Malucrawl

A Trending Topic Malware Crawler

Week 8 Progress Seminar

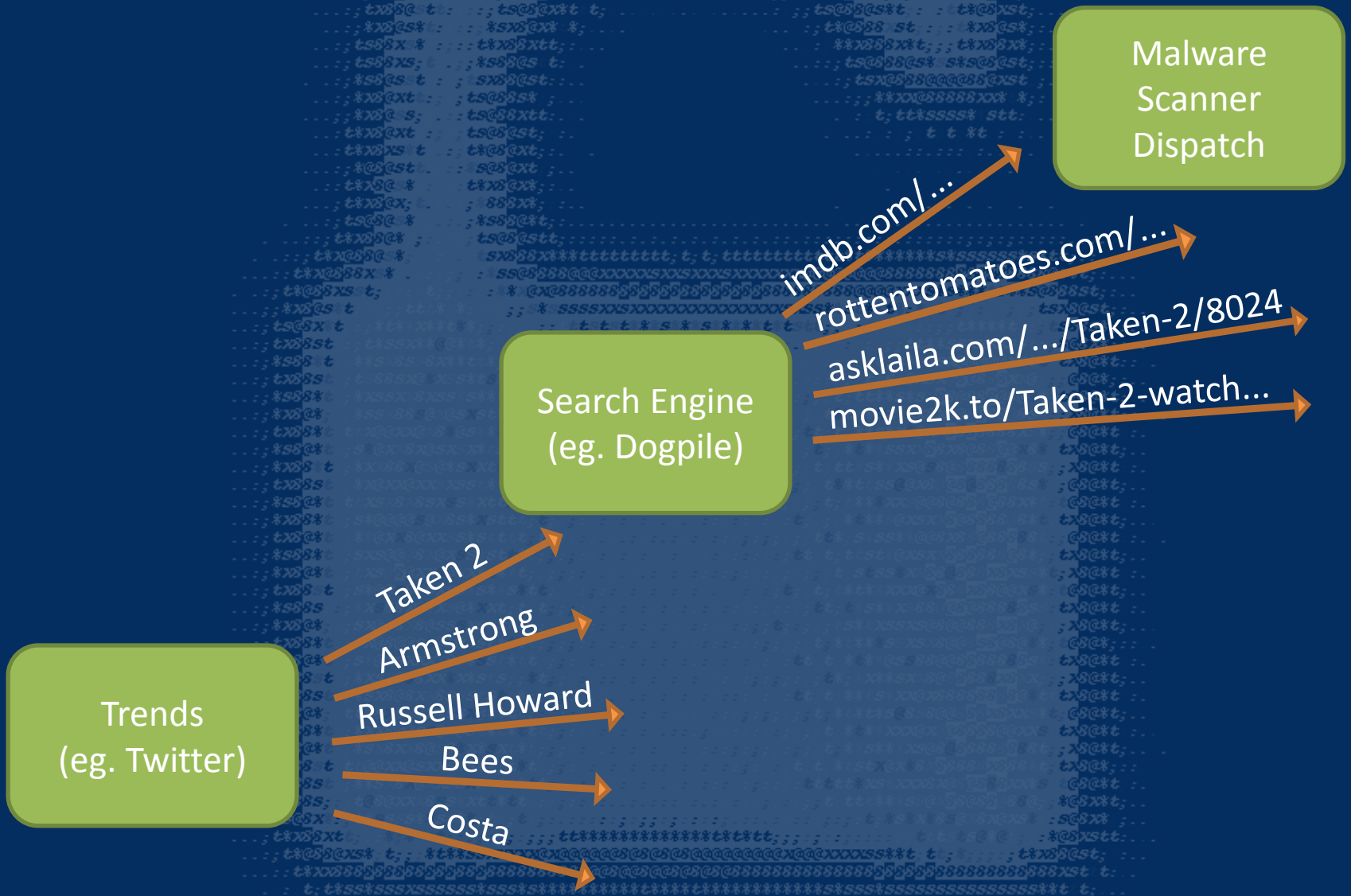
Contents

- Introduction & Architecture Recap
- Implementation of Malware Analysis
  - URL Classification
  - HTML Scanning
  - Internet Explorer under Wine
  - ClamAV HTML Scanning
  - Capture-HPC
  - Good/Bad URL Lists
- Integration and Reporting Database
- Remaining Work
- Questions

Introduction & Recap

- Trending Topics
- Is Malware targeted towards trends?
- Distributed Framework
  - Gather Trends
  - Get URLs associated with trends
  - Analyse URLs for malicious content

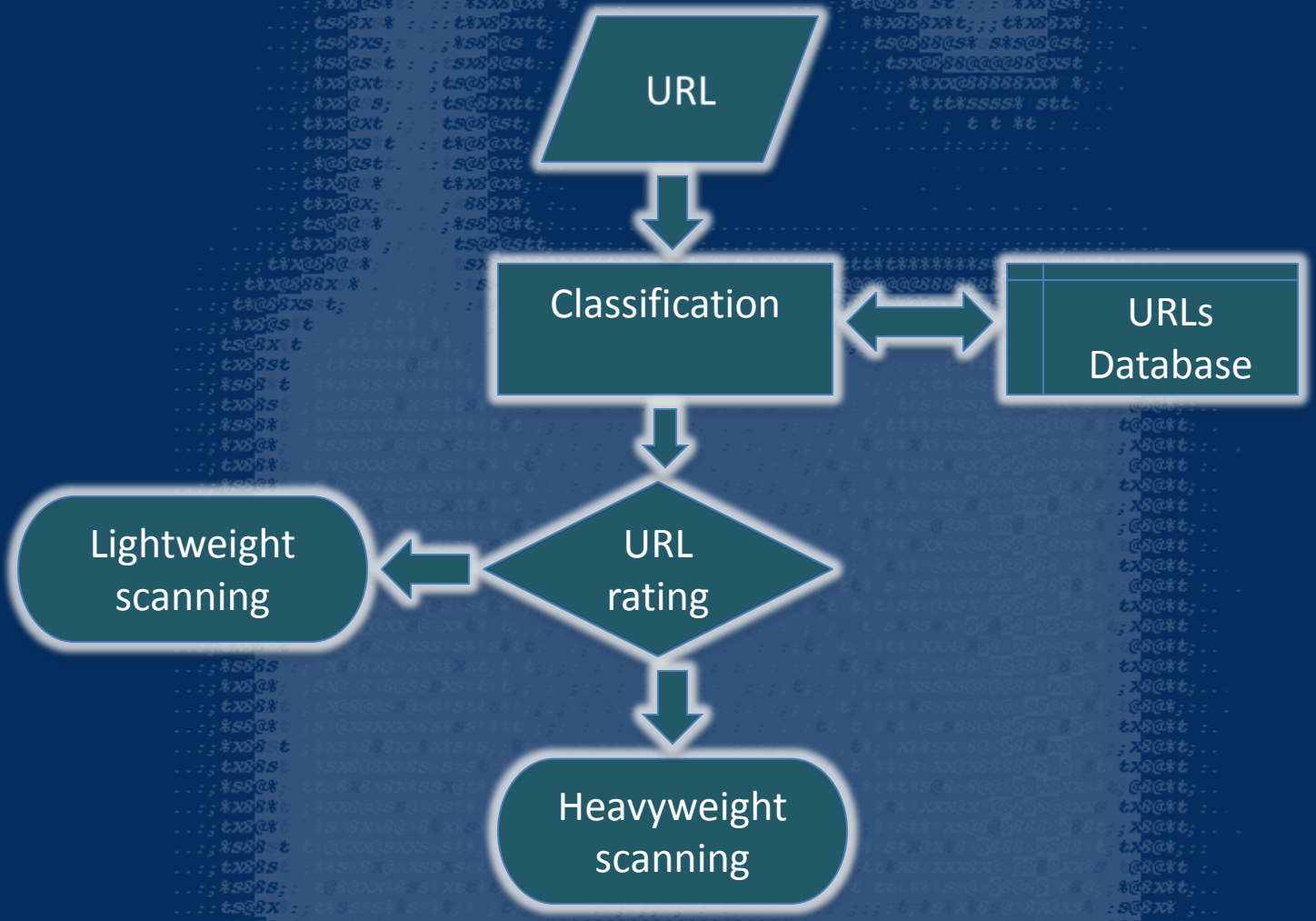
Architecture Summary



Malware Detection Methods

- No interaction malware detection
- Low interaction malware detection
  - Use lightweight or simulated clients to interact with the server.
- High interaction malware detection
  - Use emulated and virtualised clients, very closely resembling a real vulnerable system.
  - Effective at detecting unknown attacks on clients

URL Classification Algorithm



URL Classification

- Two machine-learning methods applied in classification algorithm:
- Levenshtein distance
  - String metric for measuring the difference between two sequences
  - Minimum number of edits needed to transform one string into the other
- K-mean Clustering
  - Data mining
  - Partition *n* observations into *k* clusters
  - Applicable on large amount of data

No

HTML Scanning

- Malware detection based on repetition of the trending keyword:
  - Analyse the content of the given URL
  - Also check content of each hyperlink on the webpage (Crawling)
  - Consider frequently of reappearances of trending keyword
  - Used the result of this algorithm for URL classification database

Low

Internet Explorer under Wine

- High interaction malware detection.
- Open suspicious URLs with Internet Explorer 6 under Wine.
- “Open Source Software for running Windows applications on other operating systems.”
- For each suspicious URL we create an isolated temporary Wine environment then execute an IE instance within it.
- An original copy of Wine-prefix is downloaded and unpacked(or checked via hash if local copy exists) for each URL, to ensure consistency of results.

High

Internet Explorer under Wine

- Wait for some time after opening URL then check for file system changes.
- File system changes are measured by running recursive diff, compared to the original copy before testing.
- White list is used to filter out normal system operations.
- The temporary Wine-prefix is removed after giving results.

High

Clam AV HTML Scan

- Low interaction malware detection.
- With given URL we download the HTML then scan it with ClamAV.
- Then the webpage is crawled and all links in it are also scanned.

- Concurrent execution is achieved with Celery.

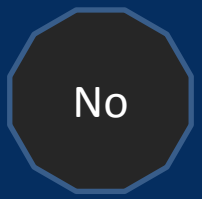
Low

Capture-HPC

- Open source high-interaction “client honeypot”
- Uses VM APIs combined with kernel drivers to automatically browse for malicious sites.
- Slow, but provides high quality emulation of a vulnerable environment
- Required customisation for ECS infrastructure.

Good/Bad URL Lists

- Alexa, Google “Safe Browsing Lookup API”, Web Of Trust “WOT” API
- Make request to Google or WOT, return reformatted result
- Alexa provides daily ZIP of top million results…



Good/Bad URL Lists

- Daily Celerybeat Task
  - Download, extract and store Alexa Zipped CSV in Redis write master
  - Redis read slaves on each worker machine mirror write master
- Tasks check Redis read slave

No

Framework Integration

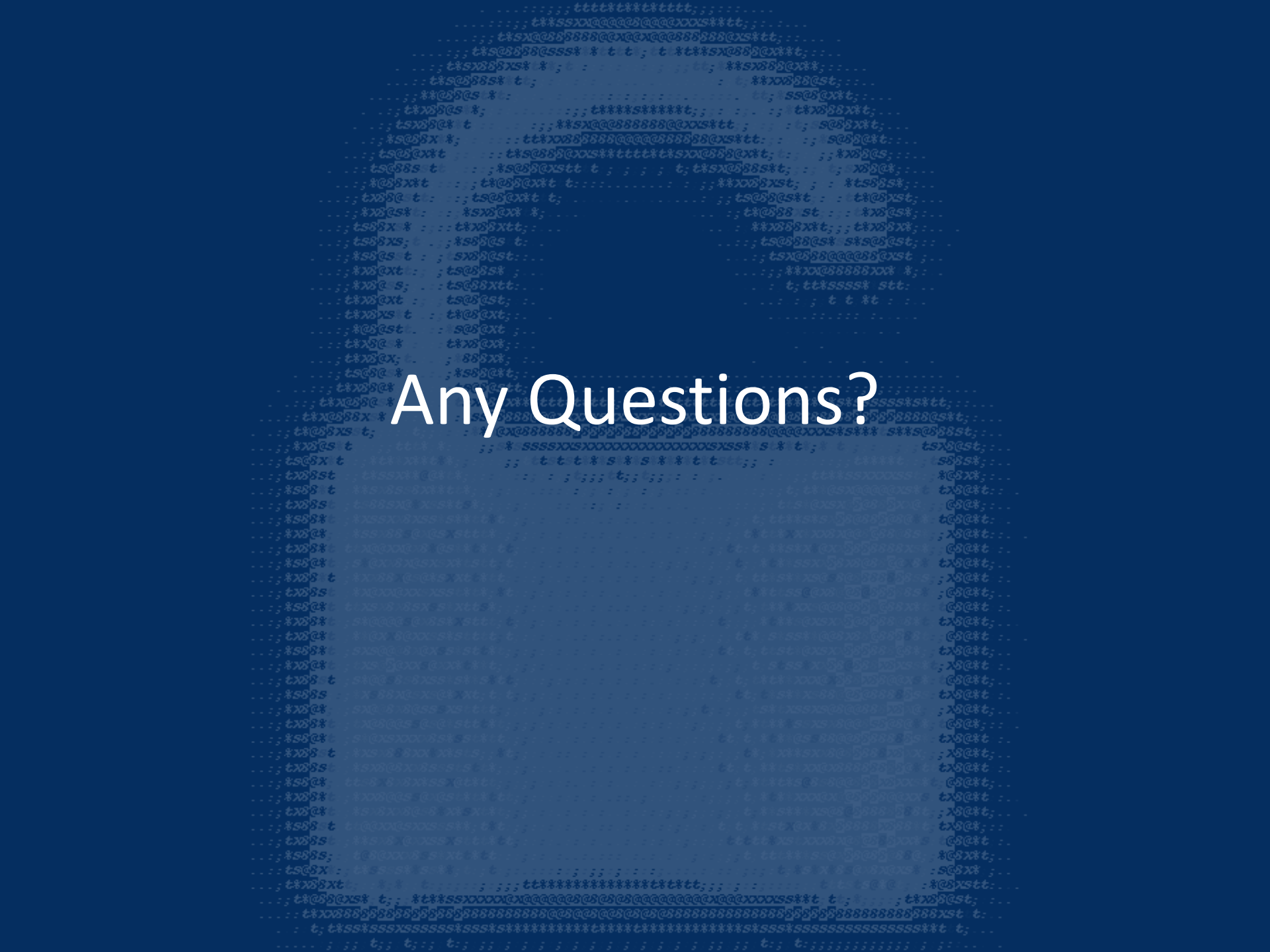
- Celery on top of RabbitMQ
- Celerybeat for daily trend scan/URL list updates
- Calls each task with the result of the previous

Storage of Analysis Results

- Want to use Django ORM for easy Analysis of Results
- Bulk Insert, and strict database schema
  - Any error will result in the entire transaction being thrown out
- Use JSON schema to validate data to be stored

Remaining Work

- Web Reporting interface
  - Graphing
- Compile sample data for report
- Produce Report
  - Latex template already produced
  - Use productivity tools to aid production.


Any Questions?