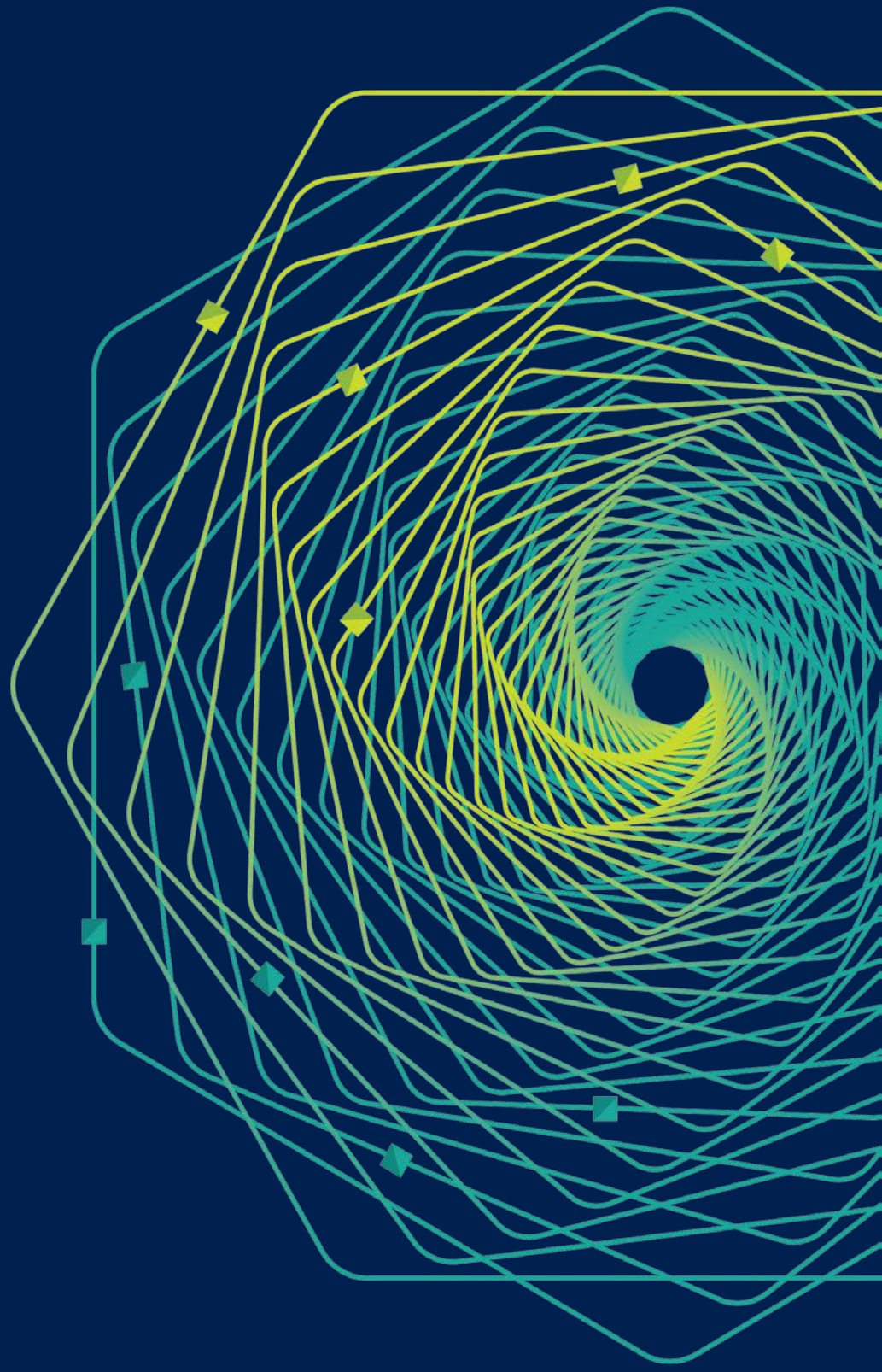
Microsoft

Research Faculty Summit 2018

Systems | Fueling future disruptions



The Good, the Bad, and the Ugly of ML for Networked Systems

bruce MacdowelL maggs

Data-driven design of networked systems

Idea: marry disjoint data sets gathered from existing systems to drive the design and simulation of new systems that have not yet been deployed

Example: use logs from Akamai's commercial real-time (live) fixed-infrastructure-based video delivery system to drive simulations of purely peer-to-peer "End System Multicast"

TLS Certificate revocation checking is broken

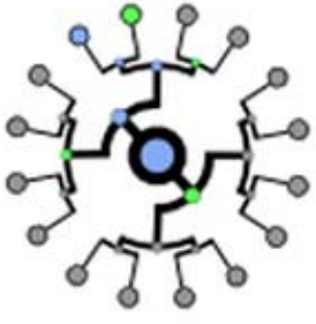
Mobile devices don’t check

- Too much bandwidth to download certificate revocation lists

- Too much latency to check on the fly with OCSP

- Too little deployment of OCSP stapling

Opportunity:
new databases now include all valid certificates!



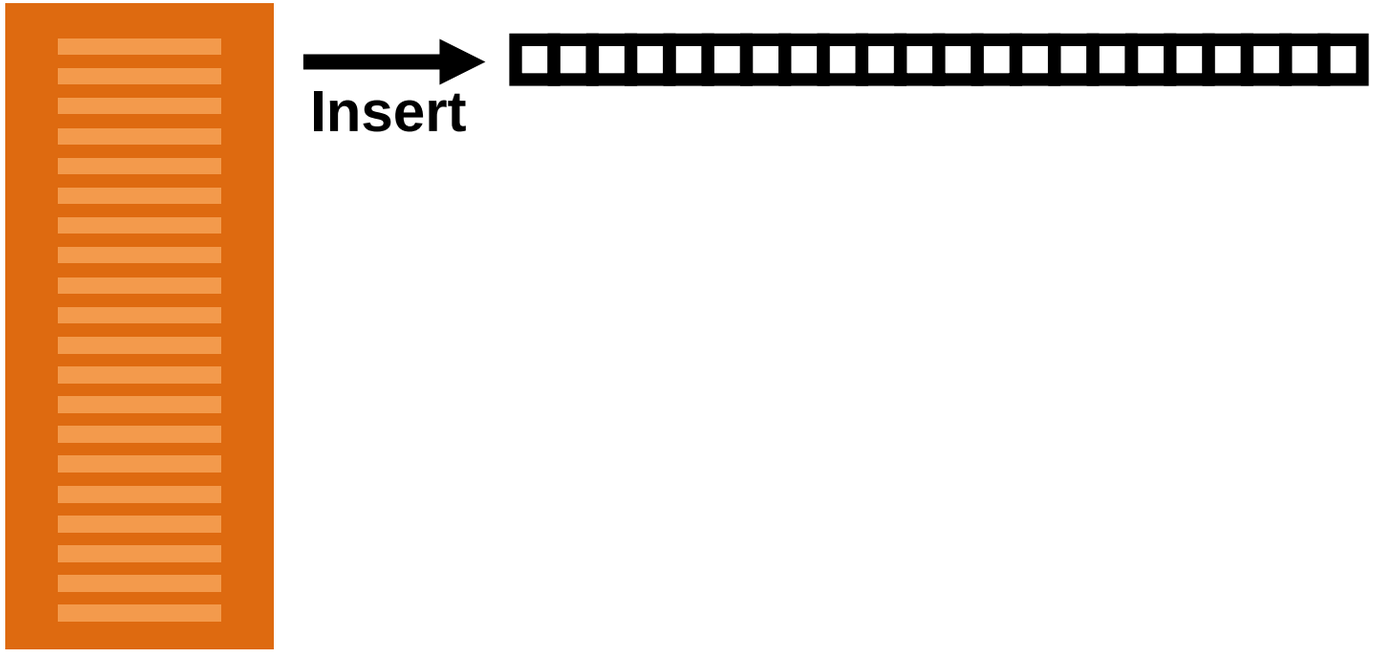
Certificate Transparency

RAPID7

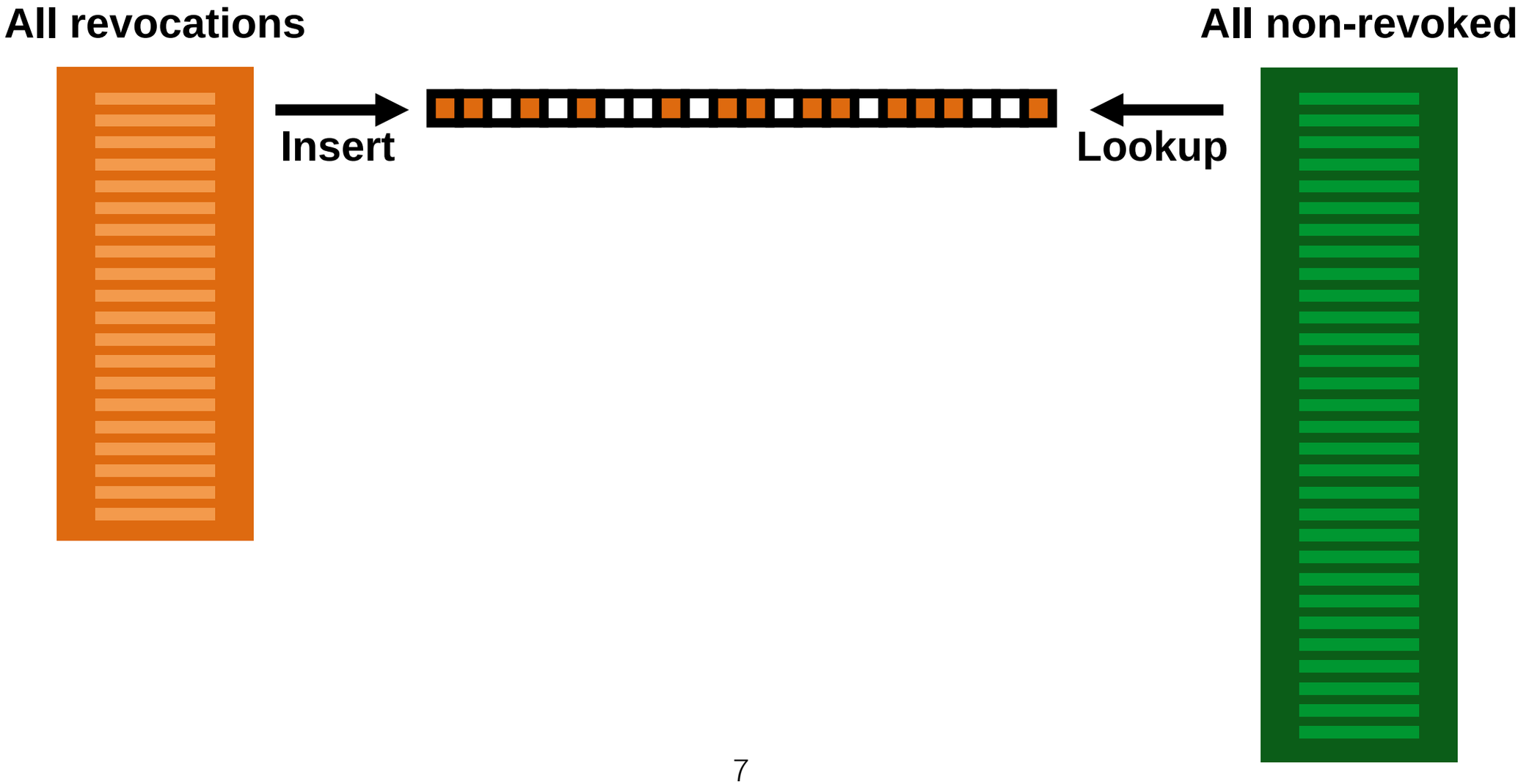
Towards a complete view of the certificate ecosystem - Vandersloot et al. **2016**

Filter Cascade

All revocations

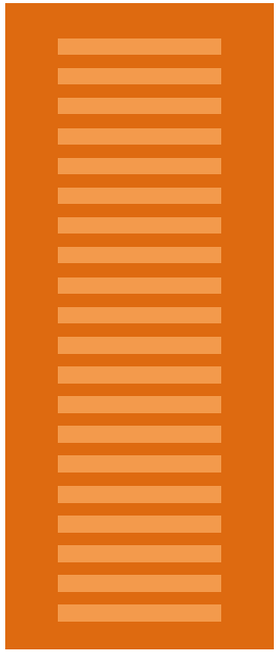


Filter Cascade



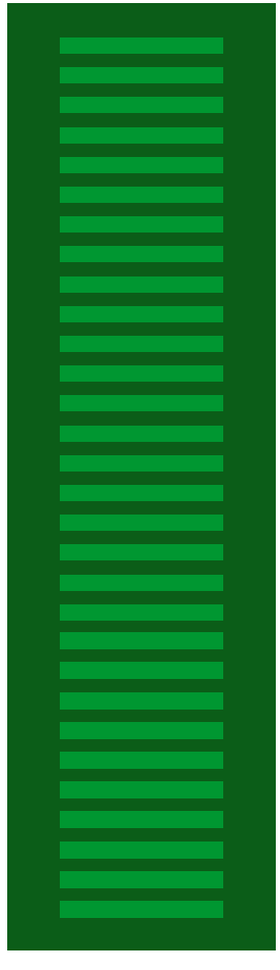
Filter Cascade

All revocations

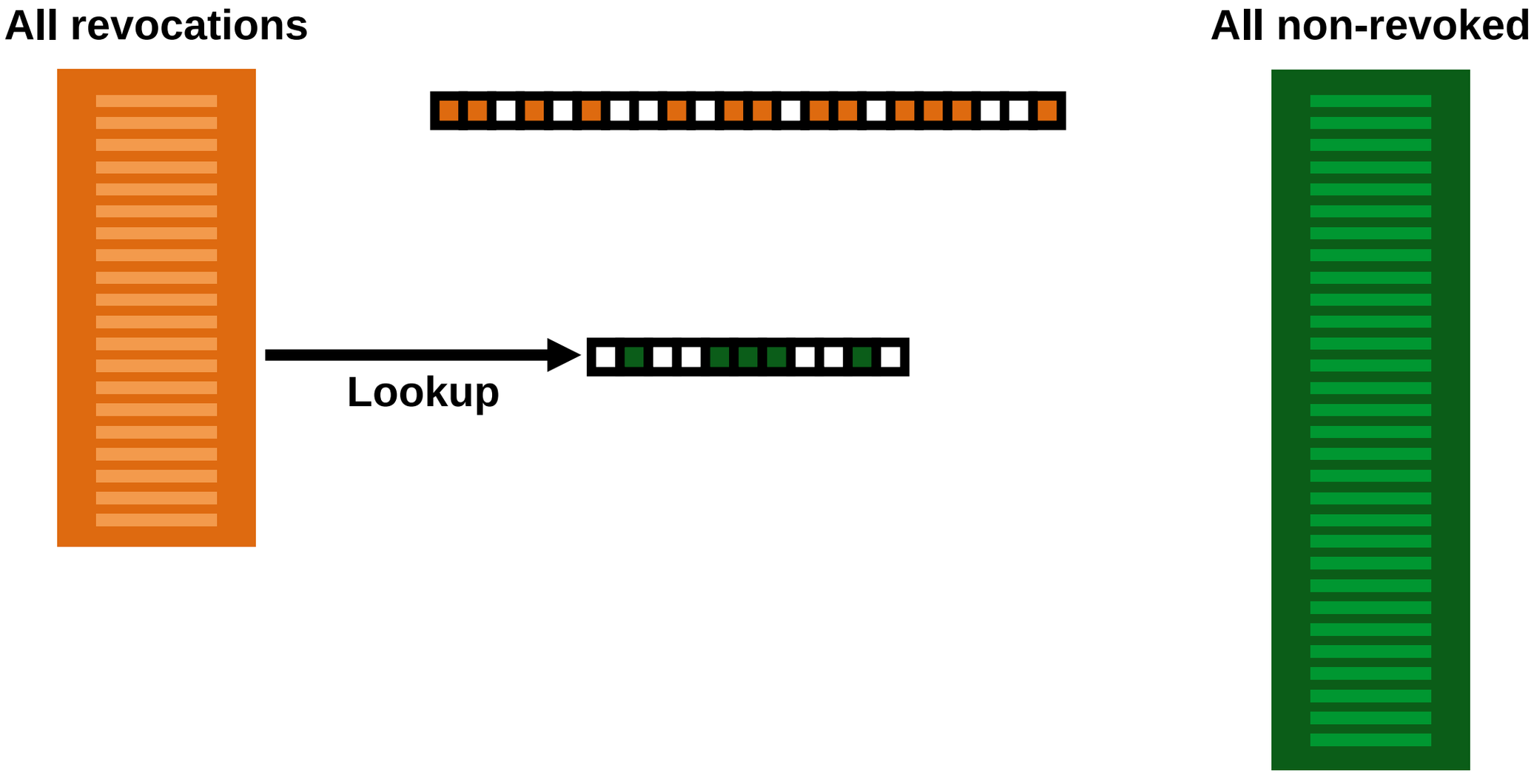


All false positives

All non-revoked

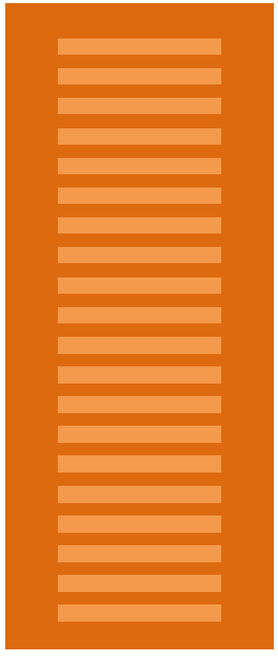


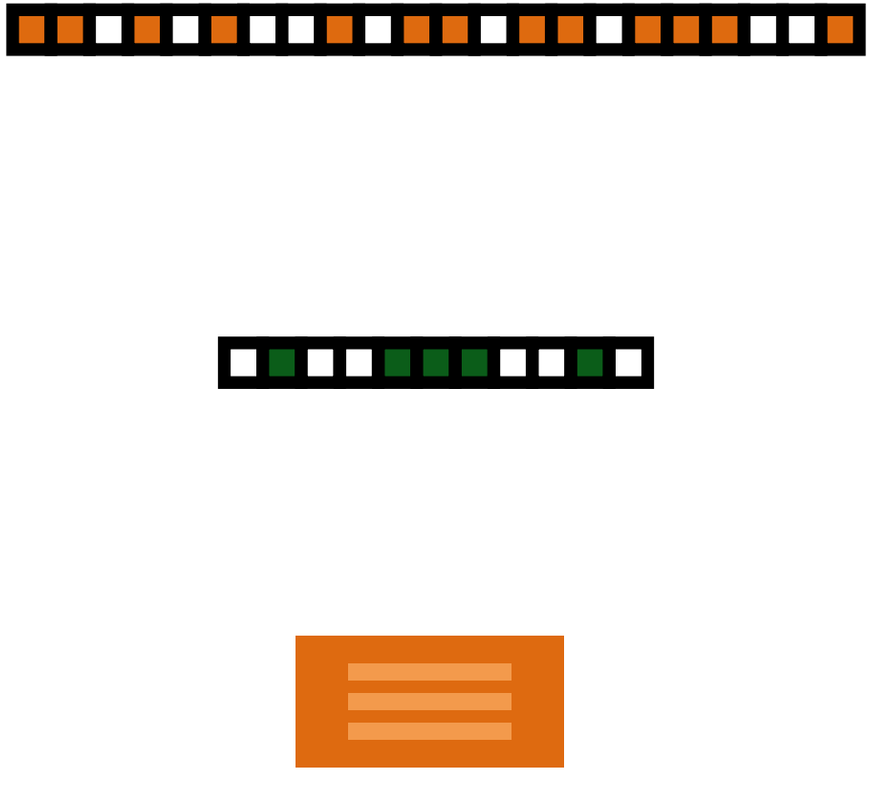
Filter Cascade



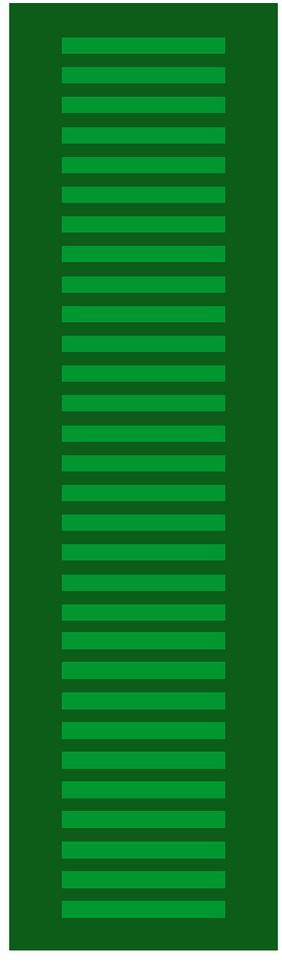
Filter Cascade

All revocations

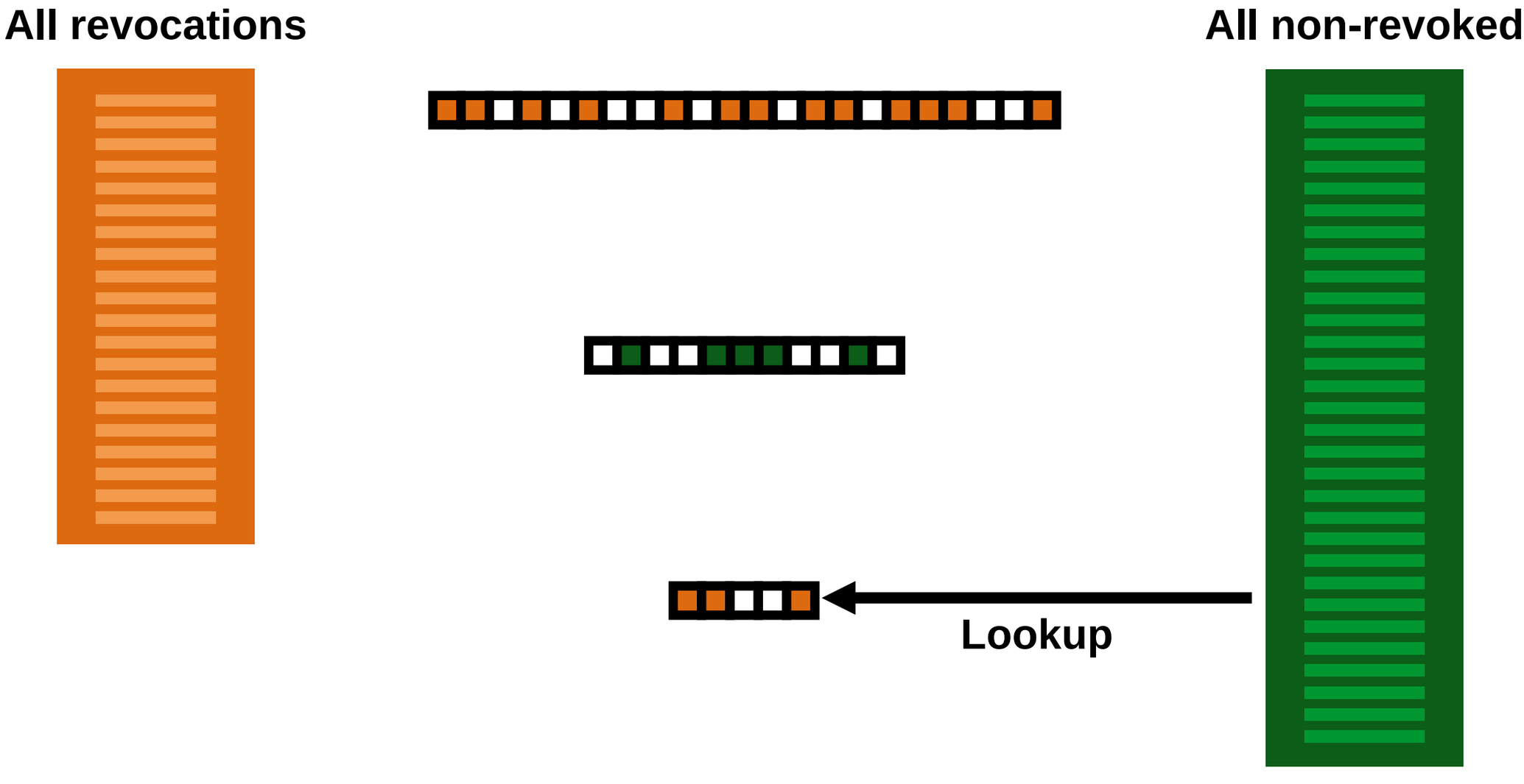




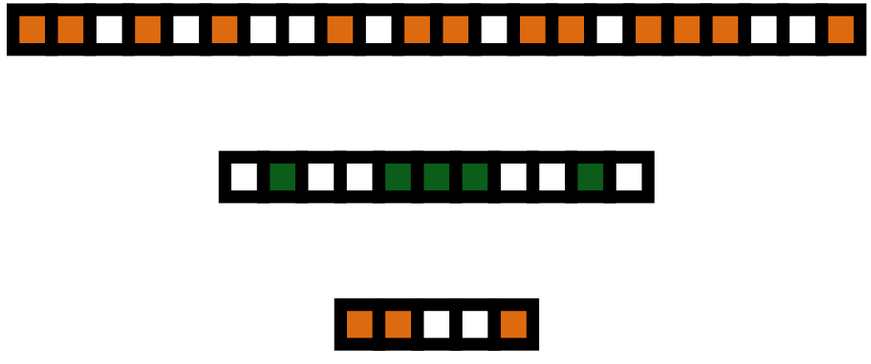
All non-revoked



Filter Cascade

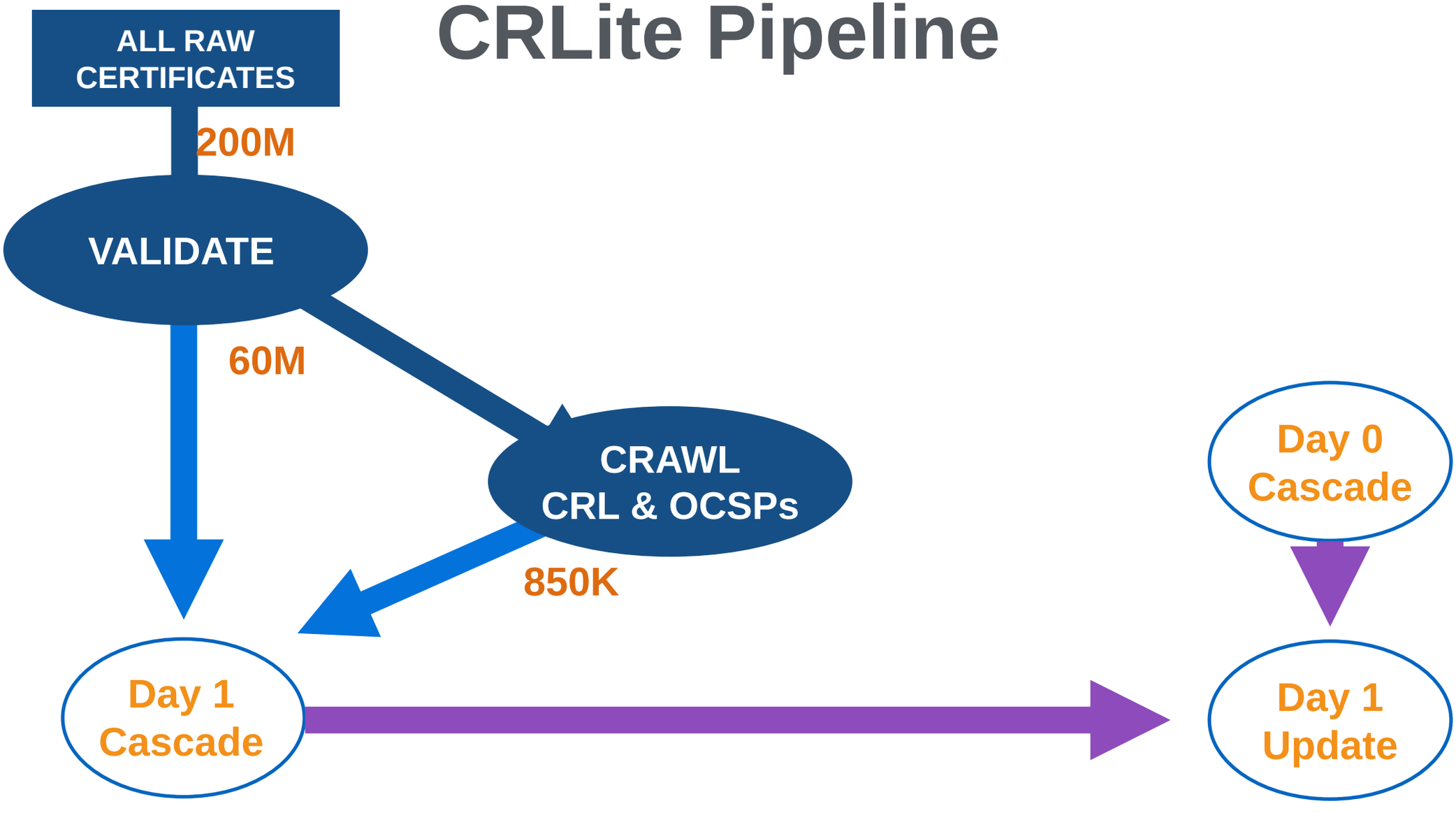


Our Filter Cascade

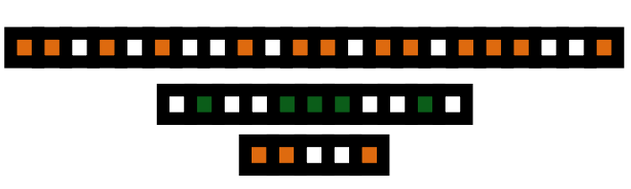


- 43 levels
- 7 hashes at first level, 1 per level thereafter
- 10 hashes expected
- 1.1 MB size (lower bound is 0.77 MB)

CRLite Pipeline



Small Enough

CRLite w/ Updates	CRLSet	OneCRL
		
40 KB / day	250KB	34KB
		
850K REVOCATIONS (100%)	**14K REVOCATIONS (1.6%)**	**400 REVOCATIONS (0.047%)**

A Survey of Techniques for Internet Traffic Classification using Machine Learning

Thuy T.T. Nguyen and Grenville Armitage

***Abstract*—The research community has begun looking for IP traffic classification techniques that do not rely on 'well known' TCP or UDP port numbers, or interpreting the contents of packet payloads. New work is emerging on the use of statistical traffic characteristics to assist in the identification and classification process. This survey paper looks at emerging research into the application of Machine Learning (ML) techniques to IP traffic classification - an inter-disciplinary blend of IP networking and data mining techniques. We provide context and motivation for the application of ML techniques to IP traffic classification, and review 18 significant works that cover the dominant period from 2004 to early 2007. These works are categorized and reviewed according to their choice of ML strategies and primary contributions to the literature. We also discuss a number of key requirements for the employment of ML-based traffic classifiers in operational IP networks, and qualitatively critique the extent to which the reviewed works meet these requirements. Open issues and challenges in the field are also discussed.**

***Index Terms*—Traffic classification, Internet Protocol, Machine Learning, Real Time, Payload inspection, Flow clustering, Statistical traffic properties.**

I. Introduction

REAL-TIME traffic classification has the potential to solve difficult network management problems for Internet service providers (ISPs) and their equipment vendors. Network operators need to know what is flowing over their networks promptly so they can react quickly in support of their various business goals. Traffic classification may be a core part of automated intrusion detection systems [1] [2] [3], used to detect patterns indicative of denial of service attacks, trigger automated re-allocation of network resources for priority customers [4], or identify customer use of network whose controlling application we wish to determine. Simple classification infers the controlling application's identity by assuming that most applications consistently use 'well known' TCP or UDP port numbers (visible in the TCP or UDP headers). However, many applications are increasingly using unpredictable (or at least obscure) port numbers [6]. Consequently, more sophisticated classification techniques infer application type by looking for application-specific data (or well-known protocol behavior) within the TCP or UDP payloads [7].

Unfortunately, the effectiveness of such 'deep packet inspection' techniques is diminishing. Such packet inspection relies on two related assumptions:

- Third parties unaffiliated with either source or recipient are able to inspect each IP packet's payload (i.e. is the payload visible)
- The classifier knows the syntax of each application's packet payloads (i.e. can the payload be interpreted)

Two emerging challenges undermine the first assumption - customers may use encryption to obfuscate packet contents (including TCP or UDP port numbers), and governments may impose privacy regulations constraining the ability of third parties to lawfully inspect payloads at all. The second assumption imposes a heavy operational load - commercial devices will need repeated updates to stay ahead of regular (or simply gratuitous) changes in every application's packet payload formats.

The research community has responded by investigating classification schemes capable of inferring application-level usage patterns without deep inspection of packet payloads. Newer approaches classify traffic by recognising statistical

… over 1000 citations

Potential applications of machine learning in the context of Content Delivery Networks

Predicting client-server performance based on past performance, network measurements, network topology, …

Determining client reputation (is this a bot?) based on past behavior

Estimating physical location of client

Best when penalty for false positives is low

Rate-limit (but don't ban) requests by suspicious clients

Kick suspicious peers out of the peer-to-peer system (fall back to CDN)

Thank you!