

Automatically Detecting Vulnerable Sites Before They Turn Malicious

Kyle Soska

Carnegie Mellon University

ECE / Cylab

ksoska@cmu.edu

Nicolas Christin

Carnegie Mellon University

ECE / Cylab

nicolasc@cmu.edu

Problem Setting

- Adversaries compromise websites
 - Economically Rational – monetize compromises
 - Neutral to victims
 - Maximize volume, efficiency, profits
 - Hacktivist – promote social, political, or religious agenda
 - Targeted attacks
 - Low volume



Economically Rational Adversary

- Decisions always maximize profit
- Probabilistic Polynomial Time
 - Cannot break standard crypto, session cookies, hashes, etc.
- Does not control significant portion of web
 - Cannot perform adversarial machine learning attacks by poisoning a random sample of the web
- Able to exploit vulnerable web software

Mode of Operation

Step 1: Find bug or vulnerability in popular web software or content management system (CMS)

Step 2: Enumerate sites containing vulnerability

Step 3: Exploit vulnerable sites

Step 4: Monetize and profit

Google



Google Search

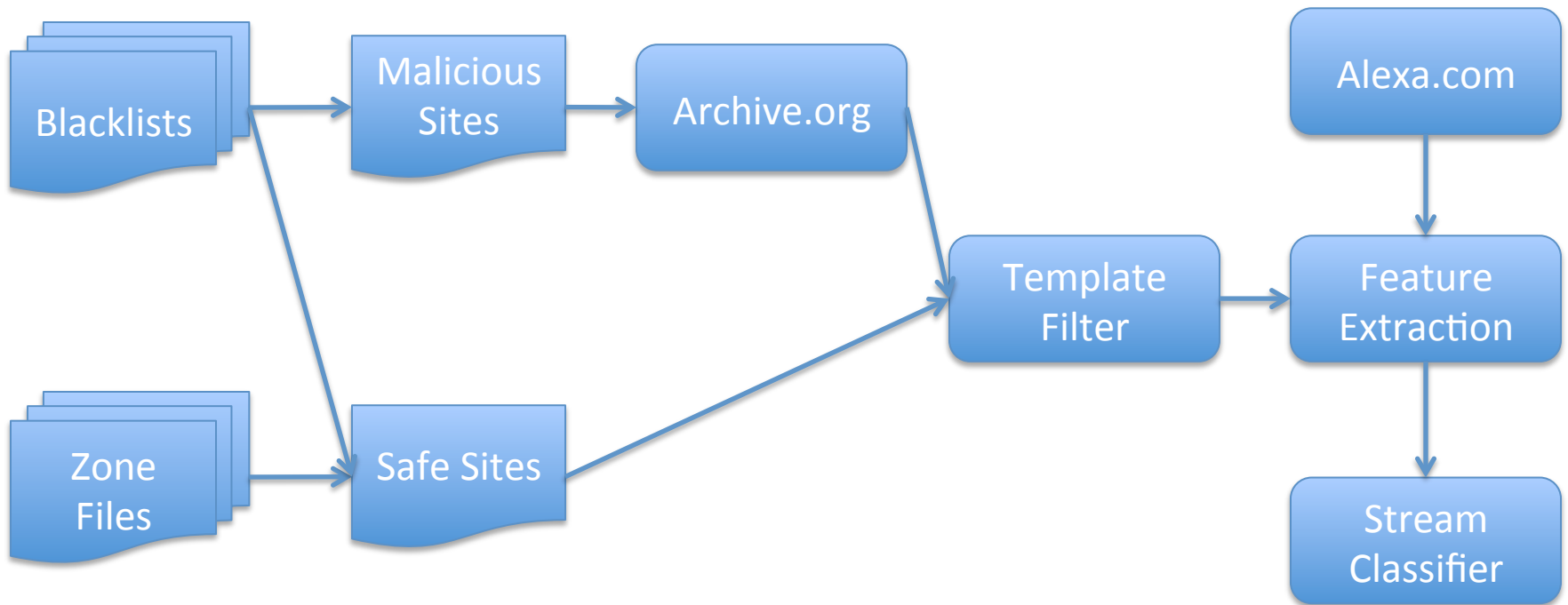
I'm Feeling Lucky

Problem and Goal

- Existing approaches detect if a **webpage** is **already** malicious

- Is it possible to predict if a non-malicious **website** will become malicious in the **future**?
 - What would such a system look like?
 - What requirements are imposed on such a system?
 - What are the fundamental limitations?

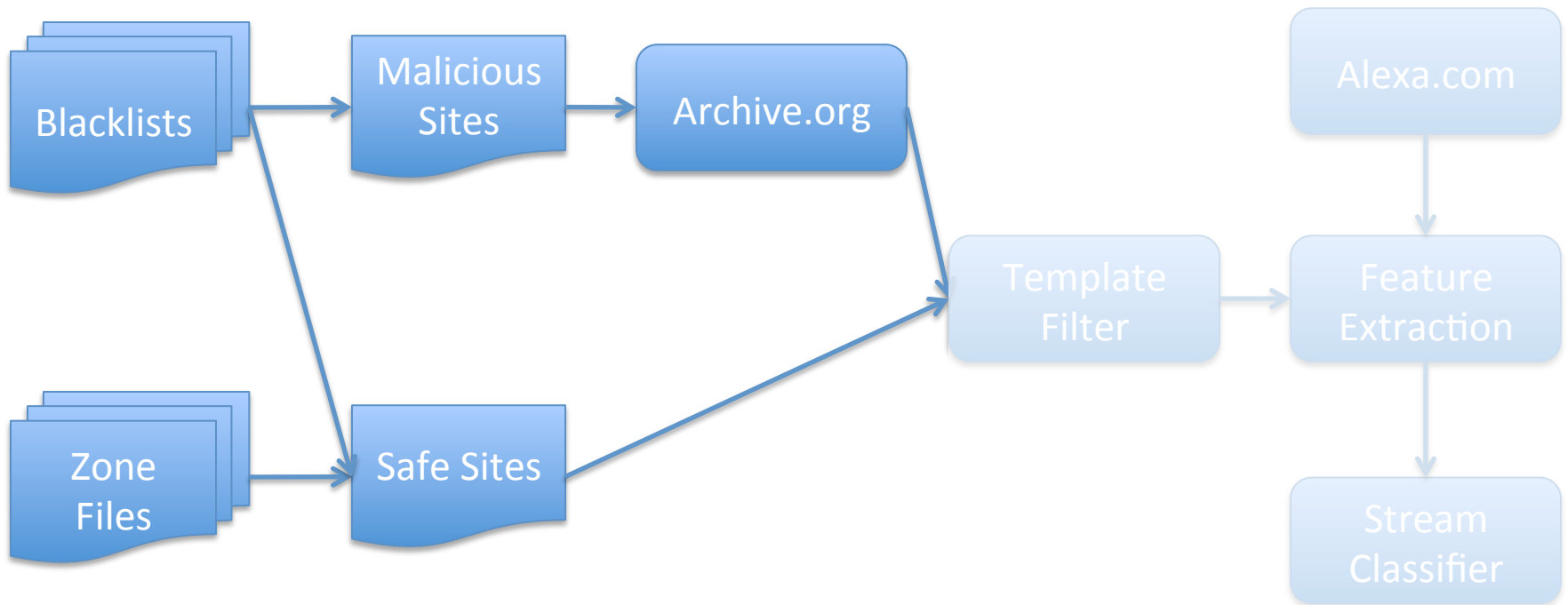
System Design



System Properties

- Efficiency
 - Internet dataset
- Interpretability
 - Need to build intuition about why the site will become compromised
- Robustness to Imbalanced Data
 - Far more benign examples than malicious ones
- Robustness To Mislabeled Data
 - Blacklists may contain errors or be incomplete
- Adaptive
 - Internet is a concept drifting, requires active adaptation

Dataset



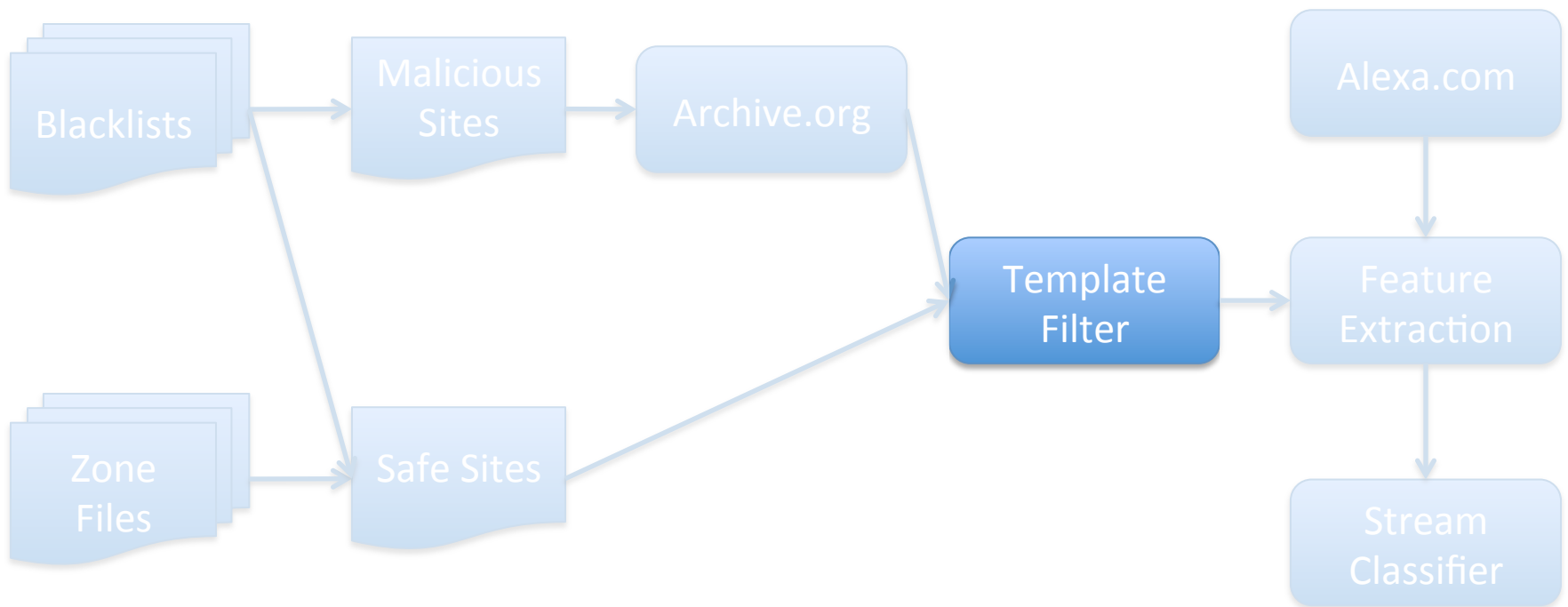
Dataset

Type	Instances	Archived Instances	% Archived
PhishTank	91,555	34,922	38.1
Search Redirection*	16,173	14,425	89.2
.com Zone Files	336,671	336,671	N/A

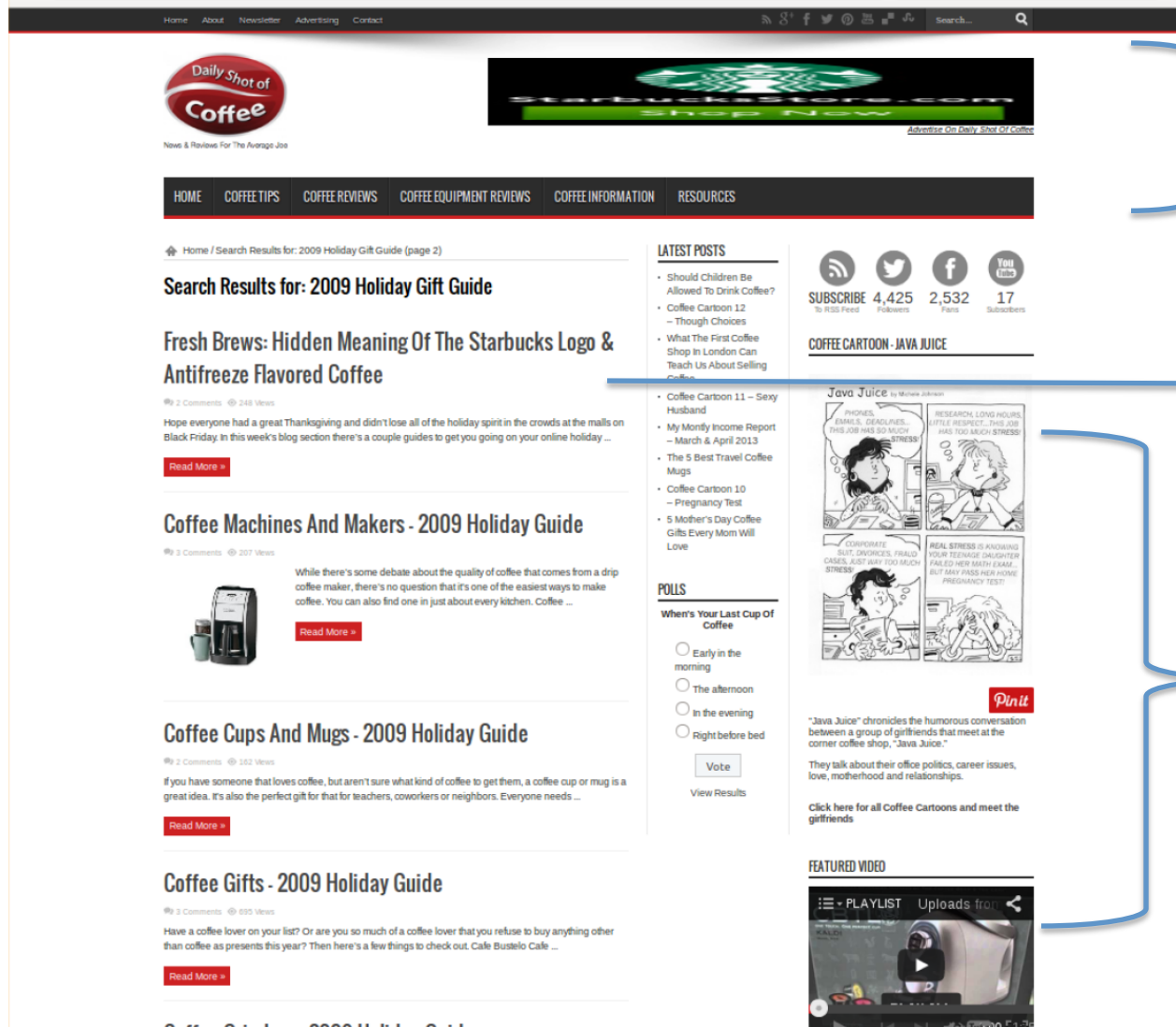
- PhishTank: Feb 2013 – Dec 2013
- Search Redirection: Oct 2011 – Sept 2013
- Zone Files: Feb 2010 – Sept 2013

* Leontiadis et al., 2014

Filtering



Filtering



The screenshot shows the homepage of 'Daily Shot of Coffee'. At the top, there is a navigation bar with links for Home, About, Newsletter, Advertising, and Contact. Below this is a search bar and a main menu with categories like HOME, COFFEE TIPS, COFFEE REVIEWS, etc. The main content area displays search results for '2009 Holiday Gift Guide', featuring several articles such as 'Fresh Brews: Hidden Meaning Of The Starbucks Logo & Antifreeze Flavored Coffee', 'Coffee Machines And Makers - 2009 Holiday Guide', 'Coffee Cups And Mugs - 2009 Holiday Guide', and 'Coffee Gifts - 2009 Holiday Guide'. On the right side, there are sections for 'LATEST POSTS', 'SOCIAL MEDIA LINKS' (RSS, Twitter, Facebook, YouTube), 'COFFEE CARTOON - JAVA JUICE', and a 'POLL' titled 'When's Your Last Cup of Coffee?'. At the bottom right, there is a 'FEATURED VIDEO' section with a video player.

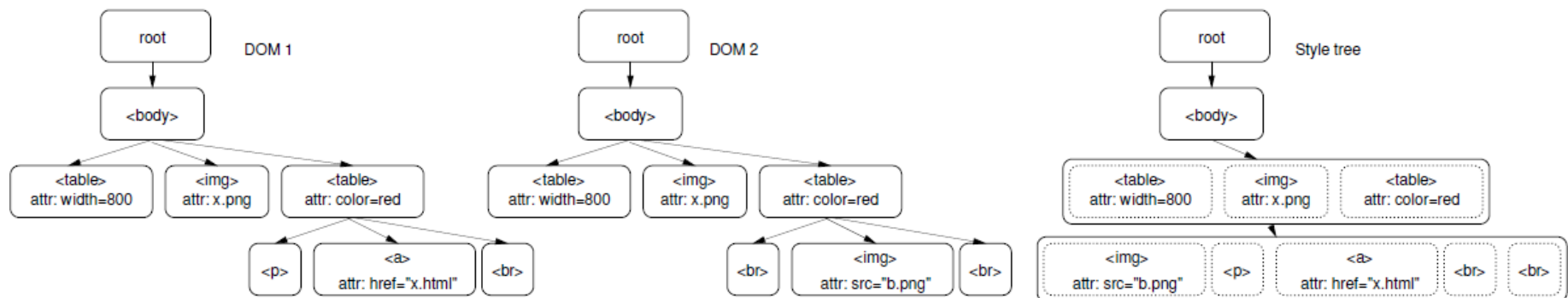
Navigation

User Content

Social Media Links



Filtering – Based on [Yi et al., 2003]

- Compute entropy-like heuristic “Composite Importance” ($CmpImp \in [0, 1]$) for each element on a page
- Remove elements above a fixed threshold



Original Page

Home About Newsletter Advertising Contact
Search...

Advertise On Daily Shot of Coffee

HOME
COFFEE TIPS
COFFEE REVIEWS
COFFEE EQUIPMENT REVIEWS
COFFEE INFORMATION
RESOURCES

Home / Search Results for: 2009 Holiday Gift Guide (page 2)

Search Results for: 2009 Holiday Gift Guide

Fresh Brews: Hidden Meaning Of The Starbucks Logo & Antifreeze Flavored Coffee


2 Comments @ 248 Views

Hope everyone had a great Thanksgiving and didn't lose all of the holiday spirit in the crowds at the malls on Black Friday. In this week's blog section there's a couple guides to get you going on your online holiday ...

Read More >

Coffee Machines And Makers - 2009 Holiday Guide

3 Comments @ 207 Views



While there's some debate about the quality of coffee that comes from a drip coffee maker, there's no question that it's one of the easiest ways to make coffee. You can also find one in just about every kitchen. Coffee ...

Read More >

Coffee Cups And Mugs - 2009 Holiday Guide

2 Comments @ 182 Views

If you have someone that loves coffee, but aren't sure what kind of coffee to get them, a coffee cup or mug is a great idea. It's also the perfect gift for that for teachers, coworkers or neighbors. Everyone needs ...

Read More >

Coffee Gifts - 2009 Holiday Guide

3 Comments @ 695 Views

Have a coffee lover on your list? Or are you so much of a coffee lover that you refuse to buy anything other than coffee as presents this year? Then here's a few things to check out. Cafe Bustelo Cafe ...

Read More >

LATEST POSTS

- Should Children Be Allowed To Drink Coffee?
- Coffee Cartoon 12 – Tough Choices
- What The First Coffee Shop In London Can Teach Us About Selling Coffee
- Coffee Cartoon 11 – Sexy Husband
- My Monthly Income Report – March & April 2013
- The 5 Best Travel Coffee Mugs
- Coffee Cartoon 10 – Pregnancy Test
- 5 Mother's Day Coffee Gifts Every Mom Will Love

POLLS

When's Your Last Cup Of Coffee

Early in the morning
 The afternoon
 In the evening
 Right before bed

Vote

View Results

RSS

SUBSCRIBE 4,425 To RSS Feed

Twitter

FOLLOWERS 2,453

Facebook



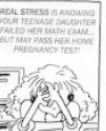
FANS 17

YouTube

SUBSCRIBERS 17

COFFEE CARTOON - JAVA JUICE

Java Juice by Mark Johnson

 <p>PHONES, EMAILS, SKYPE... THIS JOB HAS SO MUCH STRESS</p>	 <p>RESEARCH, LONG HOURS, LITTLE RESPECT... THIS JOB HAS TOO MUCH STRESS</p>
 <p>CORPORATE SUIT, DRINKING, FRUSTRATED CASES, JUST WAY TOO MUCH STRESS</p>	 <p>REAL STRESS IS KNOWING YOUR TEENAGE DAUGHTER FAILED HER MATH EXAM... BUT SAYS PASS HER PHONE PREGNANCY TEST</p>

Pin it

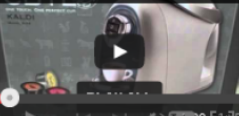
"Java Juice" chronicles the humorous conversation between a group of girlfriends that meet at the corner coffee shop, "Java Juice."

They talk about their office politics, career issues, love, motherhood and relationships.

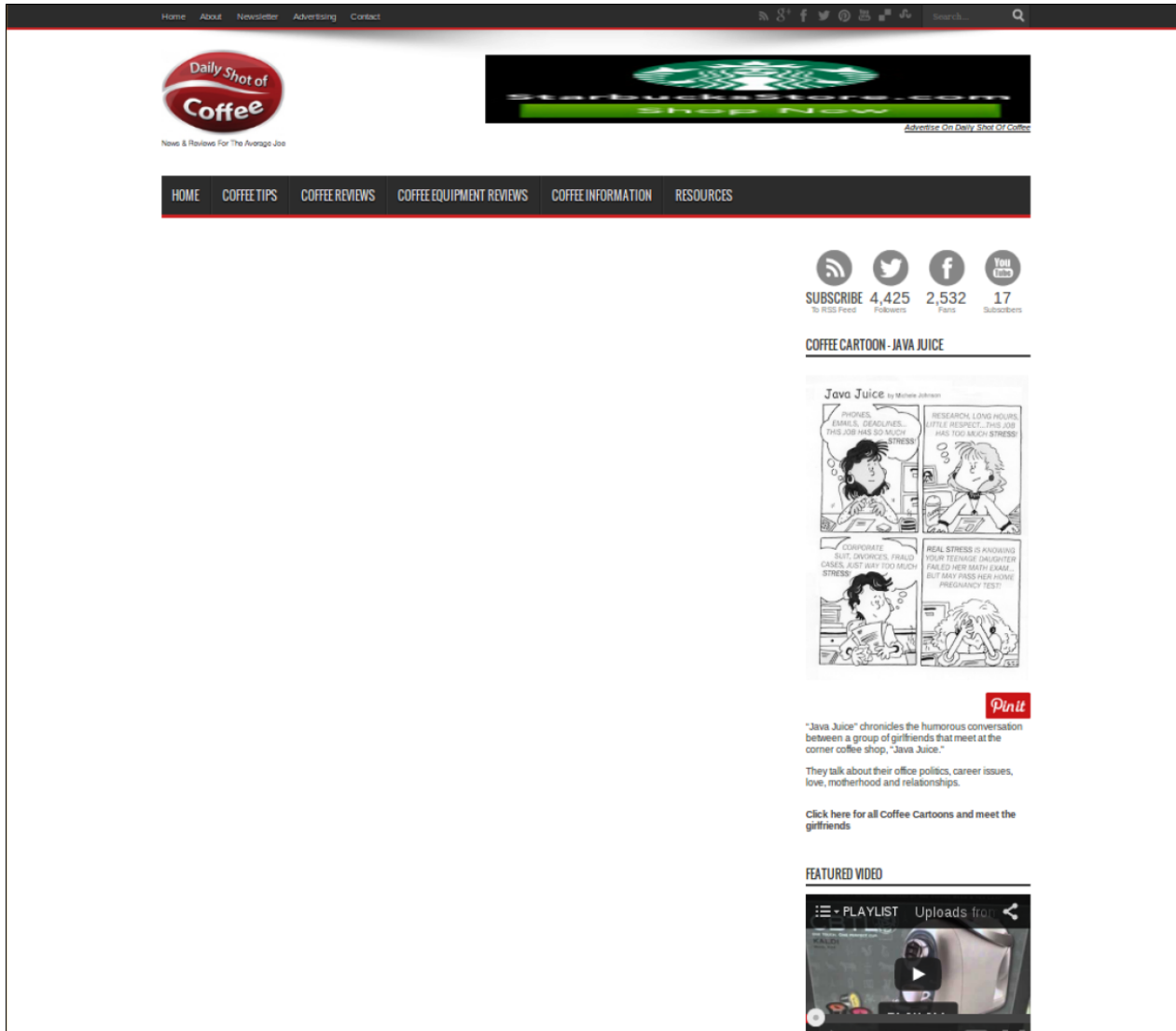
Click here for all Coffee Cartoons and meet the girlfriends

FEATURED VIDEO

PLAYLIST Uploads from



Threshold = 0.99



The screenshot shows the homepage of the 'Daily Shot of Coffee' website. At the top, there is a navigation bar with links for Home, About, Newsletter, Advertising, and Contact. Below this is a search bar and social media icons. The main header features the 'Daily Shot of Coffee' logo and a banner image of a coffee machine. A navigation menu below the header includes links for Home, Coffee Tips, Coffee Reviews, Coffee Equipment Reviews, Coffee Information, and Resources. On the right side, there are social media statistics for RSS (4,425), Twitter (2,532), Facebook (17), and YouTube (17). The main content area features a section titled 'COFFEE CARTOON - JAVA JUICE' with a comic strip by Mike Johnson. The comic strip consists of four panels where two women discuss various stressors in their lives, such as work, research, and pregnancy. Below the comic is a 'Pin it' button and a description of the cartoon. At the bottom, there is a 'FEATURED VIDEO' section with a video player showing a coffee machine.

Home About Newsletter Advertising Contact

Search

Daily Shot of Coffee
News & Reviews For The Average Joe

Advertisement On Daily Shot of Coffee

HOME COFFEE TIPS COFFEE REVIEWS COFFEE EQUIPMENT REVIEWS COFFEE INFORMATION RESOURCES

SUBSCRIBE 4,425 To RSS Feed
FOLLOWERS 2,532
FANS 17
SUBSCRIBERS 17

COFFEE CARTOON - JAVA JUICE

Java Juice by Mike Johnson

PHONE, EMAIL, DEADLINES, THIS JOB HAS SO MUCH STRESS

RESEARCH, LONG HOURS, LITTLE RESPECT... THIS JOB HAS TOO MUCH STRESS

CORPORATE SUIT, DIVORCES, PRALD CASES, JUST WHY TOO MUCH STRESS

REAL STRESS IS KNOWING YOUR TEENAGE DAUGHTER FAILED HER MATH EXAM... BUT SAYS PASSES HER PREGNANCY TEST!

Pin it

"Java Juice" chronicles the humorous conversation between a group of girlfriends that meet at the corner coffee shop, "Java Juice."

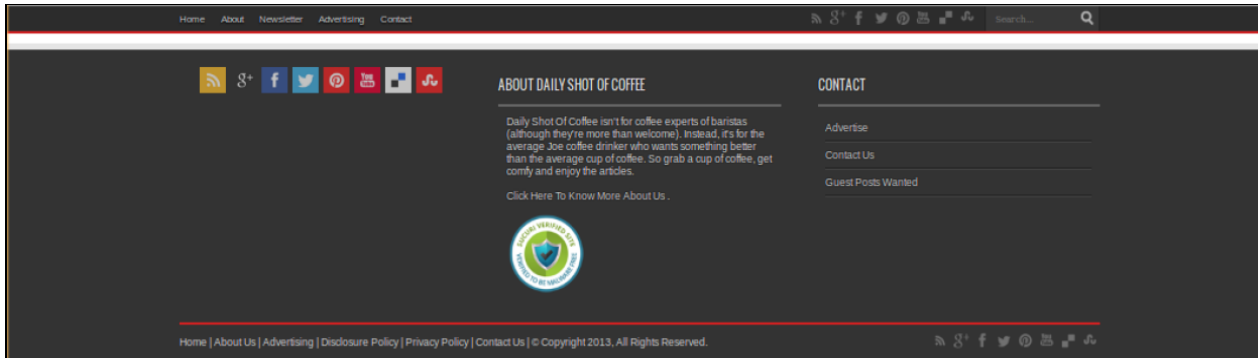
They talk about their office politics, career issues, love, motherhood and relationships.

Click here for all Coffee Cartoons and meet the girlfriends

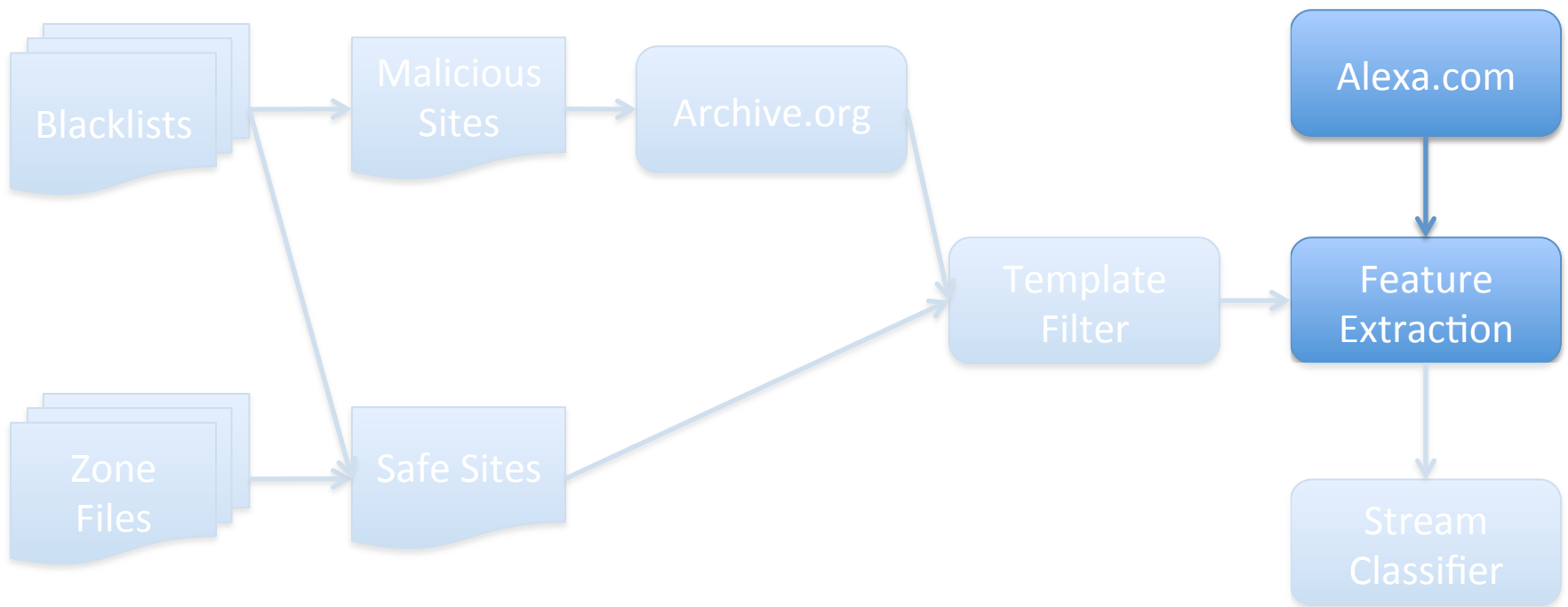
FEATURED VIDEO

PLAYLIST Uploads from

Threshold = 0.1



Feature Extraction



Feature Set

- Traffic Features
 - Site Rank
 - Links into site
 - Load Percentile
 - ...more
- Content Features
 - HTML Tags (type, content, attributes)

Dynamic Features

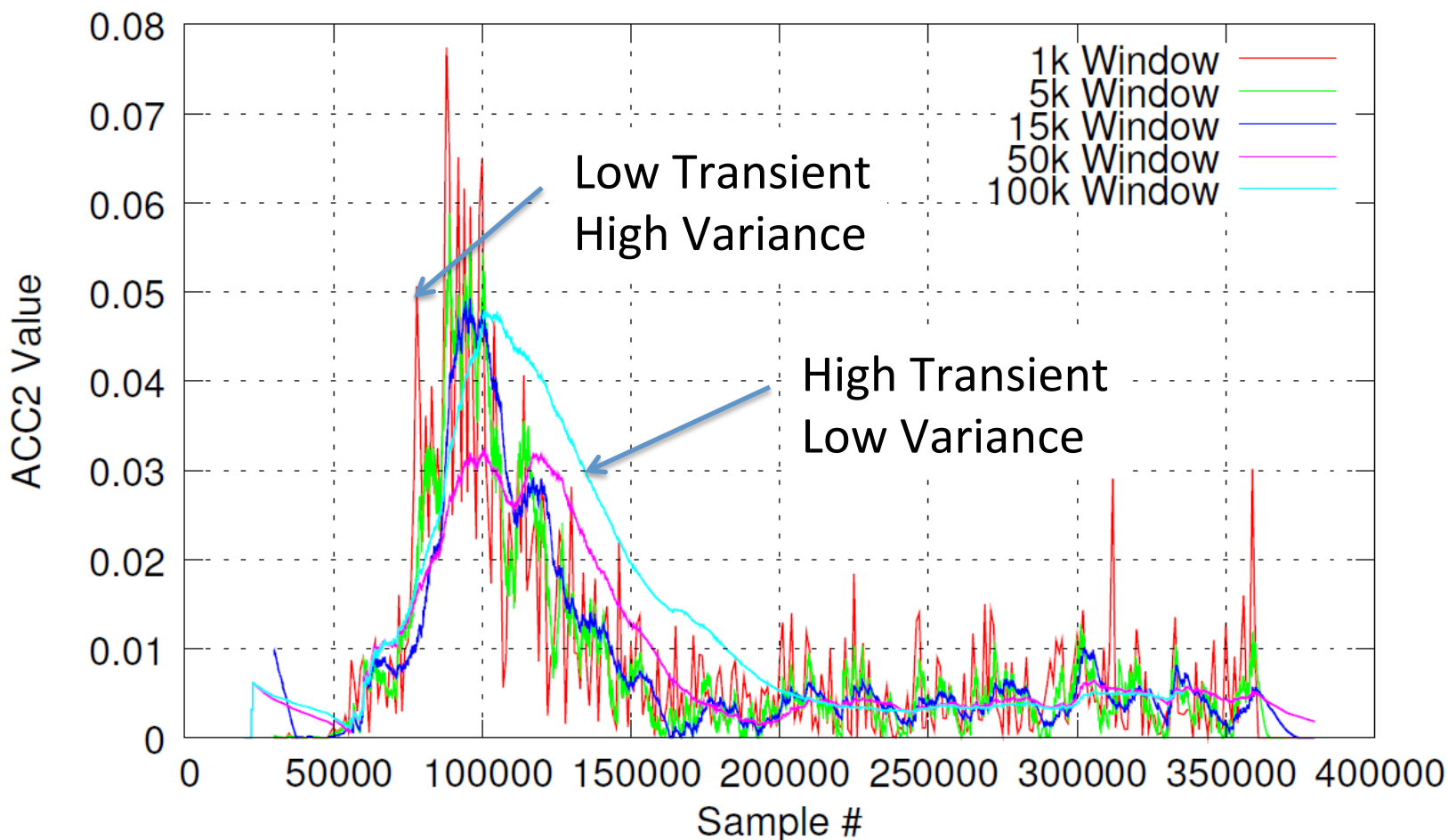
- Millions of unique HTML tags (including content)
- Solution: order tags by some statistic, select top N
 - ACC2 based on [Foreman, 2003]
 - Let \mathcal{B} , \mathcal{M} denote the set of benign and malicious sites respectively, w the set of tags from a site, then ACC2 for a tag x can be defined as:

$$s(x) = \frac{|\{w: x \in w, w \in \mathcal{M}\}|}{|\mathcal{M}|} - \frac{|\{w: x \in w, w \in \mathcal{B}\}|}{|\mathcal{B}|}$$

Prominent Features After 90,000 Samples

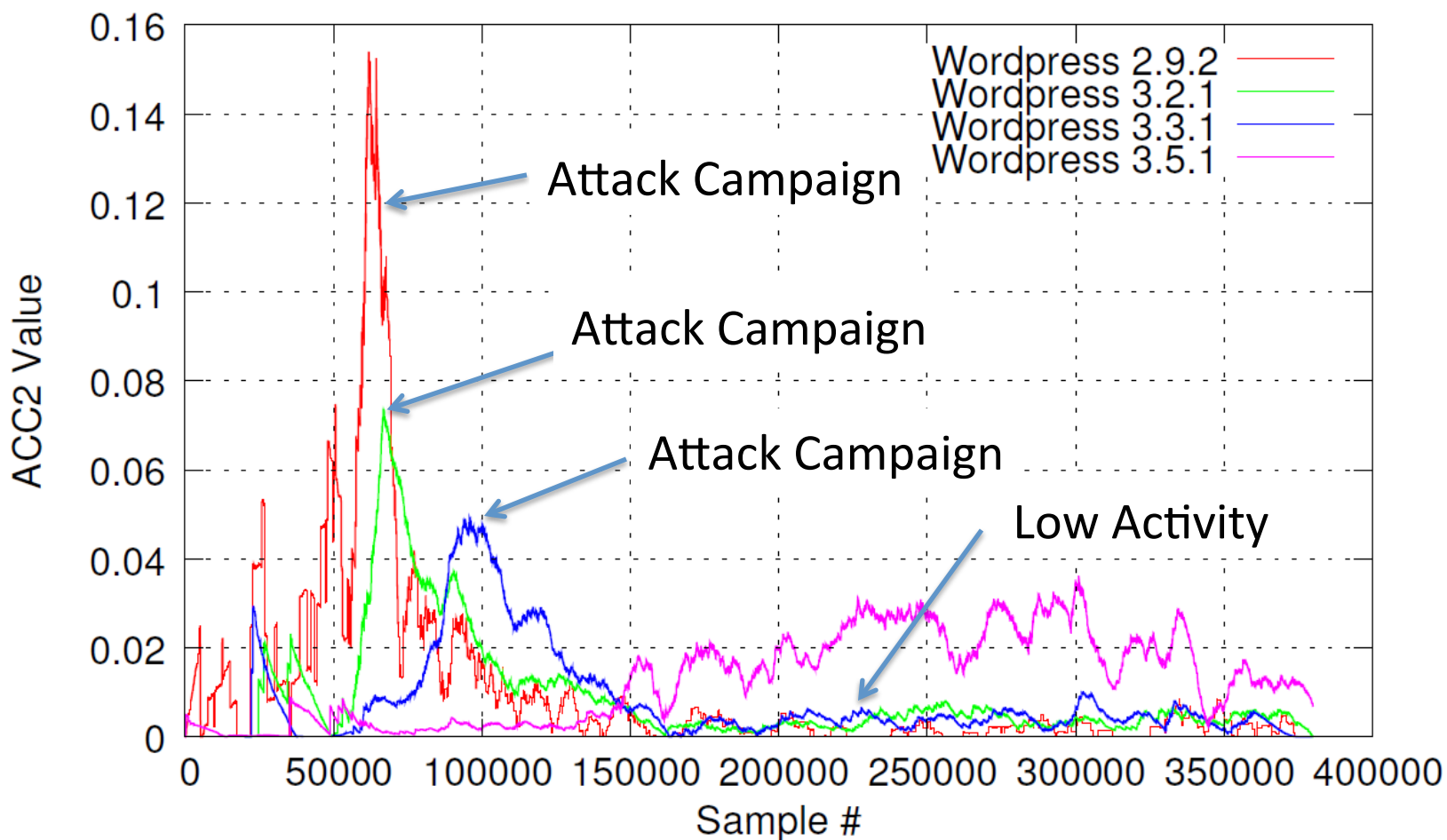
Feature	Statistic Value
meta{'content': 'Wordpress 3.2.1', 'name': 'generator'}	0.0569
ul{'class': ['xoxo', 'blogroll']}	0.0446
You can start editing here.	0.0421
meta{'content': 'Wordpress 3.3.1', 'name': 'generator'}	0.0268
/all in one seo pack	0.0252
span{'class': ['breadcrumbs', 'pathway']}	0.0226
If Comments are open, but there are no comments.	0.0222
div{'id': 'content_disclaimer'}	0.0039

Varying Window Sizes

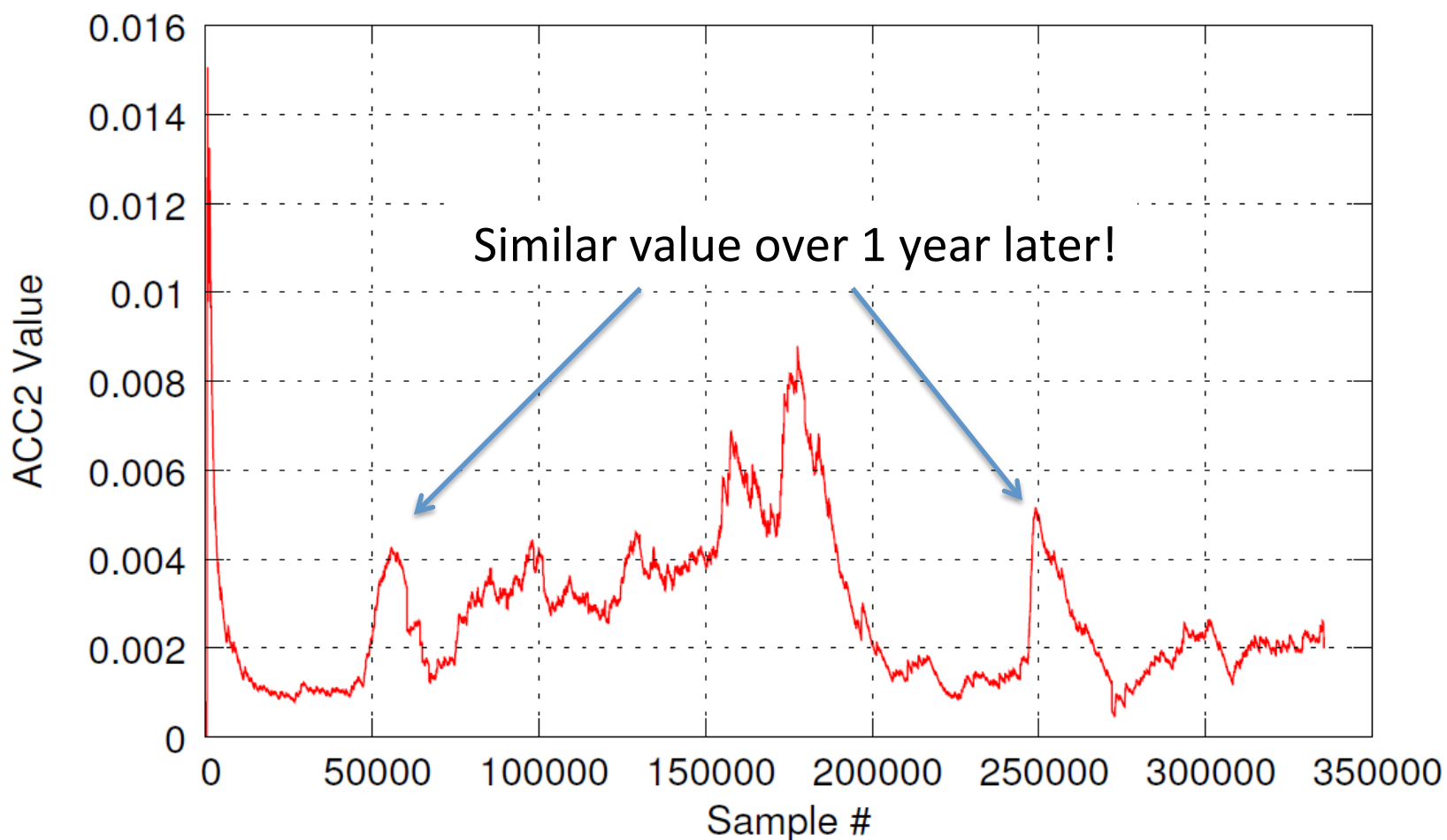


Feature: meta{'content': 'Wordpress 3.2.1', 'name': 'generator'}

CMS Evolution

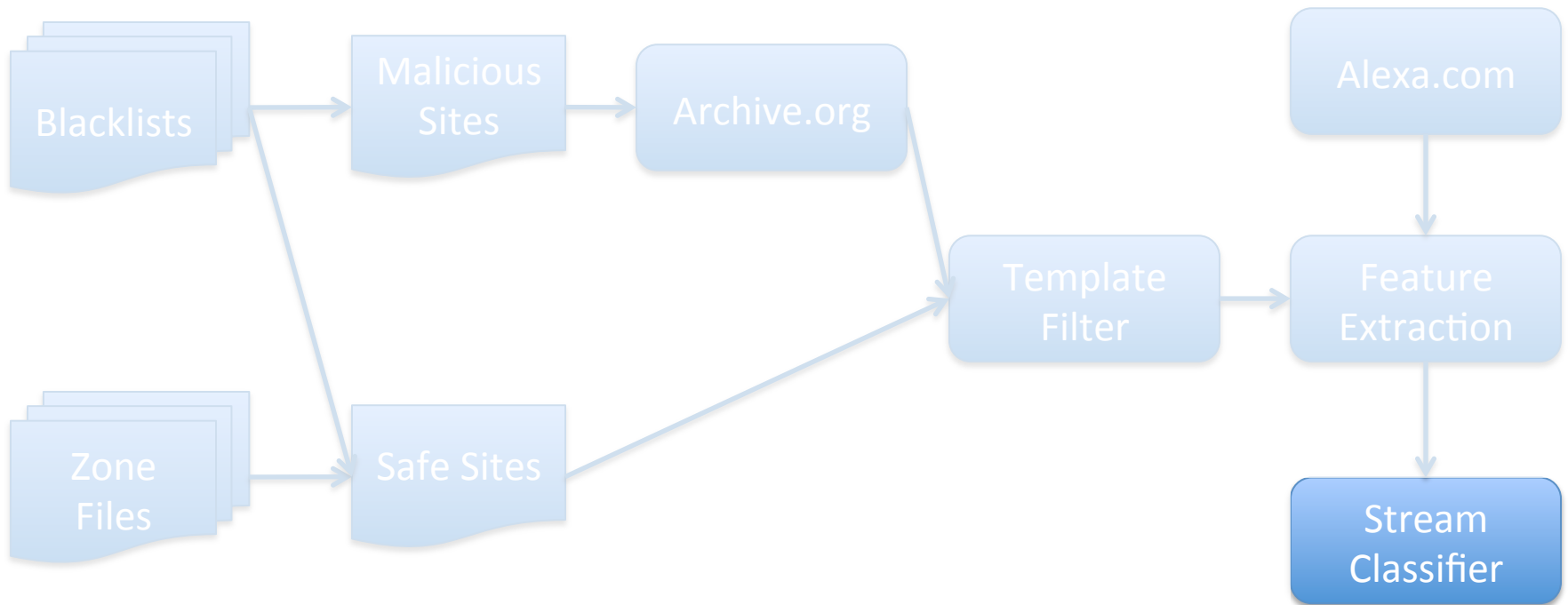


Parking Page Feature



Feature: div{'id': 'content_disclaimer'}

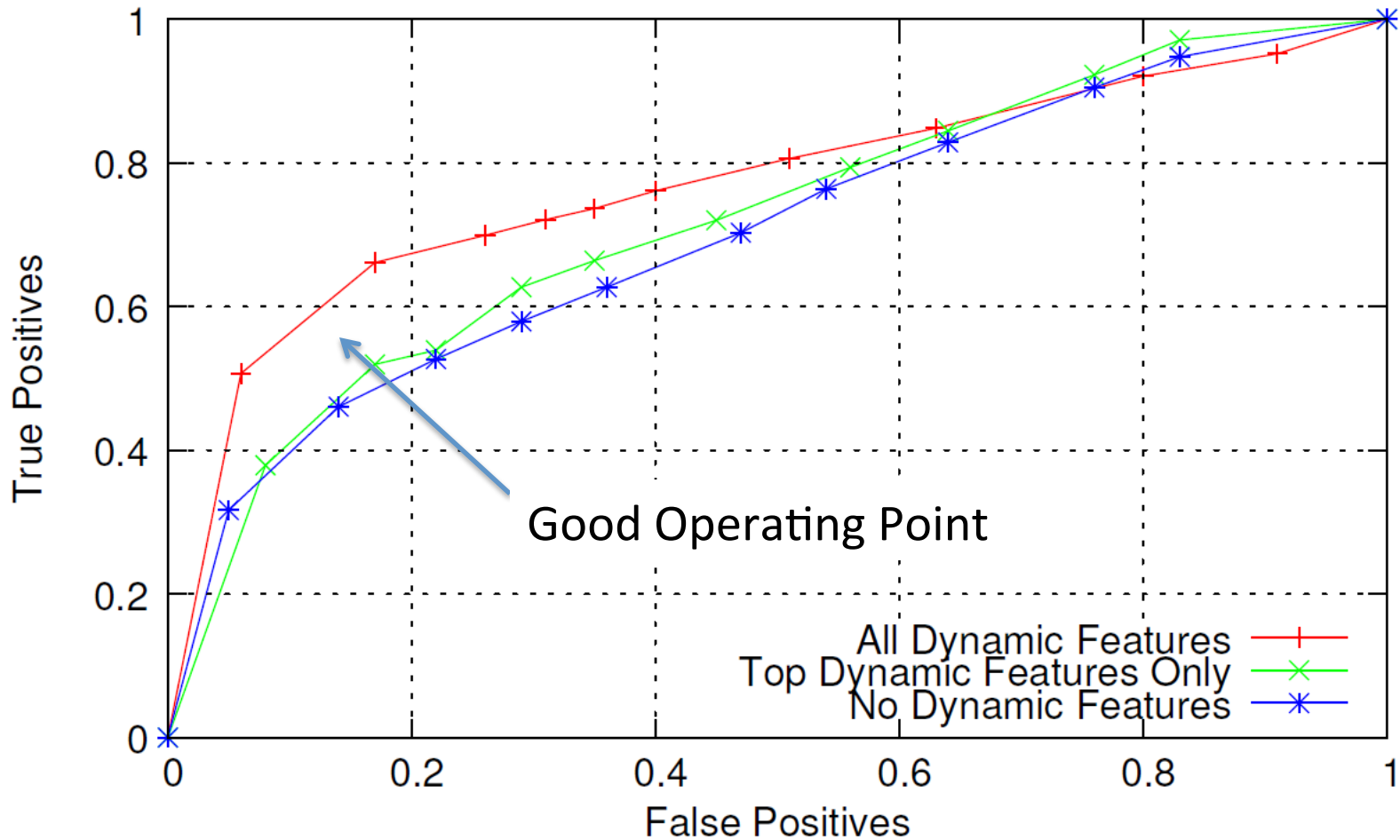
Classification



Classification

- Largely based on [Gao et al., 2007]
- Break input data stream into blocks
- Resample input blocks
- Train ensemble C4.5 decision tree classifiers using Hoeffding bounds [Domingos et al., 2000]
- Retrain periodically using new dynamic features

Classification Results



Limitations

- Only makes sense when page content and traffic statistics are risk factors of malice
 - Sites hacked via weak passwords or via social engineering attacks violate this
 - Sites that are maliciously hosted may violate this
- Requires some sites to become compromised in order to make predictions

Conclusions

- Predicting websites that become malicious in the future is possible!
- Acceptable performance can be achieved even on our modest dataset

Kyle Soska – ksoska@cmu.edu