Cyber Security

Data Analysis Project

DC-DAT-10 Course

Problem Statement

Apply Machine Learning to Cyber Security data

- Analysts perform triage, analysis, and forensics on security events
- Signature-based approaches vs zero-day attacks
- Rapid advances in technologies for massive volumes of data

How do we stay ahead of the game?

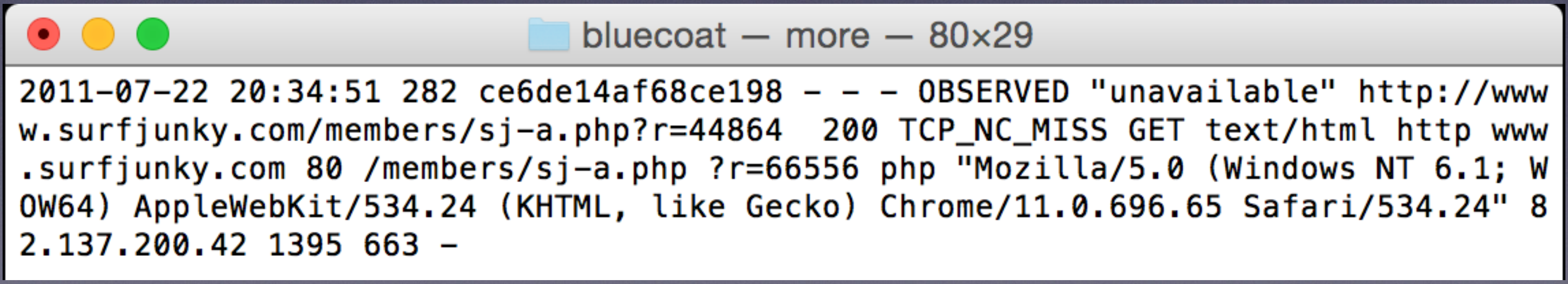
Background

- What constitutes 'cybersecurity data? Some common examples:
  - audit logs (e.g. fw, router, o/s, application…)
  - flow data
  - alerts (ids/ips, virus alerts, etc)
  - BUT also- inventory data, h/r databases, acquisition info, etc. Basically anything that gives you the context for making decisions actionable
- Proposed Project data sets:
  - Bluecoat Proxy data → unsupervised
  - IDS and FW logs (VAST challenge) → supervised
  - Malware (Kaggle competition) → supervised

Data Set

- Bluecoat Proxy Logs
- Security appliance serves multiple functions, depending on configuration
- Large data set acquired by hacker group Telecomix (sp?)
- Redacted for privacy reasons
- Large data set some of which must be processed outside of IPython Notebook

Example log entry:



```
2011-07-22 20:34:51 282 ce6de14af68ce198 - - - OBSERVED "unavailable" http://www
w.surfjunky.com/members/sj-a.php?r=44864  200 TCP_NC_MISS GET text/html http www
.surfjunky.com 80 /members/sj-a.php ?r=66556 php "Mozilla/5.0 (Windows NT 6.1; W
OW64) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.65 Safari/534.24" 8
2.137.200.42 1395 663 -
```

Questions

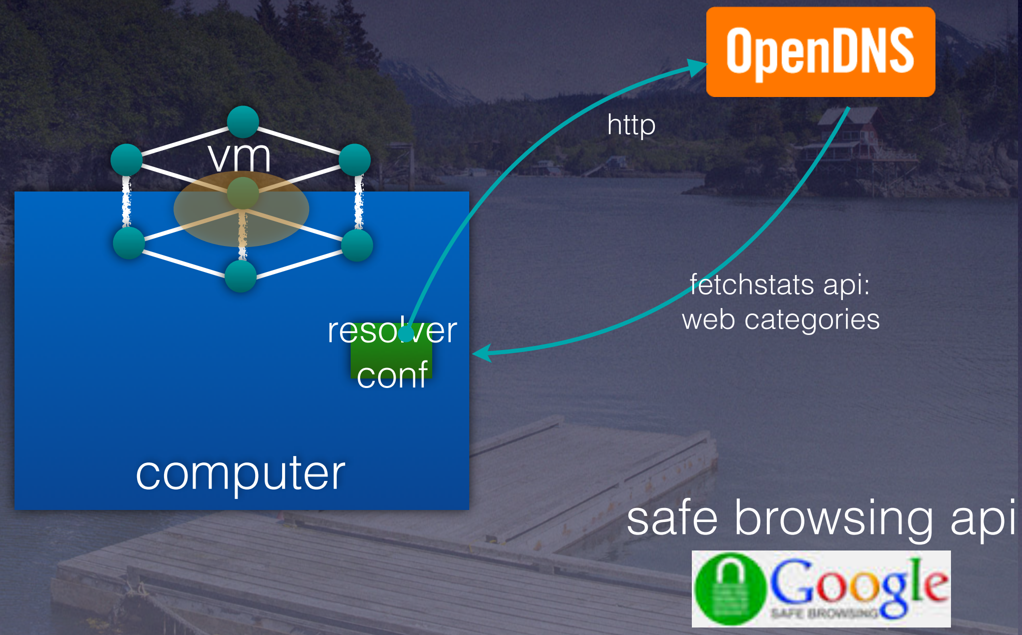
- Can we identify what are the censorship policies thru the logs:
  - by IP address or subnet?
  - by topic?
  - by domain?
  - by services?
- Can we identify how policies change over time?
- Can we identify unique clients and determine traffic patterns?

New questions may arise after going thru EDA..

Plans and Progress Status

- Data extraction and cleaning [In Progress]
- Data enrichment [In Progress]
  - Web categorization: OpenDNS API
  - Google SafeBrowsing API
- Exploratory data analysis [In Progress]
  - Censored vs. Proxied vs. Observed requests
  - Censored Requests Trending
  - Top 20 Ports associated with Censored Requests
  - Top 20 Domains associated with Censored Requests
  - Top 20 Domains associated with Allowed Requests
- M/L model ideas [Not Yet Started]
  - Cluster clients based browsing behavior
  - Time series analysis of specific clusters of users
  - Analysis based on web topics (categories)

Data Enrichment Setup


OpenDNS
http
vm
resolver
conf
fetchstats api:
web categories
computer
safe browsing api
Google
SAFE BROWSING