Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs

Justin Ma, Lawrence Saul, Stefan Savage, Geoff Voelker
Computer Science & Engineering
UC San Diego

Presentation for KDD 2009
June 30, 2009

UCSDCSE
Computer Science and Engineering

Detecting Malicious Web Sites

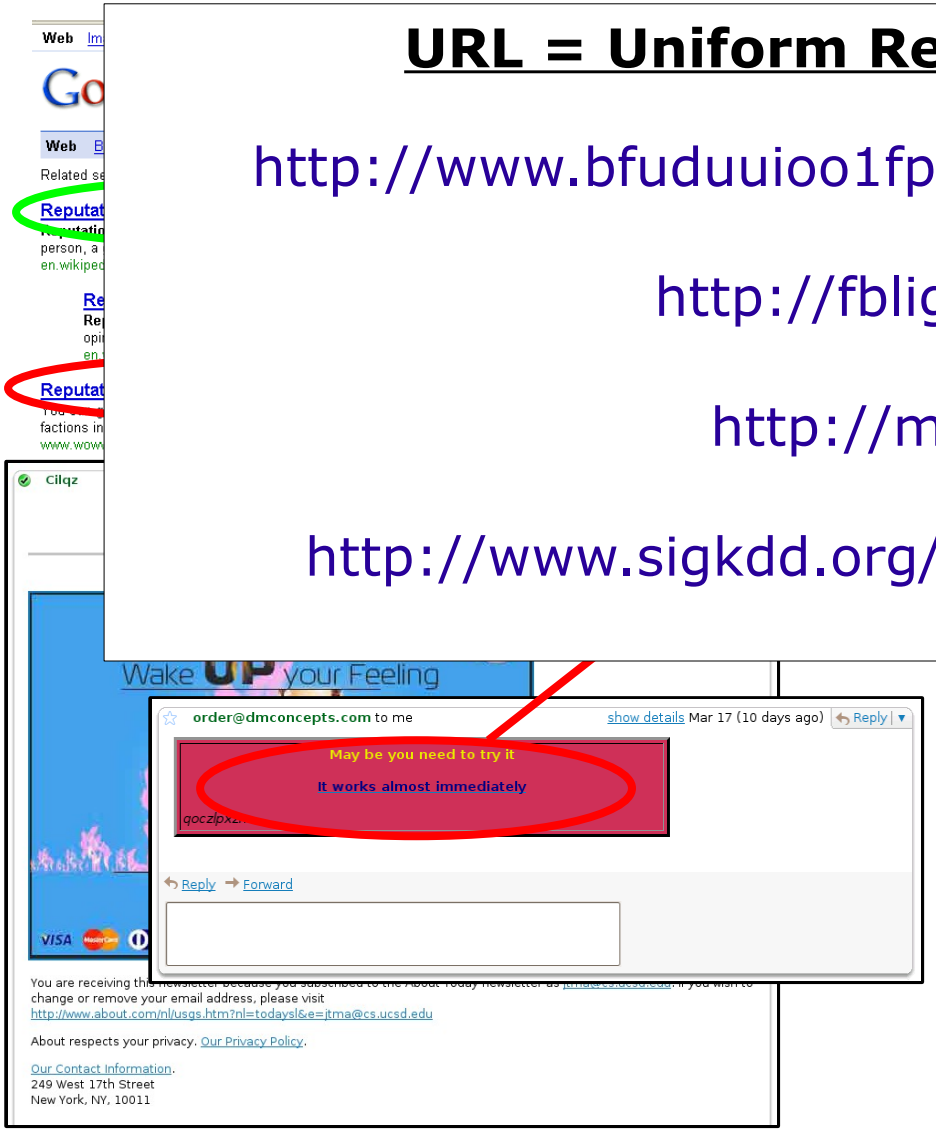
URL = Uniform Resource Locator

http://www.bfuduuioo1fp.mobi/ws/ebayisapi.dll

http://fblight.com

http://mail.ru

http://www.sigkdd.org/kdd2009/index.html

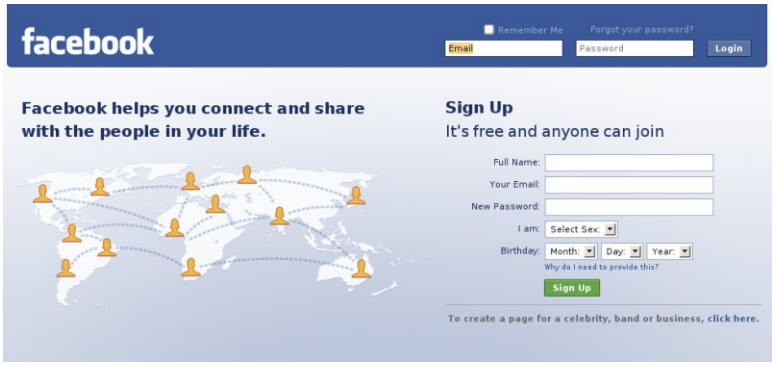


safe without committing to risky actions

Problem in a Nutshell

- URL features to identify malicious Web sites
  - No context, no content
- Different classes of URLs
  - Benign, spam, phishing, exploits, scams…
  - For now, distinguish benign vs. malicious

facebook.com



fblight.com

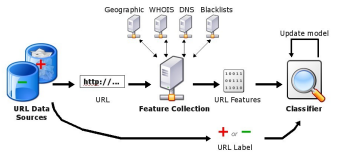
State of the Practice

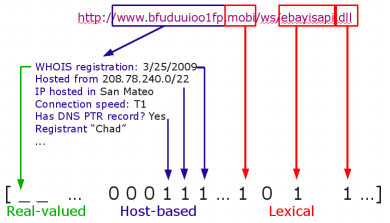
- Current approaches
  - Blacklists [SORBS, URIBL, SURBL, Spamhaus]
  - Learning on hand-tuned features [Garera et al, 2007]
- Limitations
  - Cannot predict unlisted sites
  - Cannot account for new features
- Arms race

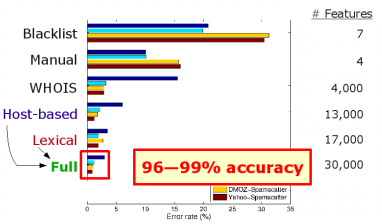
More automated approach?

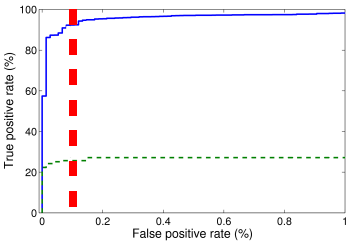
Today's Talk

- Motivation
- System overview
  - Training data
  - Algorithms
  - **Features** ← focus of today's talk
- Experimental results
- Conclusion

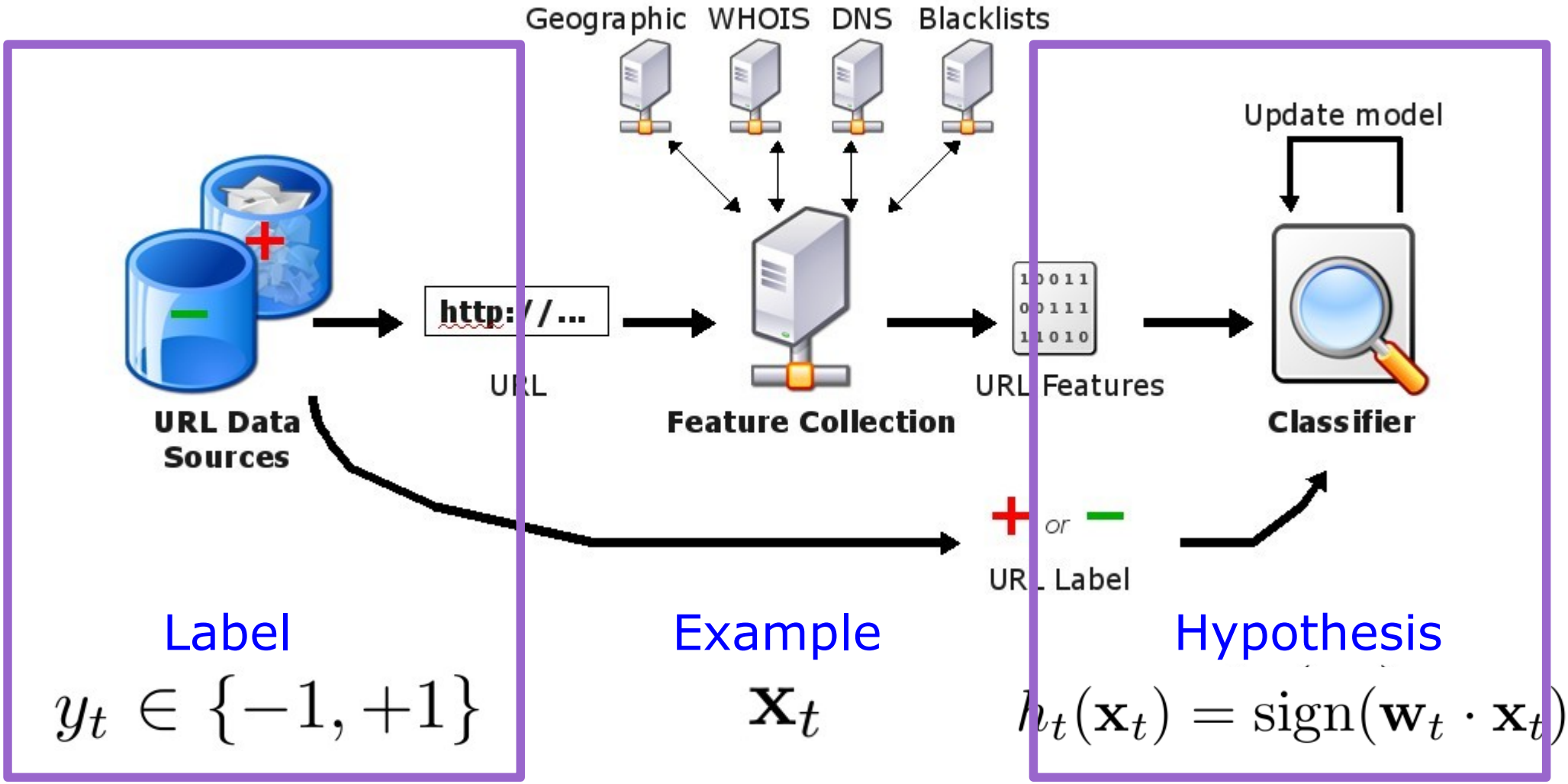








URL Classification System



Data Sets

- Malicious URLs
  - 5,000 from PhishTank (phishing)
  - 15,000 from Spamscatter (spam, phishing, etc)
- Benign URLs
  - 15,000 from Yahoo Web directory
  - 15,000 from DMOZ directory
- Malicious x Benign → 4 Data Sets
  - 30,000 – 55,000 features per data set

Algorithms

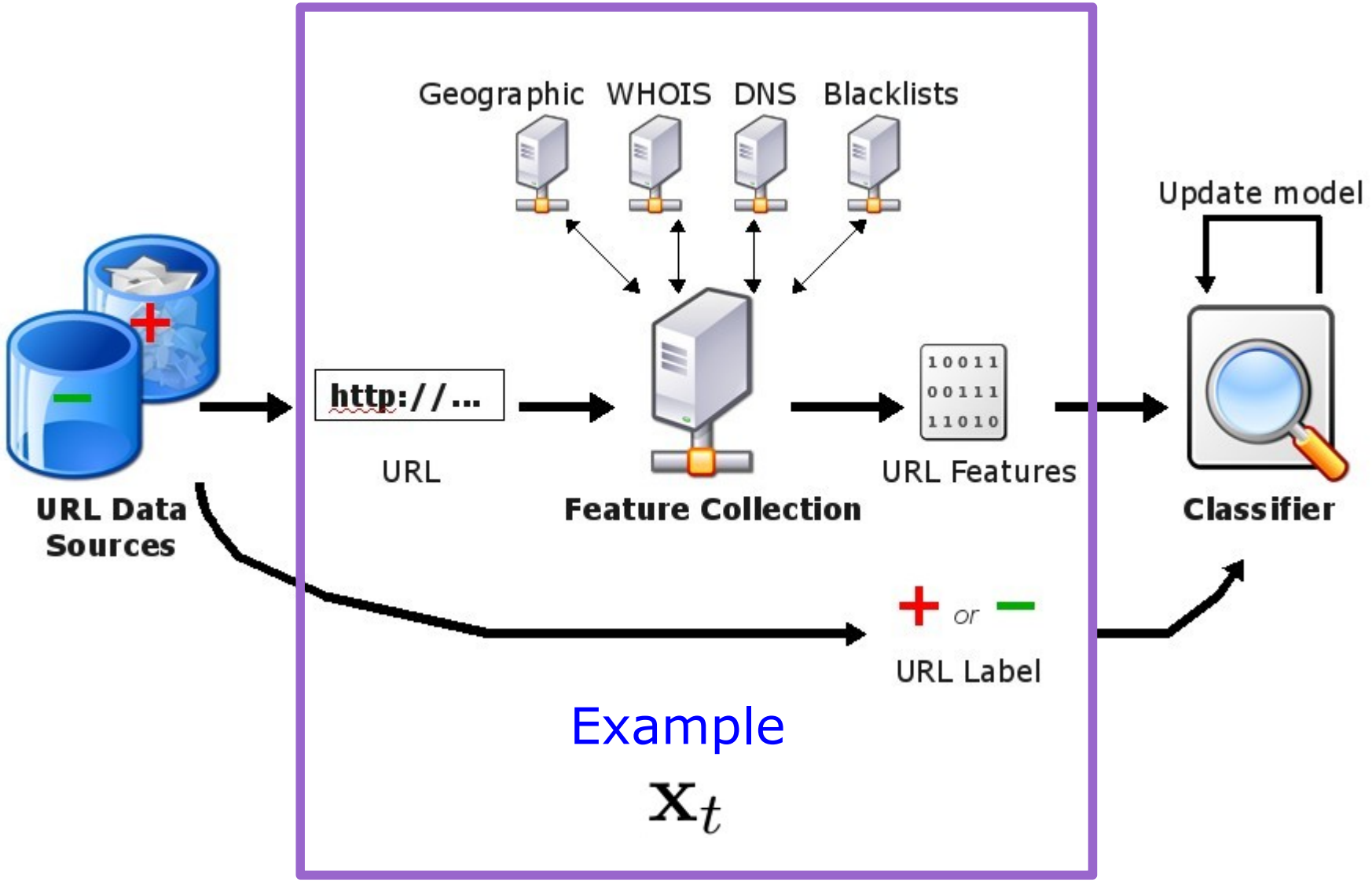
- Logistic regression w/ L1-norm regularization

$$L(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \log P(y_i | \mathbf{x}_i, \mathbf{w}) - \lambda \|\mathbf{w}\|_1$$

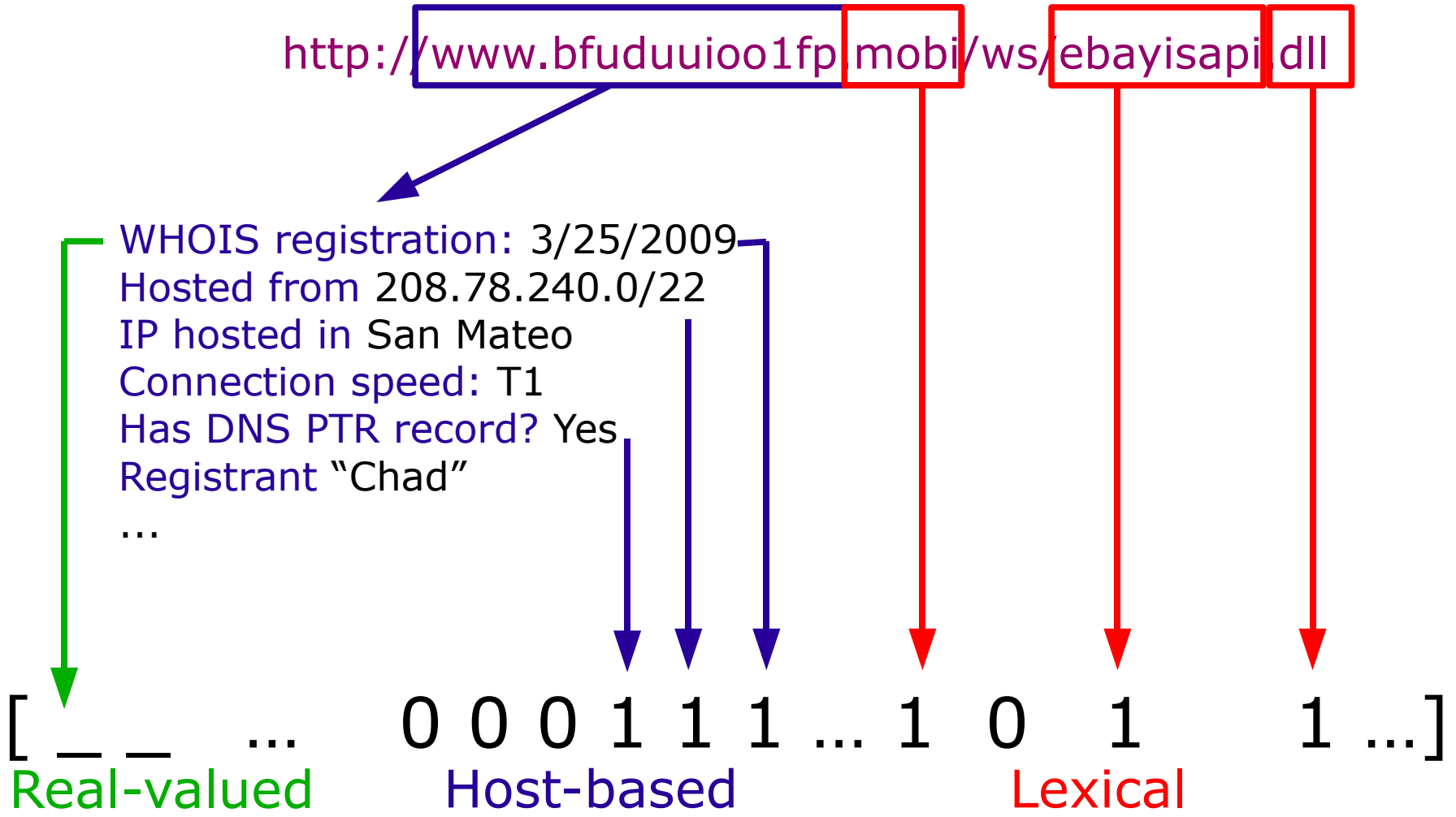
  - Implicit feature selection
  - Easier to interpret

- Other models
  - Naive Bayes
  - Support vector machines (linear, RBF kernels)

Today's Focus



Feature vector construction



Features to consider?

1) Blacklists 

2) Simple heuristics 

3) Domain name registration 

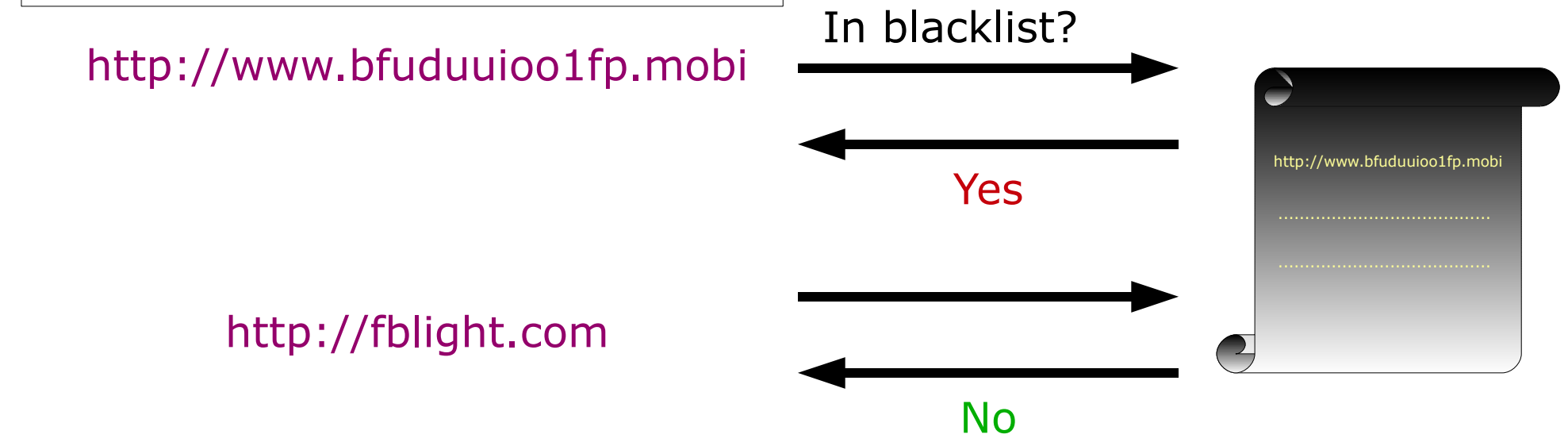
4) Host properties 

5) Lexical 

(1) Blacklist Queries

- List of known malicious sites
- Providers: SORBS, URIBL, SURBL, Spamhaus

Blacklist queries as features



(2) Manually-Selected Features

[Fette et al., 2007][Zhang et al., 2007][Bergholz et al., 2008]

- Considered by previous studies
  - IP address in hostname?
  - Number of dots in URL
  - WHOIS (domain name) registration date

http://72.23.5.122/www.bankofamerica.com/

http://www.bankofamerica.com.qytrpbcw.stopgap.cn/

stopgap.cn registered
28 June 2009

(3) WHOIS Features

- Domain name registration
  - Date of registration, update, expiration
  - Registrant: Who registered domain?
  - Registrar: Who manages registration?

http://yammeringyellowtail.com

http://angryalbacore.com

http://sleazysalmon.com

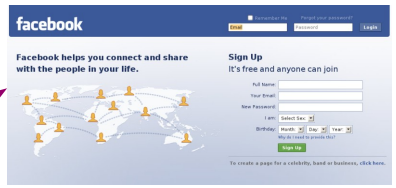
http://mangymackerel.com

Registered on
29 June 2009
By SpamMedia

(4) Host-Based Features

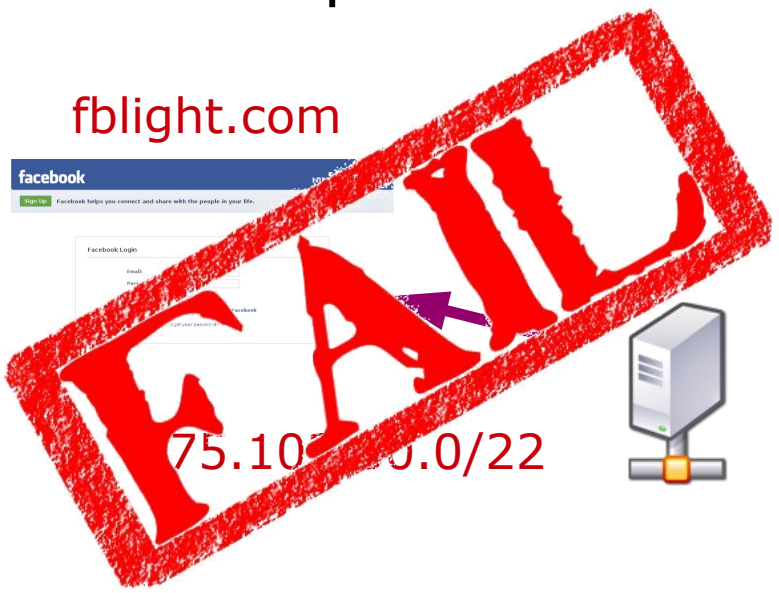
- Blacklisted? (SORBS, URIBL, SURBL, Spamhaus)
- WHOIS: registrar, registrant, dates
- IP address: Which ASes/IP prefixes?
- DNS: TTL? PTR record exists/resolves?
- Geography-related: Locale? Connection speed?

facebook.com



69.63.176.0/20

fblight.com



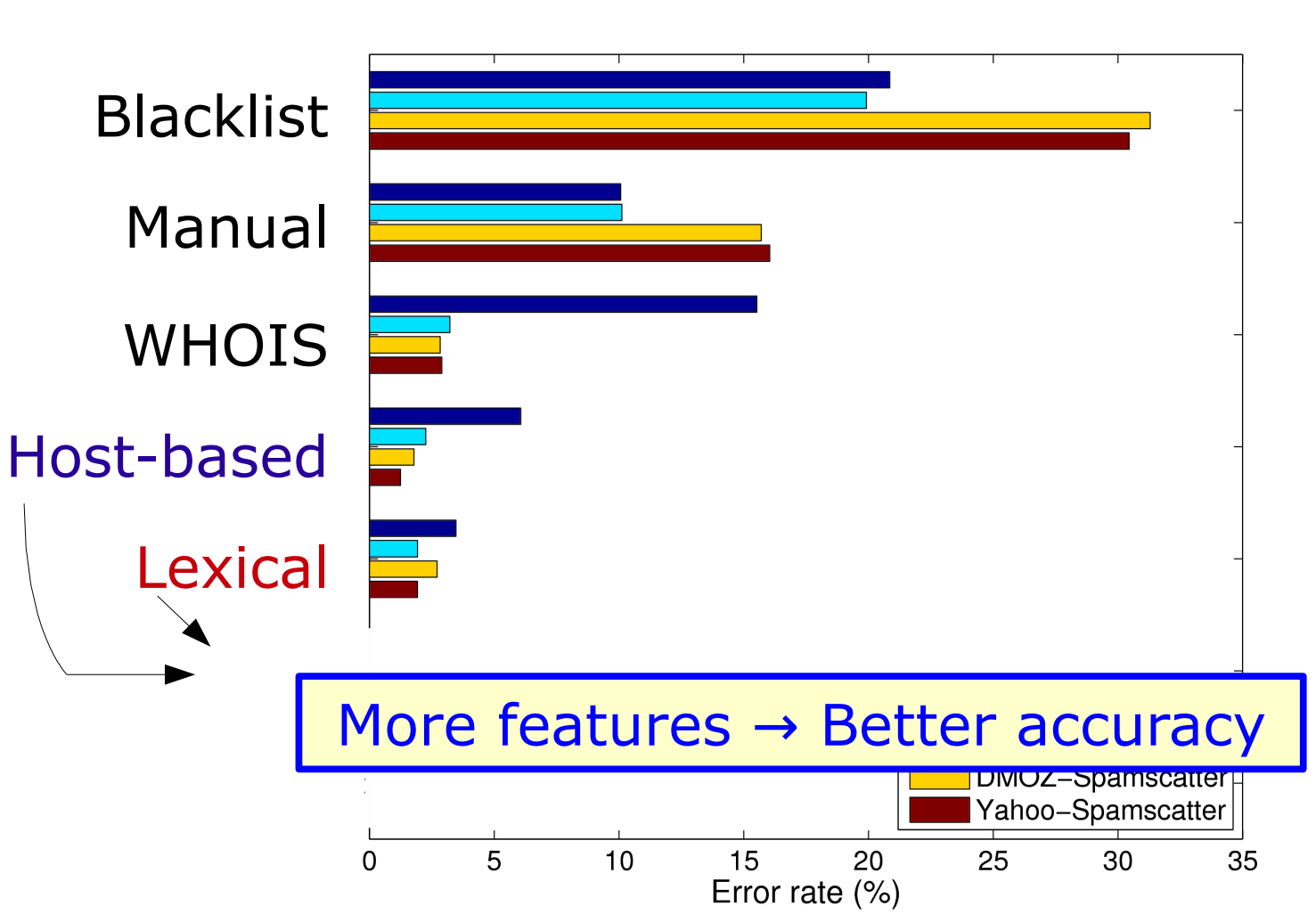
75.10[illegible].0.0/22

(5) Lexical Features

- Tokens in URL hostname + path
- Length of URL
- Number of dots

http://www.bfuduuioo1fp.mobi/ws/ebayisapi.dll

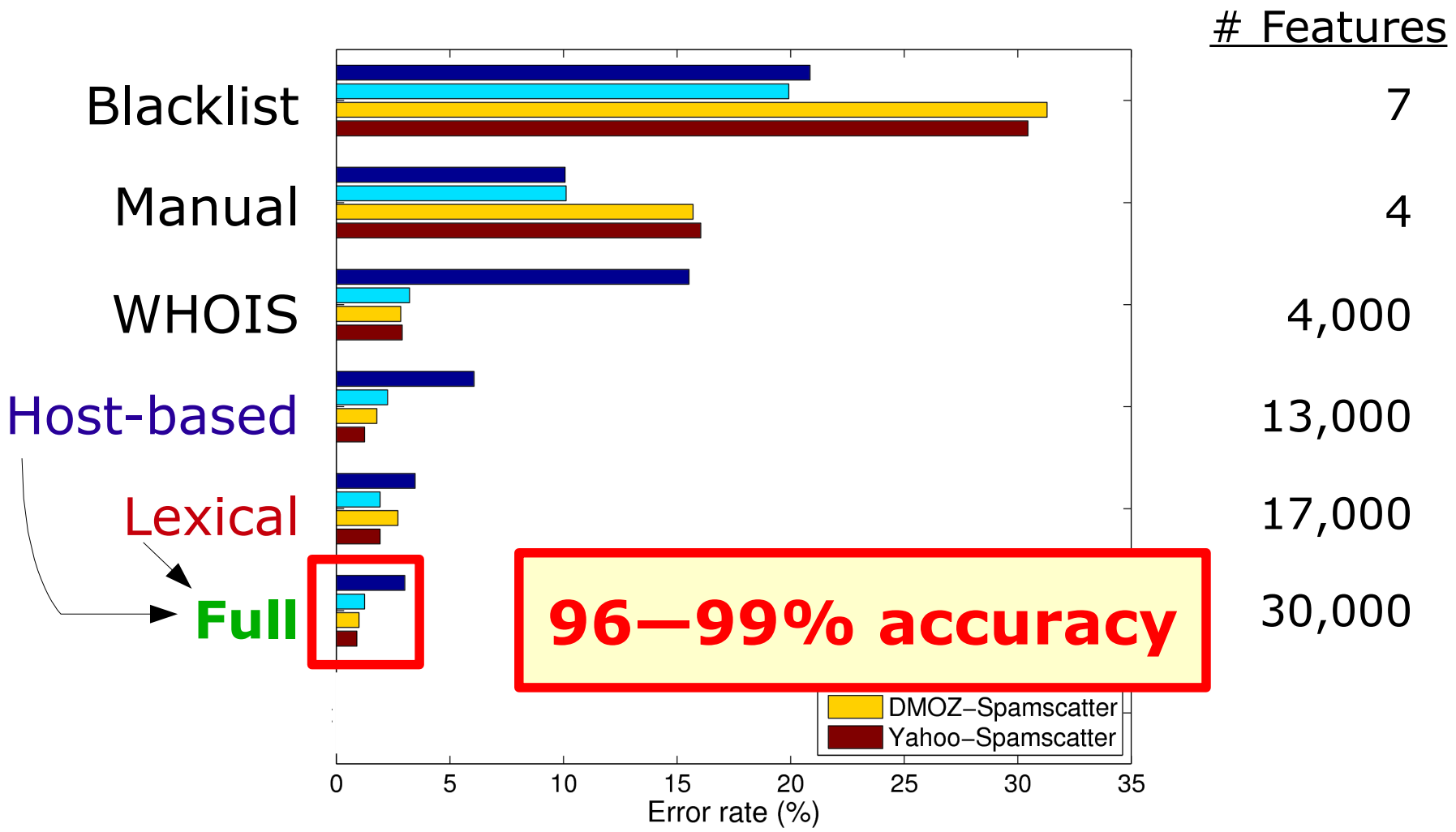
Which feature sets?



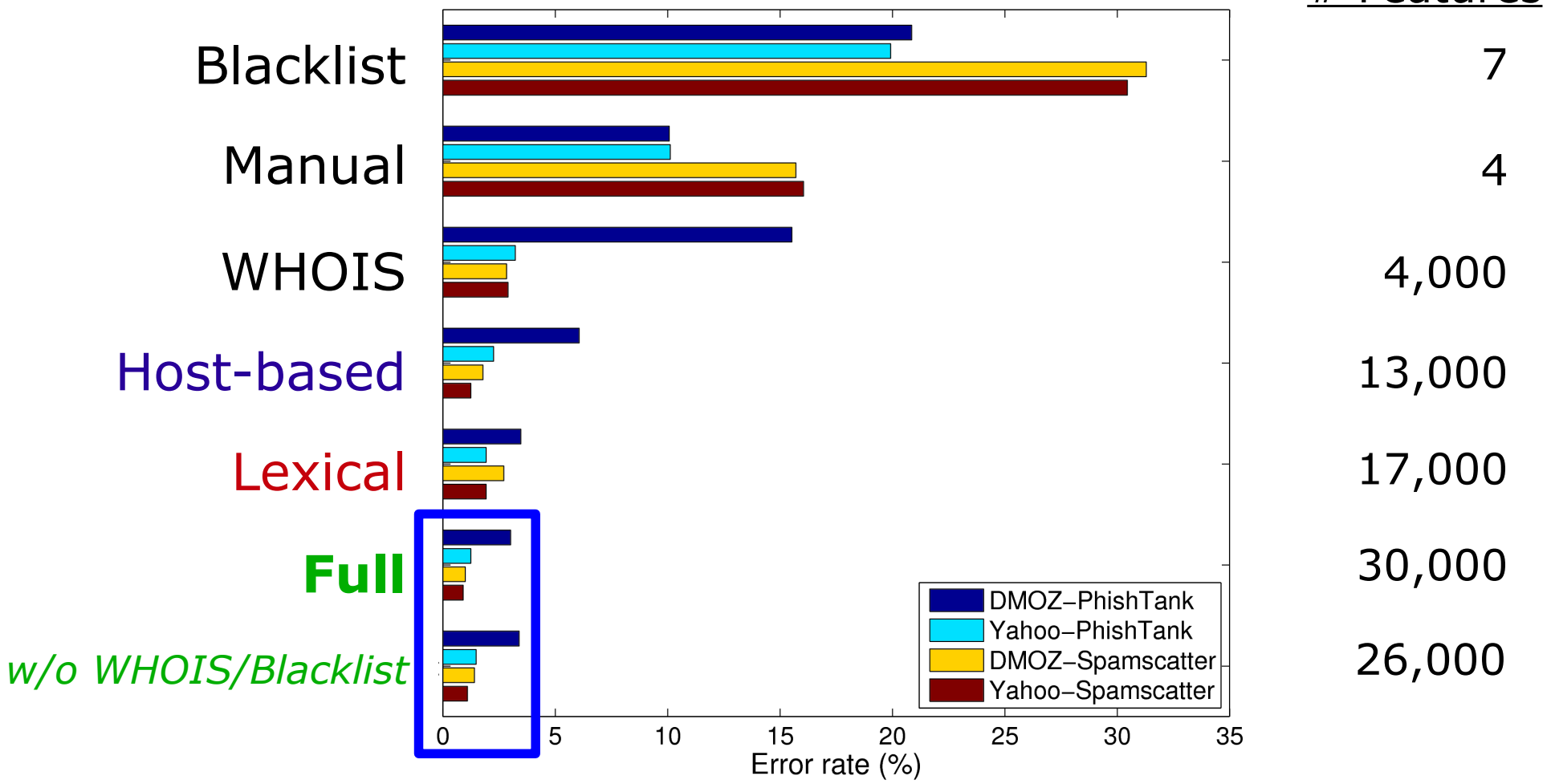
	# Features
Blacklist	7
Manual	4
WHOIS	4,000
Host-based	13,000
Lexical	17,000

More features → Better accuracy

Which feature sets?



Which feature sets?

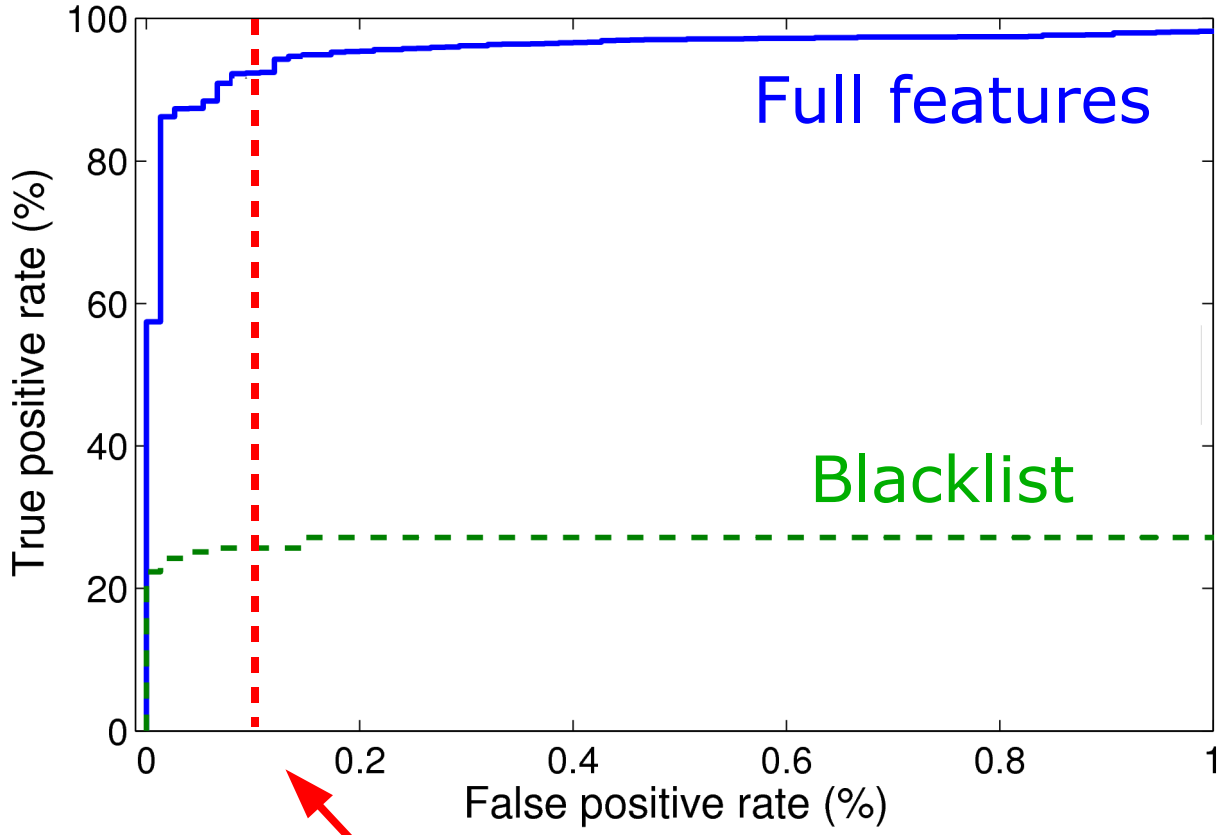


	# Features
Blacklist	7
Manual	4
WHOIS	4,000
Host-based	13,000
Lexical	17,000
Full	30,000
w/o WHOIS/Blacklist	26,000

UCSDCSE
Computer Science and Engineering

Beyond Blacklists

Yahoo-PhishTank



Higher detection rate for given false positive rate

Limitations

- False positives
  - Sites hosted in disreputable ISP
  - Guilt by association
- False negatives
  - Compromised sites
  - Free hosting sites
  - Redirection (but we consider TinyURL malicious :)
  - Hosted in reputable ISP
- Future work: Web page content

Conclusion

- Detect malicious URLs with high accuracy
  - Only using URL
  - Diverse feature set helps: 99% w/ 30,000+ features
  - Model analysis (more in paper)
- Our related efforts
  - Online learning for URL reputation [ICML 2009]
- Future work
  - Scaling up for deployment