Using Decision Trees to Identify Phishing Sites

Lherisson Medina

Using Decision Trees to Identify Phishing Sites

- Every day more and more people are susceptible to phishing attempts through email and sites they may visit.
- The affects of a successful phishing attempt are long term, and can ruin an individual’s economic and social life for years.

Hackers Hide Cyberattacks in Social Media Posts - The New York Times

The New York Times

TECHNOLOGY

Hackers Hide Cyberattacks in Social Media Posts

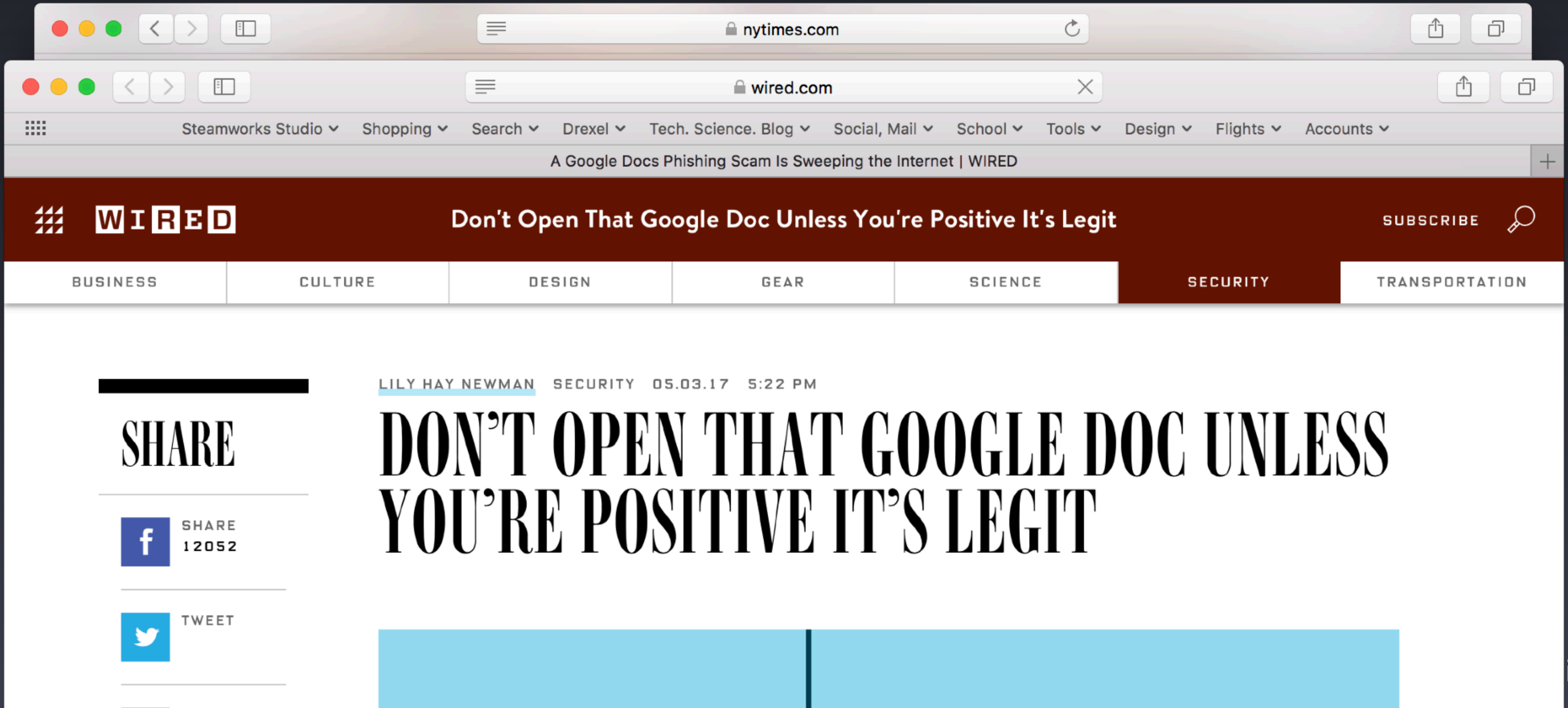
By SHEERA FRENKEL MAY 28, 2017

RELATED COVERAGE

With New Digital Tools, Even Nonexperts Can Wage Cyberattacks

Using Decision Trees to Identify Phishing Sites

- In May, millions of Google Gmail users were hit with a sophisticated phishing attack which took all day to identify and fix
- Estimated 85% of companies are hit with phishing attacks in recent years



nytimes.com

wired.com

A Google Docs Phishing Scam Is Sweeping the Internet | WIRED

Don't Open That Google Doc Unless You're Positive It's Legit

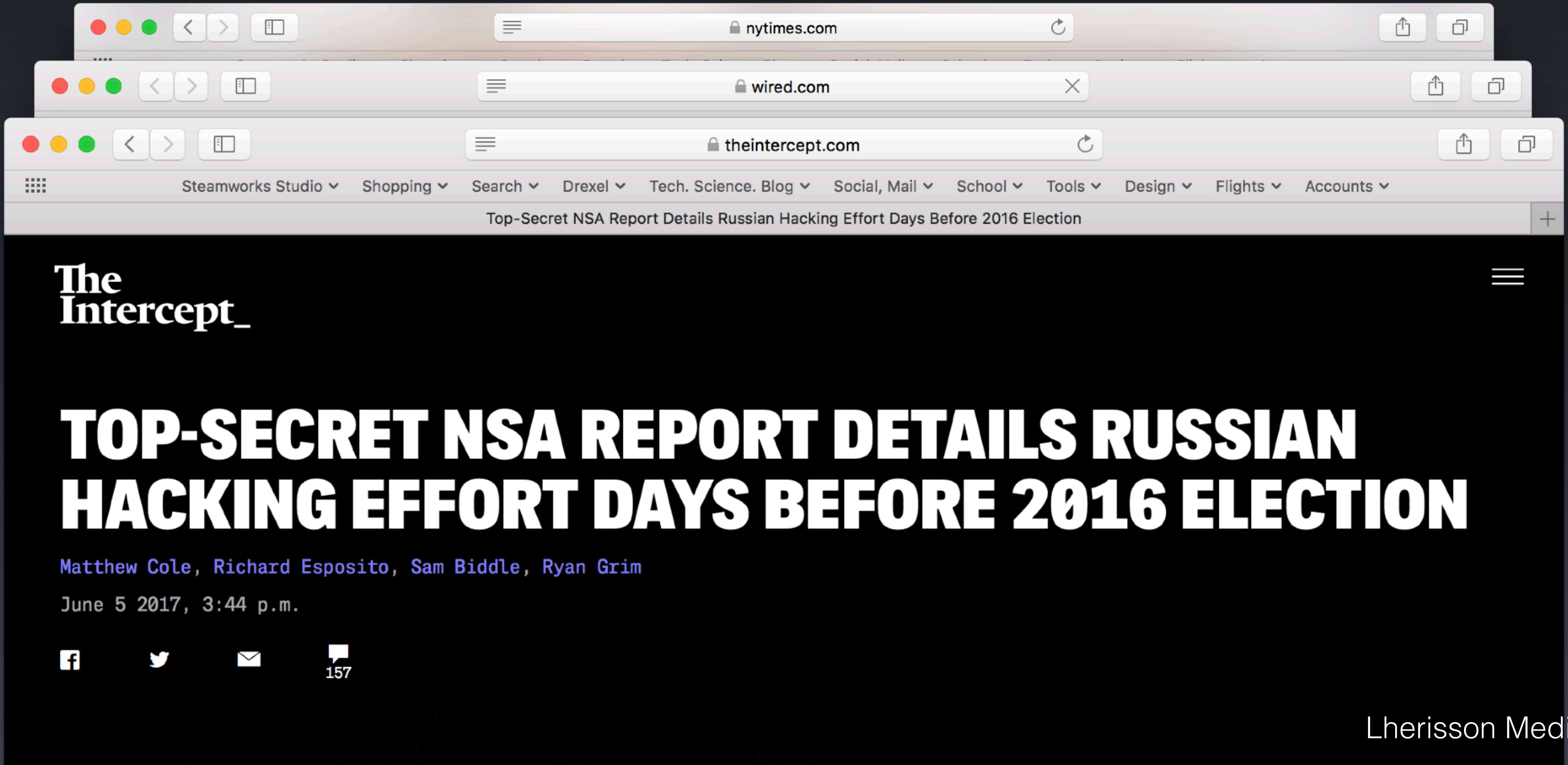
LILY HAY NEWMAN SECURITY 05.03.17 5:22 PM

DON'T OPEN THAT GOOGLE DOC UNLESS YOU'RE POSITIVE IT'S LEGIT

ina

Using Decision Trees to Identify Phishing Sites

- Because of its consequences, it is important to recognize when a site may be attempting to phish data
- These sites are the subject of my term project



Top-Secret NSA Report Details Russian Hacking Effort Days Before 2016 Election

TOP-SECRET NSA REPORT DETAILS RUSSIAN HACKING EFFORT DAYS BEFORE 2016 ELECTION

Matthew Cole, Richard Esposito, Sam Biddle, Ryan Grim

June 5 2017, 3:44 p.m.

Lherisson Medina

The Data

- The data was downloaded form the Machine Learning Repository at UC Irvine

- Based on attributes gathered by Auckland Institute of Studies

9 Available Features

SFH	HasPopUp	SSL	Request URL	Anchor URL
Web Traffic	URL Length	Domain Age	IP Address in URL	

- Feature values have been normalized to [-1, 0, 1] where

  -1 - Phishy
  0 - Suspicious
  1 - Legitimate

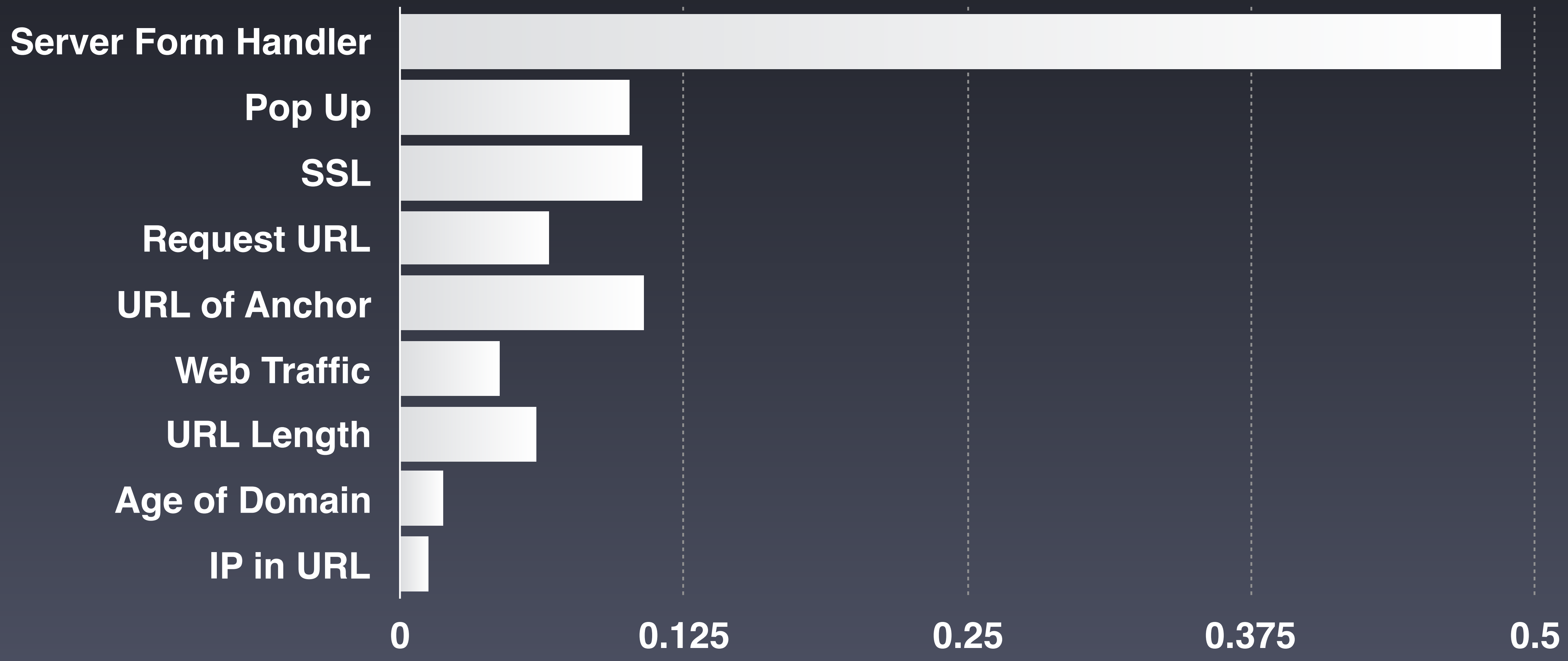
Approach

- I used a supervised learning approach
- Decision Trees used to create a predictive model based on the data
- Out of 1,353 Data Points, a random 25% used as test size
    - 75% used for training

Approach (Feature Selection)

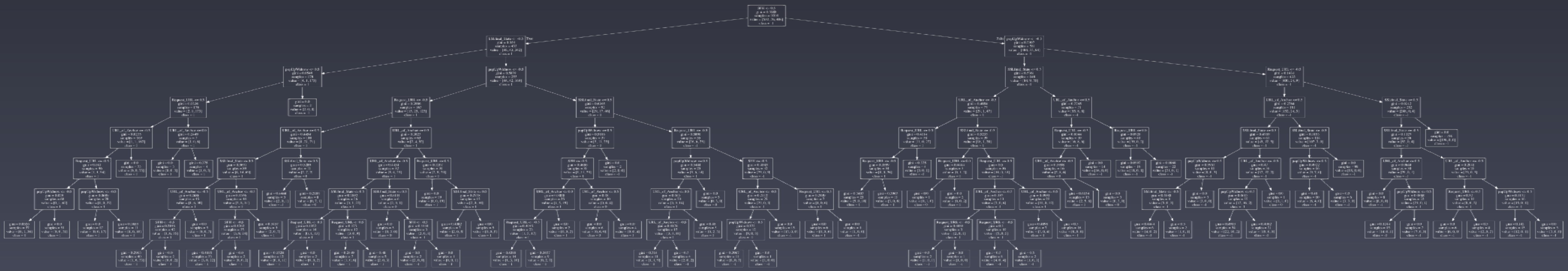
- The Server Form Handler is the most informative feature
- ‘Domain age’ and ‘IP in URL’ are the least informative

Feature Importance



Results

- With top 5 most informative features selected
  - ~87% Accuracy
- Can be improved
  - Encountered overfitting
- Regression Trees also tested



Further Work

- Classification Trees only one approach
- Consider comparing with other approaches
  - Neural Net, Nearest Neighbor, etc
- Optimize tree (Pruning?) to increase Accuracy and decrease overfitting

Similar Work

- Black-Lists and White-Lists are more generally used to track Phishing and Legitimate Sites
- Norman Sadeh et al. looked at Emails and URLs to try and classify whether they are Phishing or Legitimate using Random Forest Classifier
- May 31st Google started delaying and flagging emails having predictable patterns to Phishing emails
  - No info on Algorithm

Questions

Thank You