# Information Exposure From Consumer IoT Devices: A Multidimensional Network-Informed Approach

Jingjing Ren, Daniel J. Dubois, David Choffnes - *Northeastern University*
**Anna Maria Mandalari**, Roman Kolcun, Hamed Haddadi - *Imperial College London*

# IoT Challenges: Privacy in a World without Walls

20.4 billion IoT devices by 2020 (Gartner Inc.)

- Closed systems and lack of ground truth
  - MITM fails most of the time
- Lack of automation and emulation tools
- Lack of standard testbed and controlled experiments that enable comparisons across IoT deployment sites

# Privacy Concerns

- Personal Information: Stored, Sensor, or Activity data

**What information is exposed?**

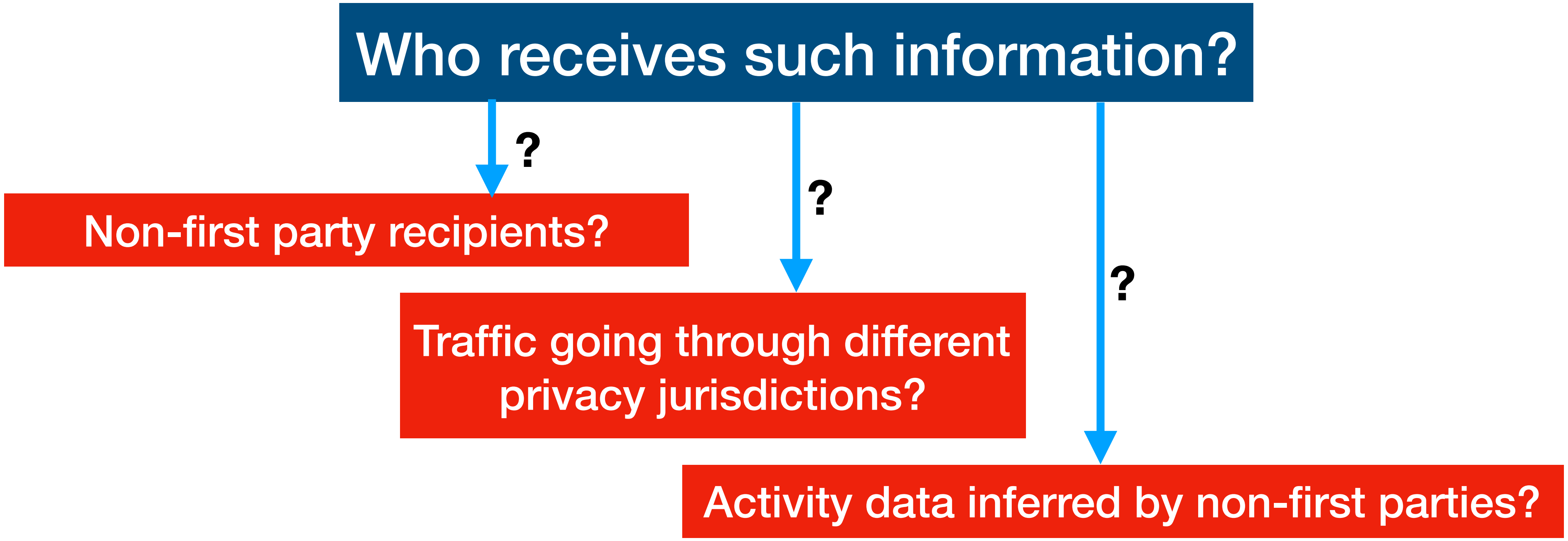
IoT vendor

Cloud/CDN

Other intended destinations

Unintended destinations

- Destination Parties: First, Support, Third, Eavesdroppers



**Who receives such information?**

?

Non-first party recipients?

?

Traffic going through different privacy jurisdictions?

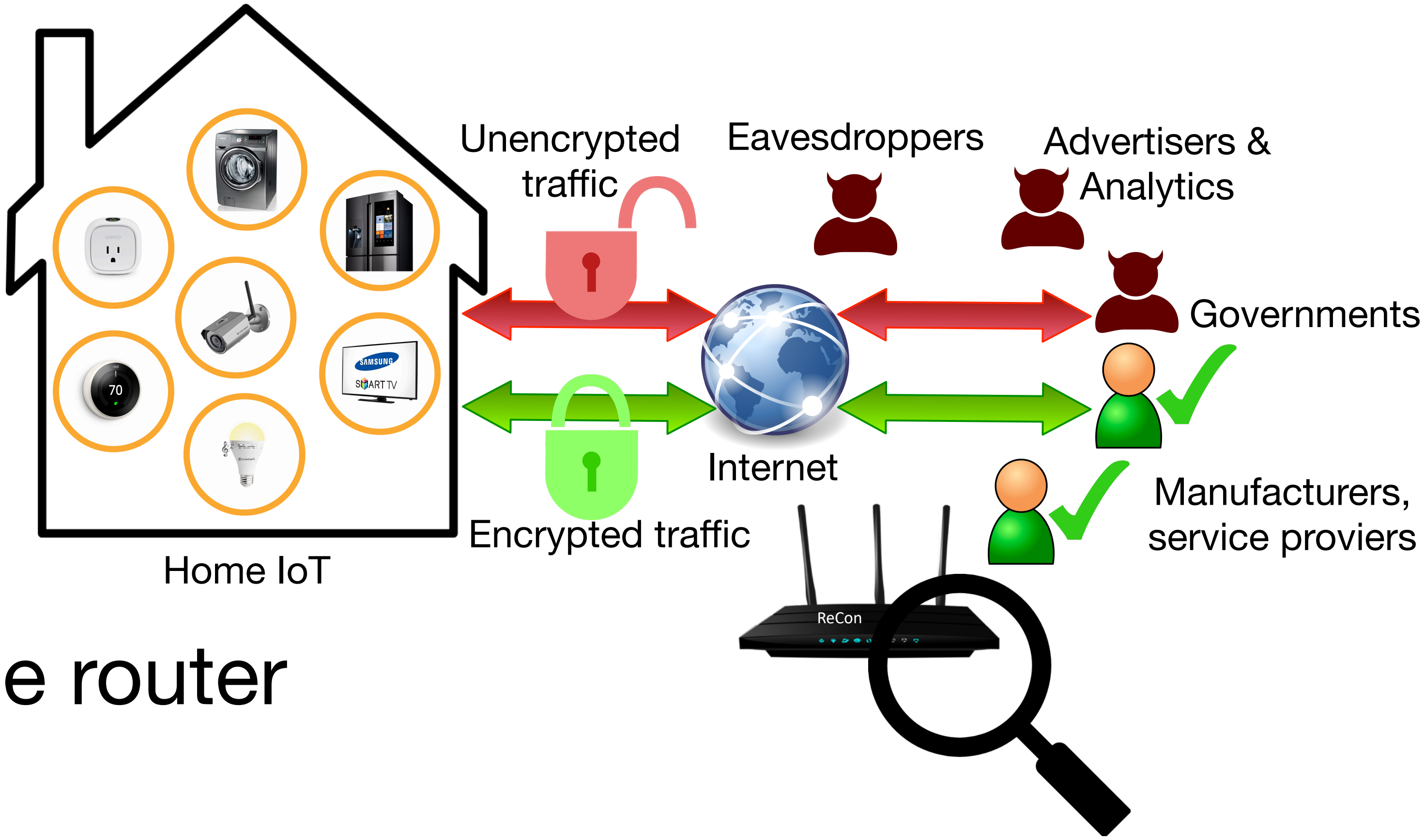
?

Activity data inferred by non-first parties?

# Research Questions

- What is the destination of network traffic?
- To what extent is the traffic encrypted?
- What content is sent?
- Does a device expose information unexpectedly?

# Data Collection Methodology

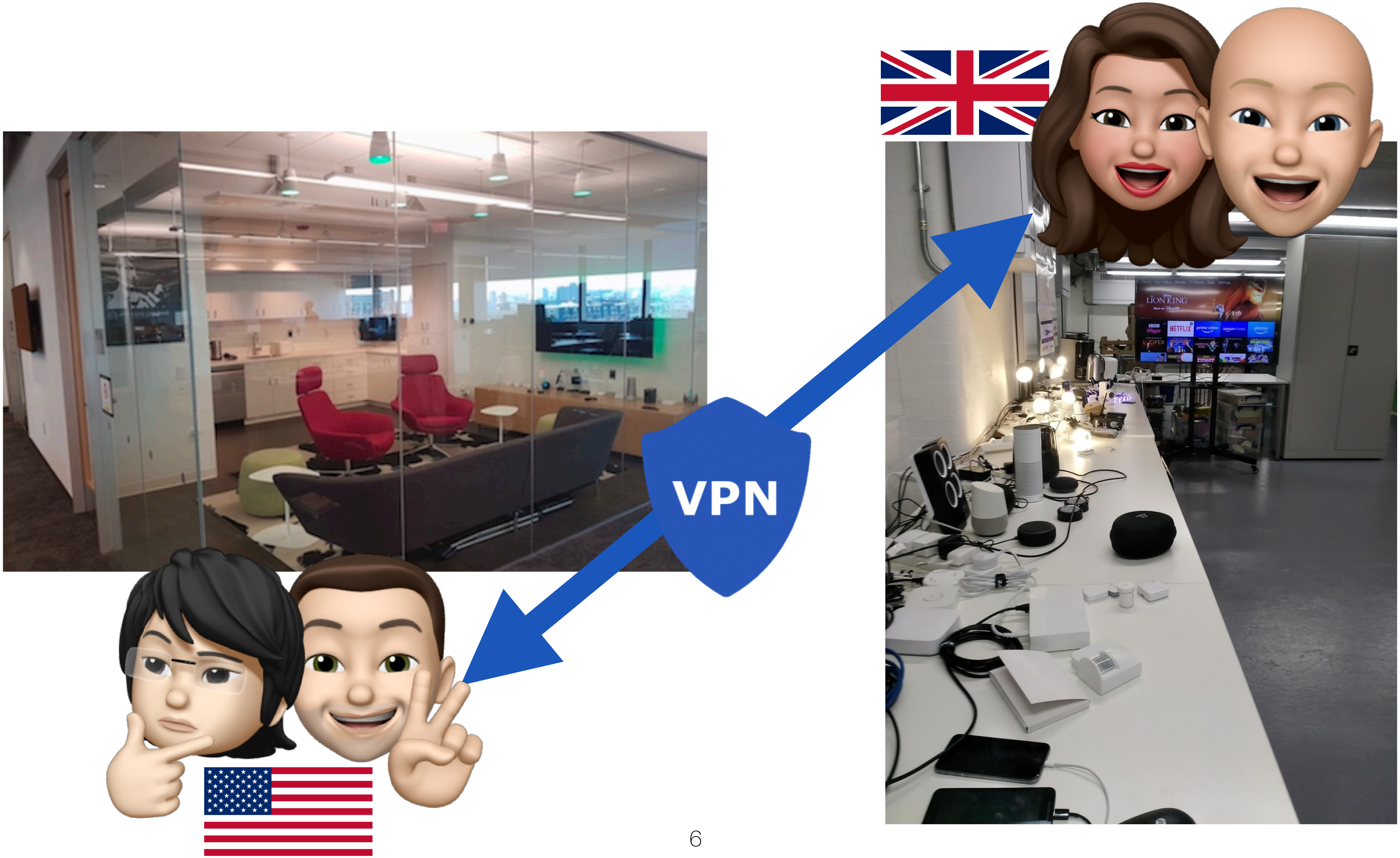


- Monitor all traffic at the router
    - per-device
    - per-experiment
- Labs: US and UK (GDPR)

# Testbeds

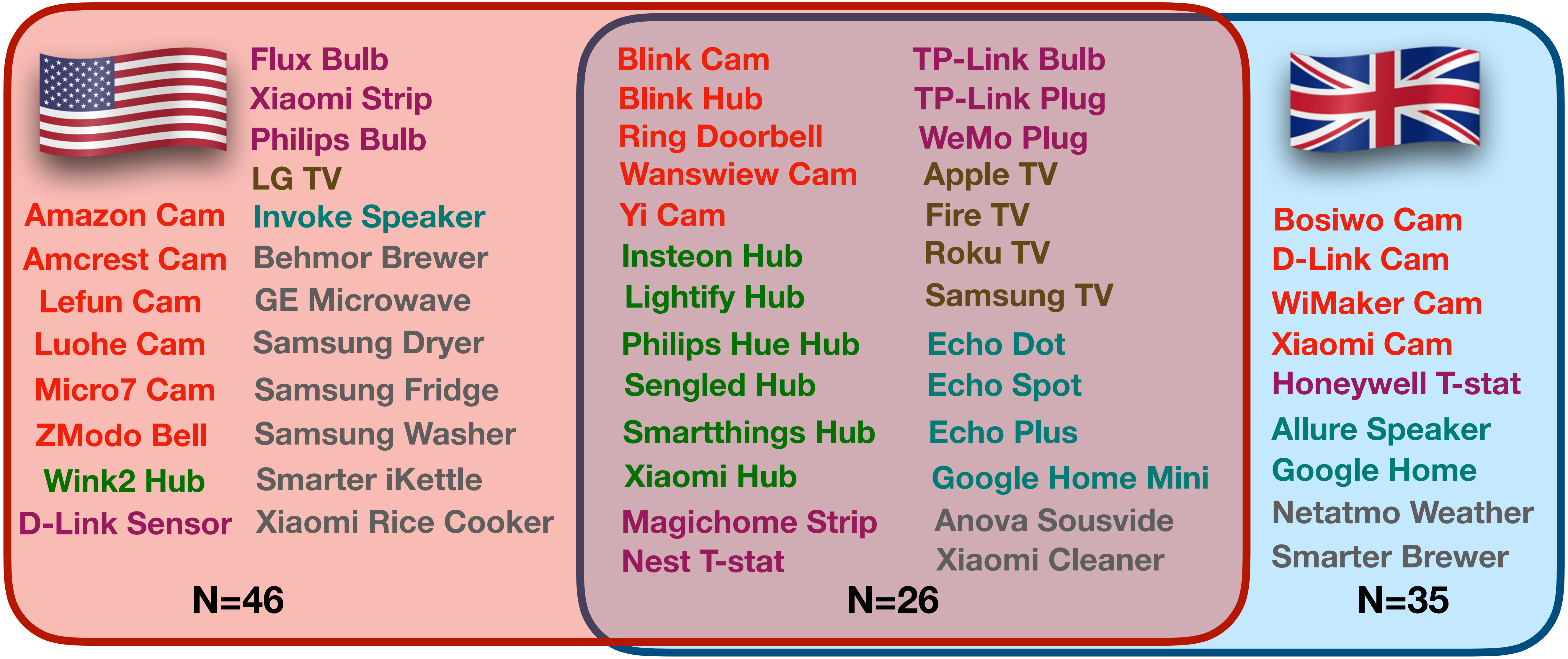
US: Northeastern University

UK: Imperial College London



# Selecting IoT Devices

- **Criteria**: category; features; popularity; US & UK markets



Flux Bulb
Xiaomi Strip
Philips Bulb
LG TV

Amazon Cam
Amcrest Cam
Lefun Cam
Luohe Cam
Micro7 Cam
ZModo Bell
Wink2 Hub
D-Link Sensor

Invoke Speaker
Behmor Brewer
GE Microwave
Samsung Dryer
Samsung Fridge
Samsung Washer
Smarter iKettle
Xiaomi Rice Cooker

N=46

Blink Cam
Blink Hub
Ring Doorbell
Wanswiew Cam
Yi Cam
Insteon Hub
Lightify Hub
Philips Hue Hub
Sengled Hub
Smartthings Hub
Xiaomi Hub
Magichome Strip
Nest T-stat

TP-Link Bulb
TP-Link Plug
WeMo Plug
Apple TV
Fire TV
Roku TV
Samsung TV
Echo Dot
Echo Spot
Echo Plus
Google Home Mini
Anova Sousvide
Xiaomi Cleaner

N=26

Bosiwo Cam
D-Link Cam
WiMaker Cam
Xiaomi Cam
Honeywell T-stat
Allure Speaker
Google Home
Netatmo Weather
Smarter Brewer

N=35

| 20 Cameras | 13 Smart Hubs | 15 Home Automation | 9 TVs | 11 Speakers | 13 Appliances | 81 Total |
|---|---|---|---|---|---|---|

# Design of Experiments

- **Idle: ~112 hours**
- **Controlled interactions**

34,586 experiments (92.6% automated)

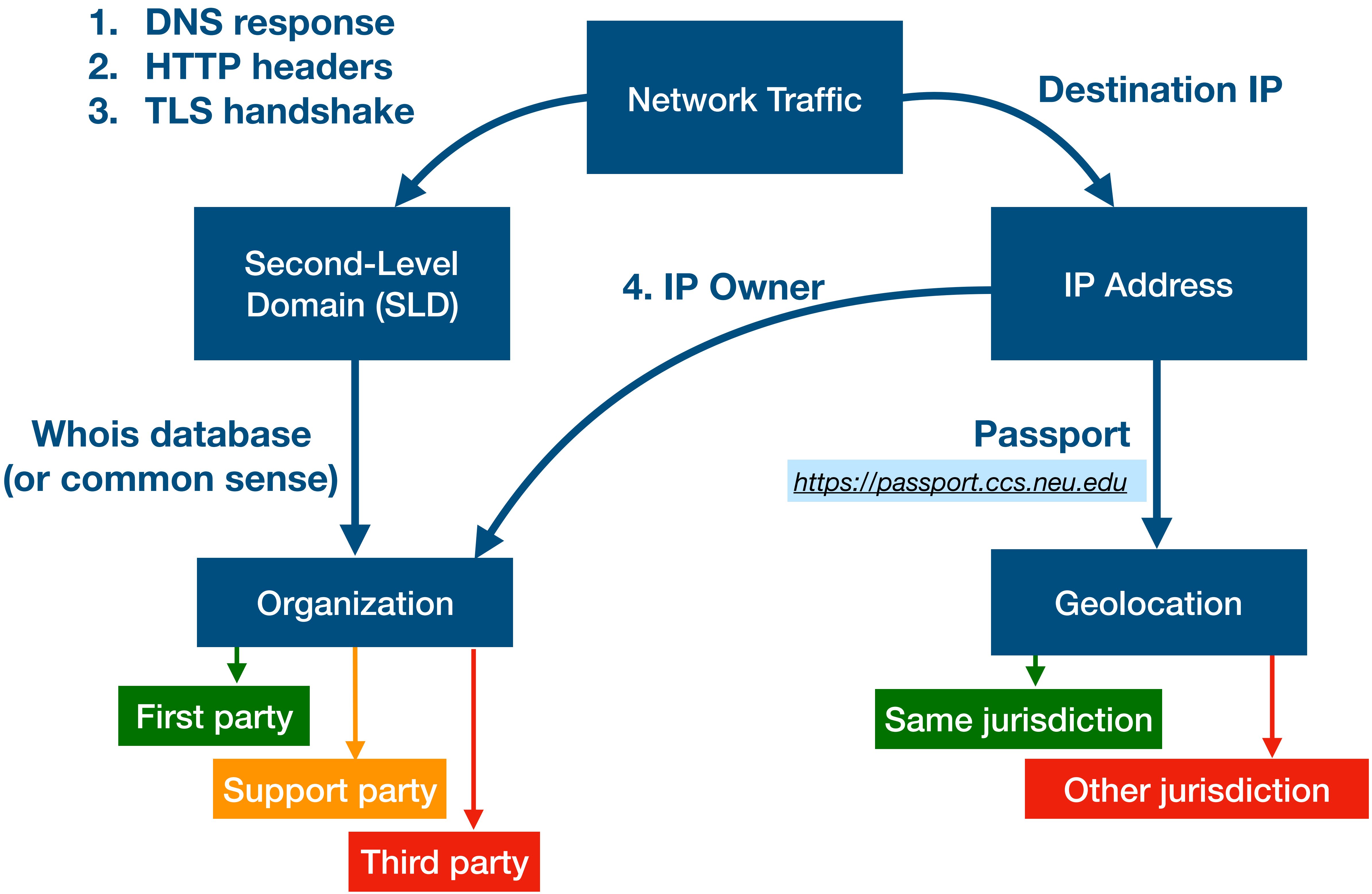
  - Manual (repeated 3 times)
  - Automated (repeated 30 times)
    - Text-to-speech to smart assistants (Alexa/Google/Cortana/Bixby)
    - Monkey instrumented control from Android companion apps
- **Uncontrolled interactions**
  - IRB-approved user study
  - 36 participants, 6 months
    Sep/2018 to Feb/2019

| Activity | Description |
|---|---|
| Power | power on/off the device |
| Voice | voice commands for speakers |
| Video | record/watch video |
| On/Off | turn on/off bulbs/plugs |
| Motion | move in front of device |
| Others | change volume, browse menu |

# Research Questions

- What is the destination of network traffic?
- To what extent is the traffic encrypted?
- What content is sent?
- Does a device expose information unexpectedly?

# What is the Destination?



1. DNS response
2. HTTP headers
3. TLS handshake

Network Traffic

Destination IP

Second-Level Domain (SLD)

4. IP Owner

IP Address

Whois database (or common sense)

Passport

https://passport.ccs.neu.edu

Organization

Geolocation

First party

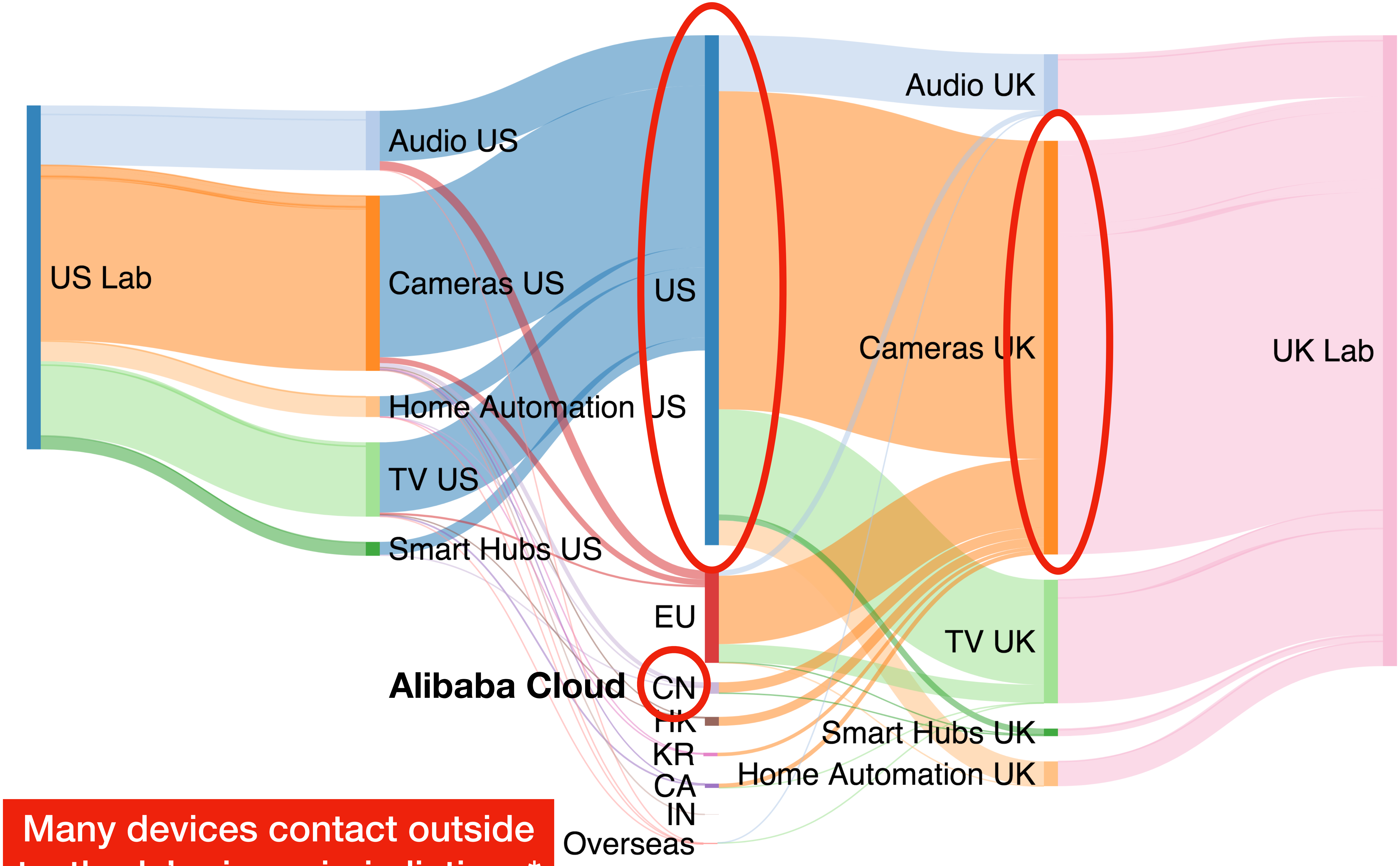
Support party

Third party

Same jurisdiction

Other jurisdiction

# Destination Characterization



Many devices contact outside testbeds' privacy jurisdictions*

*__BBC World News: “__*Who Has My Data? - BBC Click”. https://www.bbc.co.uk/news/technology-48434175*

# Who is Contacted by Many Devices?

High reliance on AWS, followed by Google, Microsoft for hosting

| Organization | US 46 | UK 35 | US Common 24 | UK Common 24 |
|---|---|---|---|---|

Nearly all TVs contact Netflix w/o it being logged in or used

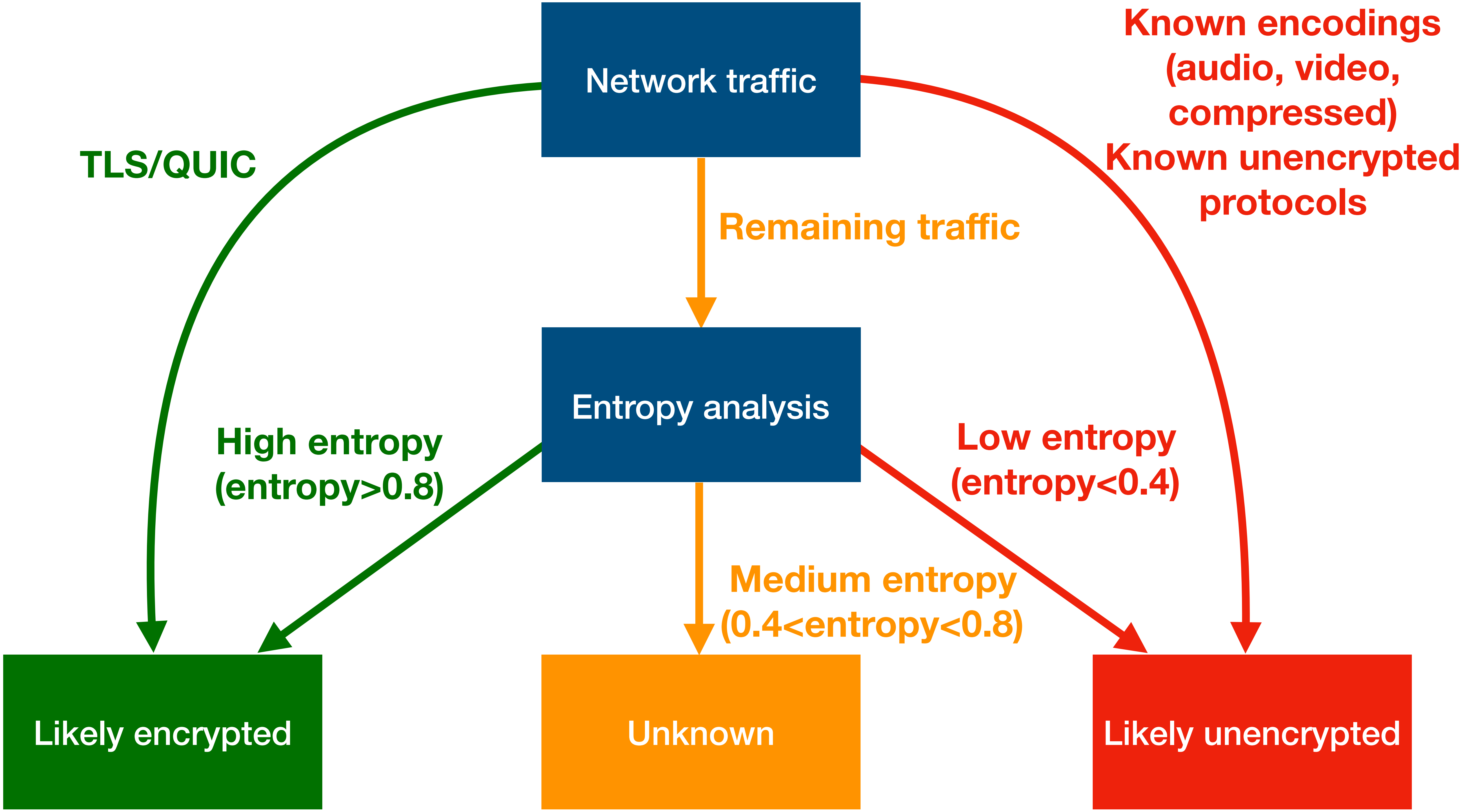
Chinese cloud providers

- Non-first party organizations receive information from many IoT devices
- US devices tends to contact more

# Research Questions

- What is the destination of network traffic?
- To what extent is the traffic encrypted?
- What content is sent?
- Does a device expose information unexpectedly?

# Is the Traffic Encrypted?


Network traffic
TLS/QUIC
Known encodings (audio, video, compressed)
Known unencrypted protocols
Remaining traffic
Entropy analysis
High entropy (entropy>0.8)
Low entropy (entropy<0.4)
Medium entropy (0.4<entropy<0.8)
Likely encrypted
Unknown
Likely unencrypted

Entropy thresholds calculated using min and max over 5311 randomly sampled IoT traffic traces

# How Many Devices Do Encrypt Their Traffic?

|  | Range (%) | US<br>46 | UK<br>35 | US<br>Common<br>24 | UK<br>Common<br>24 |
|---|---|---|---|---|---|
| Unencrypted | >75 |  |  |  |  |
|  | 50-75 |  |  |  |  |
|  | 25-50 |  |  |  |  |
|  | <25 |  |  |  |  |
| Encrypted | >75 |  |  |  |  |
|  | 50-75 |  |  |  |  |
|  | 25-50 |  |  |  |  |
|  | <25 |  |  |  |  |
| Unknown | >75 |  |  |  |  |
|  | 50-75 |  |  |  |  |
|  | 25-50 |  |  |  |  |
|  | <25 |  |  |  |  |

Only 2/81 devices have most traffic unencrypted

26/81 devices have most traffic encrypted

43/81 devices have most traffic unknown

# How Much Traffic is Sent Unencrypted?

| | Device Type | US 46 | UK 35 | US Common 24 | UK Common 24 |
|---|---|---|---|---|---|
| Unencrypted | Appliances | | | | |
| | Speakers | | | | |
| | Hubs | | | | |
| | Automation | | | | |
| | Cameras | | | | |
| | TVs | | | | |
| Encrypted | Appliances | | | | |
| | Speakers | | | | |
| | Hub | | | | |
| | Automation | | | | |
| | Cameras | | | | |
| | TVs | | | | |
| Unknown | Appliances | | | | |
| | Speakers | | | | |
| | Hubs | | | | |
| | Automation | | | | |
| | Cameras | | | | |
| | TVs | | | | |

Cameras and TVs have the most recognizable unencrypted traffic

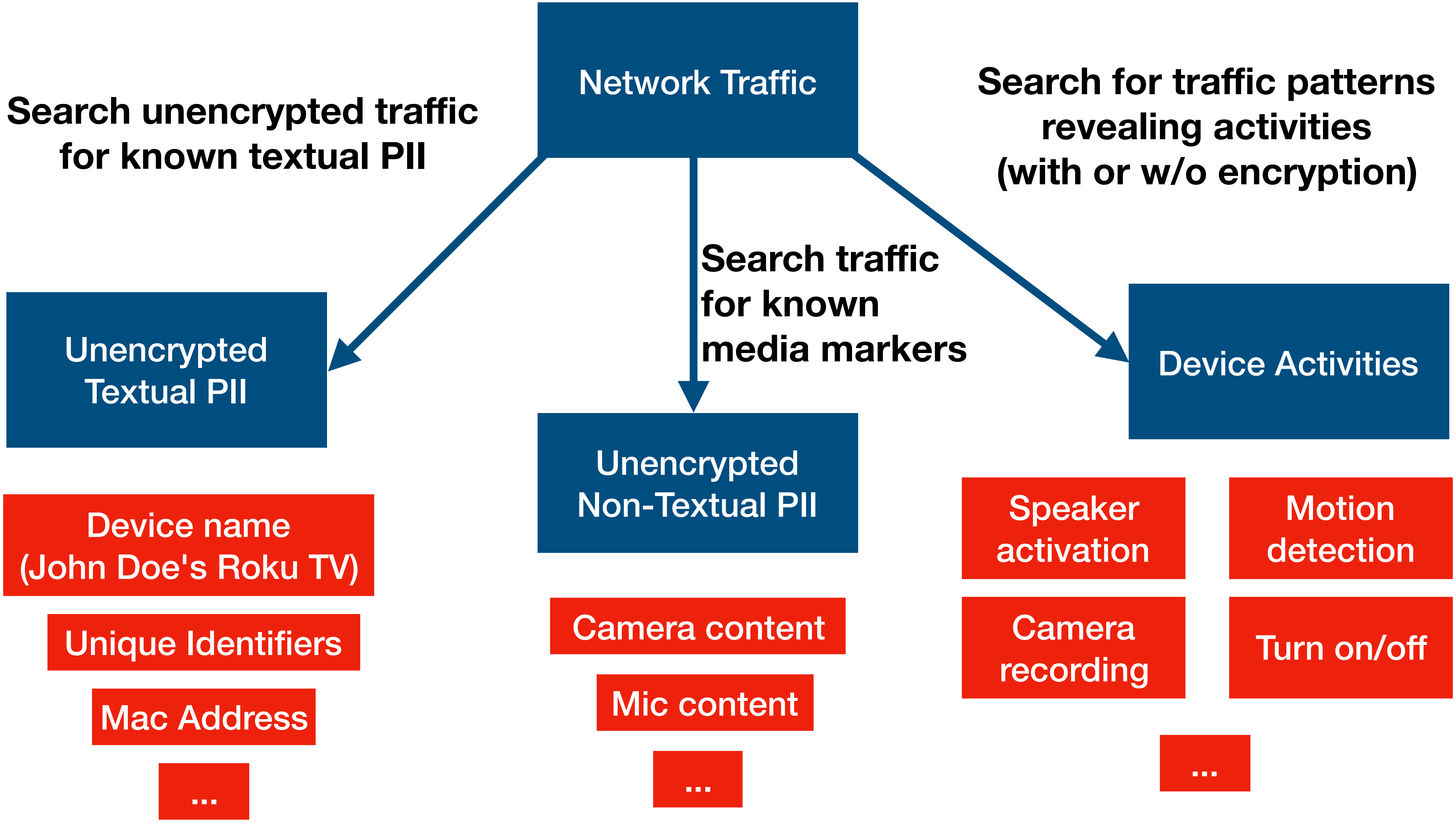
Speakers and TVs have the most recognizable encrypted traffic

26/81 devices have most traffic unrecognizable

# Research Questions

- What is the destination of network traffic?
- To what extent is the traffic encrypted?
- What content is sent?
- Does a device expose information unexpectedly?

# What Information is Sent?



Network Traffic

Search unencrypted traffic for known textual PII

Search traffic for known media markers

Search for traffic patterns revealing activities (with or w/o encryption)

Unencrypted Textual PII

- Device name (John Doe's Roku TV)
- Unique Identifiers
- Mac Address
- ...

Unencrypted Non-Textual PII

- Camera content
- Mic content
- ...

Device Activities

- Speaker activation
- Motion detection
- Camera recording
- Turn on/off
- ...

# Unencrypted Content Leakage

MagicHome LED

Samsung Fridge

Insteon Hub

PII: MAC Address unencrypted!

PII: MAC Address and Timestamped Video unencrypted each time a motion is detected!

Xiaomi Camera

## Other unencrypted content

- Device toggle actions (e.g., on-off)
- Firmware updates
- Metadata pertaining to initial device set up

# Can we Infer an Activity from Network Traffic?

**Hypothesis:**

Eavesdroppers may infer **activity information** even from encrypted traffic

Interaction method
(local, app, or voice?)

Functionality
(e.g., toggling a light)

**Idea:** Given the traffic patterns of an activity, look for similar patterns

**Solution:** use supervised machine learning

**ML APPROACH**

- Random Forest Tree Classifier
- Features (*assuming encrypted*):
  - packet size, inter-arrival times
  - min, max, mean, deciles, …

**ML VALIDATION**

- Cross validation:
  - 7/3 split, averaged across 10 times
- F1 score (val=[0,1]):
  - 0 is the worst, 1 is the best

# Device Activity Inference

- An activity is predictable when F1-score is >0.75

**Number of predictable devices by activity**

| Activity | Total devices | US | UK | US common | UK Common |
|---|---|---|---|---|---|
| Power | | | | | |
| Voice | | | | | |
| Video | | | | | |
| On/Off | | | | | |
| Movement | | | | | |
| Others | | | | | |

Power is the most predictable activity

**Number of predictable devices by category**

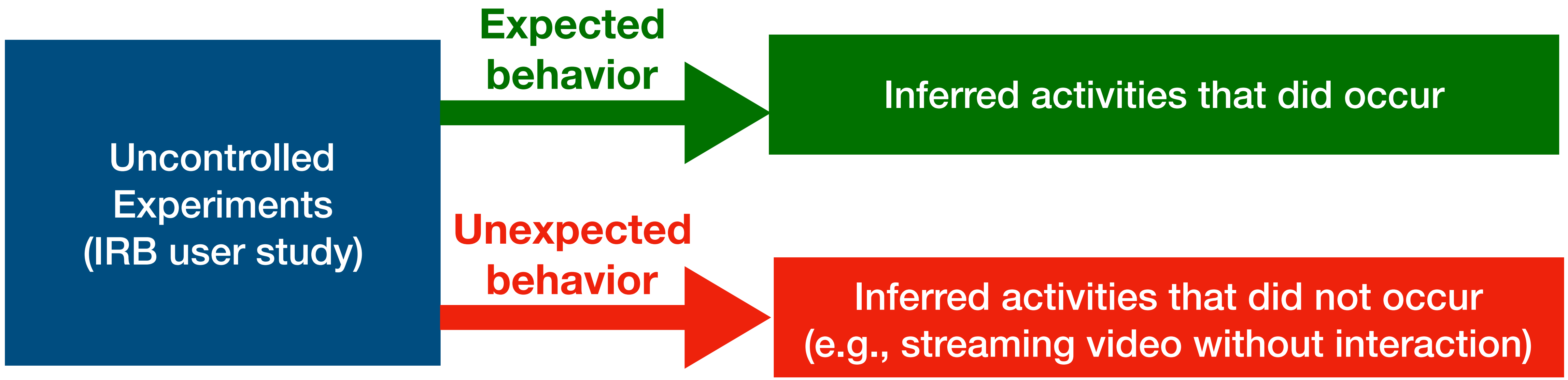
| Category | Total devices | US | UK | US common | UK Common |
|---|---|---|---|---|---|
| Appliances | | | | | |
| Speakers | | | | | |
| Cameras | | | | | |
| Home Automation | | | | | |
| Smart Hubs | | | | | |
| TVs | | | | | |

An activity/device is more predictable when it generates more traffic

# Research Questions

- What is the destination of network traffic?
- To what extent is the traffic encrypted?
- What content is sent?
- Does a device expose information unexpectedly?

# Does a Device Expose Information Unexpectedly?



Uncontrolled Experiments (IRB user study)

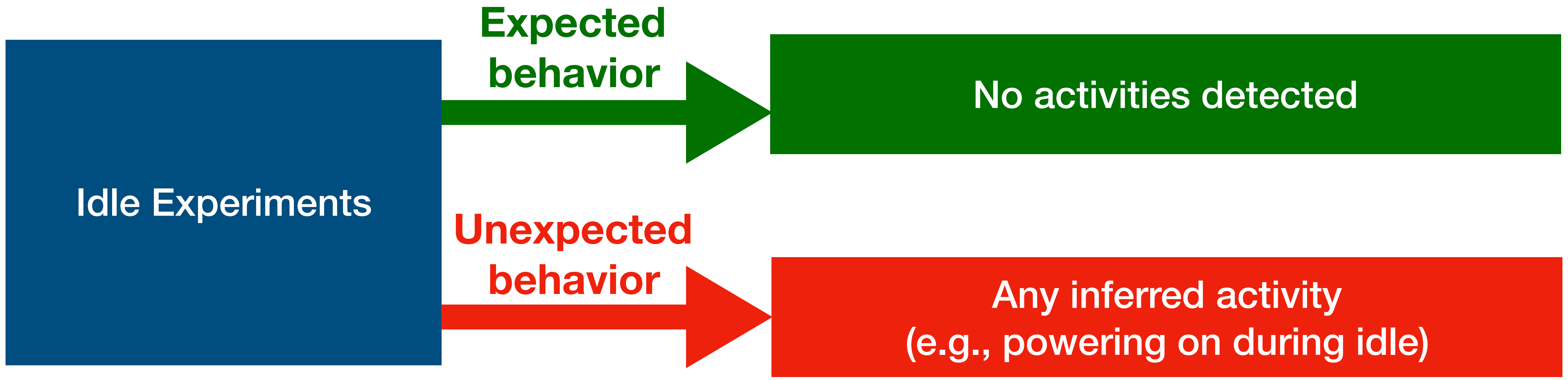
Expected behavior

Inferred activities that did occur

Unexpected behavior

Inferred activities that did not occur (e.g., streaming video without interaction)

**Ground truth provided by camera and user interviews.**
**Activity inference models provided by the controlled experiments.**



Idle Experiments

Expected behavior

No activities detected

Unexpected behavior

Any inferred activity (e.g., powering on during idle)

# Cases of Unexpected Behavior

**Popular doorbells**

Video recording on detected motion (cannot  be disabled)

**Popular smart TVs**

Contact **Netflix, Google**, and **Facebook** unexpectedly

**Alexa-enabled devices**

Frequently falsely triggered (e.g. "**I like S**tar Trek")

***Financial Times***: *"Smart TVs sending private data to Netflix and Facebook".* *https://www.ft.com/content/23ab2f68-d957-11e9-8f9b-77216ebe1f17*

- Other notable cases of activities detected when idle
  - **local move**: cameras triggered "falsely"
  - **power**: devices frequently (dis)connect from WiFi

# Conclusion

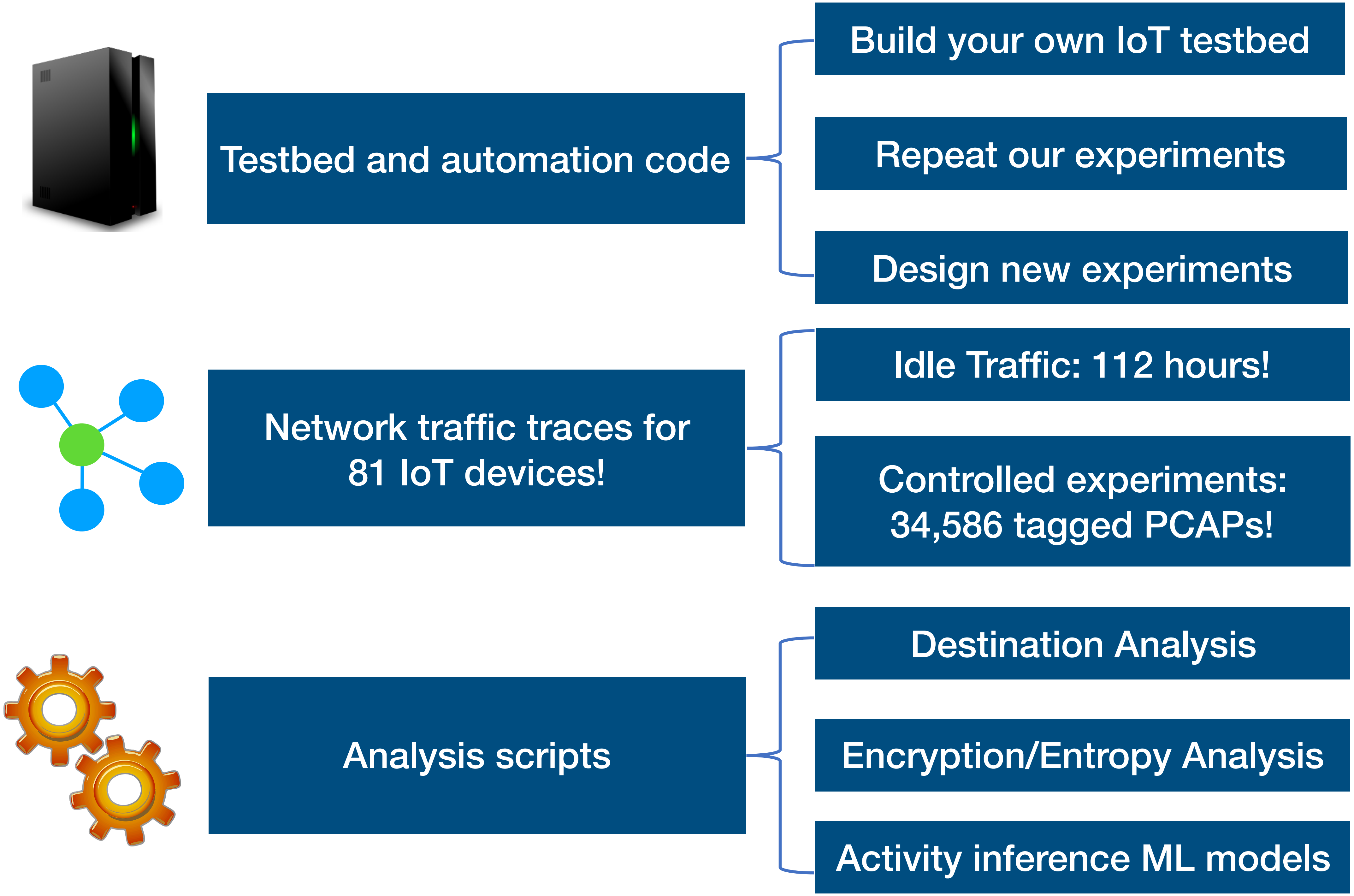
- First step towards more large-scale IoT measurements

- Non-first parties are contacted by many devices
- Some (24/81) devices are vulnerable to activity inference
- Inference models to identify *unexpected* activities

- **Testbed framework, data, and analysis scripts are publicly available at:**

https://moniotrlab.ccis.neu.edu/imc19/

# Fostering Further IoT Privacy Research



Testbed and automation code

- Build your own IoT testbed
- Repeat our experiments
- Design new experiments

Network traffic traces for 81 IoT devices!

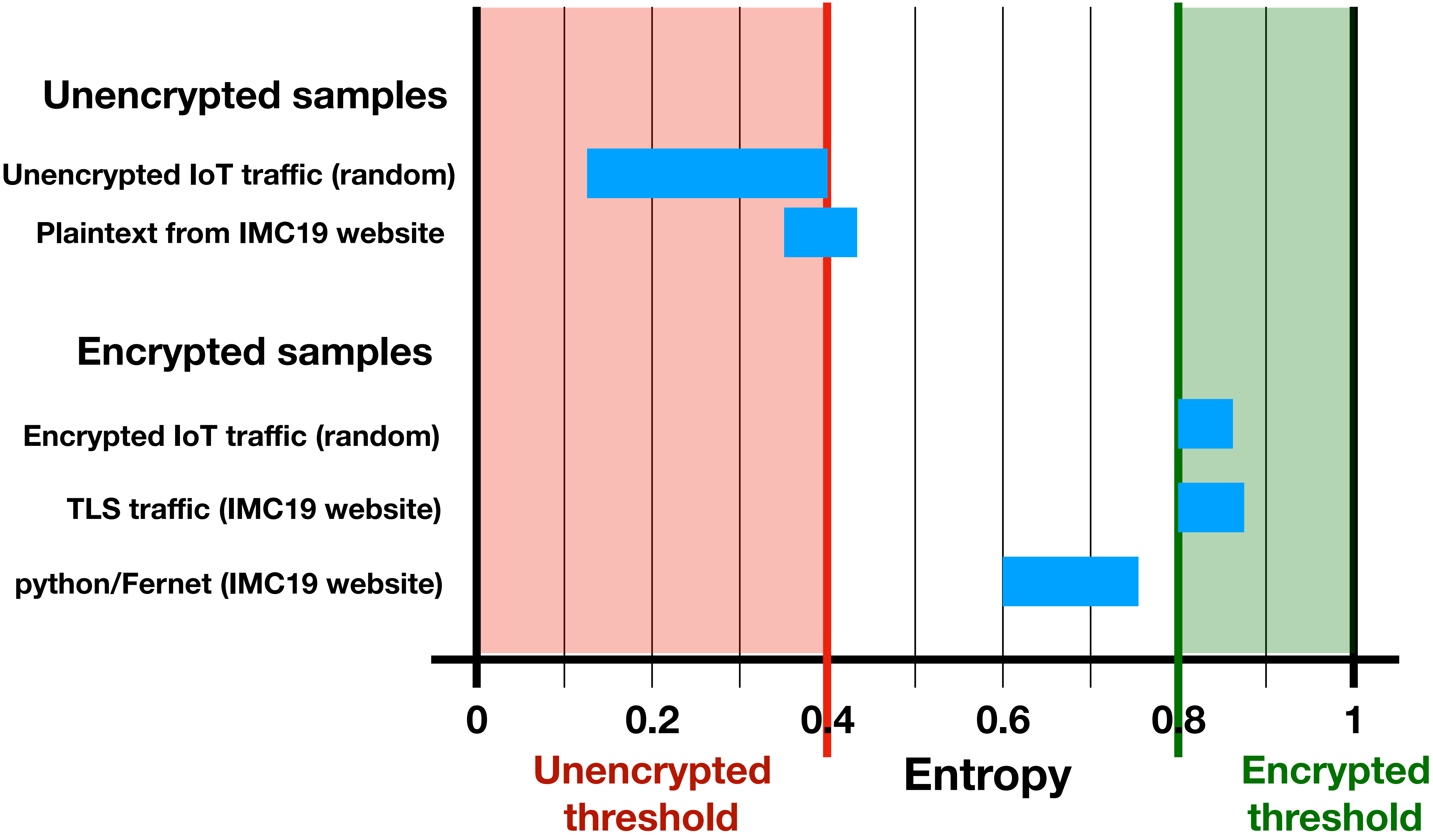
- Idle Traffic: 112 hours!
- Controlled experiments: 34,586 tagged PCAPs!

Analysis scripts

- Destination Analysis
- Encryption/Entropy Analysis
- Activity inference ML models

https://moniotrlab.ccis.neu.edu/imc19/

# Entropy Threshold


Unencrypted samples
Unencrypted IoT traffic (random)
Plaintext from IMC19 website
Encrypted samples
Encrypted IoT traffic (random)
TLS traffic (IMC19 website)
python/Fernet (IMC19 website)
Unencrypted threshold
Entropy
Encrypted threshold