DNS backscatterの紹介

国立情報学研究所
福田 健介

Detecting malicious activity with DNS backscatter

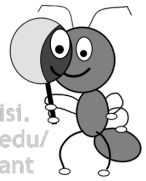
Kensuke Fukuda (NII/Sokendai)

John Heidemann (USC/ISI)

Appeared in IMC2015
http://www.fukuda-lab.org/publications/FH_imc2015.pdf

National Institute of Informatics

USC Viterbi School of Engineering Information Sciences Institute



isi.edu/ant

- 2014: Heartbleed
  - A bug in critical millions of Internet hosts
  - Security researchers scanned to find unpatched servers
- Question: Who else was scanning?
  - Criminals?
  - Black hats?
  - Others?

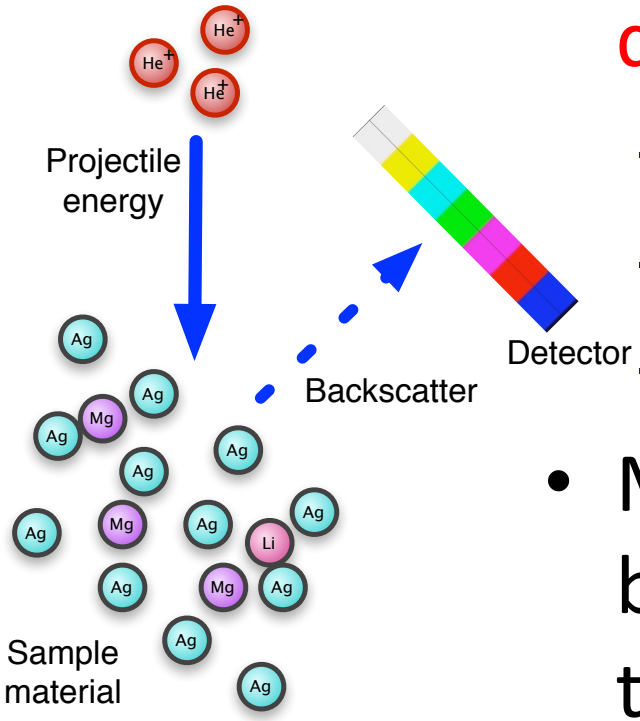
Goal: Finding Originators of Network-wide events

- Large-scale event involves many Internet hosts
  - Malicious: Scan, Attacks, Spams
  - Benign: CDN, Web crawler, DNS, NTP, Updates
  - Border: Ad tracking
- Importance of monitoring those events
  - Malicious: security consideration
  - Benign: stability of infrastructure

Our contribution

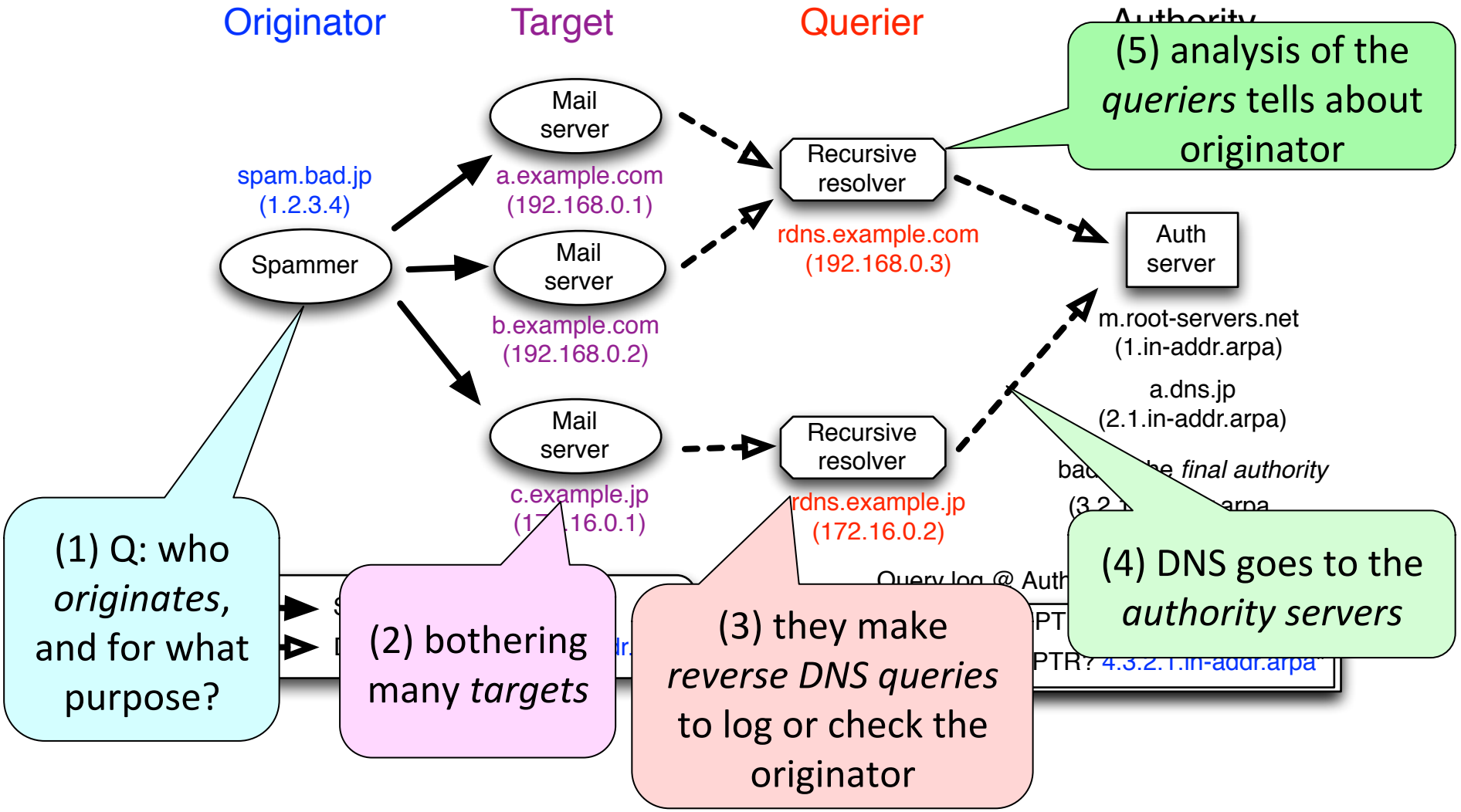
- New data source - DNS backscatter - to identify network-wide activity
  - Deployable
  - Privacy friendly
  - Robust against malicious source
- Validate with several DNS authoritative servers
- Evaluate over time: 6 months dataset

Key idea of DNS backscatter



- Large event triggers reverse DNS queries near target automatically
  - SMTP server: hostname of spammer
  - Firewall: hostname of scanner
  - Web server: hostname of web crawler
- Many reverse DNS queries (DNS backscatter) at auth server are hint to identify events

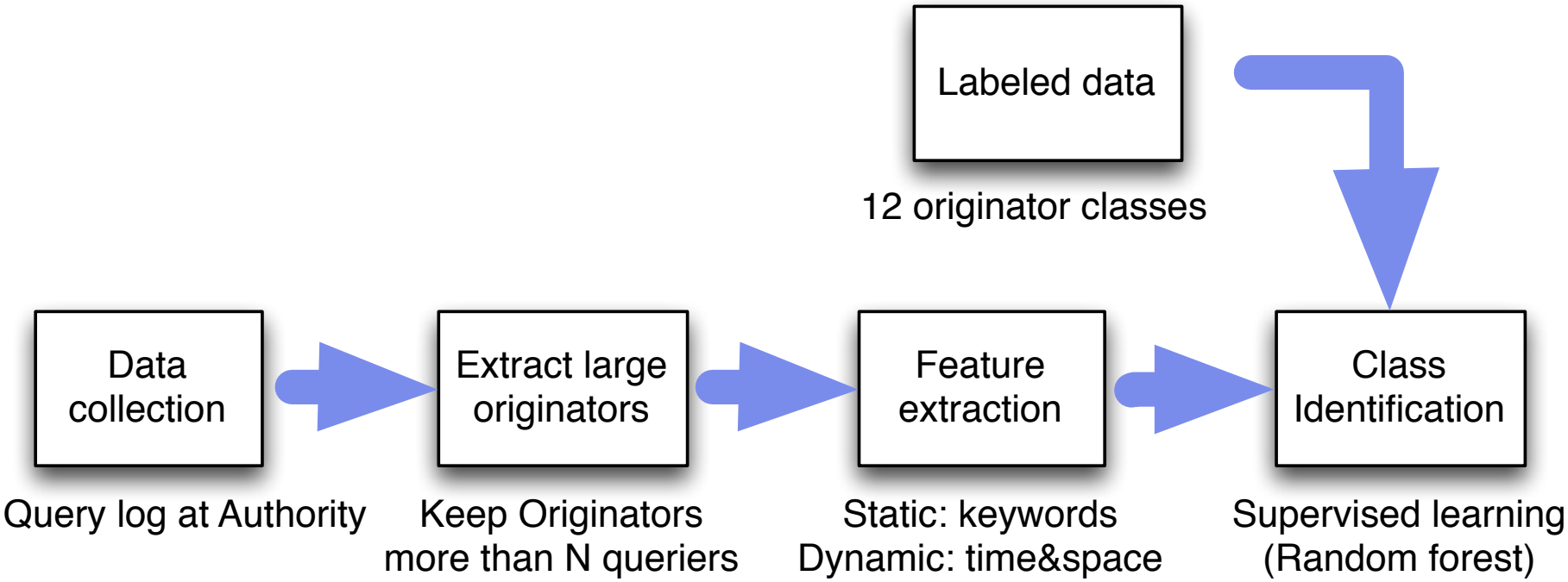
Detecting Events through DNS backscatter



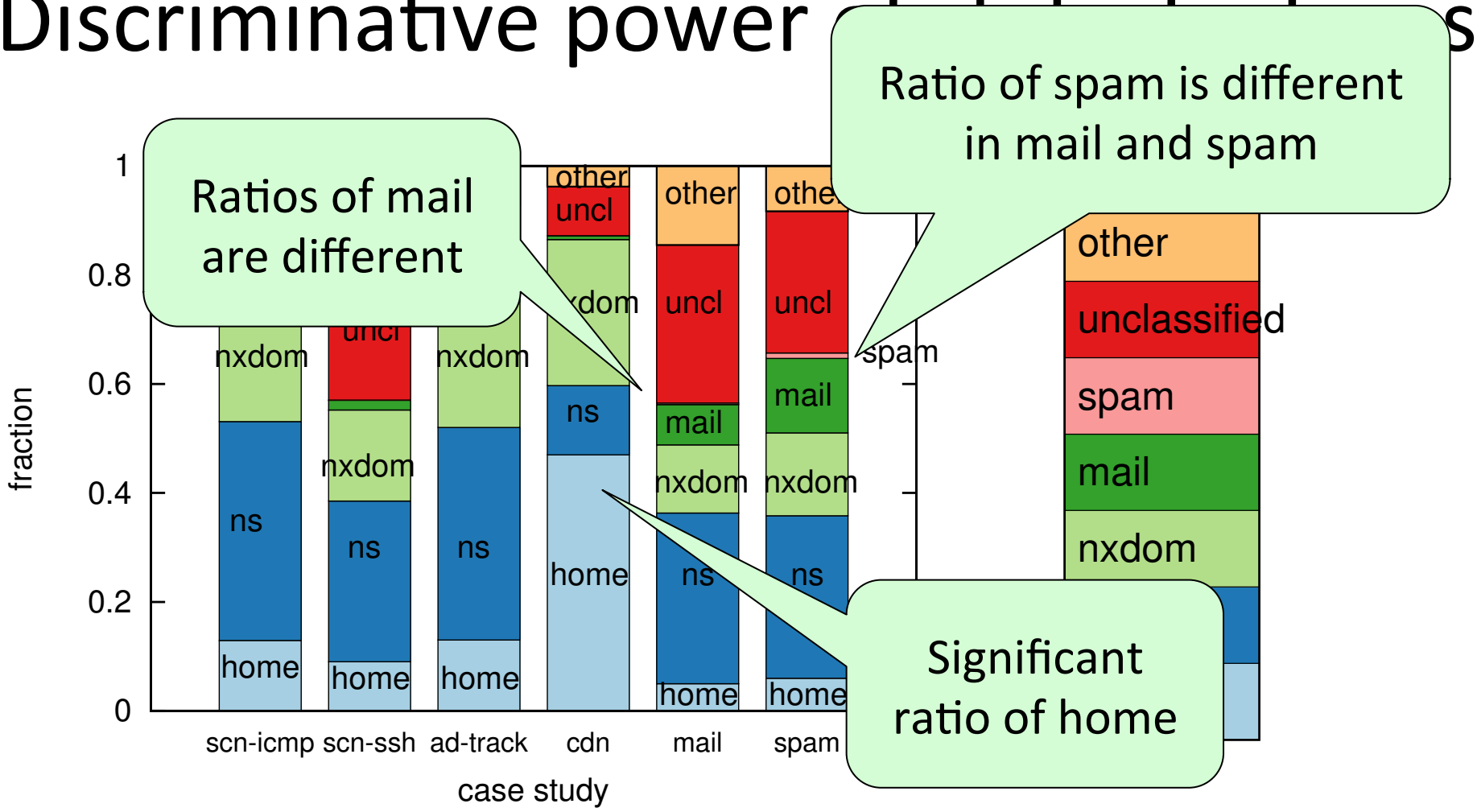
Advantages

- Deployable
  - Centralized monitoring at DNS authority
- Privacy friendly
  - Information is on queriers NOT originators
  - Reverse queries are generated automatically
  - Focus on large events (ignore small users)
- Robust against malicious originators
- Can infer different class of originator (e.g., scanner) with Machine Learning

Identification process



Discriminative power of statistical features



Different mixes of features allow distinguishing different classes of events

Picking the best ML algorithm

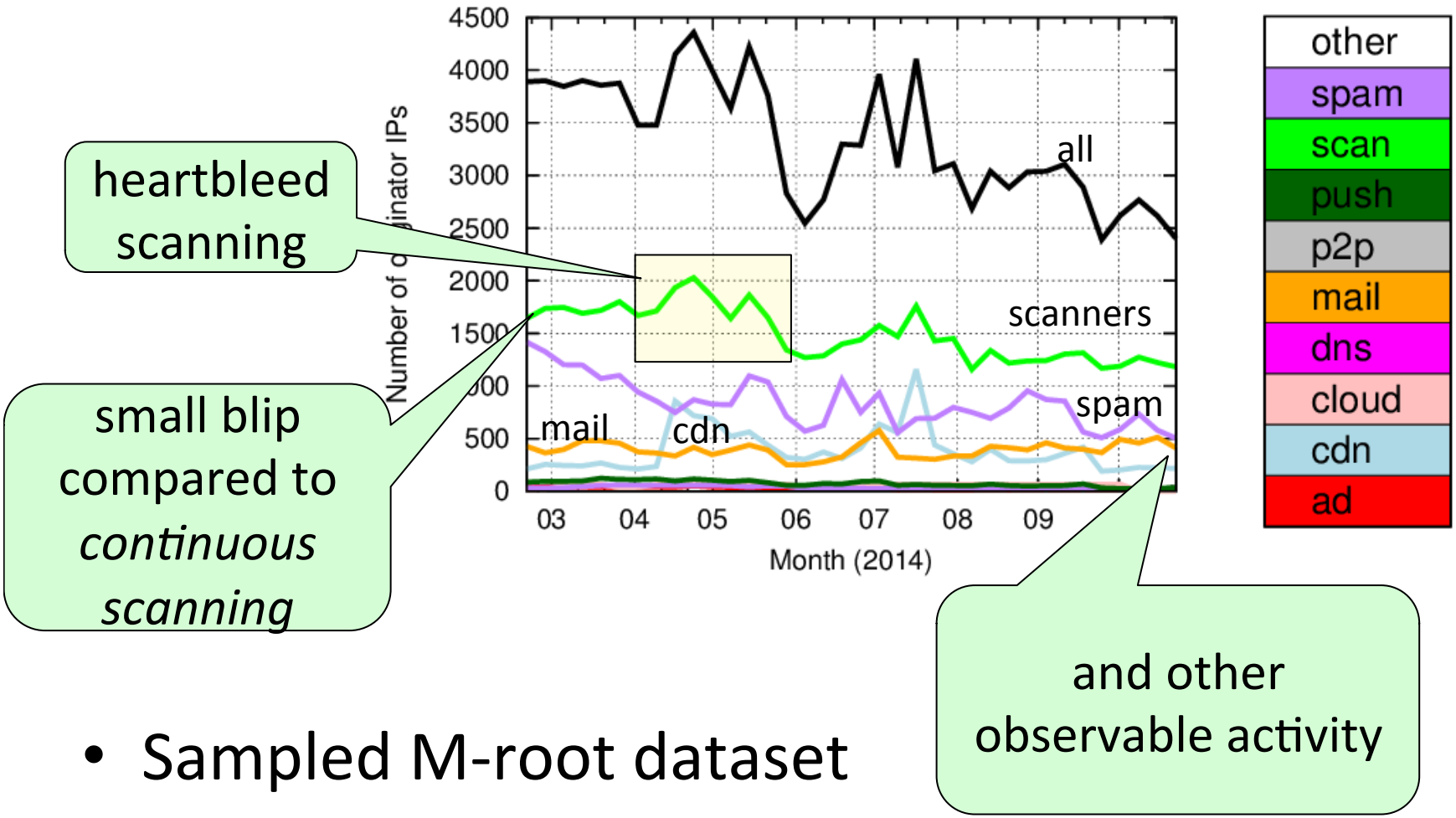
dataset	algorithm	accuracy	precision	recall	F1-score
	CART	0.66	0.63	0.60	0.61
JP	**RF**	**0.78**	**0.82**	**0.76**	**0.79**
ditl	SVM	0.73	0.74		
B	CART	0.48	0.48	0.45	
post-	**RF**	**0.62**	**0.66**	**0.60**	
ditl	SVM	0.38	0.50	0.32	
	CART	0.53	0.52	0.49	0.51
M	**RF**	**0.68**	**0.74**	**0.63**	**0.68**
ditl	SVM	0.60	0.6		
	CART	0.61	0.6		
M	**RF**	**0.79**	**0.8**		
sampled	SVM	0.72	0.7		

RandomForest is best

Hope to improve with better training data

- Cross validation with 3 ML algorithms
- Num classes: 12, labeled data:200-800
- Precision: 70-80% (imbalanced dataset problem)

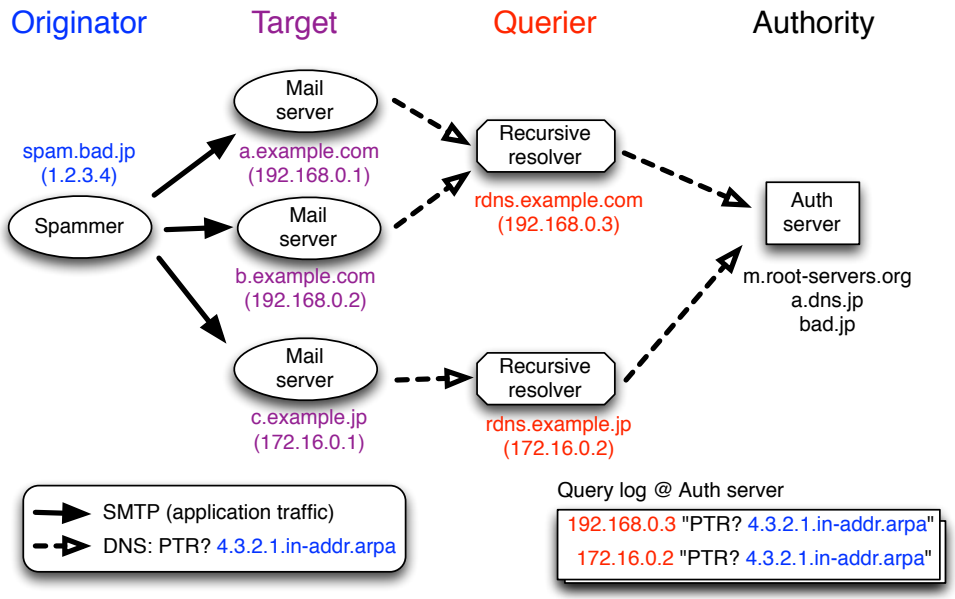
Finding Network-wide events over time



- Sampled M-root dataset

Conclusion

- DNS backscatter - a new data source for Internet-wide events
- Advantages:
  - Deployable
  - Privacy-friendly
  - Reasonable accuracy
- Longitudinal results



DNS operators may apply this to detect large events!