

# PhishDef: URL Names Say It All



Anh Le, Athina Markopoulou (UC Irvine)  
Michalis Faloutsos (UC Riverside)

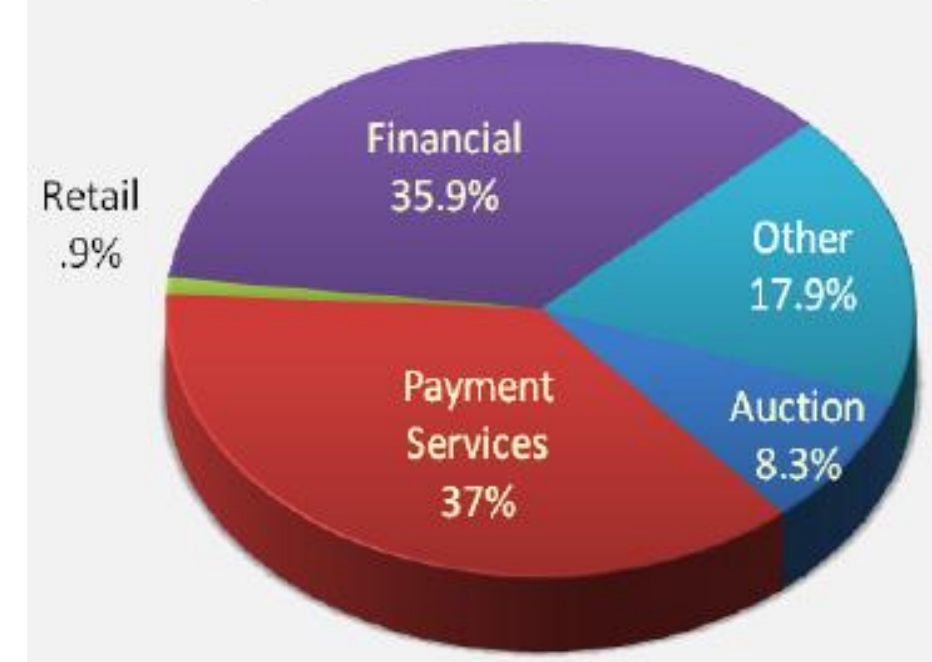
## Phishing Background

### What is Phishing?



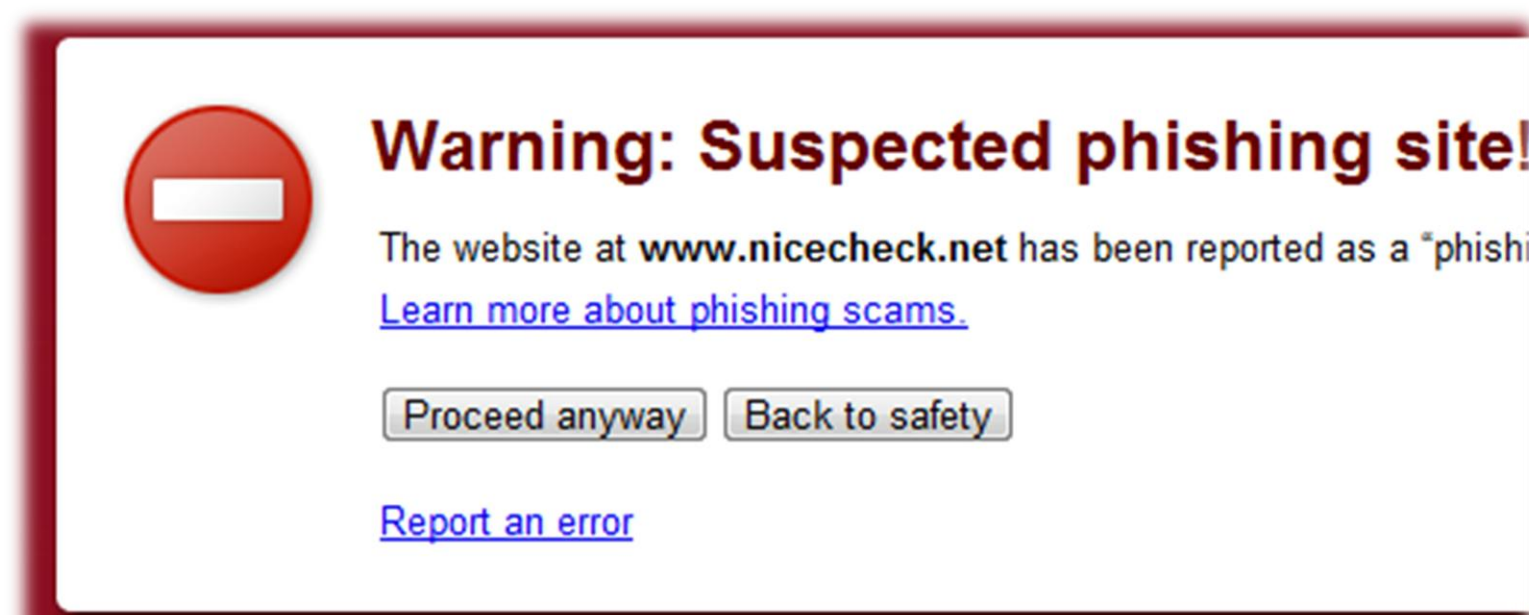
- Criminal mechanism employing both **social engineering** and **technical means** to steal consumers' personal identity data and financial account credentials
- Causes **billions of dollars** loss annually

Most Targeted Industry Sectors 1st Quarter '10

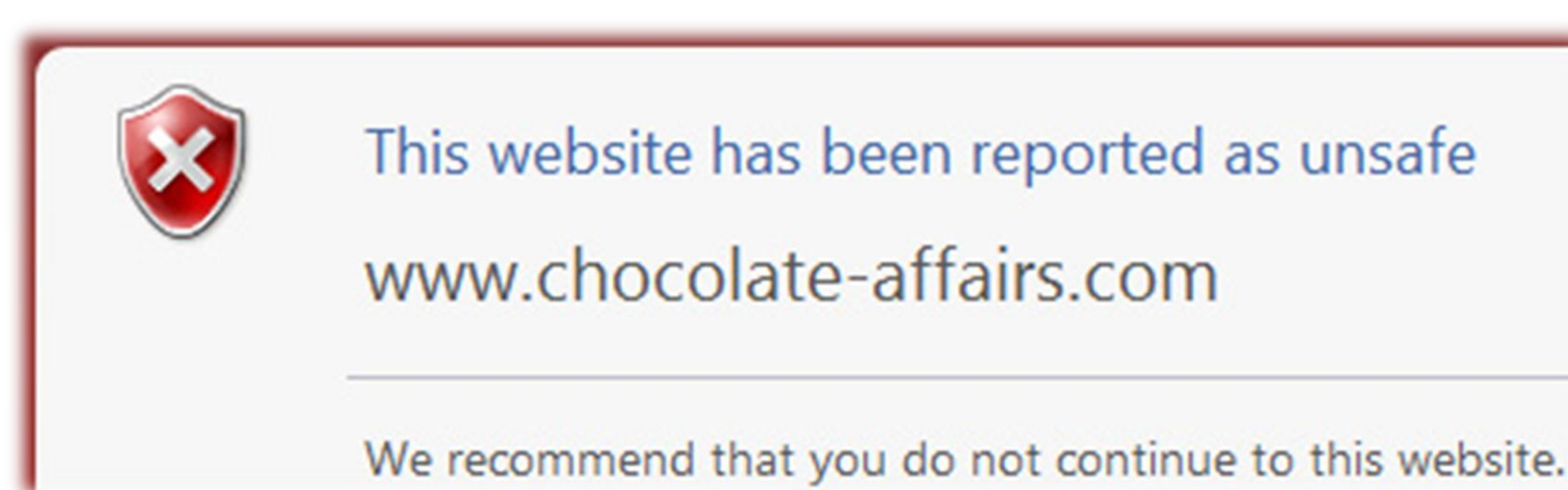


### How are you currently protected?

o Google Safe Browsing



o Microsoft Smart Screen



o Third-Party

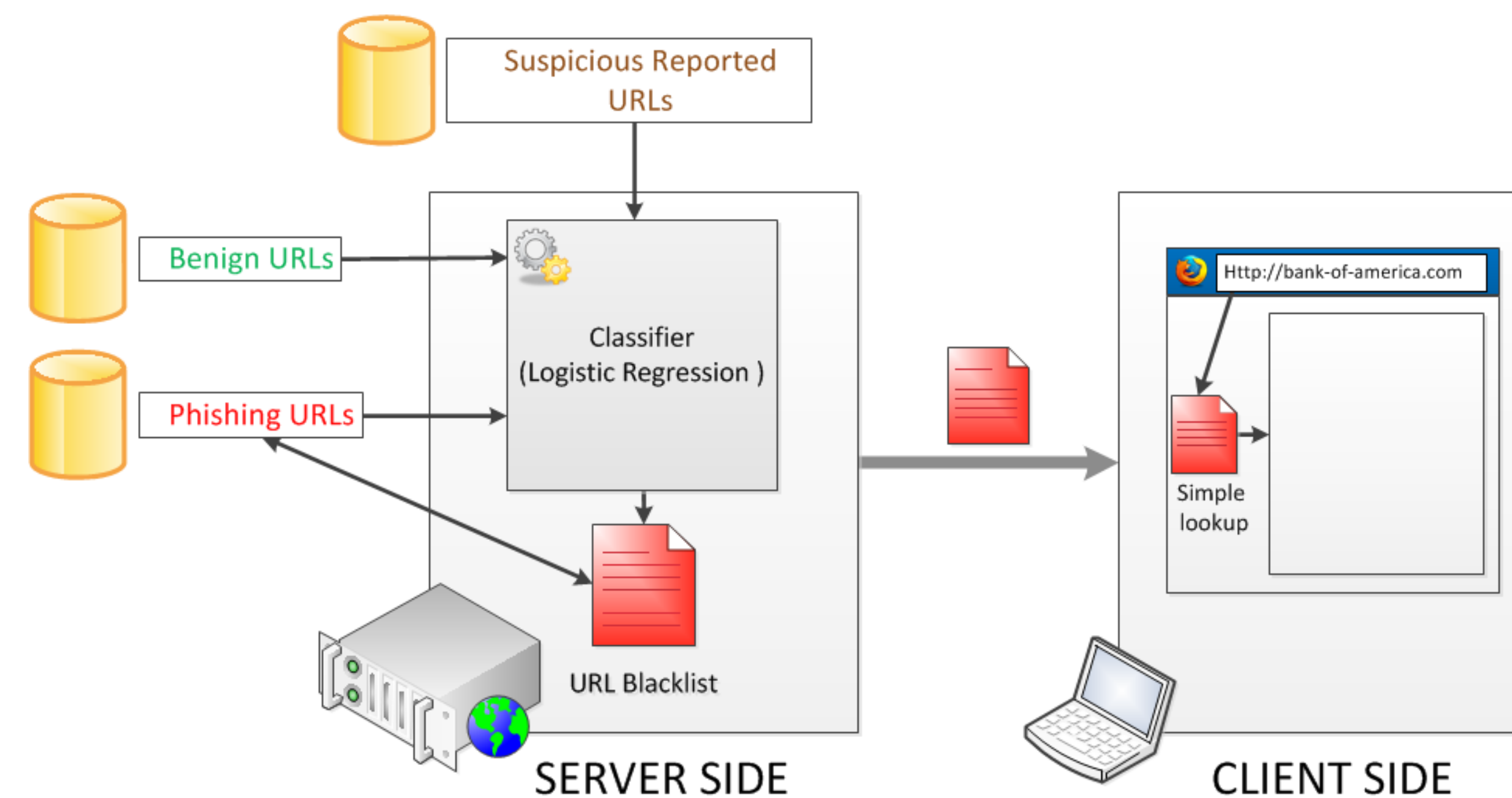


## Machine Learning Techniques for Detecting Phishing URLs

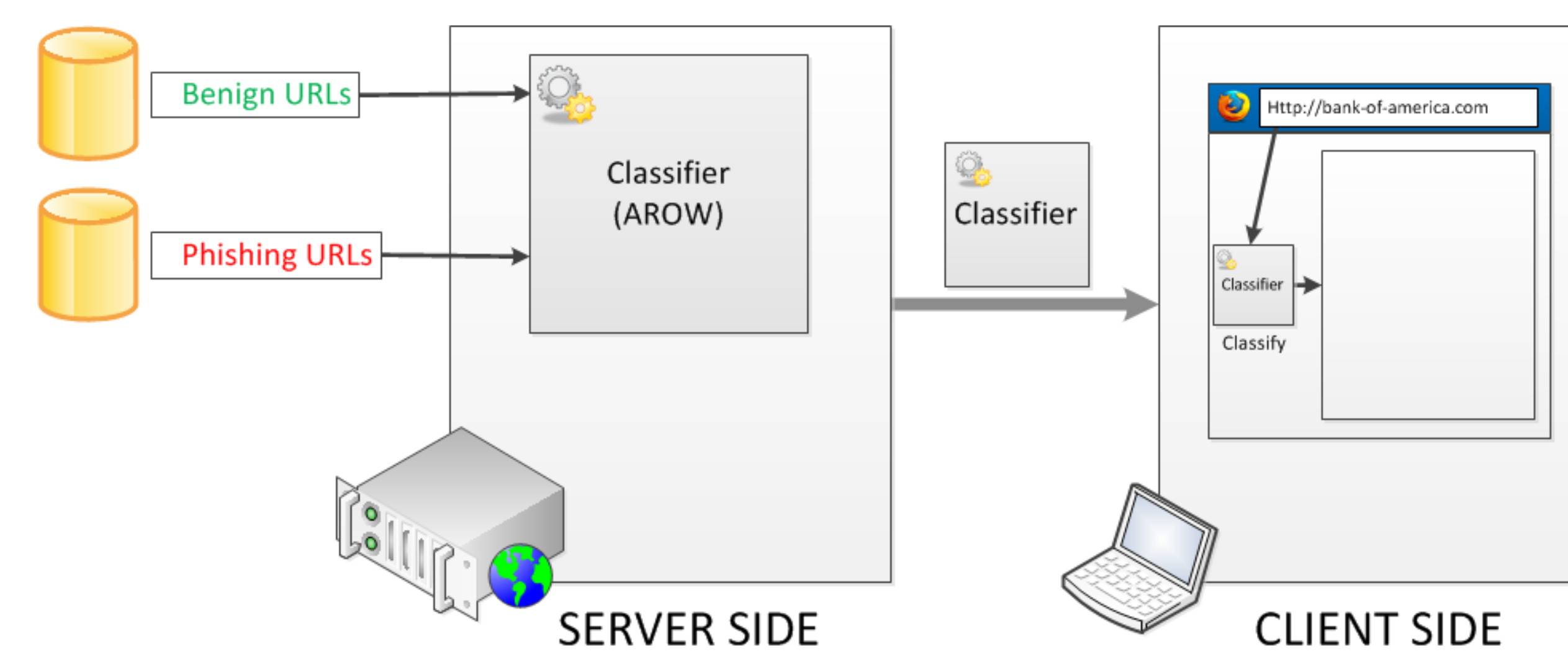
### What's going wrong?

- None of the existing protection mechanisms protect against **zero-day phishing**

### Current Reactive Protection



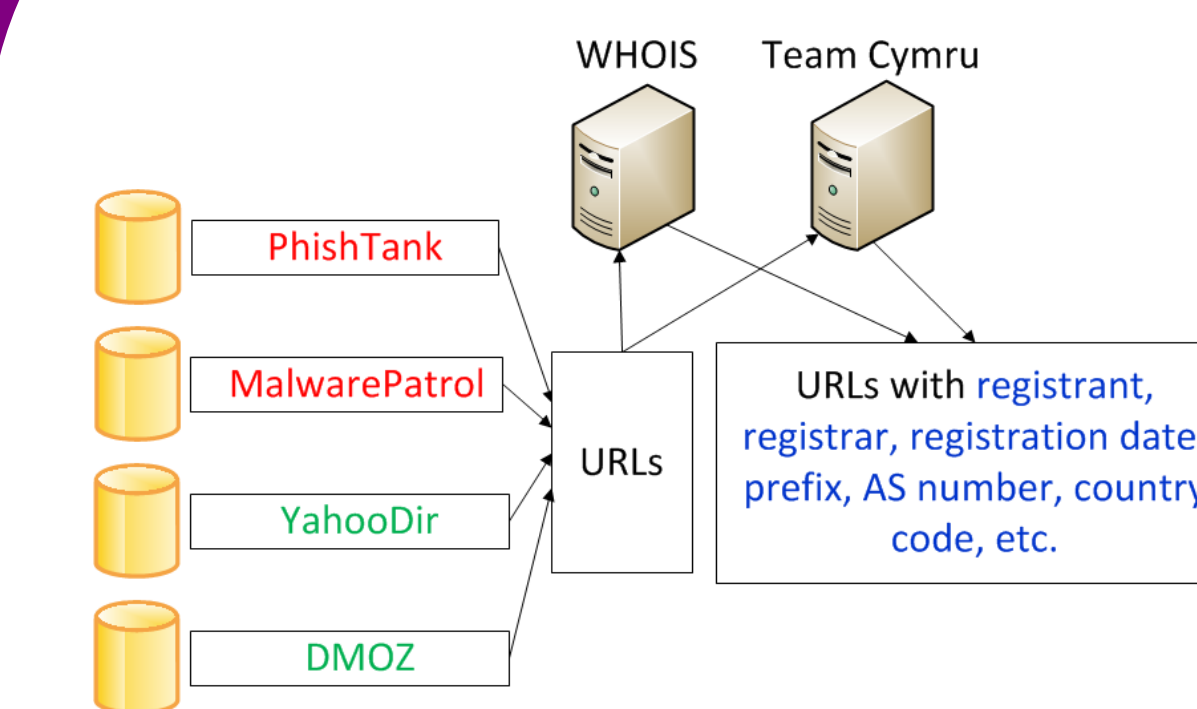
### PhishDef Proactive Protection



- Main challenge: Page loading time
- Classifier must be light weight! (**URL names**)
- Which set of (full or **lexical**) features works best?
- Which classification algorithm works best?

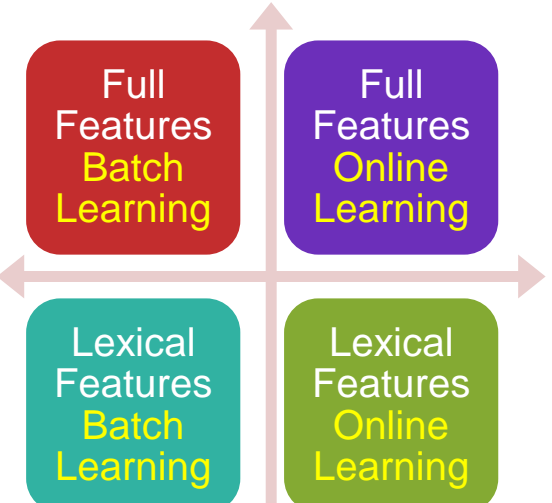
## PhishDef Study

### Data



### Learning Algorithms

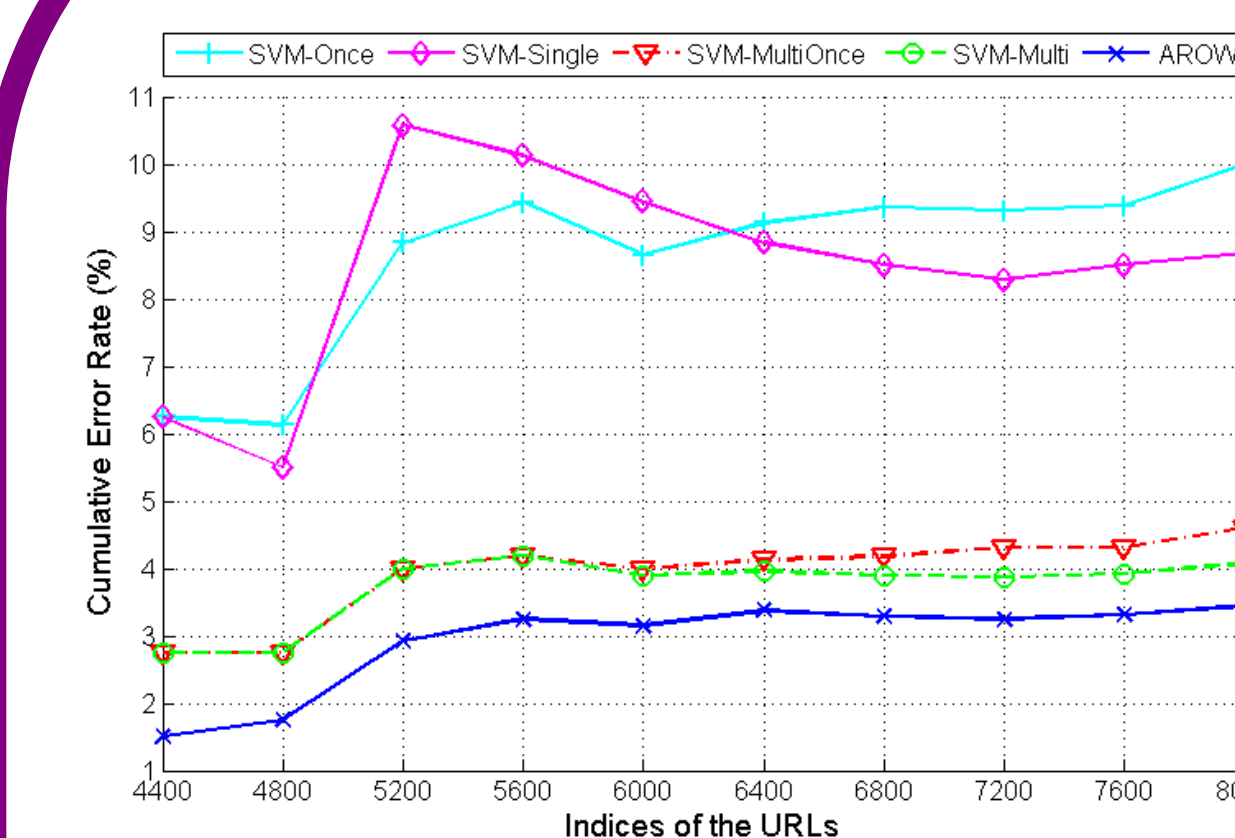
- Support Vector Machine
- Online Perceptron
- Confident Weighted
- Adaptive Regularization Of Weights



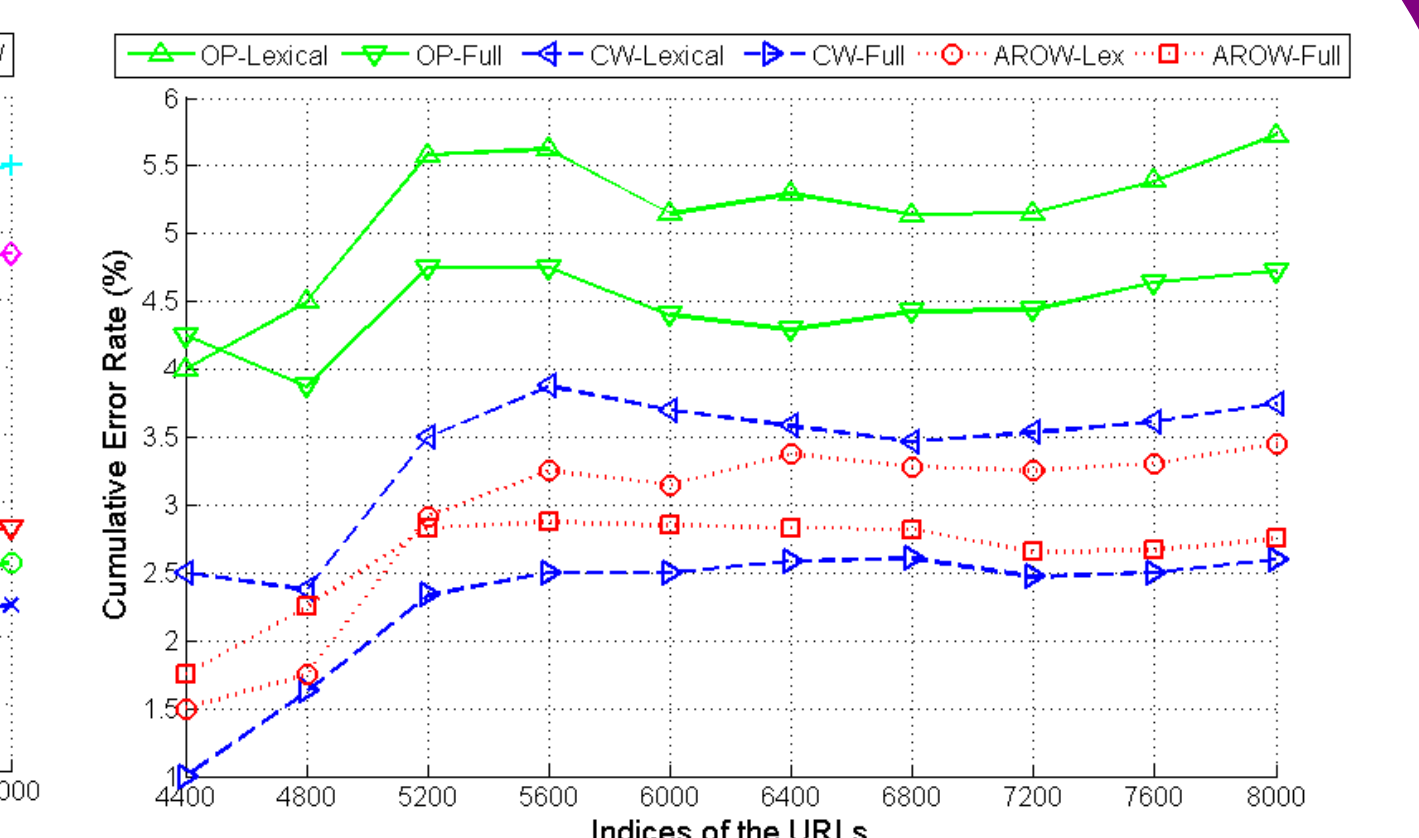
### Lexical Features

URL	www.naturenilai.com/form2/paypal/webscr.php?cmd=_login
<b>Auto-Selected</b>	name=www, name=naturenilai, tld=com, dir=form2, dir=paypal, file=webscr, ext=php, arg=cmd, arg=login
<b>Obfuscation-Resistant</b>	URL len=54, n_dot=3, blacklist=1
	Domain Name len=19, IP=0, port=0, n_token=3, n_hyphen=0, max_len=11
	Directory len=14, n_subdir=2, max_len=6, max_dot=0, max_delim=0
	File Name len=10, n_dot=1, n_delim=0
	Argument len=11, n_var=1, max_len=6, max_delim=1

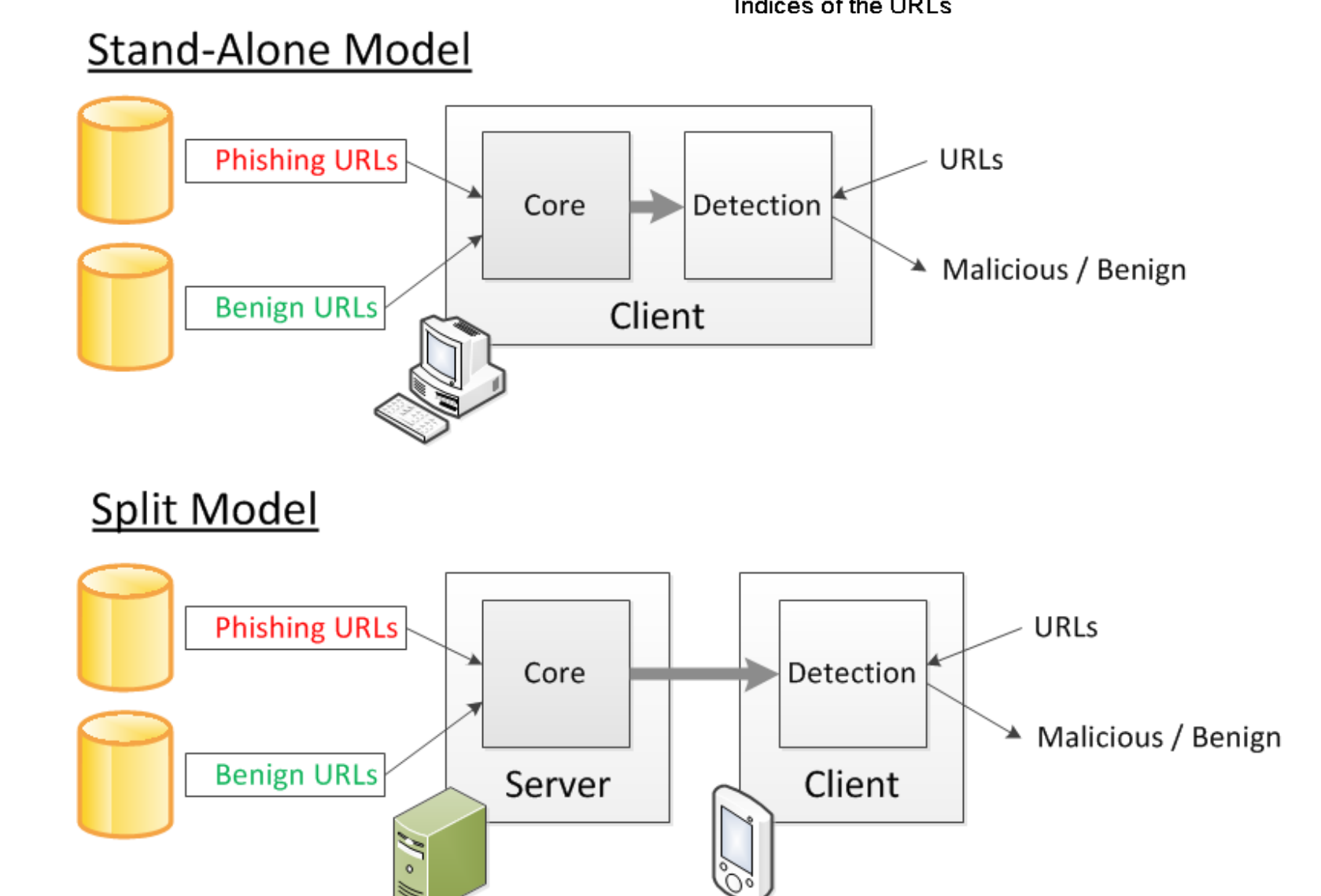
### Batch vs Online



### Full vs Lexical



### Deployment Options



Reference: A. Le, A. Markopoulou, M. Faloutsos, "PhishDef: URL Names Say It All," in Proc. of INFOCOM mini-conference 2011, <http://arxiv.org/abs/1009.2275>