Website Fingerprinting

HTTP Workshop
Amsterdam
Tommy Pauly & Chris Wood

Privacy Considerations

Networking privacy always has (at least) two sides:

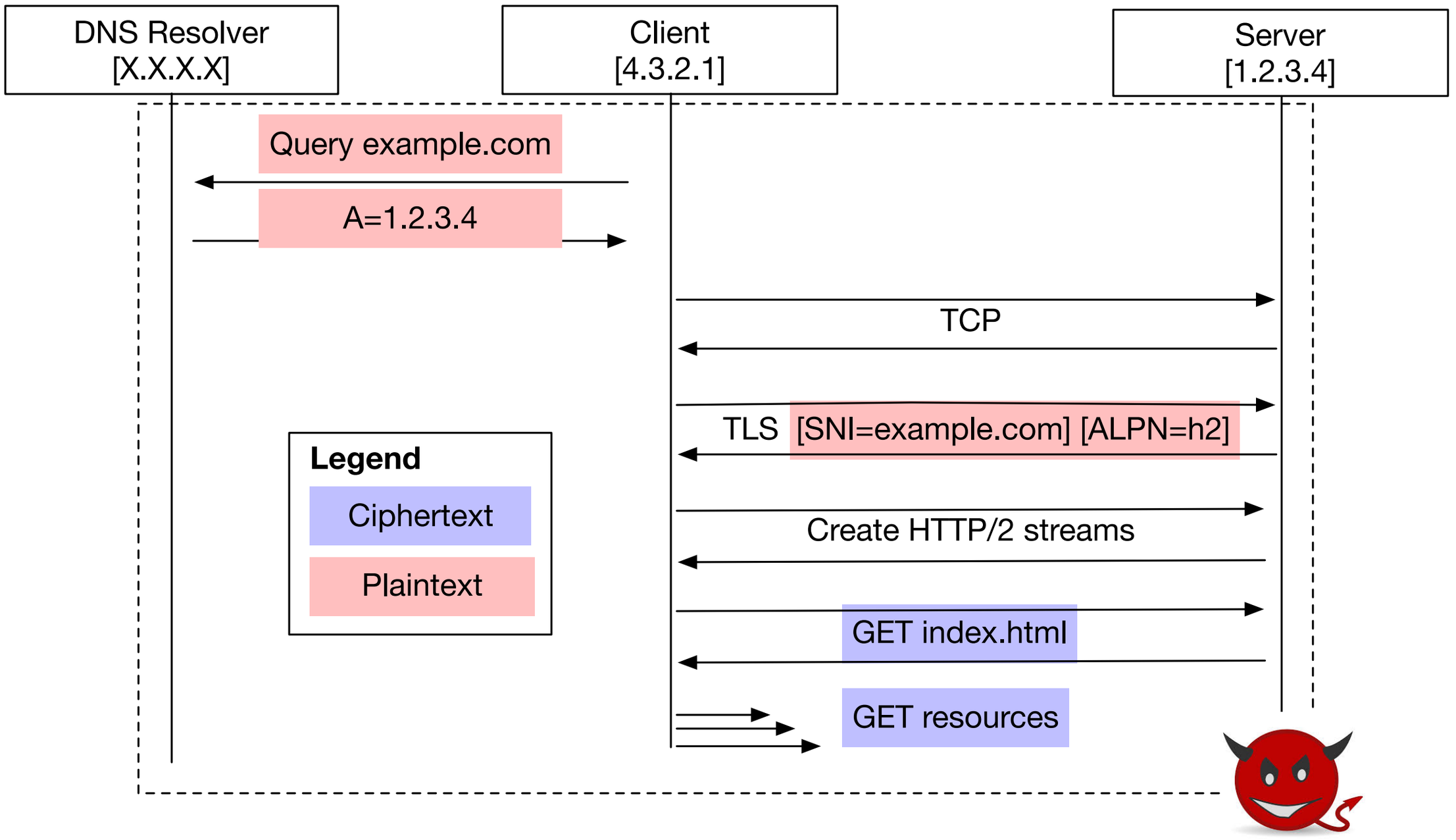
- Who is accessing the resource?
    - Browser fingerprinting
    - Client IP address
- What resource are they accessing?
    - Contents of connections
    - Page load analysis

Connection Fingerprinting

TLS (HTTPS) *ESNI* *DoT/DoH*

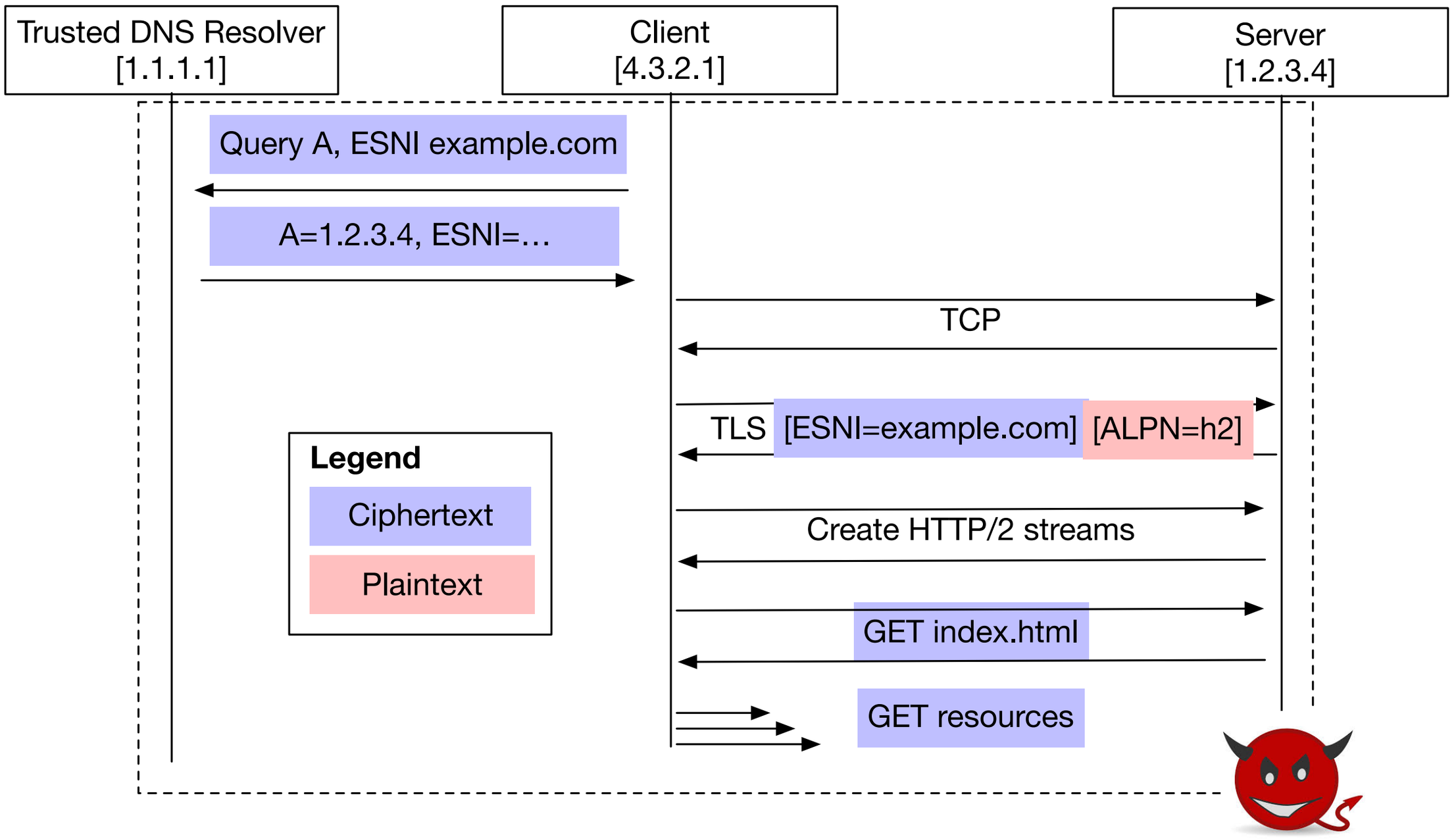
Connection Fingerprinting

Current State



Connection Fingerprinting

DoT + ESNI



Connection Fingerprinting

SNIs are being used for censorship

> https://www.bleepingcomputer.com/news/security/south-korea-is-censoring-the-internet-by-snooping-on-sni-traffic

“Deep fingerprinting” can be 99% effective at identifying websites, even in an **open world** scenario

> https://arxiv.org/pdf/1801.02265.pdf

Approaches like TOR using padding techniques such as WTF-PAD and Walkie-Talkie

Page Load Fingerprinting

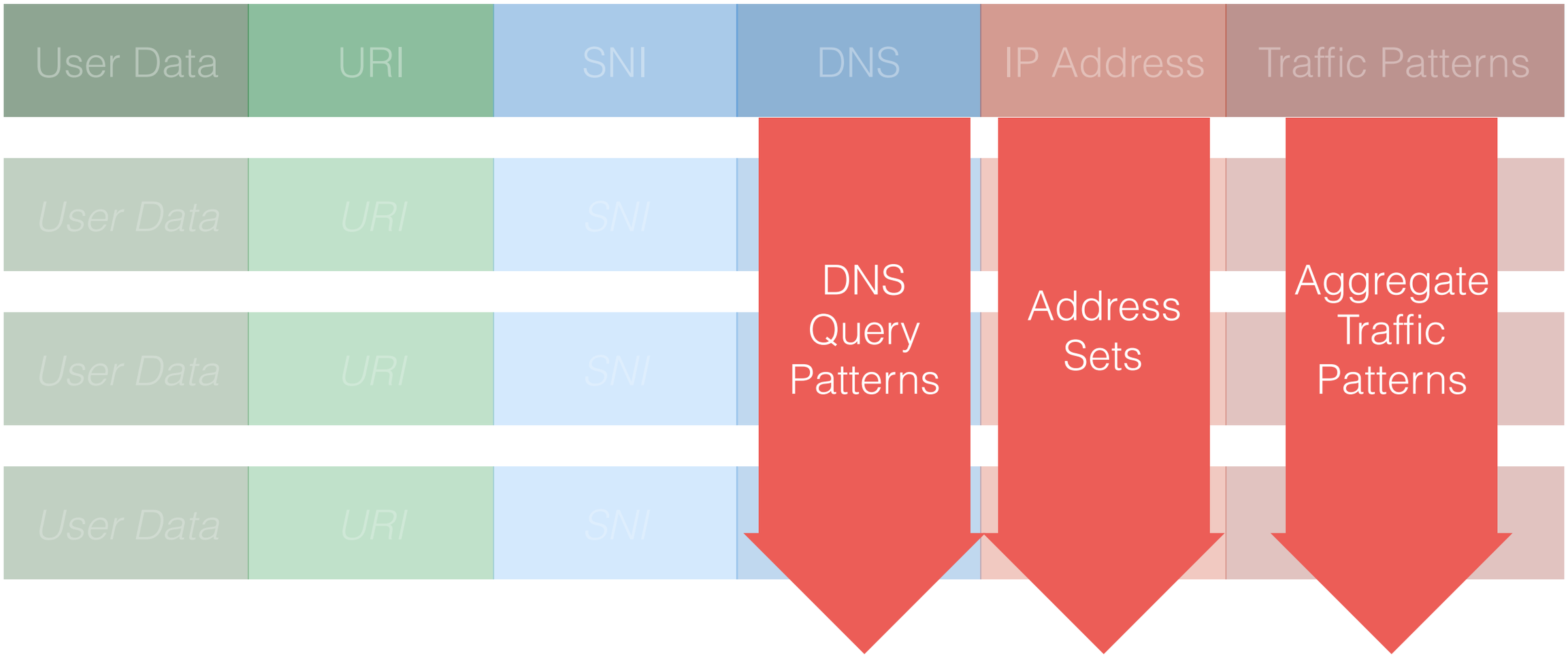
Page load fingerprints (PLFs) contain the set of connections and their traffic associated with a page load event

- DNS query patterns
- TCP/TLS connection patterns

Even with all elements encrypted, the patterns are often uniquely identifying

Example: Loading https://nytimes.com in Safari

Page Load Fingerprinting



Page Load Fingerprinting

Possible Mitigations

HTTP/2 connection coalescing can decrease connection information from the PLF

CDN consolidation can make a PLF simpler and less unique

DNS-based load balancing may redirect clients to different servers, or even different providers

Happy Eyeballs and connection racing may make results less predictable, but may also expand fingerprint

Connection encryption

Connection privacy

Page load privacy

Discussion

Which of these attacks should we address?

If we do DoT/DoH and ESNI, it follows that we'll want to address some of these issues next

Which mitigations are worth the effort and trade-offs?

How can we research PLF attacks more?