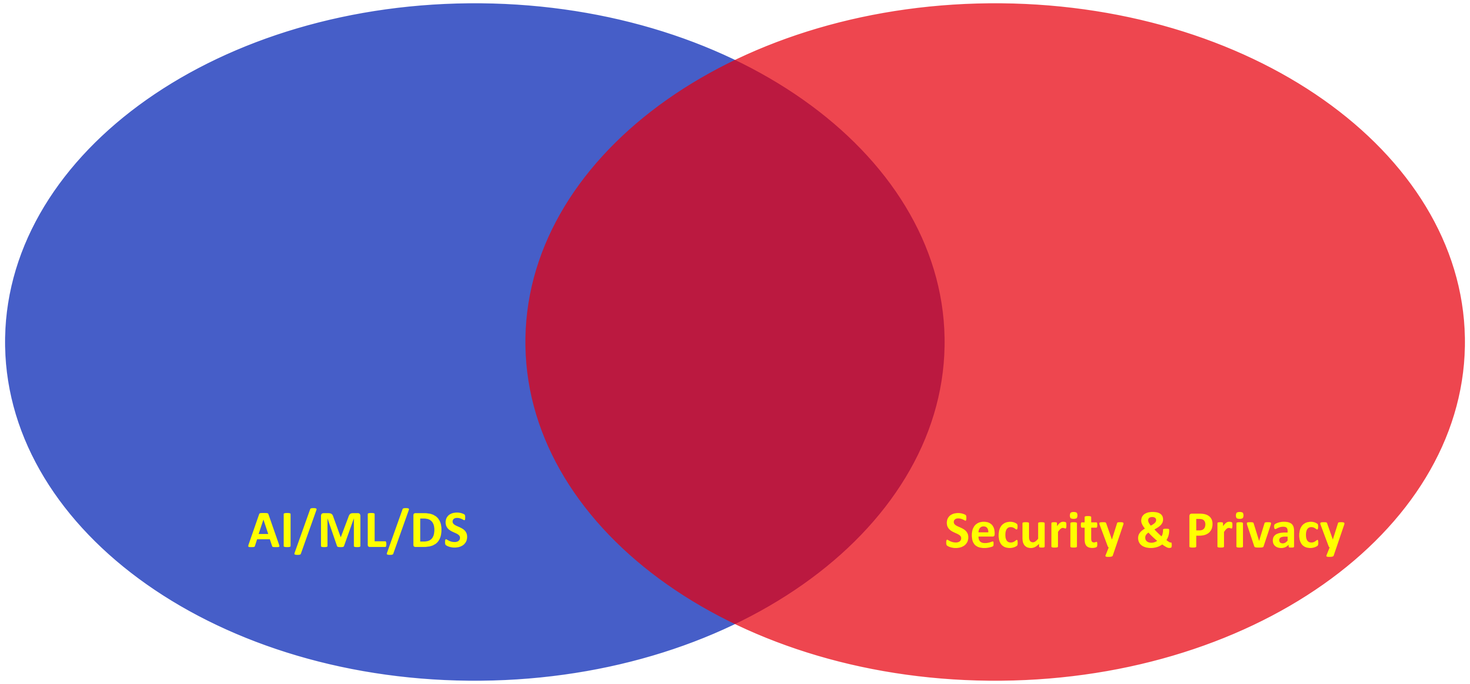
MESA团队：网络流处理与网络空间测绘

针对网络技术迅猛发展给网络空间安全带来的挑战，研究**分布式计算**、**高性能网络数据获取**、**网络空间测绘的**基本理论、架构、模型、算法，融合SDN、BigData，AI等相关技术，研制高性能防火墙设备、网络空间测绘系统，支持网络空间安全治理各类需求。



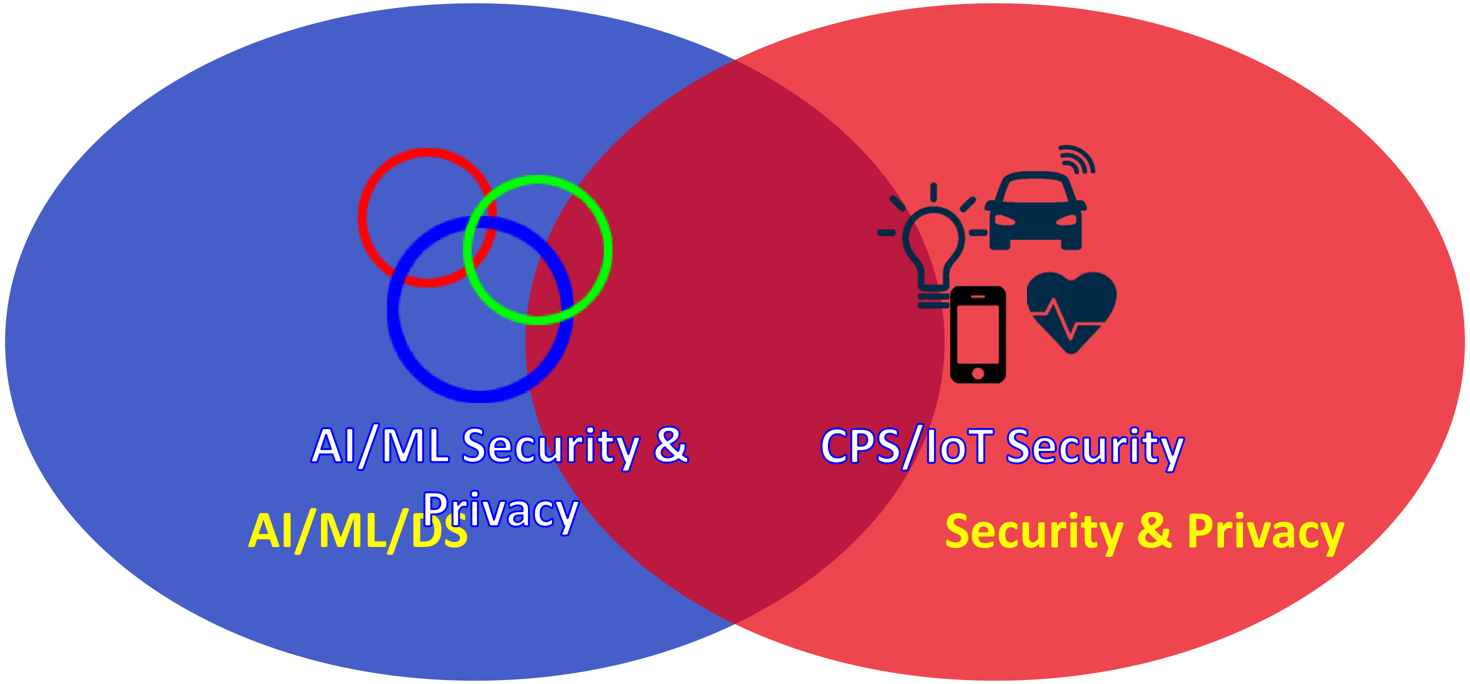


The InfoSec Group at KU



Contact: Bo Luo bluo@ku.edu

The InfoSec and KUPS Groups at KU



Contact: Bo Luo bluo@ku.edu

Thanks for listening！ Q&A

email:yulingjing@iie.ac.cn