PhishAri
Automatic Realtime Phishing Detection on Twitter

Anupama Aggarwal[+], **Ashwin Rajadesingan***
Ponnurangam Kumaraguru[+]
[+]Indraprastha Institute of Information Technology, New Delhi, IN
*Arizona State University, AZ, USA

Motivation: Some Statistics

- $520 million were lost worldwide from phishing attacks in H1 of 2011 alone.

- In 2012, around 20% of all phishing attacks targeted Facebook

- Social network phishing has jumped by 221% during Q1 of 2012

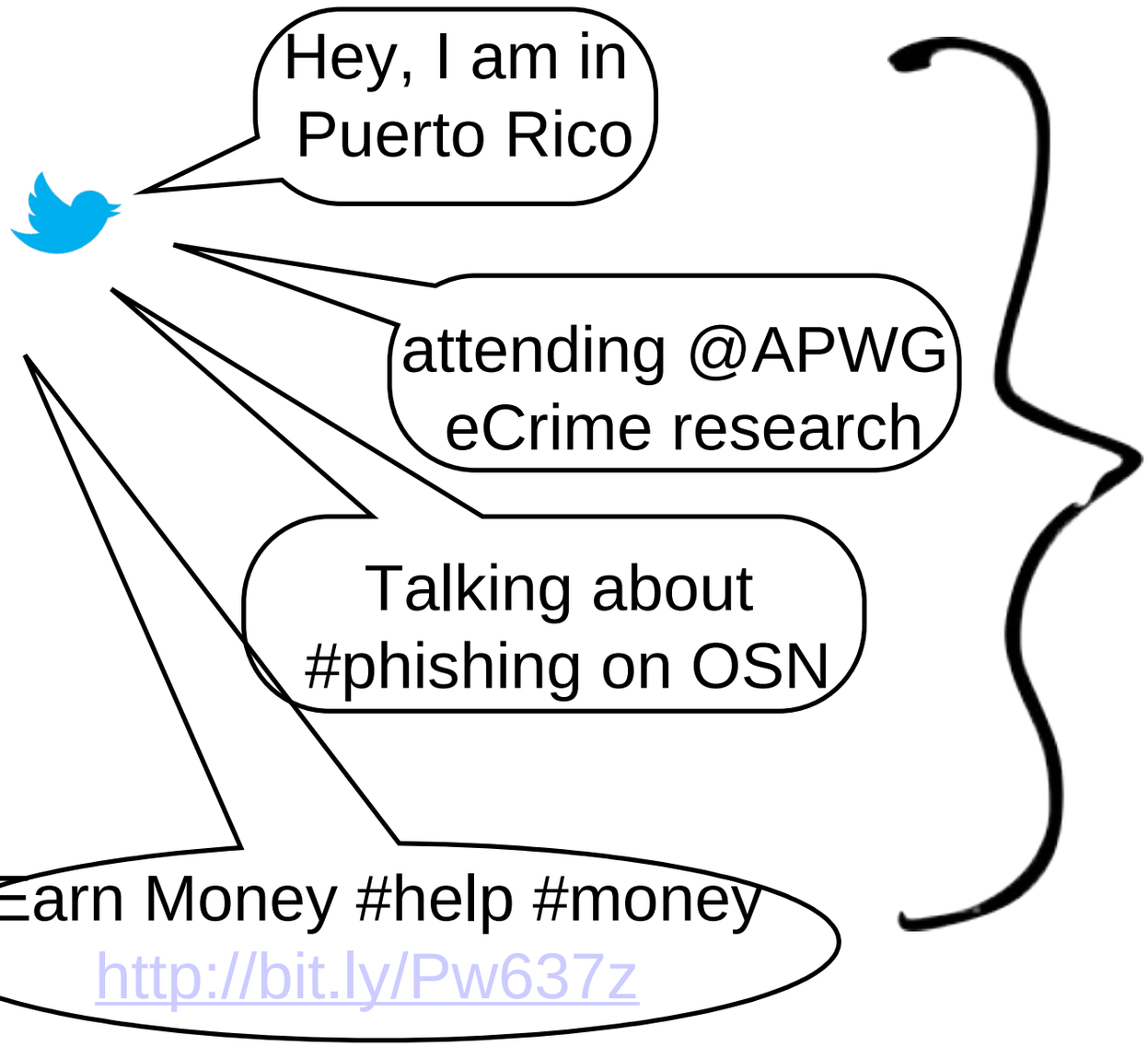
Current Scenario

- Offline spam characterization & detection studies
- No characterization of phishing on OSM
- Dependence on spam / phishing blacklists
- Absence of end-user solution deployed

What Did We Do to Fill the Gap?

- Built a system to automatically detect phishing on Twitter in realtime
- No dependency on blacklists
- Deployed end-user system for Twitter users - Chrome extension

Twitter 101


Hey, I am in Puerto Rico
attending @APWG eCrime research
Talking about #phishing on OSN
Earn Money #help #money
http://bit.ly/Pw637z
Tweets
<140 char

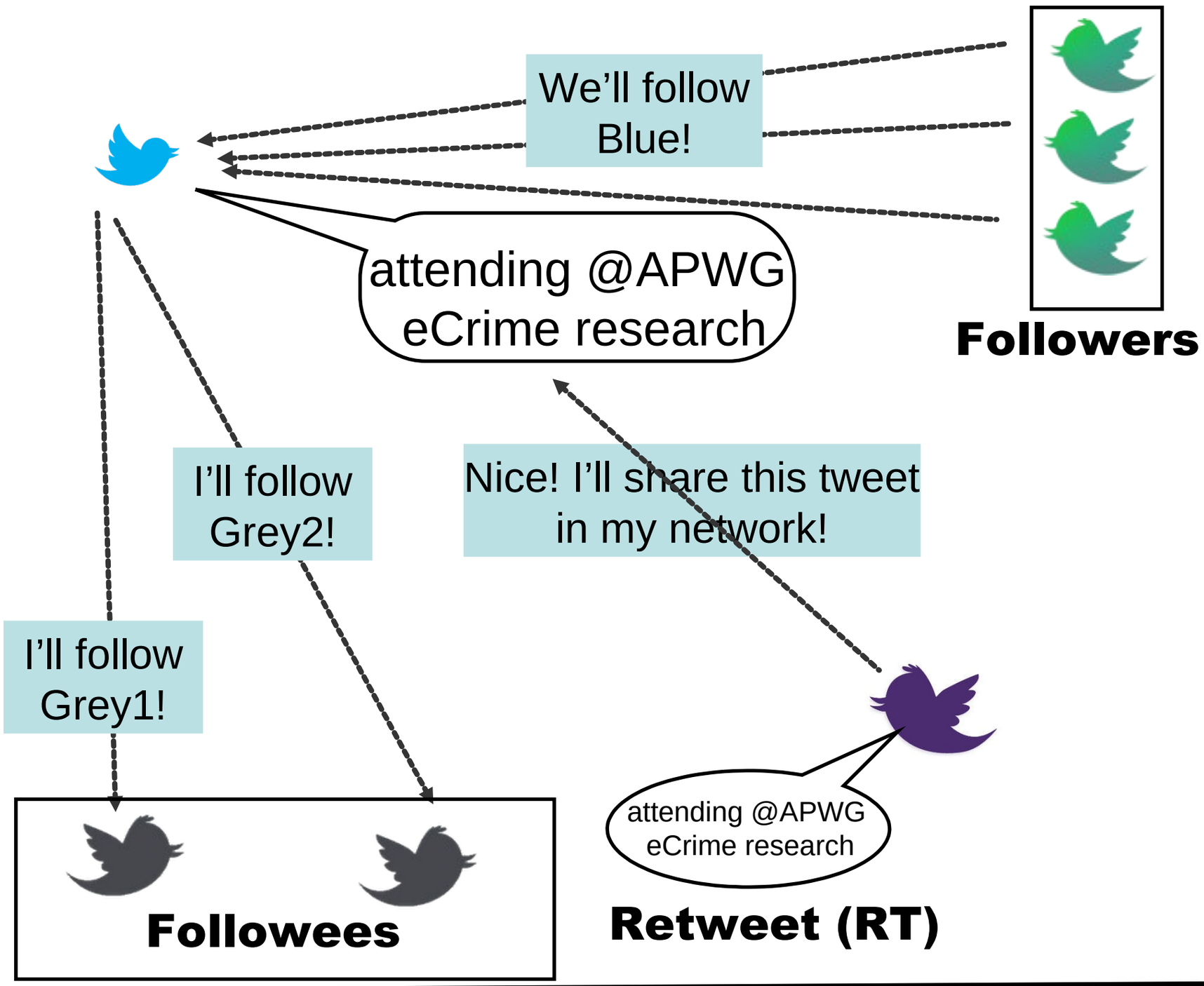
Twitter 101

@Tag → To mention/reply to a Twitter user

#Tag → To mention a topic

URL in Tweet → To link external media

Twitter 101



Major Challenges

- Only 140 characters - very less information
- 180,000 tweets per minute - quick spread
- Phishing blacklists are slow – less reliable

PhishAri

- Detecting phishing in Twitter
- Realtime, zero-hour
- Easy user interface, Google Chrome extension

Methodology

Two phases:

- Building classification model for phishing detection
- Realtime Detection

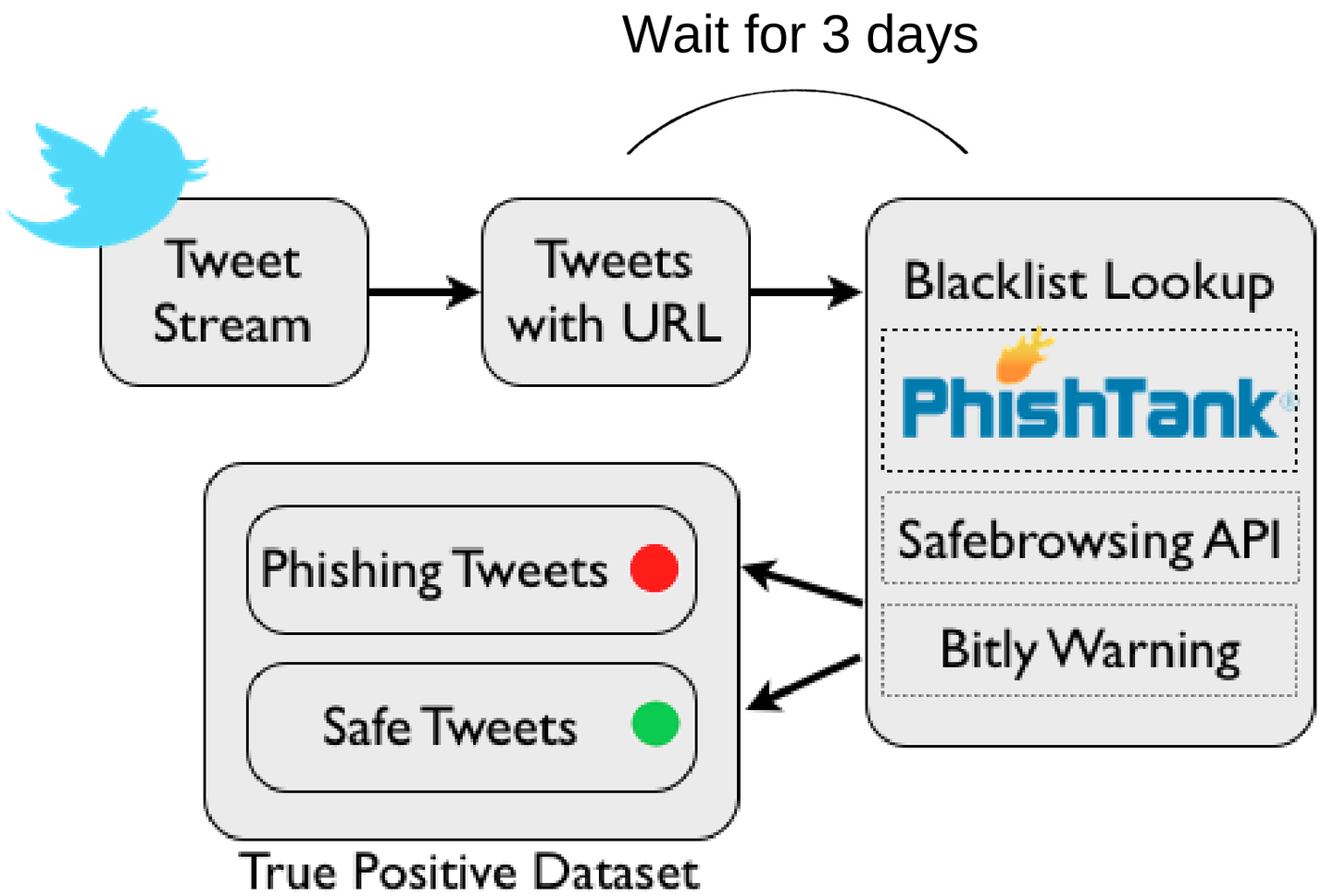
Building Classification Model

- Data Collection
- Feature Extraction
- Classification

Data Collection

- 309,321 total tweets with URL
- 1,589 Phishing Tweets
- 903 Unique phishing URLs

Data Collection (contd.)



Feature Extraction

Types of Features Used

- URL Features
- WHOIS Features
- Tweet Features
- Network Features

Major URL Based Features

- Length of URL
- Number of subdomains
- Number of redirections
- Presence of conditional redirects

Major WHOIS Based Features

- Registrar's name
- Ownership period
- Time difference between phishing domain creation and Twitter account creation

Major Tweet Based Features

- Number of #tags
- Number of @tags
- Presence of trending #tags
- Number of RTs
- Length of Tweet
- Position of #tags

Major Network Based Features

- Number of followers
- Number of followees
- Ratio of followers / followees
- Part of public lists
- Age of account
- Biography information
- Number of tweets

Classification Results

Evaluation metric	Naive Bayes	Decision Tree	Random Forest
Accuracy	87.02%	89.28%	92.52%
Precision (phishing)	89.21%	88.05%	95.24%
Precision (safe)	92.12%	94.15%	97.23%
Recall (phishing)	68.32%	74.51%	92.21%
Recall (safe)	85.67%	89.20%	95.54%

Evaluation

Comparison with blacklists

80.6% more phishing tweets detected by PhishAri at zero hour which were caught by blacklists after 3 days.

Comparison with Twitter's detection mechanism

84.6% more phishing tweets detected by PhishAri at zero hour which were marked as suspicious by Twitter after 3 days

Time required for the feature extraction & classification of a tweet is a maximum of 0.522 seconds (Min: 0.167 sec, Avg: 0.425 sec, Median 0.384 sec)

Realtime Detection

- Using trained classification model
- PhishAri API (Server Side)
- Google Chrome Browser Extension (Client Side)

PhishAri: RESTful API

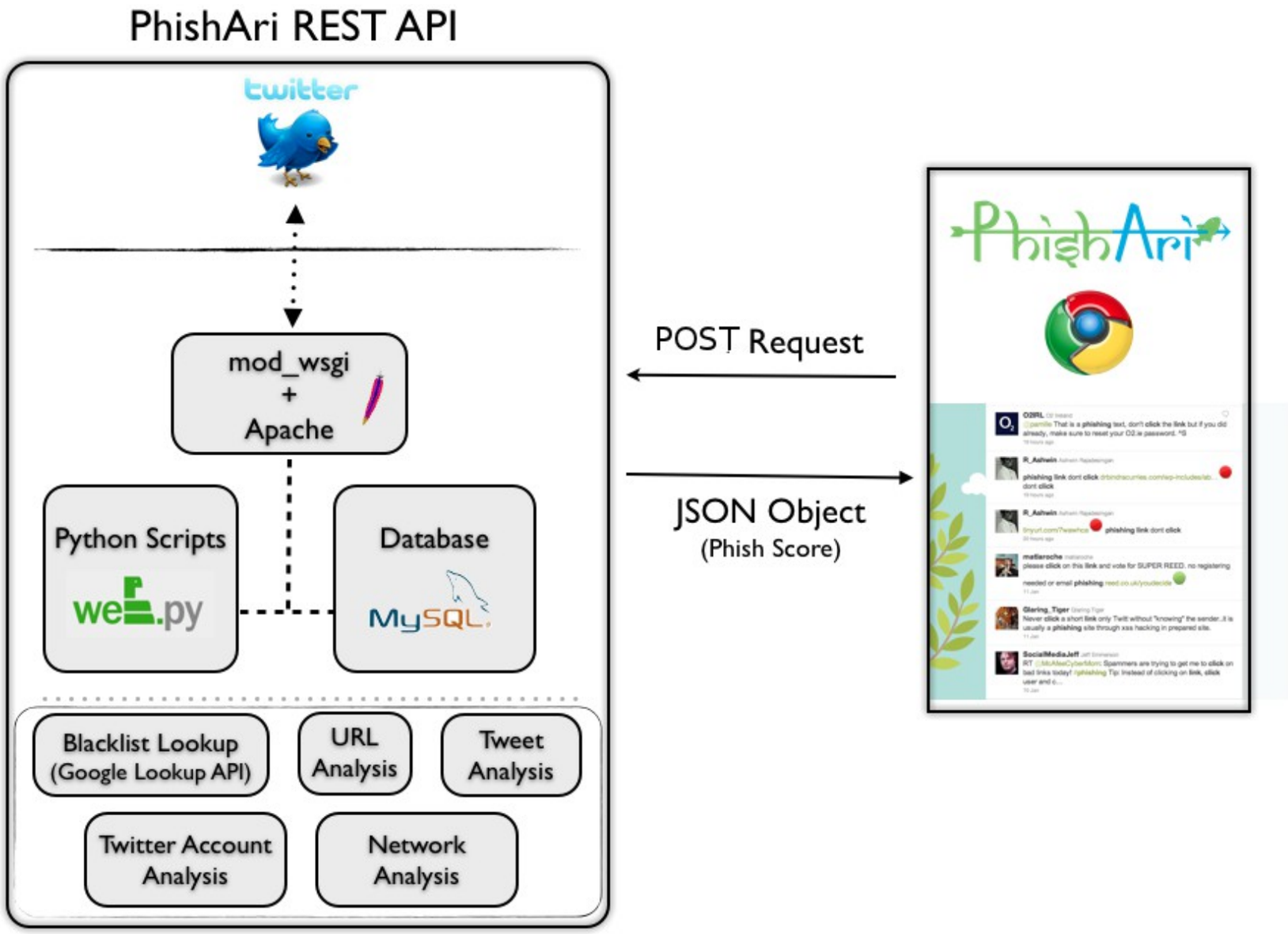
- Use above classification model to build a RESTful API
- POST requests containing list of urls and tweet ids can be made to API.
- Returns if phishing or not

PhishAri Chrome Extension

- Red indicators - phishing links
- Green indicators – safe links

How does PhishAri Work?

- Integration of API with the browser extension



Demo

User Experience and Statistics

- 74 active users
- User study shows that
  - Users want support for other browsers, mobile apps
  - Found it to be efficient and easy to use
  - Asked to improve on response time

Conclusion

- Realtime automatic phishing detection
- High accuracy of 92.52% obtained with Random Forest Classifier
- Time-efficient - takes only 0.522 seconds for indicator to appear
- More efficient than Blacklists
- More efficient than Twitter's detection mechanism

Future Work

- Implement a backend database cache for faster lookup
- Increase the scope of PhishAri from public to all tweets
- Decrease response time of PhishAri and appearance of indicators
- Support for other browsers and mobile apps

Thank You!

Questions?

e-mail: pk@iiitd.ac.in

Backup Slides

Text Analysis

Legitimate Tweets

Phishing Tweets

Features Used

- URL Features - Length, number of dots, characters, redirections
- WHOis Features - domain name, ownership period
- Tweet Features - Number of #tags, @mentions, length, trending topics
- Network Features - Follower/Followee ratio, Age of account, Number of Tweets

All Features Used

URL Based (F1)	Length of URL	Length of expanded URL in number of characters
	Number of dots	Number of dots (.) used
	Number of subdomains	Number of subdomains (marked by /) in the expanded URL
	Number of Redirections	Number of hops between the posted URL and the Landing page
	Levenshtein distance between redirected hops	Avg Levenshtein distance between length of redirected URLs between original & final URL
	Presence of conditional redirects	Whether the URL is redirected to different landing page for browser or an automated program
WHOIs Based (F2)	Registering domain name	Name of the domain provider
	Ownership period	Age of the domain
	Time taken to create Twitter account	How much time lapsed between creation of domain and the Twitter account
Tweet Based (F3)	Number of #tags	Number of topics mentioned in tweet
	Number of @tags	Number of Twitter users mentioned in tweet
	Presence of trending #tags	Number of topics mentioned which were trending at that time
	Number of RTs	Number of times the tweet was reposted
	Length of Tweet	Length of tweet in number of characters
	Position of #tags	Number of characters of tweets after which the #tag appears
Network Based (F4)	Number of Followers	Number of Twitter users who follow this Twitter user
	Number of Followees	Number of Twitter users who are being followed by this Twitter user
	Ratio of Followers-Followees	Number of Followers / Number of Followees
	Part of Lists	Whether the Twitter user is part of a public list
	Age of account	How old the Twitter account is
	Presence of description	Whether the Twitter account has a profile description
	Number of Tweets	Number of tweets posted by the Twitter user

Detailed Evaluation

Evaluation metric	Naive Bayes	Decision Tree	Random Forest
Accuracy	87.02%	89.28%	92.52%
Precision (phishing)	89.21%	88.05%	95.24%
Precision (safe)	92.12%	94.15%	97.23%
Recall (phishing)	68.32%	74.51%	92.21%
Recall (safe)	85.67%	89.20%	95.54%

		Predicted	
		Phishing	**Safe**
Actual	**Phishing**	**92.31%**	7.78%
	Safe	9.60%	**94.41%**

Feature Set wise Results

Feature Sets	Precision (Phishing)	Precision (Safe)	Recall (Phishing)	Recall (Safe)	Accuracy
F1	81.27%	88.21%	79.25%	91.34%	82.22%
F1 + F2	86.11%	89.92%	85.21%	92.21%	87.31%
F1 + F2 + F3	91.10%	94.66%	88.32%	92.88%	90.03%
F1 + F2 + F3 + F4	95.24%	97.23%	92.21%	95.54%	92.52%