Predicting Malicious URLs

Proofpoint, Inc.

James Best
Aidan Cheng
Kevin Herrera
Carli Lessard
Vidushi Ojha

Advisor: Elizabeth Sweedyk
Liaisons: Thomas Lynam, Mike Morris

Activity!

apexgames.org/ykxj6/par/factura.zip	*Malicious*
oyunlar1.com/minigames.asp	*Not Malicious*
trac.cs.hmc.edu	*Not Malicious*

proofpoint™

Provide cyber security for companies!

- *Email screening*

Problem

Detecting malicious URLs, and fast!

Clinic Project

Use machine learning to detect malicious URLs — and fast!

Existing Solution

- Filtration technique:
  - Number of appearances in time period
  - Domains passed through
- Sandboxing: sped-up virtual environment

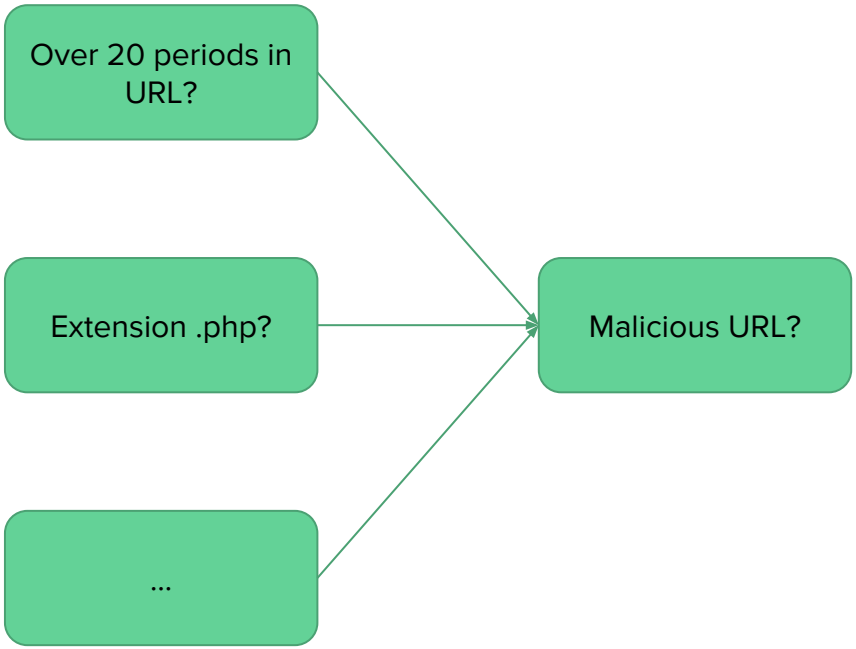
Solution Requirements

- Improve filtration with ML
- Accuracy > 70%
- Process vast numbers of URLs

Classifier Requirements

- Continuous score, 0 to 1
- Cutoffs for malicious or not
- Provide reasoning for score

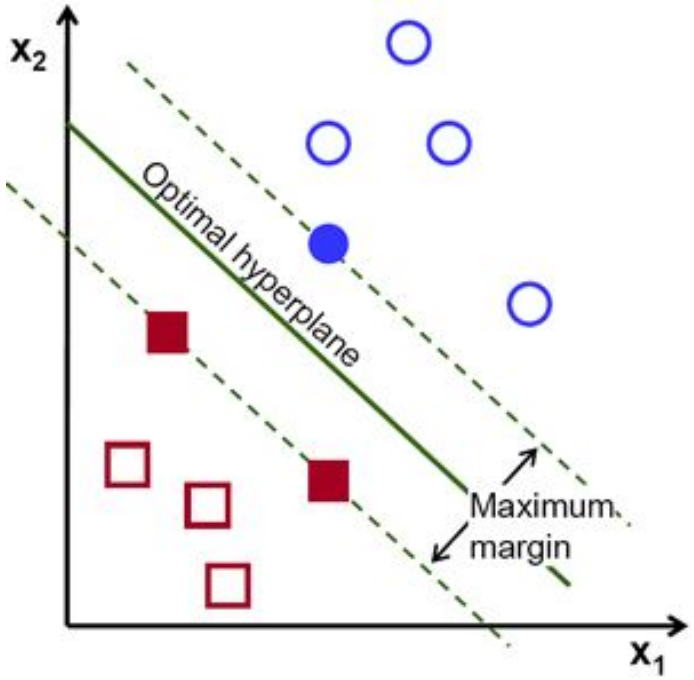
Classifier: Naive Bayes



- Bayes' net encodes *conditional dependence*
- Independent features make it *naive*
- Use probability rules to compute likelihood of URL being malicious

Classifier: Support Vector Machine

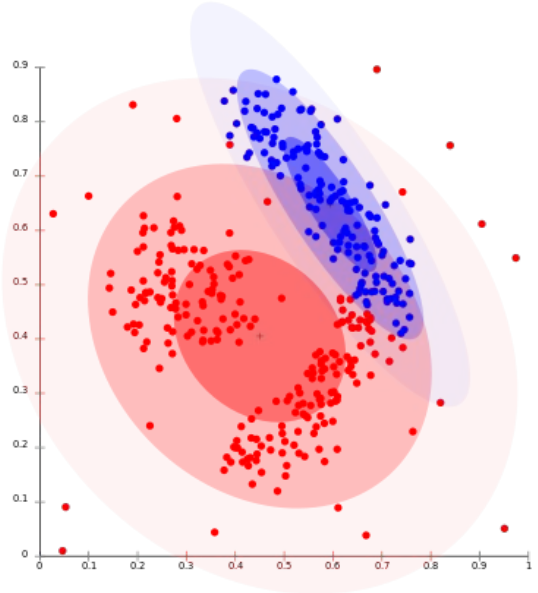
- Binary classifier
- Predicts whether a URL is malicious based on which side of the hyperplane it falls
- Relaxed online SVM has been successful in spam filtering



https://goo.gl/images/x6xoSt

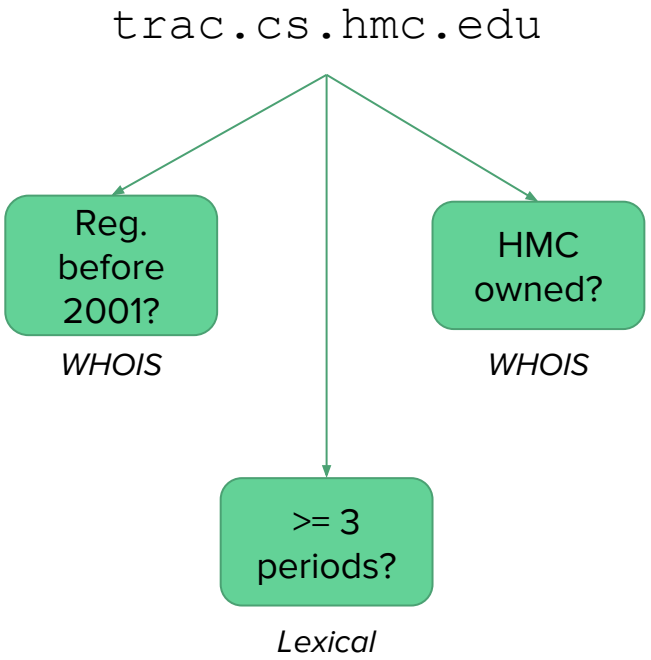
Classifier: Clustering

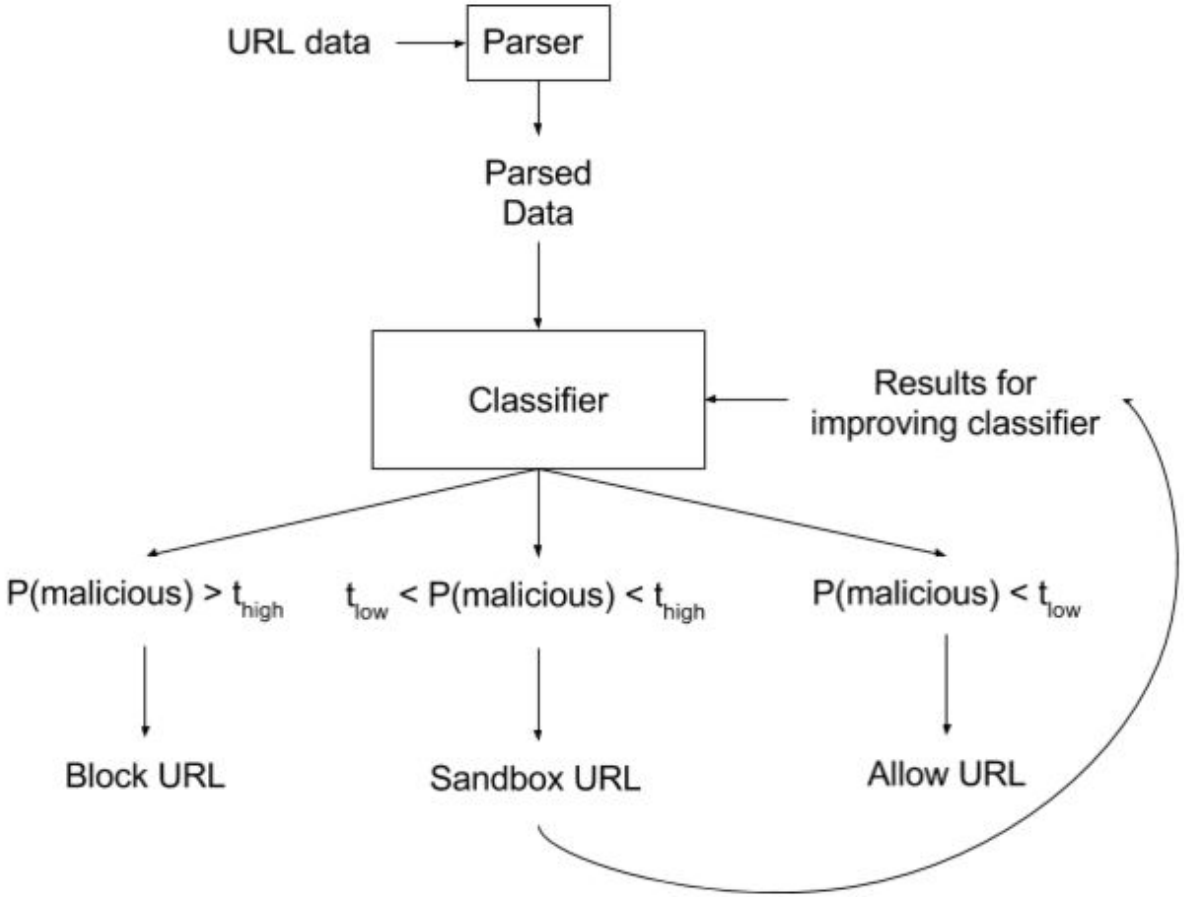
- Partition Data
- Soft Clustering (Fuzzy Clustering)
- Fuzzy C-Means (FCM) Algorithm



Feature Selection

- Existing research: lexical and WHOIS features
- More features = better results, but diminishing marginal returns
- Emphasis on relevant core features




URL data
Parser
Parsed Data
Classifier
Results for improving classifier
P(malicious) > t_high
t_low < P(malicious) < t_high
P(malicious) < t_low
Block URL
Sandbox URL
Allow URL

Questions?
Feedback?