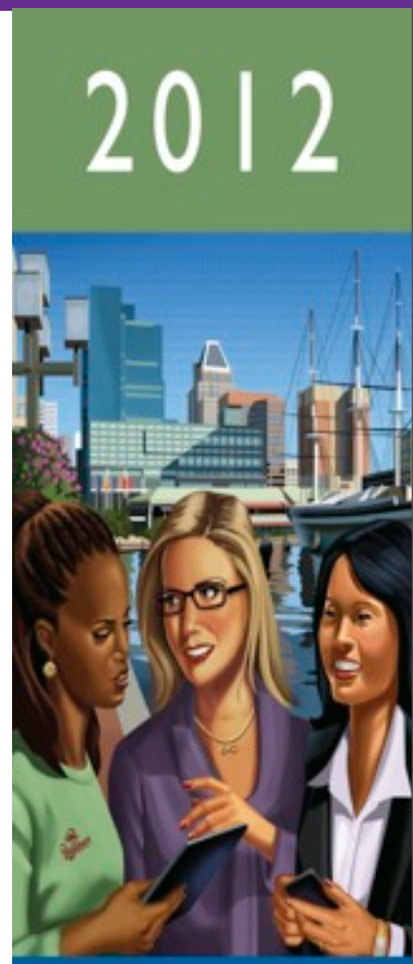
On Detecting Deception

Sadia Afroz
Privacy, Security and Automation Lab (PSAL)
Drexel University



What is Deception?

- Deception: An adversarial behavior that disrupts regular behavior of a system

Deception in Different Areas

- Deception in Writing Style
- Deception in Website (Phishing)
- Deception in Blog Comment

Deception in Writing Style

- Writing by changing regular writing style

A Gay Girl In Damascus

A blog by
Amina Arraf

Facts about Amina:

A Syrian-American activist

Lives in Damascus

A Gay Girl in Damascus becomes a heroine of the Syrian revolt

Blog by half-American 'ultimate outsider' describes dangers of political and sexual dissent

Katherine Marsh in Damascus
guardian.co.uk, Friday 6 May 2011 11.24 BST
Article history

facebook

Search

Wall
Info
Photos
Notes

About

We demand the release of Amina Abdallah Arraf al Omari

14,866
people like this

Create a Page
Add to my page's favourites
Subscribe via RSS

Free Amina Abdalla | Syrian Blogger

Like

Writer

Wall

Free Amina Abdalla | Syrian Blogger

Questions about Amina's identity are surfacing. However, we think it is possible that the writer of the blog is indeed in custody, in which case, it is important to continue to support her. Many people in Syria are forced to use alternative identities to protect themselves. However, administrators of this site cannot verify

A Gay Girl In Damascus

A Gay Girl In Damascus

Fake picture
(copied from Facebook)

A Gay Girl In Damascus

Fake picture
(copied from Facebook)

The real “Amina”

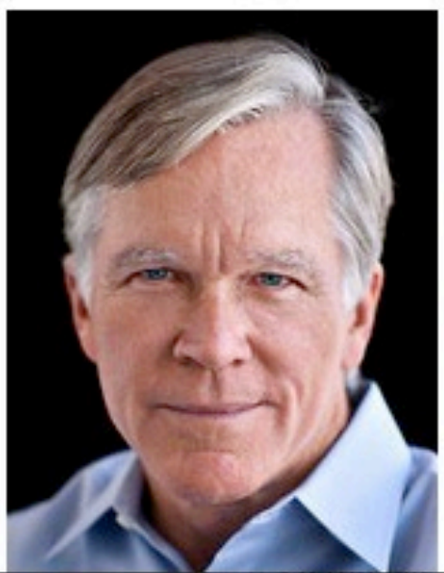
Thomas MacMaster
A 40-year old American male

OP-ED COLUMNIST

WikiLeaks, a Postscript

By BILL KELLER
Published: February 19, 2012

THIS is apparently the revenge of Julian Assange: everyone who runs afoul of the rock-star leaker is condemned to spend eternity discussing the cosmic meaning of WikiLeaks. As the editor of The Times during our publication of many articles based on that treasury of military and diplomatic secrets, and as the lucky man the WikiLeaks founder singled out as his Least Favorite Journalist, I have participated in half a dozen panel discussions, and turned down at least that many. I can't complain about the one in Madrid, where, after holding forth in a packed auditorium, the American, British, German, French and Spanish editors who broke news based on WikiLeaks commemorated the collaboration with an after-hours prowl through the Prado Museum and a 27-course meal cooked by master chef Ferran Adrià. (If Europe is dying, Spain is where I plan to go for the wake.) Unforgettable in a different way was the retrospective in Berkeley, where Assange himself, then as now awaiting an extradition ruling in England, was Skyped in on a giant screen, like the mighty Oz, to pontificate on Western media's failure to turn the files into a kind of Nuremberg trial of American imperialism. About half the audience seemed on the verge of tossing their underwear at the screen.



Add to that the three or four documentaries on the WikiLeaks adventure, the dozen books — including, weirdly, Assange's unauthorized *auto*biography — and a couple speculative Hollywood projects, in which I have a twofold interest. (1. The very slight possibility that I might make some money for my small piece of the story. 2. The exceedingly remote chance that a director will take up my wife's brilliant idea that Assange be played by Tilda Swinton.)

It's amazing they keep inviting me to these things, since I'm a bit of a spoilsport. My consistent answer to the

www.opinion-nytimes.com/2012/07/29/opinion/keller-a-post-postscript.html

Yahoo! YouTube Wikipedia News Popular Productivity hints, ti The Turing Test (Sta

Hoax

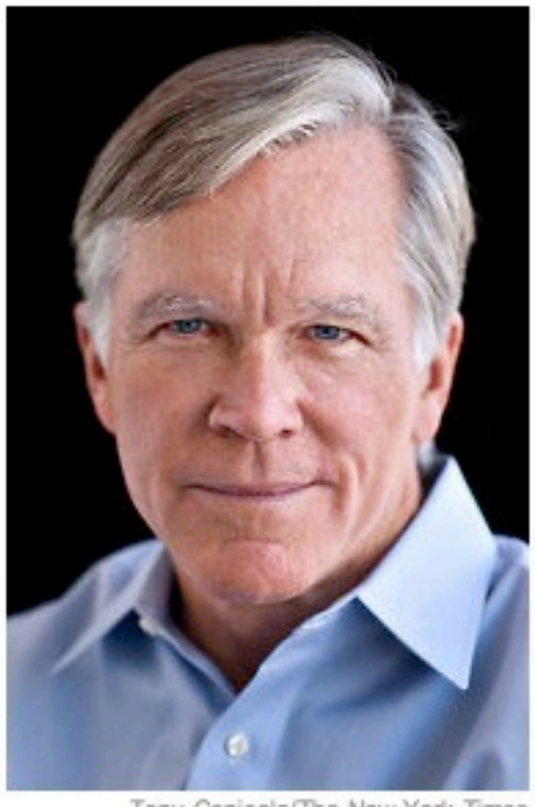
OP-ED COLUMNIST

WikiLeaks, A Post Postscript

By BILL KELLER
Published: July 29, 2012

Tweet

AS rumors build about the potential financial blockade against the New York Times by Visa, Mastercard, and American Express for hosting U.S. government cables published by WikiLeaks, I find myself in the awkward position of having to defend WikiLeaks. During the House Judiciary Subcommittee hearing on July 11th, several Republicans made it clear they also want New York Times journalists charged under the Espionage Act for their recent stories on President Obama's 'Kill List' and secret US cyber attacks against Iran.



Tony Cenicola/The New York Times

Bill Keller

As those of you who have followed my turbulent relationship with WikiLeaks and its Guru-In-Chief Julian Assange know, I am first in line when it comes to distancing myself from his brand of transparency without government checks and balances. You don't have to embrace Assange as a kindred spirit to believe that what he did in publishing those cables falls under the protection of the First Amendment. The backroom pressures by the Obama Administration's State Department to expand its financial blockade targeting WikiLeaks to include news organizations that host information from their trove of pilfered documents goes too far.

I've said repeatedly, in print and in a variety of public

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE
PRINT
SINGLE PAGE
REPRINTS

Deception in Writing Style

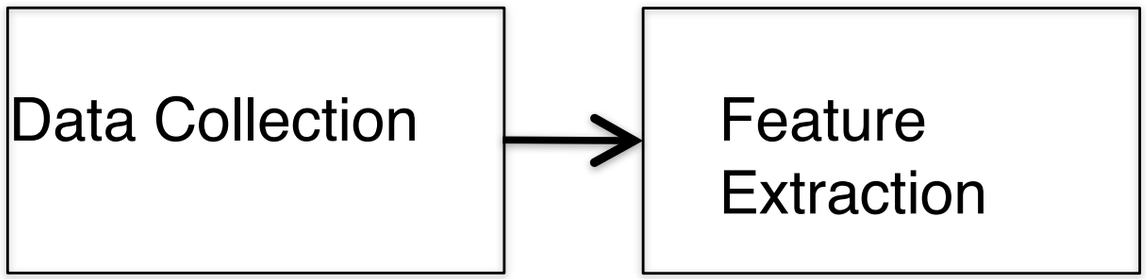
- Goal:
  - Distinguish regular writing from deceptive writings

Approach

Approach

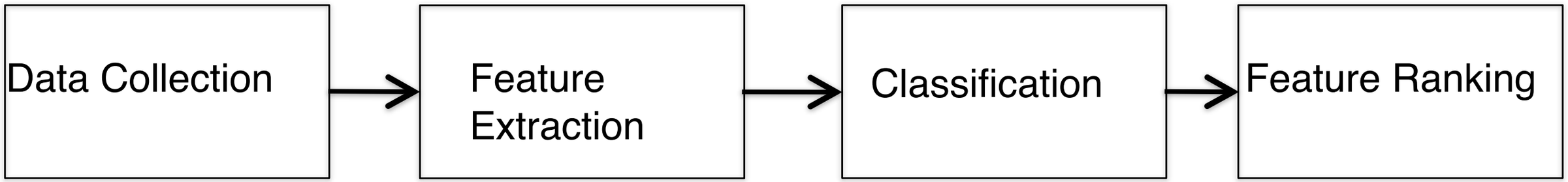
Data Collection

Approach



Approach

Approach


Data Collection
Feature Extraction
Classification
Feature Ranking

Data collection

- Short-term deception:
  - –Extended-Brennan-Greenstadt Corpus
    - Regular
    - Imitation
    - Obfuscation
  - –Hemingway-Faulkner Imitation corpus
    - Regular
    - Imitation
- Long-term deception:
  - -Thomas-Amina Hoax corpus
    - Regular
    - Deceptive

Classification

- We used WEKA for machine learning.

- Classifier:
  - –Experimented with several classifiers
  - –Choose the best classifier for a feature set

- 10-fold cross-validation
  - –90% of data used for training
  - –10% of data used for testing

Feature sets

- We experimented with 3 feature sets:
  - –Writeprints
  - –Lying-detection features
  - –9-features

Feature sets

- We experimented with 3 feature sets:
  - –Writeprints
    - 700+ features, SVM
    - Includes features like frequencies of word/character n-grams, parts-of-speech n-grams.
  - –Lying-detection features
  - –9-features

Feature sets

- We experimented with 3 feature sets:
  - –Writeprints
    - 700+ features, SVM
  - –Lying-detection features
    - 20 features, J48 decision tree
    - Previously used for detecting lying.
    - Includes features like rate of Adjectives and Adverbs, sentence complexity, frequency of self-reference.
  - –9-features

Feature sets

- We experimented with 3 feature sets:
  - –Writeprints
    - 700+ features, SVM
  - –Lying-detection features
    - 20 features, J48 decision tree
  - –9-features
    - 9 features, J48 decision tree
    - Used for authorship recognition
    - Includes features like readability index, number of characters, average syllables.

Results

- Short-term deception:
  - –Extended-Brennan-Greenstadt Corpus
    - – Regular: 98%
    - – Imitation: 85%
    - – Obfuscation: 89%
  - –Hemingway-Faulkner Imitation corpus
    - Regular: 86.2%
    - Imitation: 88.6%
- Long-term deception: -Thomas-Amina Hoax corpus
  - 14% was detected as deceptive
  - Regular authorship recognition shows inconsistency in writing style.

Deception in Website: Phishing

Real bank

Alice uses online bank

Deception in Website: Phishing

URL

Browser Indicator

SSL

Real bank

Alice uses online bank

Deception in Website: Phishing

URL

Browser Indicator

SSL

Real bank

Fake bank

Alice uses online bank

Deception in Website: Phishing

URL

Browser Indicator

SSL

Real bank

Alice uses online bank

Fake bank

Alice thinks everything that looks like her bank
Is her bank!

Approach: PhishZoo

Approach: PhishZoo

Real site

Approach: PhishZoo

Real site

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Extracts
visual elements
of the site

Images
Visible text

Approach: PhishZoo

Real site

Extracts
visual elements
of the site

Images
Visible text

Approach: PhishZoo

Real site

Extracts
visual elements
of the site

Images
Visible text

Approach: PhishZoo

Real site

Extracts
visual elements
of the site

Images
Visible text

Profile Stored

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Images
Visible text

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Images
Visible text

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Fake site

Images
Visible text

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Visual components
match but the url, ssl
don't match

Fake site

Images
Visible text

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Visual components
match but the url, ssl
don't match

Fake site

Images
Visible text

Extracts
visual elements
of the site

Approach: PhishZoo

Real site

Images
Visible text

Extracts
visual elements
of the site

Profile Stored

Visual components
match but the url, ssl
don't match

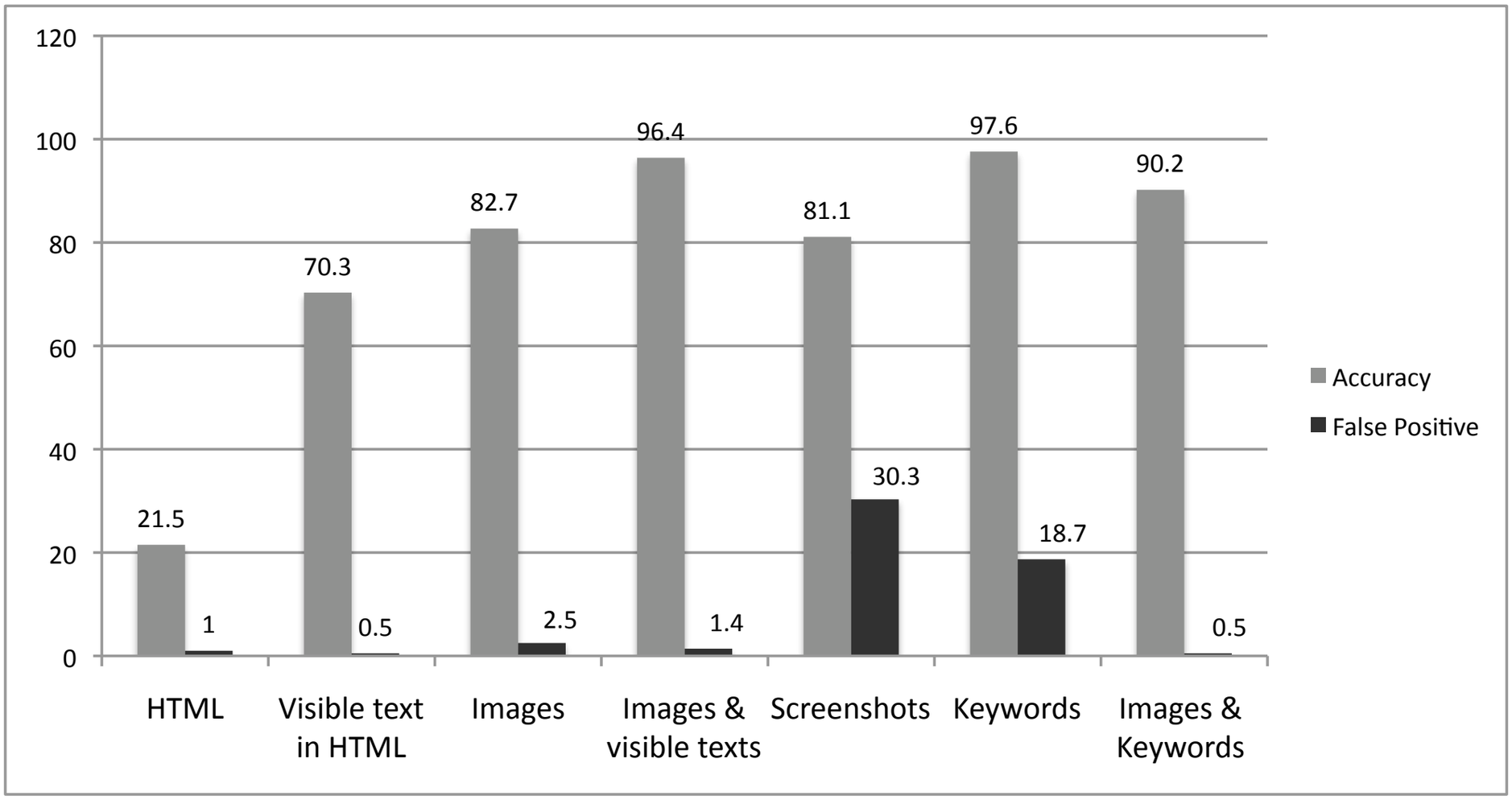
Phishing
Alert

Fake site

Images
Visible text

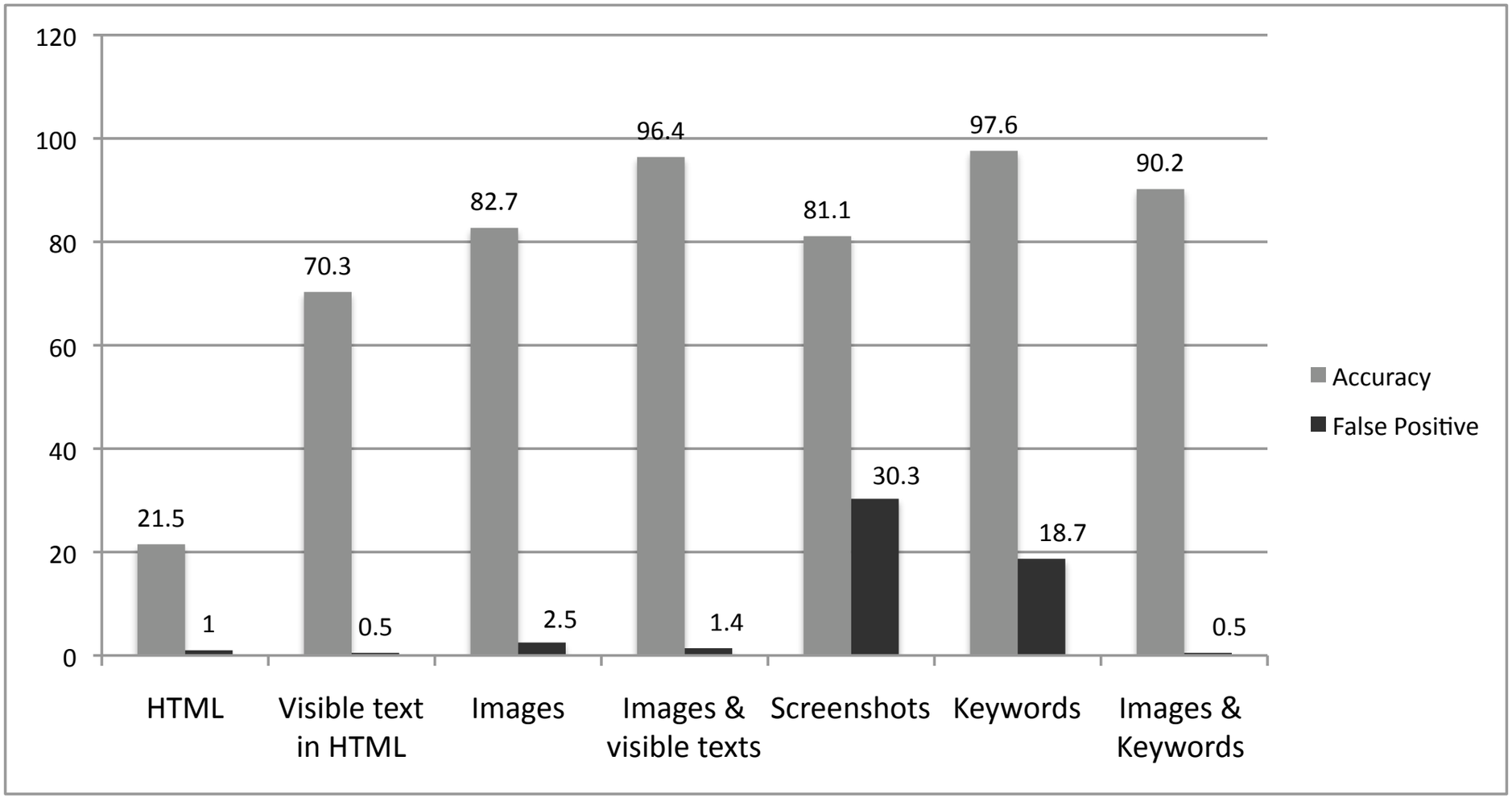
Extracts
visual elements
of the site

Result


120
100
80
60
40
20
0
70.3
82.7
96.4
81.1
97.6
90.2
21.5
1
0.5
2.5
1.4
30.3
18.7
0.5
HTML
Visible text in HTML
Images
Images & visible texts
Screenshots
Keywords
Images & Keywords
Accuracy
False Positive

Result

96.4% accurate in detecting phishing

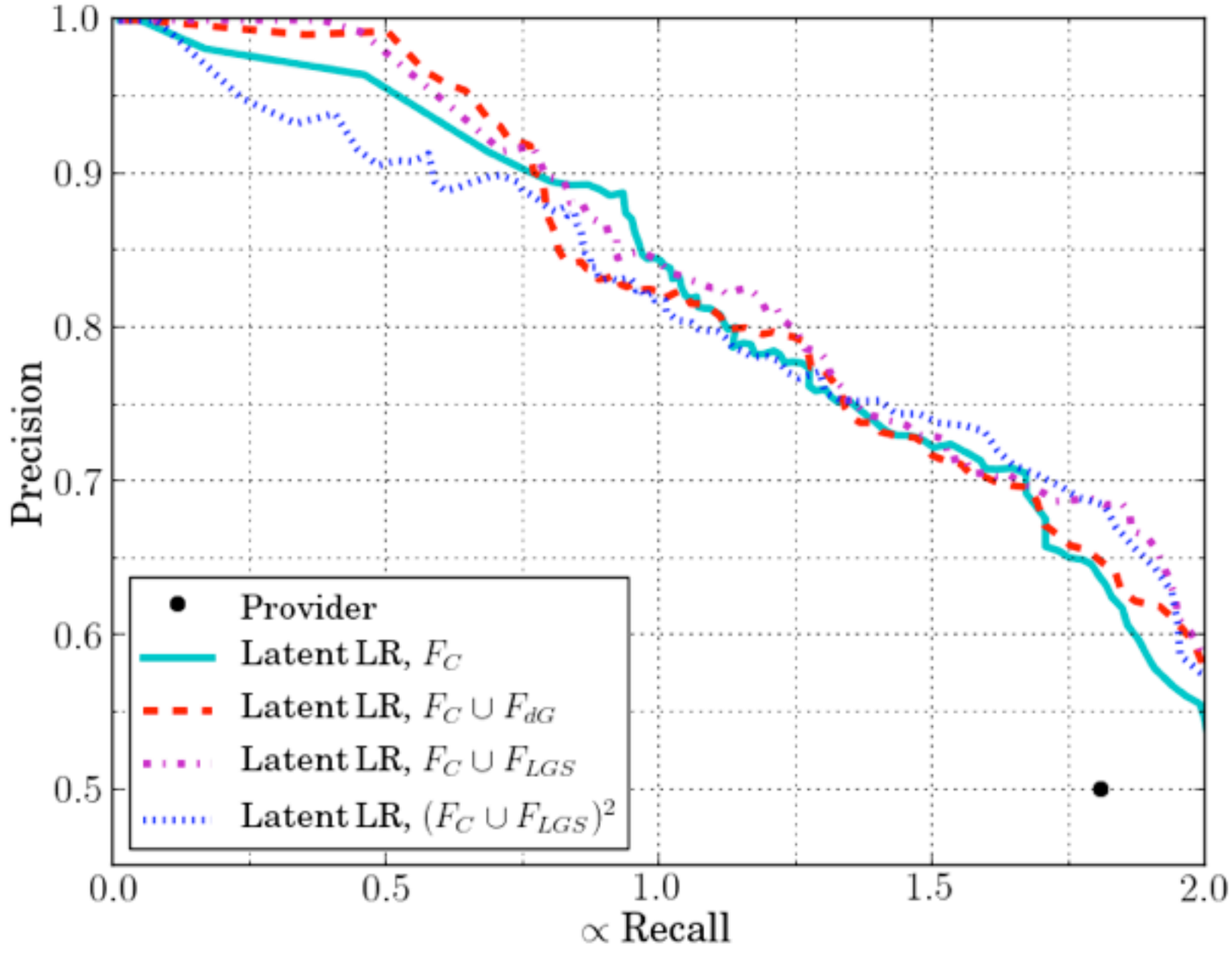


Future work: Deception in Blog Comment

Approach

- Spammers post same thing repeatedly.
- Use compression ratio (LZMA)
- Classifier: Latent logistic regression

Result


1.0
0.9
0.8
0.7
0.6
0.5
Precision
0.0
0.5
1.0
1.5
2.0
∝ Recall
Provider
Latent LR, F_C
Latent LR, $F_C \cup F_{dG}$
Latent LR, $F_C \cup F_{LGS}$
Latent LR, $(F_C \cup F_{LGS})^2$

But spammers are smart!

- There are tools for spamming: Xrumer, SEnuke, Ultimate WordPress Comment Submitter (UWCS)
- That automatically
  - create new accounts
  - Use proxy
  - Copy relevant words

Summary

- Deception in Writing Style:
  - Distinguish regular writing from deceptive writings
- Deception in Website (Phishing)
  - Detect website imitation
- Deception in Blog Comment
  - Detect spam comments

Thanks!

- Sadia Afroz: sadia.afroz@drexel.edu
- Rachel Greenstadt: greenie@cs.drexel.edu
- Michael Brennan: mb553@drexel.edu
- Ariel Stolerman: ams573@drexel.edu
- Andrew McDonald: awm32@drexel.edu
- Aylin Caliskan: ac993@drexel.edu

- Privacy, Security And Automation Lab (https://psal.cs.drexel.edu)
- Secure Computing Research for User Benefit (https://scrub.cs.berkeley.edu)