ENDGAME

John Munro / jmunro@endgame.com
Jason Trost / jtrost@endgame.com

FlonCon 2013 | January 7–10 | Albuquerque, New Mexico

Introductions

ENDGAME

- John Munro (jmunro@endgame.com)
  - Network Security Researcher and Data Scientist
- Jason Trost (jtrost@endgame.com)
  - Senior Software Engineer
  - Specializes in Hadoop/Storm/BigData

Agenda

ENDGAME

- The Problem
- Our Approach
- DGA Domain Classifier
- String Statistics as Features
- Malicious Domain Classifier
- Demo
- Real-time Streaming Platform

The Problem

ENDGAME

txmxbo.info

youtube.com

yahoo.com

Ct0u2xj5dbe4.wvw—game465.com

p4.httzd5e2ufizo.3bawhfuec45dca65.401724.s1.v4.ipv6-exp.l.google.com

abulqe.com

za6.limfoklubs.com

ns3.ohio.gov

bibz01.apple.com

docs.joomla.org

Wmk41035u3751s0bgv4n91b0b7h74v.ipcheker.com

The Problem

ENDGAME

txmxbo.info

youtube.com

yahoo.com

Ct0u2xj5dbe4.wvw—game465.com

p4.httzd5e2ufizo.3bawhfuec45dca65.401724.s1.v4.ipv6-exp.l.google.com

abulqe.com

za6.limfoklubs.com

ns3.ohio.gov

bibz01.apple.com

docs.joomla.org

Wmk41035u3751s0bgv4n91b0b7h74v.ipcheker.com

The Problem

ENDGAME

- Massive Volumes
  - Some of our partners deal with TBs per day of DNS PCAPs
- Incredible Rates
  - One partner sees 13k requests/sec
  - Another closer to 100k/sec

Our Approach: Machine Learning!

ENDGAME

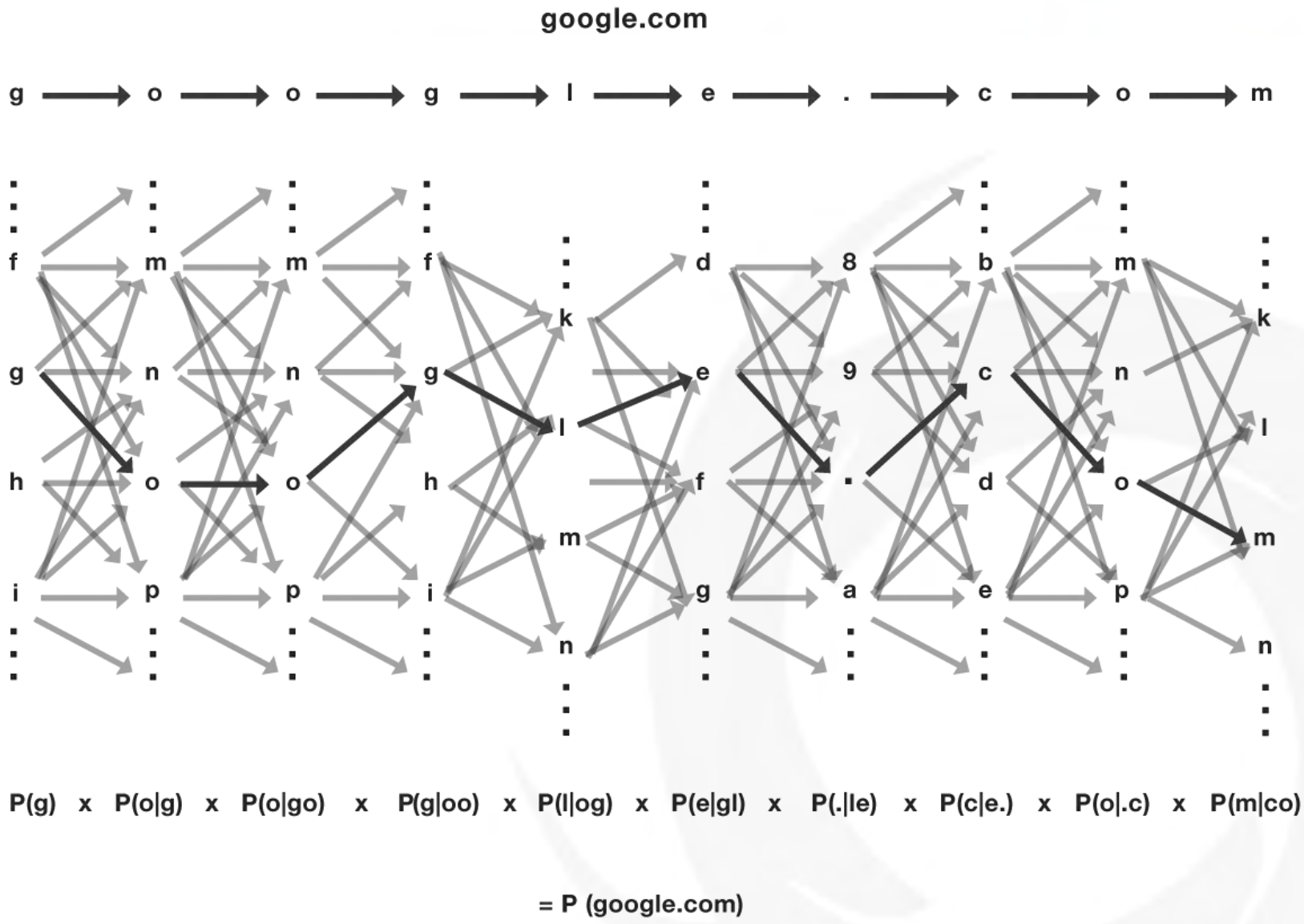
- Real-time streaming classification
  - In parallel across multiple servers
- Markov Models
  - Random Domain Generation Traffic
  - Normal Benign Traffic
- Random Forests
  - Benign vs Malicious
- Periodically retrained
  - In order to maintain accuracy

Data Sources

ENDGAME

- Benign Domains
  - Millions of popular, real domains
    - Correlated with the Alexa top 10k domains
- Malicious Domains
  - 800k domains gathered from an internal malware sandbox
  - Public blacklist domains from Conficker and Murofet Botnets

Markov Models



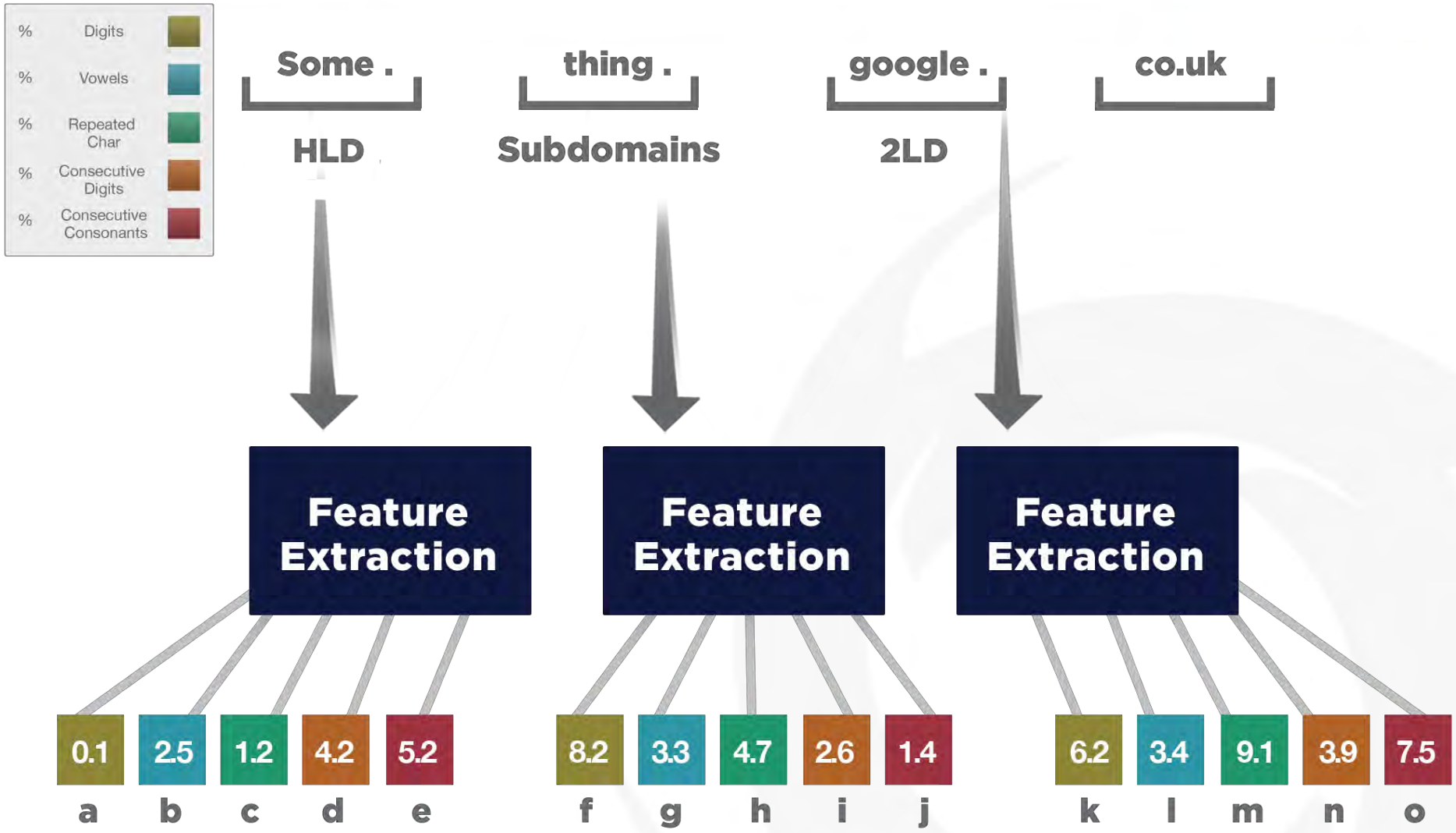
Markovian DGA Classifier

ENDGAME

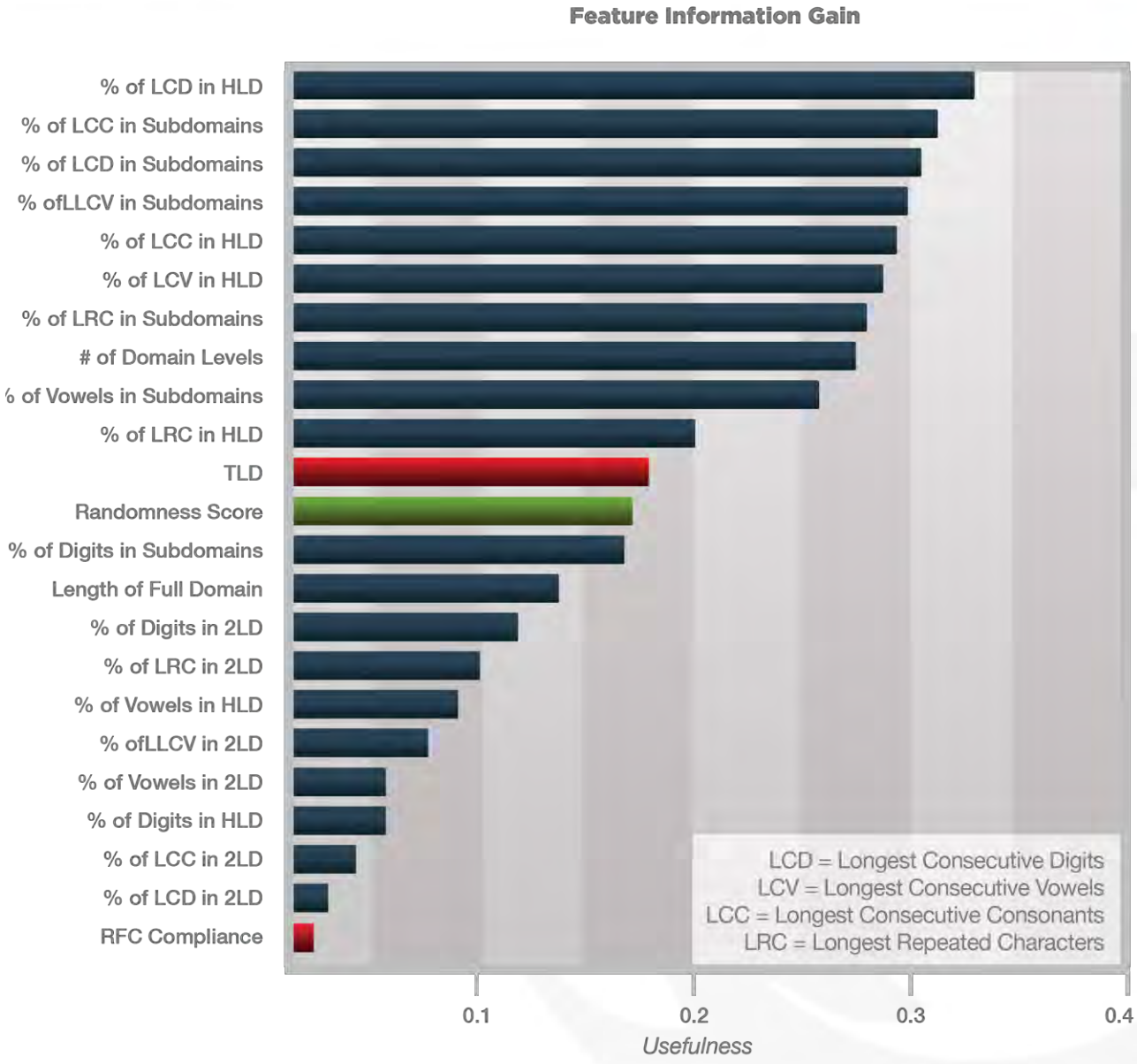
- Domain Generation Algorithm (DGA)
- Popular Domain Model
  - Trained: 258,039 domains from Day 1 of our Benign set
  - Tested: 331,359 domains from Day 2 of our Benign set
  - Accuracy: 99.40 % with 1,458 Unknown
- Randomly Generated Domain Model
  - Trained: 90,884 domains from Conficker Botnet
  - Tested: 295,306 domains from Murofet Botnet
  - Accuracy: 99.34 % with 1,923 Unknown

String Statistics as Features

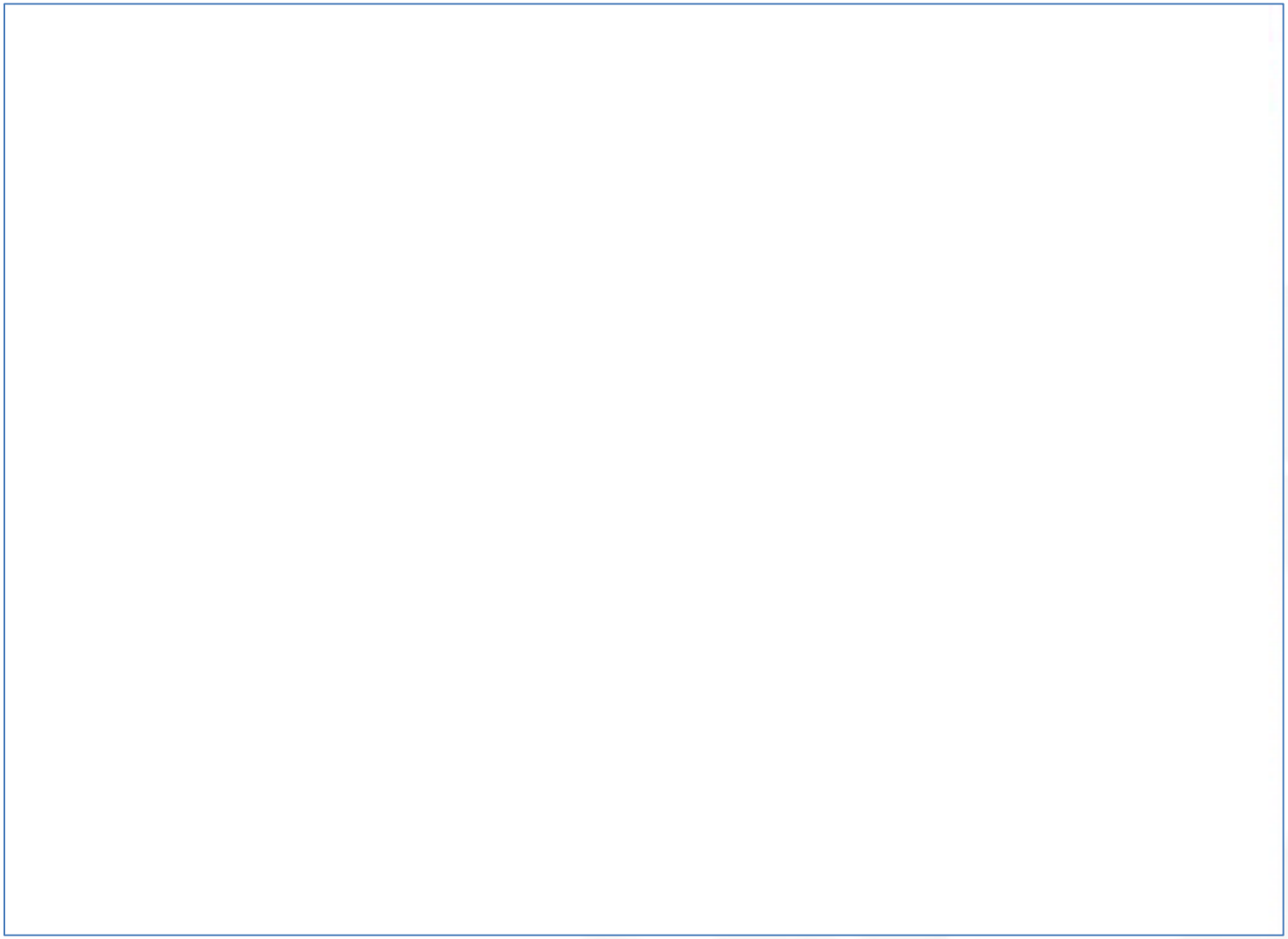
ENDGAME



Feature Usefulness



Random Forests Algorithm



Random Forests

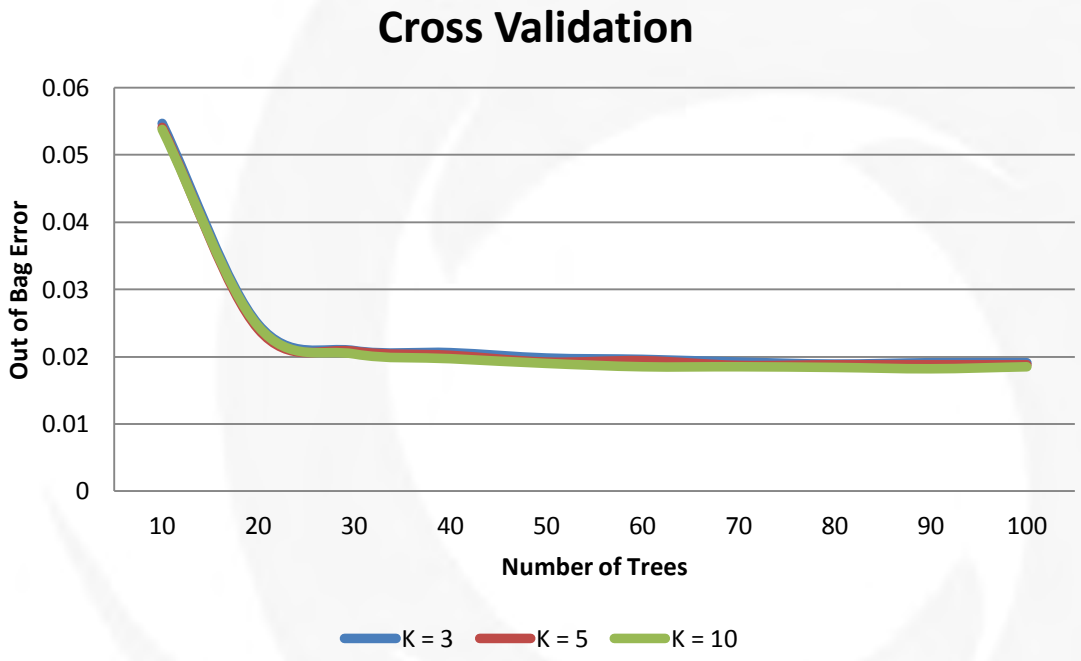
ENDGAME

- Pros:
  - Very high accuracy
  - Scalable across many nodes
  - Built-in protection from over fitting
  - Can handle very large data sets with many features
  - Robust with respect to goodness of features
  - Practical for real world use
  - Does not assume a distribution
  - Only two parameters to tune
  - Memory efficient
- Cons:
  - Not the quickest classifier, but plenty fast in practice

Malicious Domain Classifier

ENDGAME

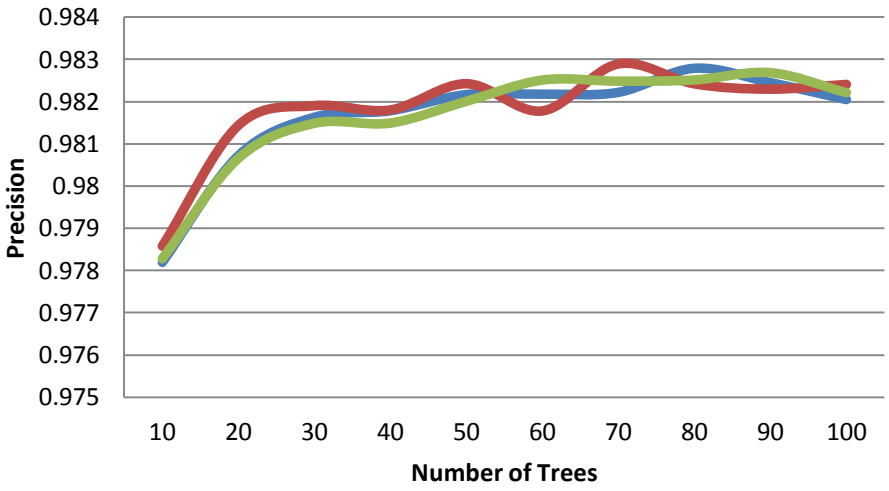
- Performance measured by 10 – fold Cross Validation
- Training Set
  - 200k Benign
  - 200k Malicious



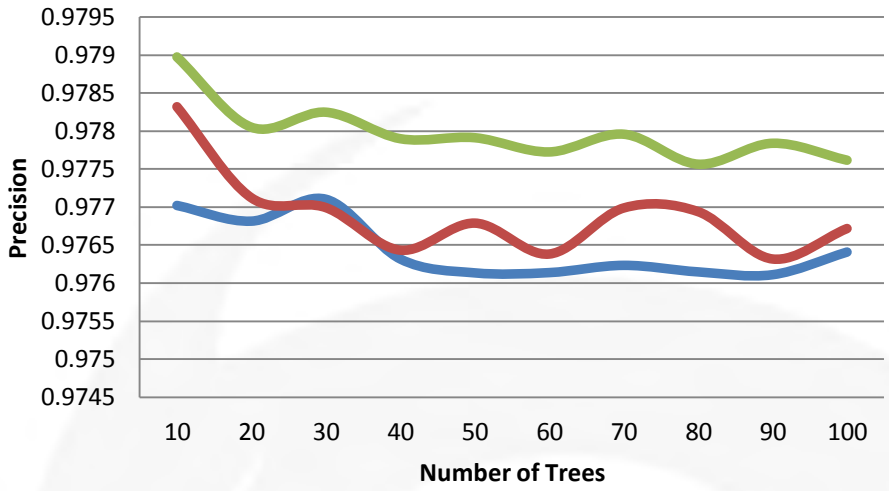
Results

ENDGAME

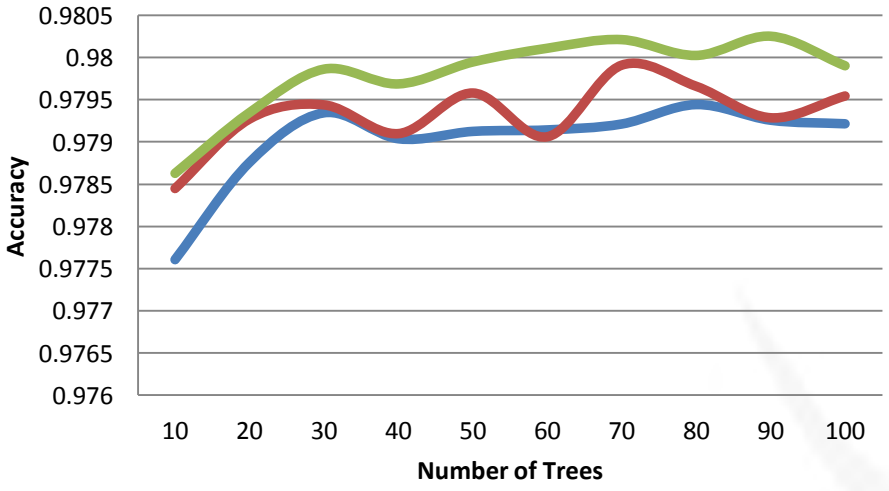
Bad Precision



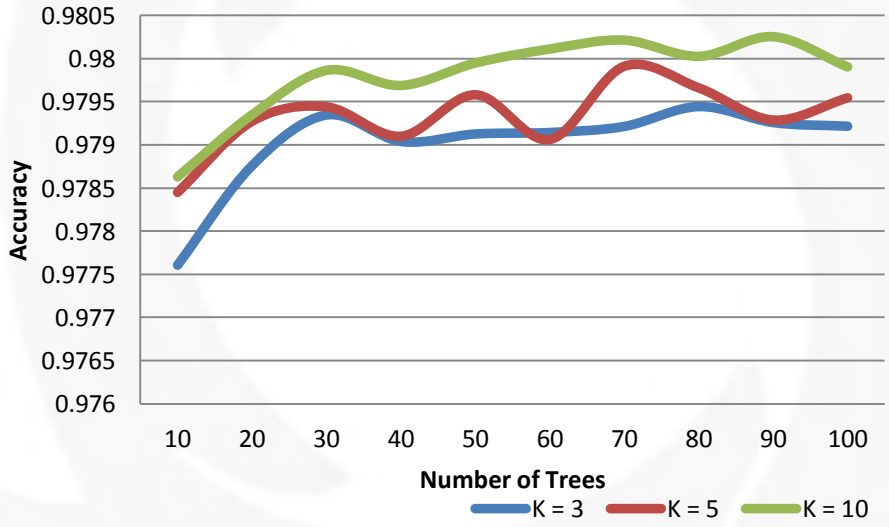
Good Precision



Bad Accuracy

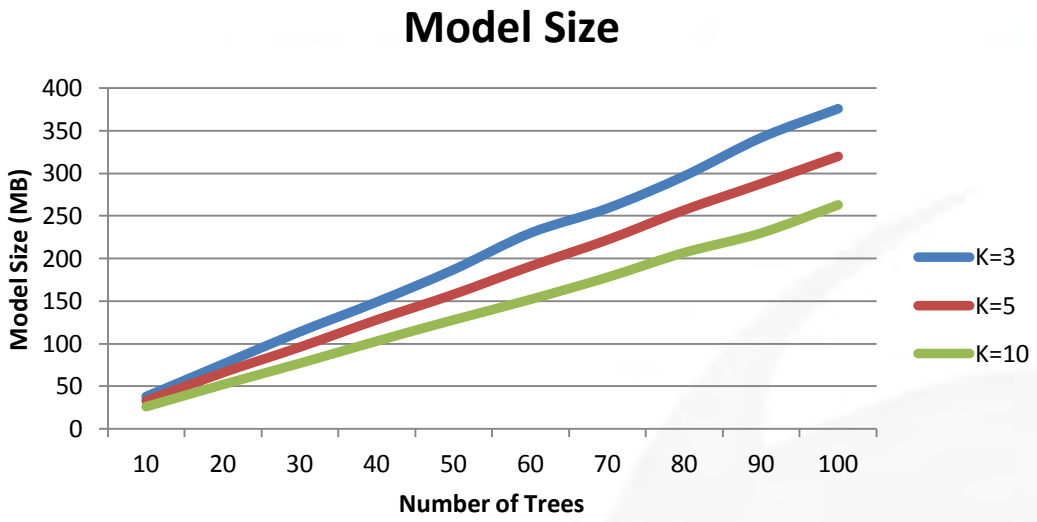


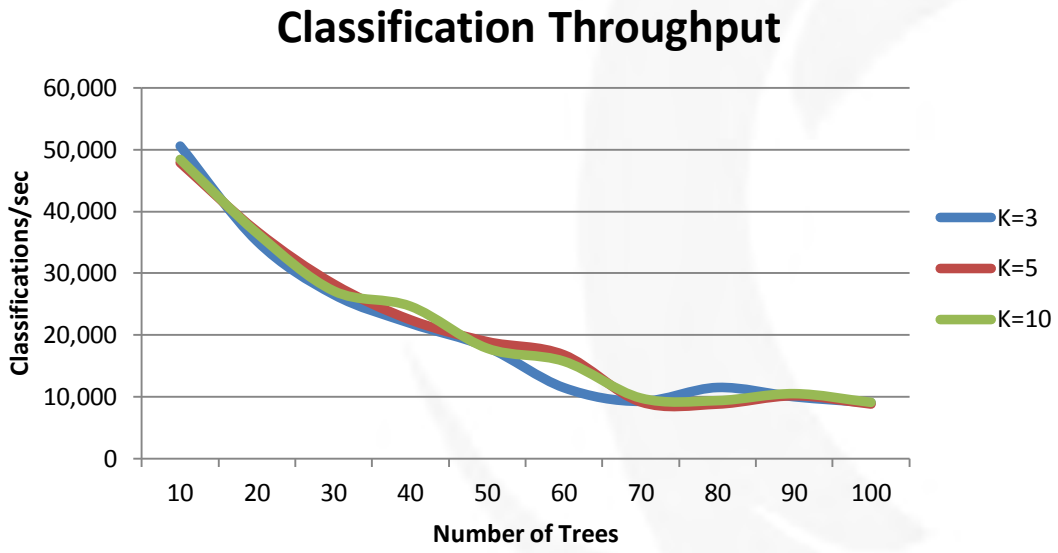
Good Accuracy



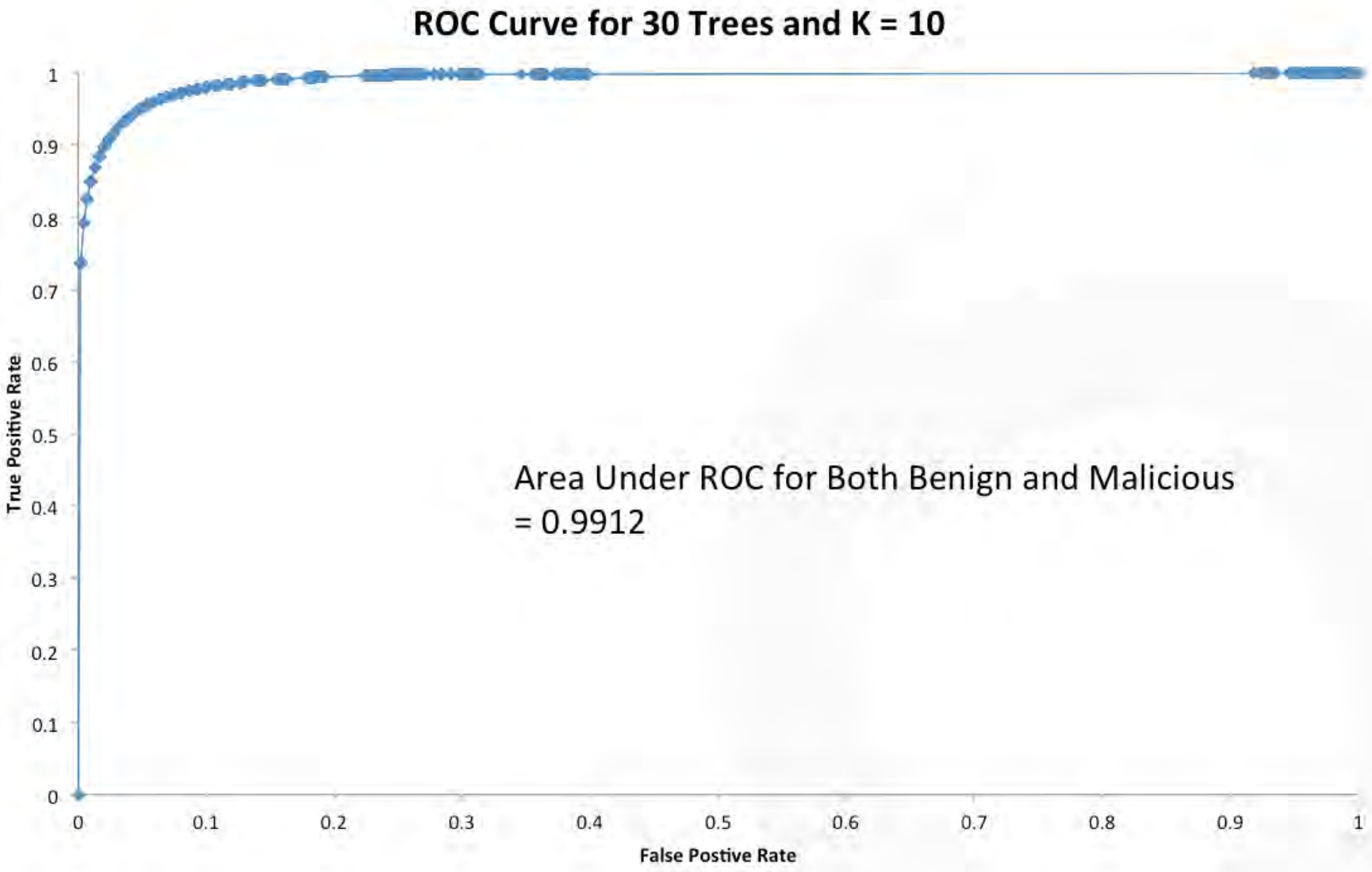
Results

ENDGAME





Results


ROC Curve for 30 Trees and K = 10
True Positive Rate
False Postive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Area Under ROC for Both Benign and Malicious
= 0.9912

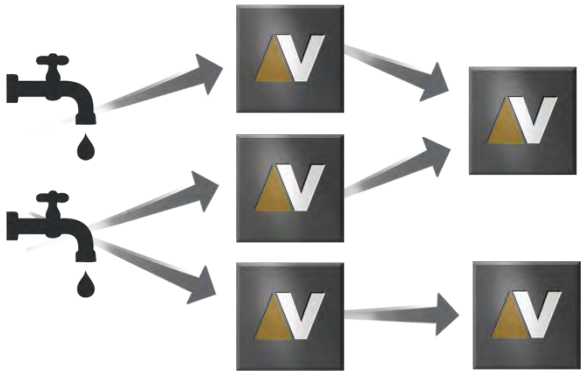
Demo

ENDGAME

Realtime Streaming Platform

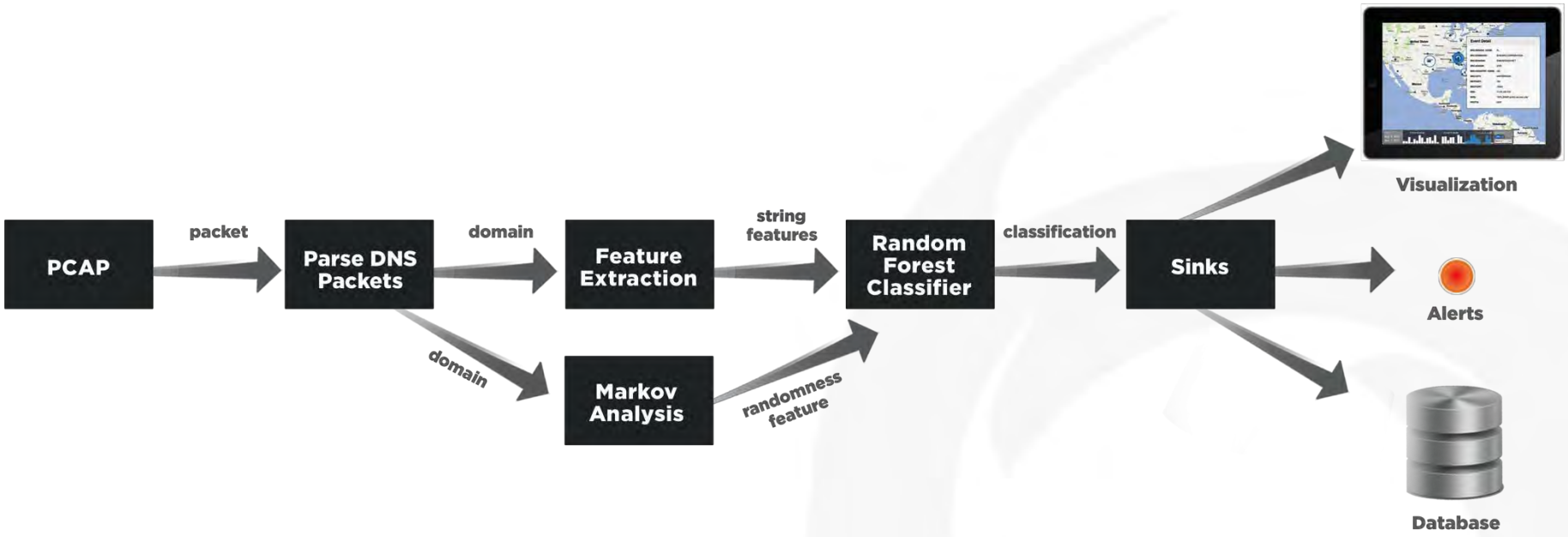
ENDGAME

- *Velocity* is a platform for processing, analyzing, and visualizing large-scale event data in real-time
- It was designed to be horizontally scalable and is built using Twitter's Storm



- It was built primarily for internal use with DNS events, IDS alerts, and netflow data, but it is in the process of being commercialized

Velocity Pipeline



Conclusion

ENDGAME

- Malicious domain classification
- DGA domain identification using Markov Models
- Summary Statistics based on domain string work well
- Random Forests are very successful at classifying domains as Benign or Malicious
- Real-time, distributed implementation

Future Work

ENDGAME

- Include more features: TTL, frequency seen, etc.
- Correlation of bad domains based on ASN, Country, Organization, etc.
- Identify subnets that are infected based on high traffic to bad domains
- Identify Content Delivery Networks
- Self Organizing Maps and other visualizations

Questions

ENDGAME

Contact Information

- John Munro
- Email: jmunro@endgame.com

- Jason Trost
- Email: jtrost@endgame.com
- Twitter: @jason_trost
- Blog: www.covert.io