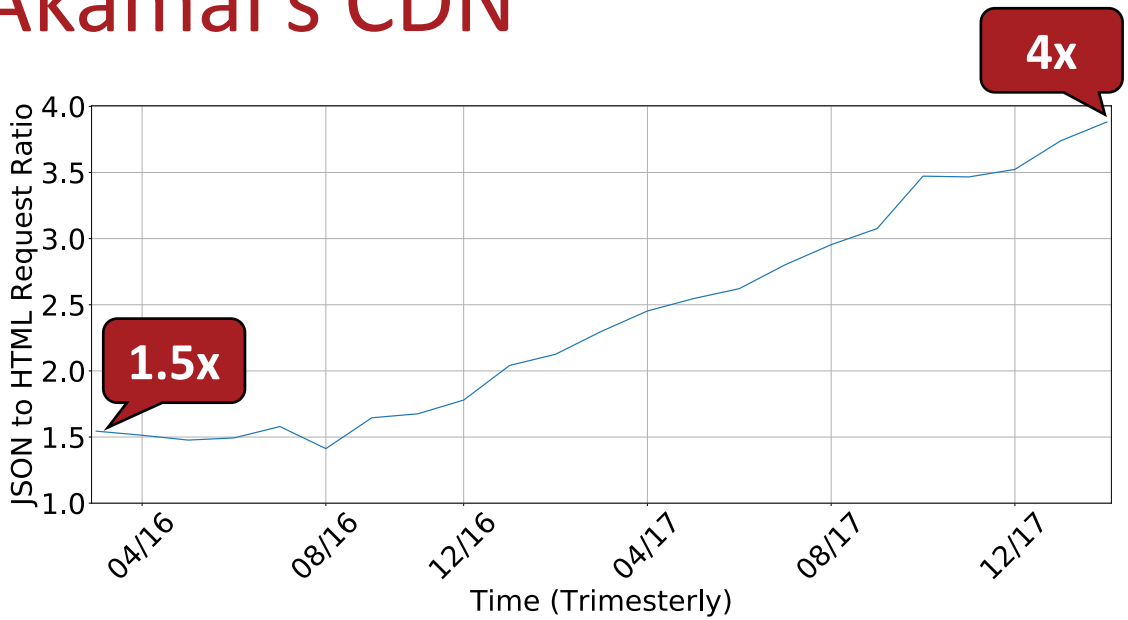
Stony Brook University

Characterizing JSON Traffic Patterns on a CDN

Santiago Vargas, *Utkarsh Goel, Moritz Steiner, Aruna Balasubramanian*

FAR BEYOND

JSON Traffic is Growing on Akamai's CDN

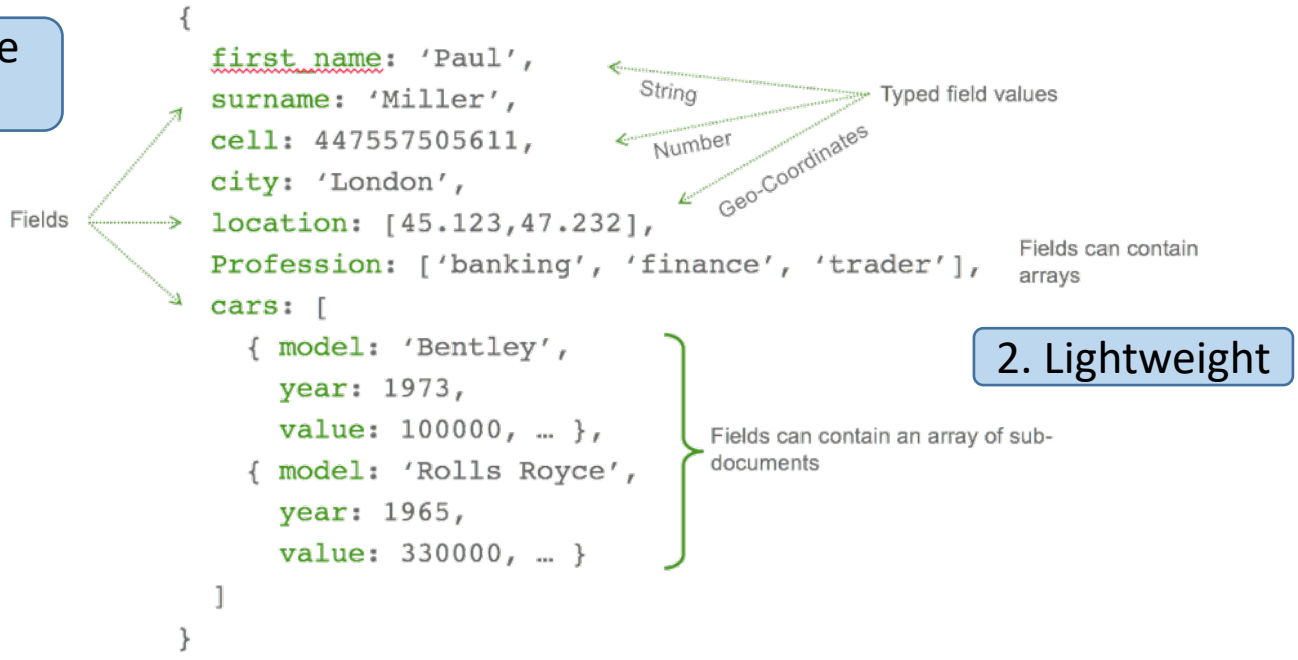


- JSON is **4x** more requested that HTML
- JSON a the leading content type on Akamai's CDN

What is JSON?

JSON is a text-based data format

1. Key-value Pairs



2. Lightweight

3. Structured

Motivation

- Little is known about JSON usage
- Optimizations exist for types of content, ie. browsing and media content. Unsure if these apply to JSON content

Akamai's Network as a Vantage Point

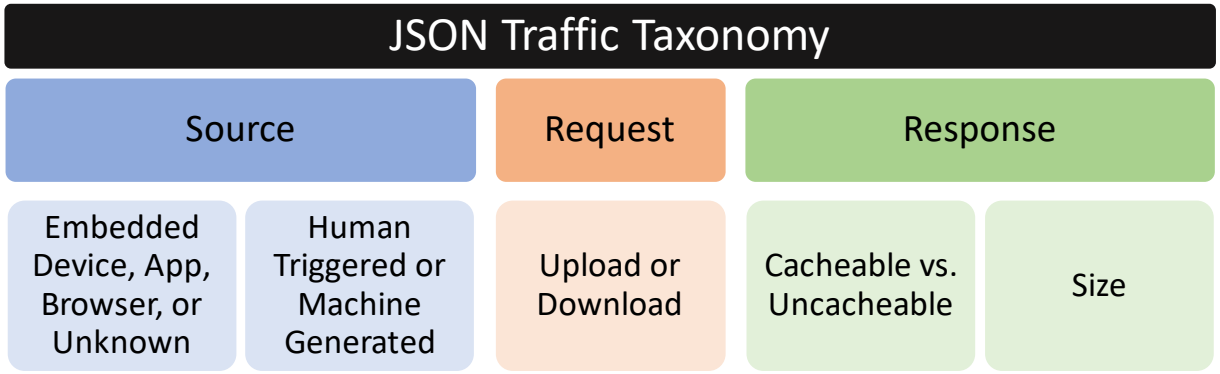
- Collect HTTP request logs from Akamai edge servers
  - Data Fields: Mime Type, Time of Request, User-Agent, Caching Information, HTTP Method, URL, Anonymized IP
- Collect 2 datasets:

Dataset	# of Logs	Duration	# of Domains
Short-term	25 million	10 mins	~5k
Long-term	10 million	24 hours	~170

  - Short-term – wide network coverage for overall characterization
  - Long-term – wide temporal coverage for pattern characterization

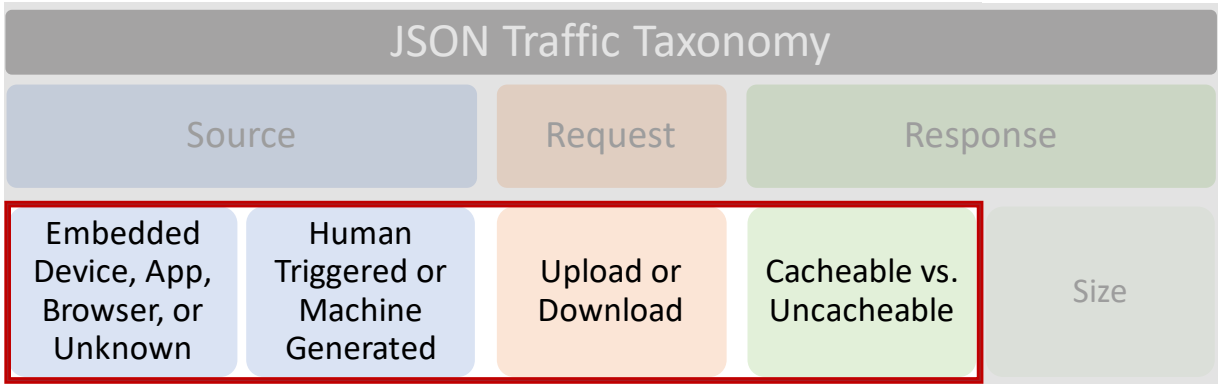
Extracting Our Data

- Develop a taxonomy to look at JSON traffic on different dimensions:



Extracting Our Data

- Develop a taxonomy to look at JSON traffic on different dimensions:



- Device/Application
- Machine-Generated
- Request Type
- Response Type

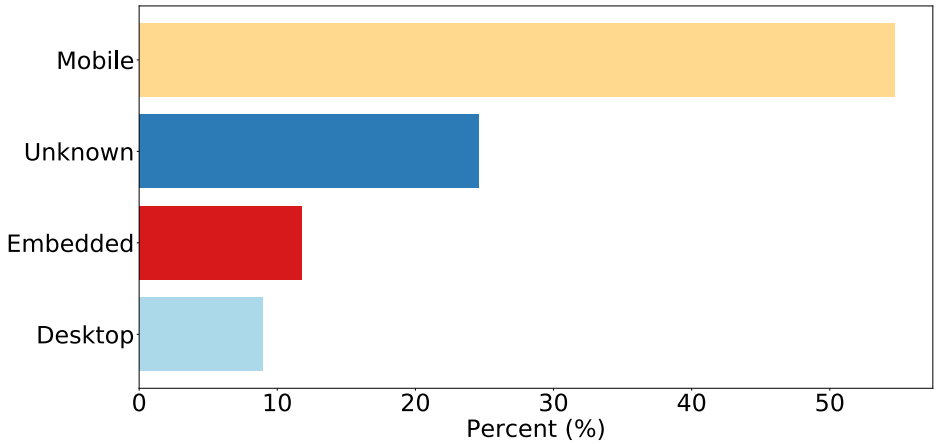
User Agent

Pattern Based

HTTP Method

Cache Labels

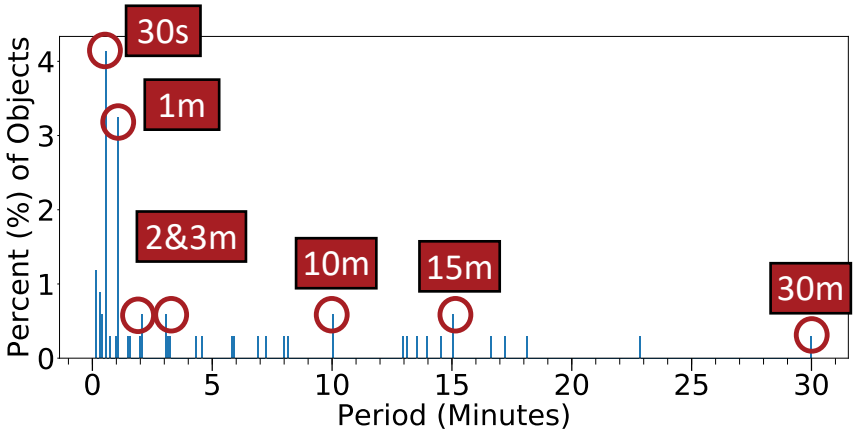
What Devices Request JSON?



- Majority of JSON is from mobile smartphones and embedded devices (>64%)
  - Embedded devices include game consoles, wearables, and smart appliances
- 88% of JSON is non-browser

Identifying Machine Traffic

- Autocorrelation techniques identify 6% of JSON traffic is requested periodically

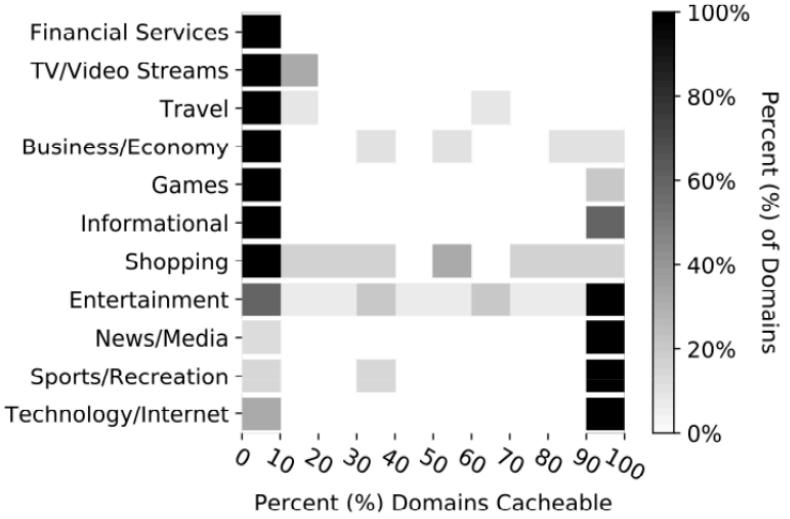


Objects are requested in **standard periods**

- Machine-to-machine traffic
  - 78% upload traffic & 56% uncacheable
  - One optimization avenue: Deprioritize machine traffic that is not QoE sensitive

Does Caching Help?

- Majority (84%) of requests are downloads making them ideal for caching
  - 55% of requests are uncacheable
  - 50% of domains don't use caching for JSON

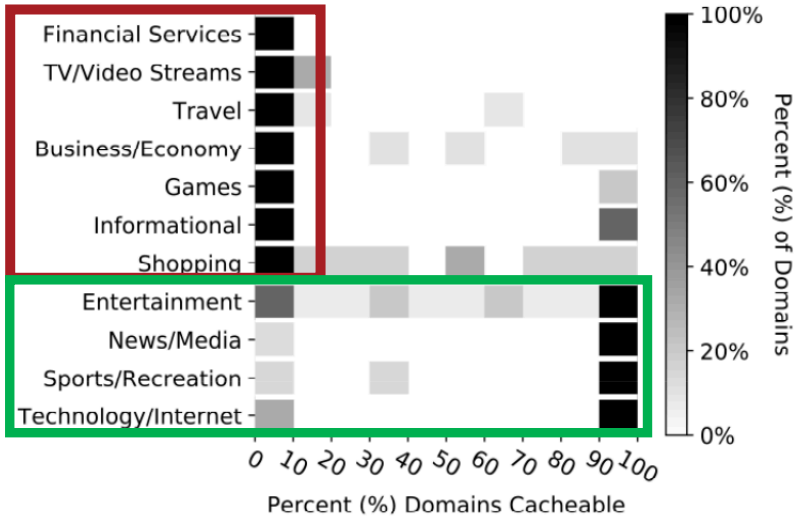


Does Caching Help?

- Majority (84%) of requests are downloads making them ideal for caching
  - 55% of requests are uncacheable
  - 50% of domains don't use caching for JSON

Uncacheable

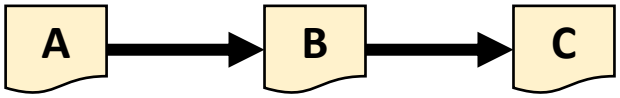
100% Cacheable



Since caching doesn't help JSON content delivery for specific industries, can other optimizations help?

Using Dependencies as avenue for optimization

- Dependencies Example:



*Requests for JSON Object **A** then **B** then **C***

Question: Given request for A, can we predict object B will be requested next?

- Methodology: Cluster objects with similar URLs

a.com/abc123/profile
a.com/def456/profile
→ a.com/*/profile *Clustered URL*

- Results: 90% accuracy using ngram model for clustered URLs*

Analyzing these patterns can inform Prefetching, Server Push, Anomaly Detection systems

*when considering top 10 predictions

Conclusions

- JSON is becoming a leading content type on the Web
- Majority of JSON is from mobile & embedded devices and is non-browser
- Avenues for optimization:
  - Periodic machine-to-machine traffic can be deprioritized
  - JSON has ordering patterns that can be predicted