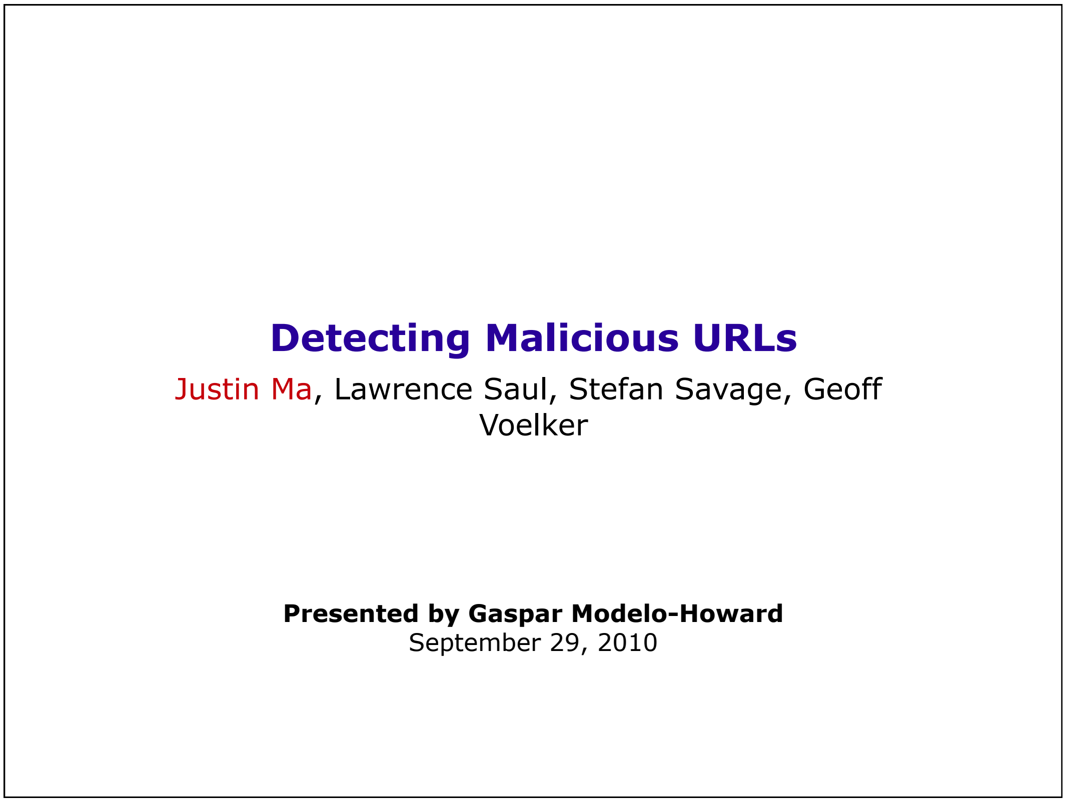
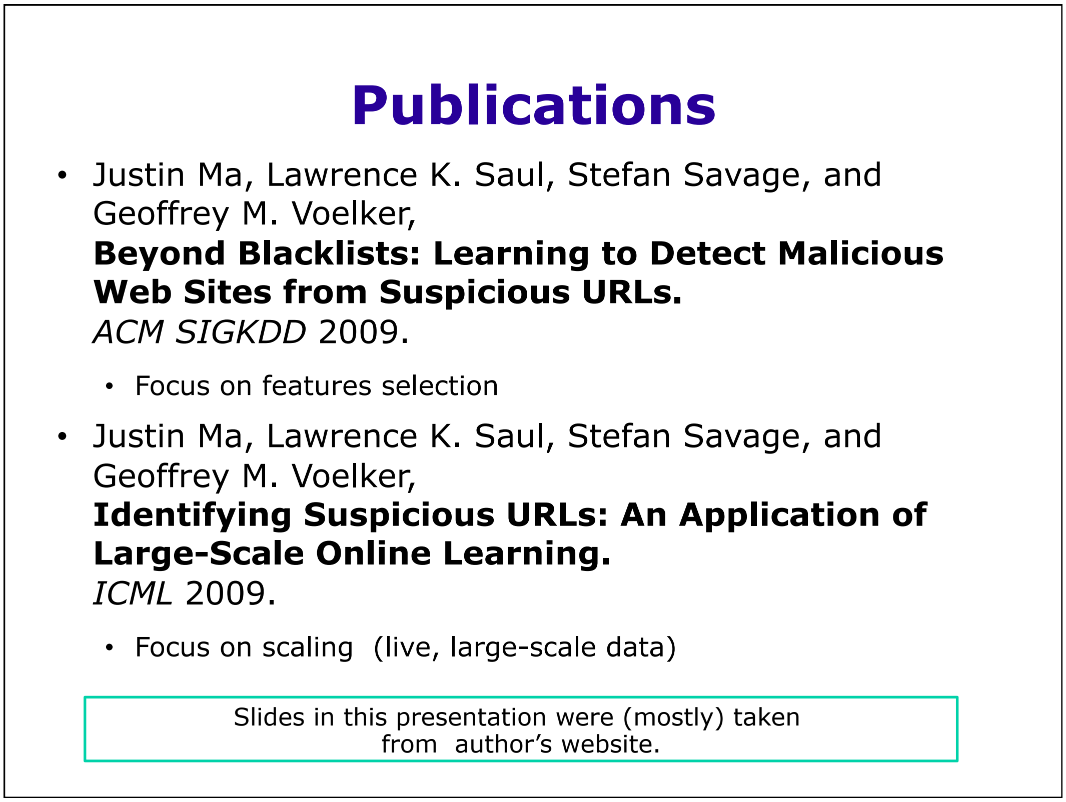


Detecting Malicious URLs

Justin Ma, Lawrence Saul, Stefan Savage, Geoff Voelker

Presented by Gaspar Modelo-Howard
September 29, 2010



Publications

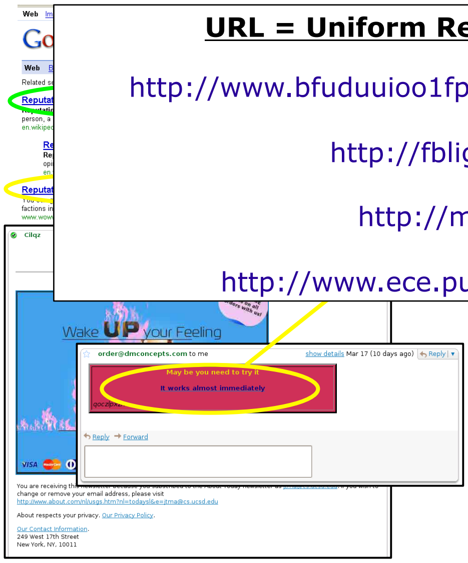
- Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker, **Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs.** *ACM SIGKDD* 2009.
  - Focus on features selection
- Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker, **Identifying Suspicious URLs: An Application of Large-Scale Online Learning.** *ICML* 2009.
  - Focus on scaling (live, large-scale data)

Slides in this presentation were (mostly) taken from author's website.

Agenda

- **Problem**
- **State of Practice**
- Beyond Blacklists
- Suspicious URLs and Large-Scale Online Learning
- Conclusion

Detecting Malicious Web Sites



order@dmconcepts.com to me show details Mar 17 (10 days ago) Reply

May be you need to try it

It works almost immediately

qoczlpx

Reply Forward

You are receiving this newsletter because you subscribed to the About Today newsletter as jtma@cs.ucsd.edu. If you wish to change or remove your email address, please visit http://www.about.com/nl/usgs.htm?nl=todaysl&e=jtma@cs.ucsd.edu

About respects your privacy. Our Privacy Policy.

Our Contact Information.
249 West 17th Street
New York, NY, 10011

URL = Uniform Resource Locator

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

http://fblight.com

http://mail.ru

http://www.ece.purdue.edu/~dcsl

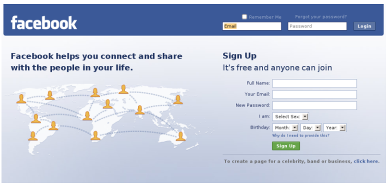
e?

Predict what is safe without committing to risky actions

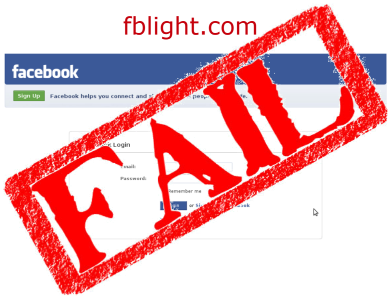
Problem in a Nutshell

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

facebook.com

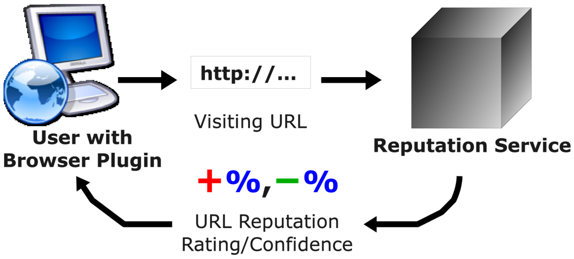

facebook
Facebook helps you connect and share with the people in your life.
Sign Up
It's free and anyone can join

fblight.com



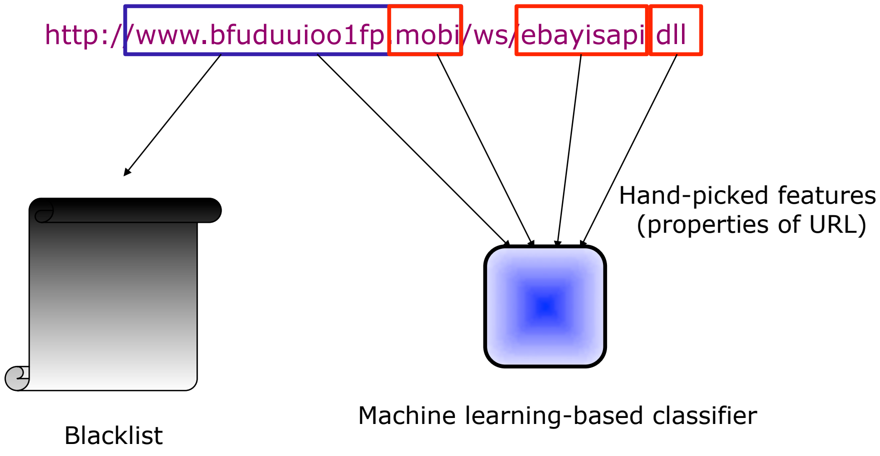
5

What we want...



6

How to build this service?



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

Hand-picked features (properties of URL)

Blacklist

Machine learning-based classifier

7

State of the Practice

- Current approaches
  - Blacklists [SORBS, URIBL, SURBL, Spamhaus, SiteAdvisor, WOT, IronPort, WebSense]
  - Learning on hand-tuned features [Kan & Thi ’05, Guan et al ’09]

- Limitations
  - Cannot learn from newest examples quickly
  - Cannot quickly adapt to newest features
- Arms race: fast feedback cycle is critical

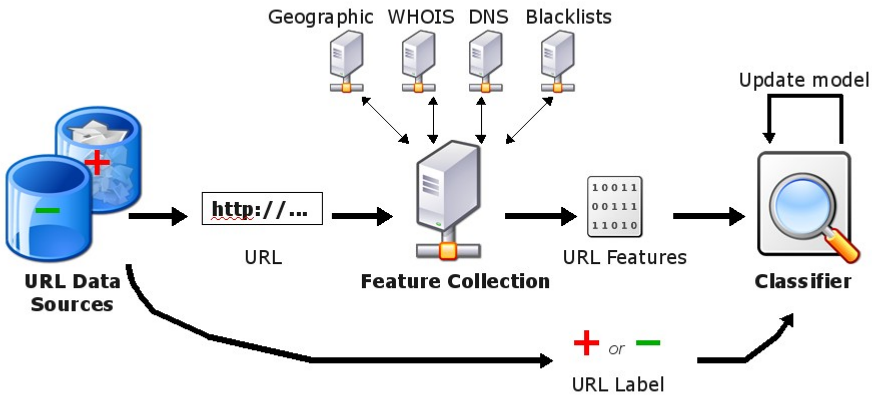
More automated approach?

8

Agenda

- Problem
- State of Practice
- **Beyond Blacklists**
- Suspicious URLs and Large-Scale Online Learning
- Conclusion

URL Classification System



Label
$y_t \in \{-1, +1\}$

Example
$\mathbf{x}_t$

Hypothesis
$h_t(\mathbf{x}_t)$

Data Sets

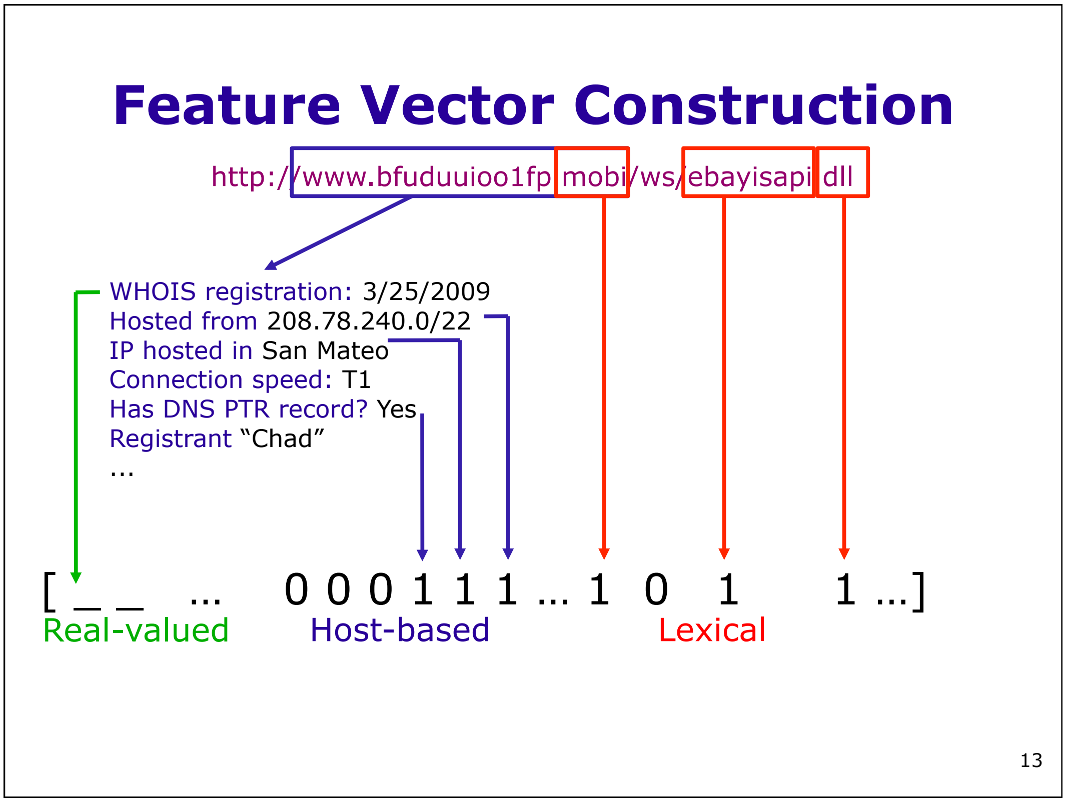
- Malicious URLs
  - 5,000 from PhishTank (phishing)
  - 15,000 from Spamscatter (spam, phishing, etc)
- Benign URLs
  - 15,000 from Yahoo Web directory
  - 15,000 from DMOZ directory
- Malicious x Benign → 4 Data Sets
  - 30,000 – 55,000 features per data set

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)



Feature Vector Construction

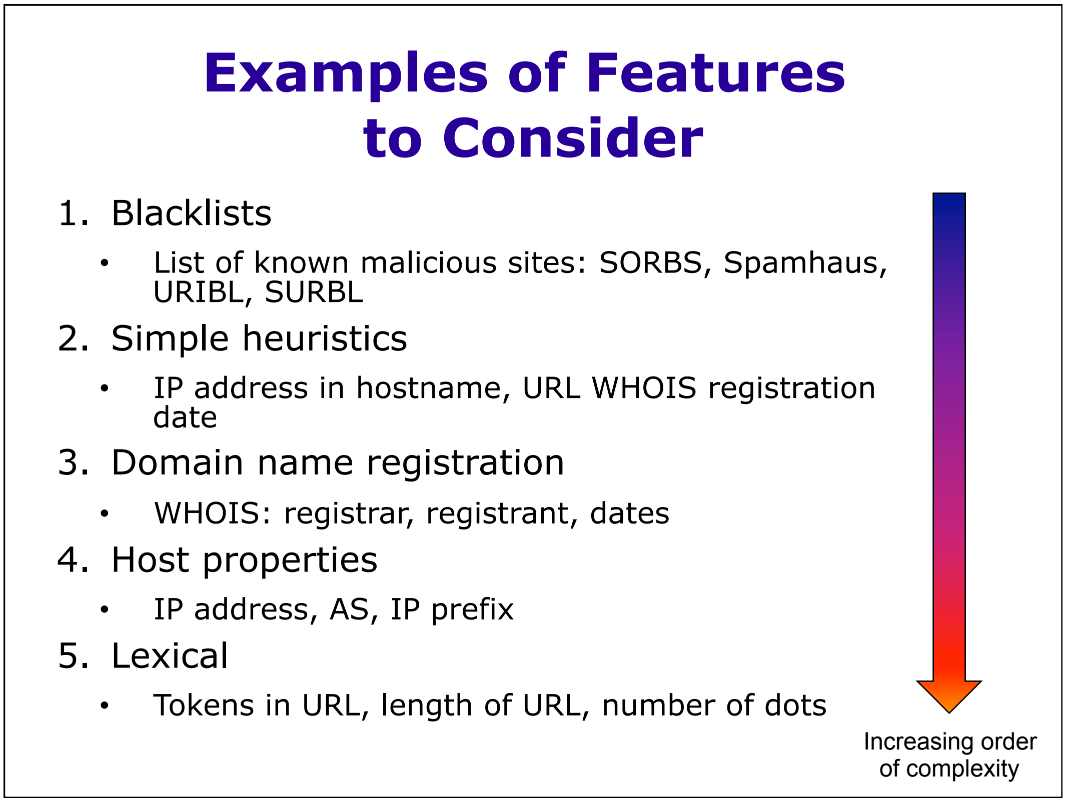
http://www.bfuduuioo1fp.mobi/ws/ebayisapi.dll

WHOIS registration: 3/25/2009
Hosted from 208.78.240.0/22
IP hosted in San Mateo
Connection speed: T1
Has DNS PTR record? Yes
Registrant "Chad"
…

[_ _ … 0 0 0 1 1 1 … 1 0 1 1 …]

Real-valued   Host-based   Lexical

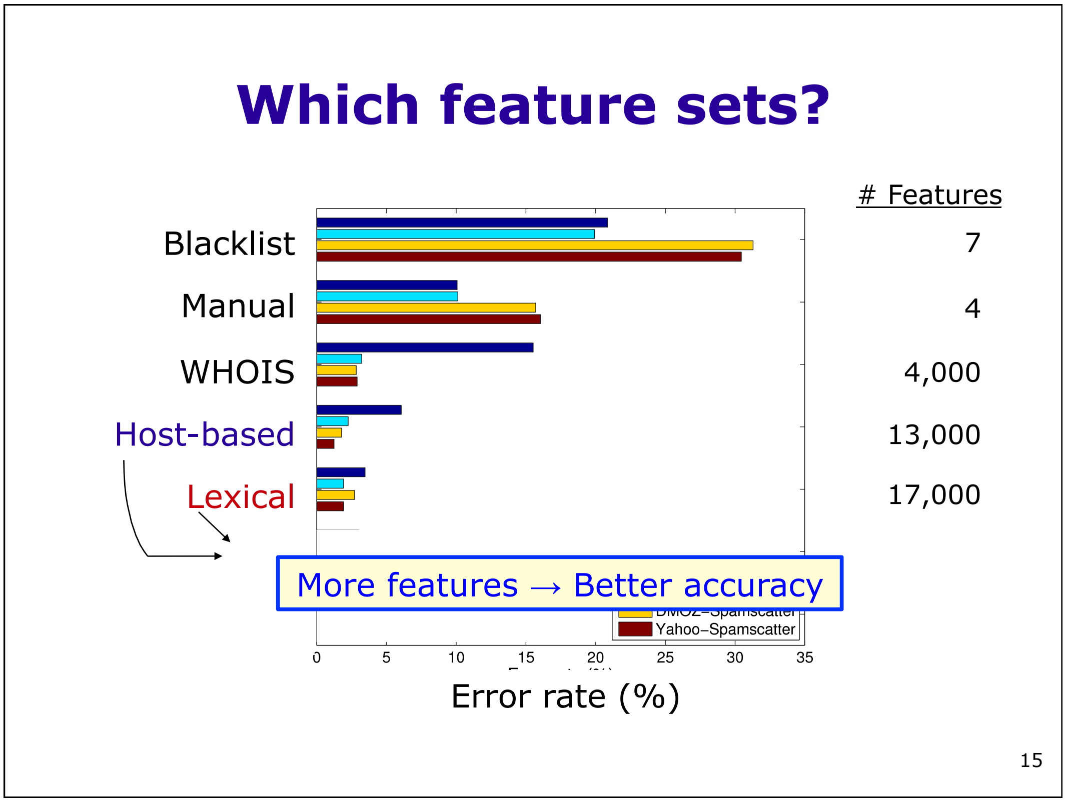
13



Examples of Features to Consider

1. Blacklists
   - List of known malicious sites: SORBS, Spamhaus, URIBL, SURBL
2. Simple heuristics
   - IP address in hostname, URL WHOIS registration date
3. Domain name registration
   - WHOIS: registrar, registrant, dates
4. Host properties
   - IP address, AS, IP prefix
5. Lexical
   - Tokens in URL, length of URL, number of dots

Increasing order of complexity

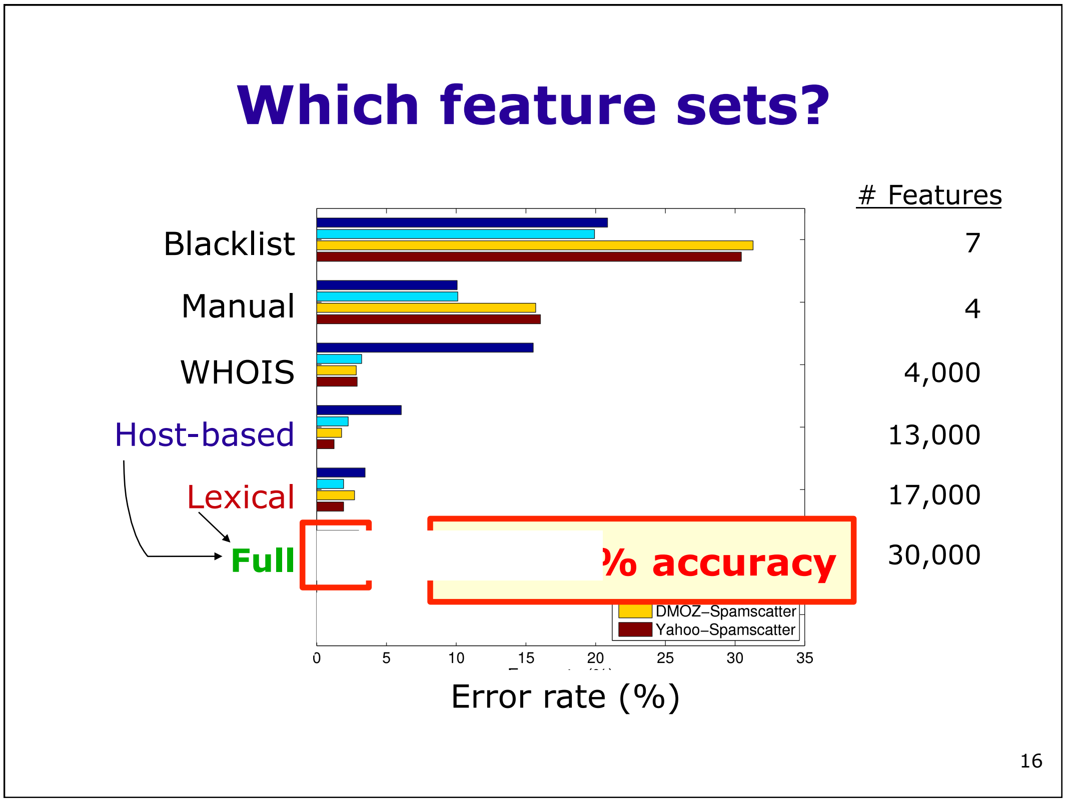


Which feature sets?

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

More features → Better accuracy

15



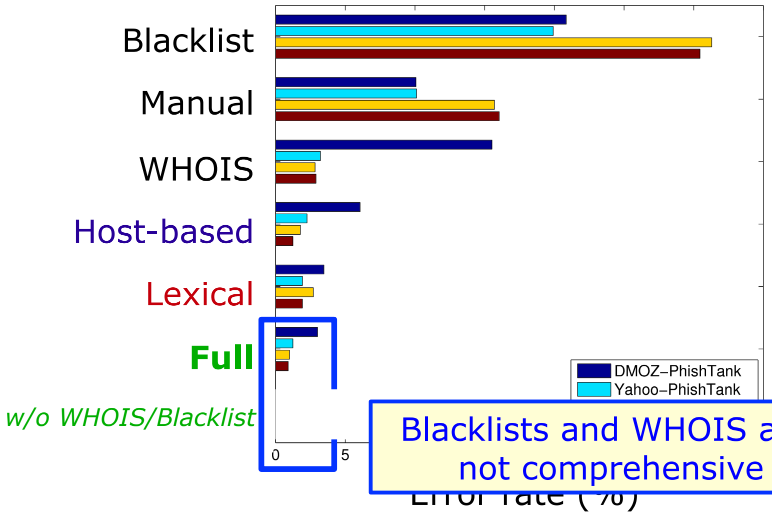
Which feature sets?

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

% accuracy

16

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

Blacklists and WHOIS are not comprehensive

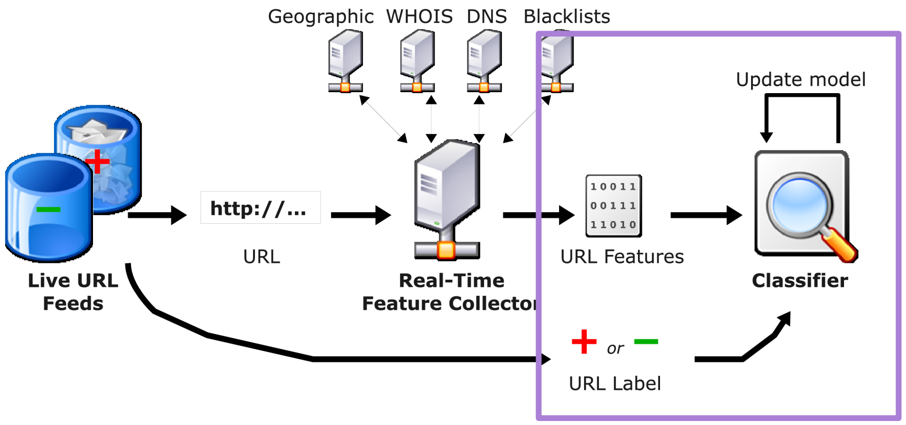
17

Agenda

- Problem
- State of Practice
- Beyond Blacklists
- **Suspicious URLs and Large-Scale Online Learning**
- Conclusion

18

Live URL Classficiation System



- Millions of examples and features → Online learning

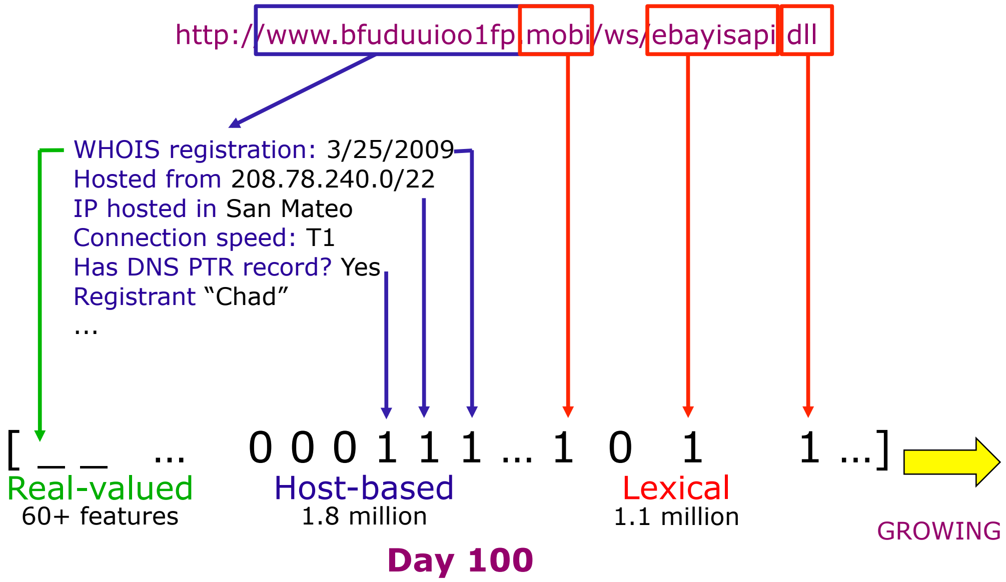
19

Live Training Feed

- Malicious URLs (spamming and phishing)
  - 6,000—7,500 per day from Web mail provider
- Benign URLs
  - From Yahoo Web directory
- Total of 20,000 URLs per day
- Live collection since Jan. 5, 2009
  - Months of data
  - Two million examples after 100 days

20

Feature vector construction



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

WHOIS registration: 3/25/2009
Hosted from 208.78.240.0/22
IP hosted in San Mateo
Connection speed: T1
Has DNS PTR record? Yes
Registrant "Chad"
…

[_ _ … 0 0 0 1 1 1 … 1 0 1 1 …]

Real-valued
60+ features

Host-based
1.8 million

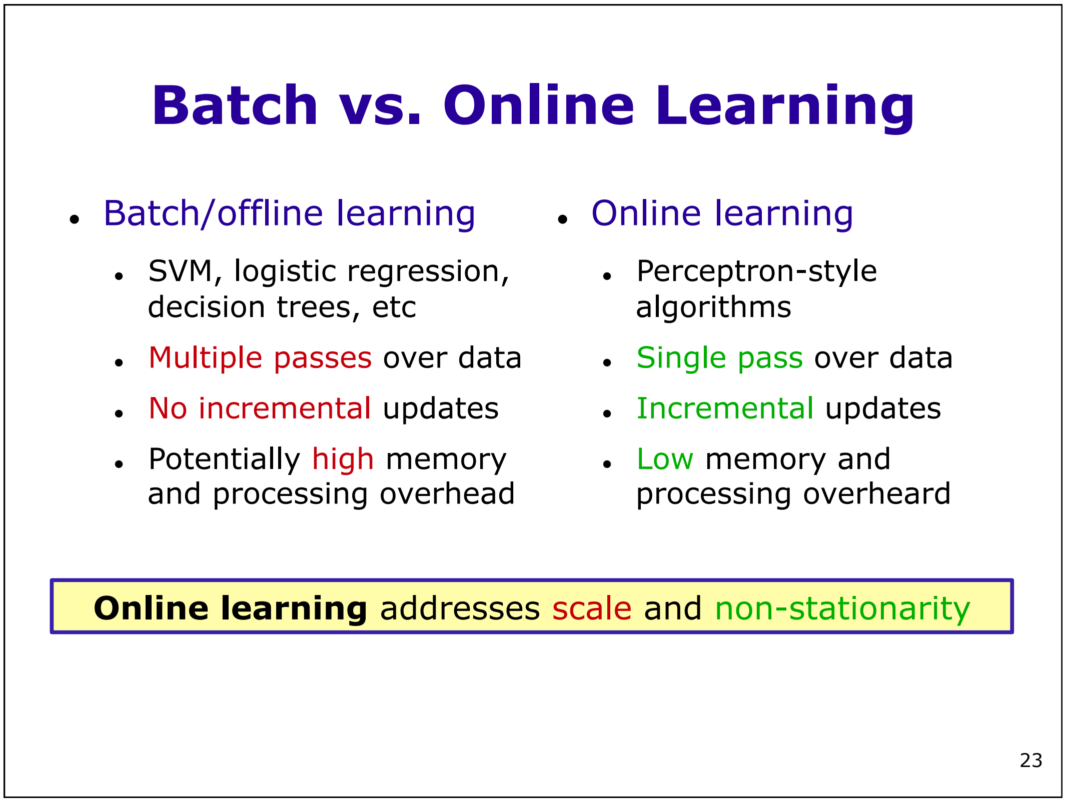
Lexical
1.1 million

GROWING

Day 100

Practical Challenges of ML in Systems

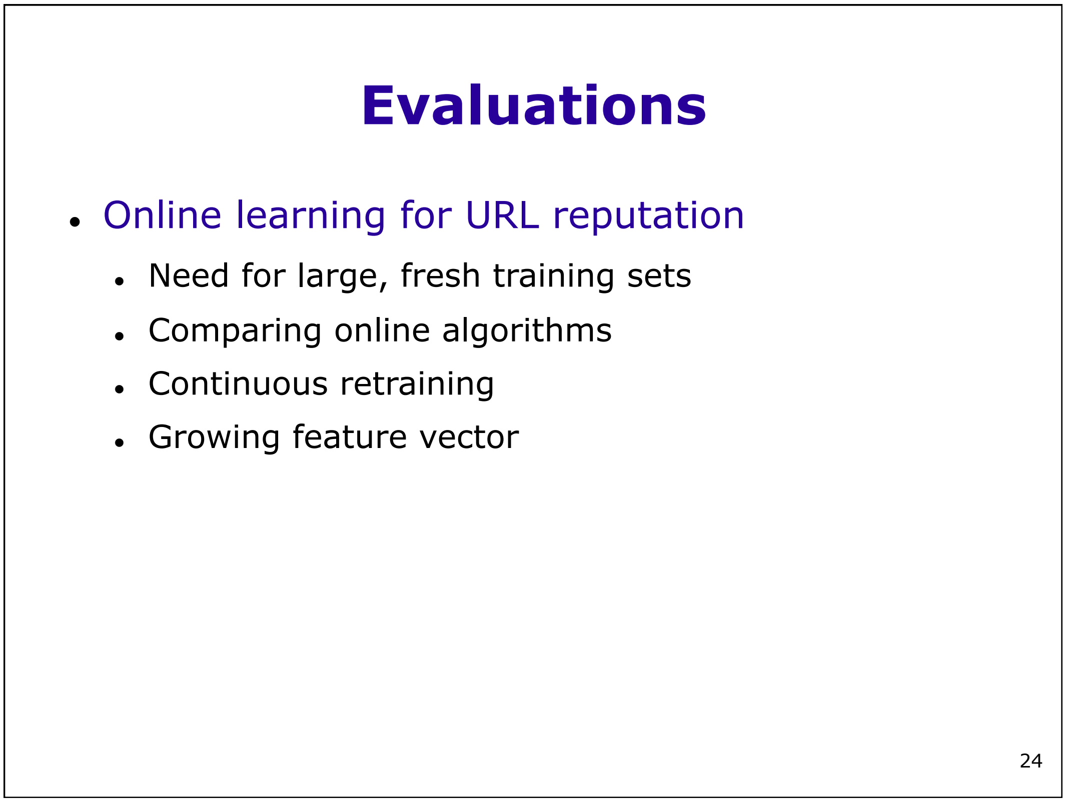
- Industrial concerns
  - Scale: millions of examples, features
  - Non-stationarity: examples change over time (arms race w/ criminals)
- Pivotal decision: batch or online?



Batch vs. Online Learning

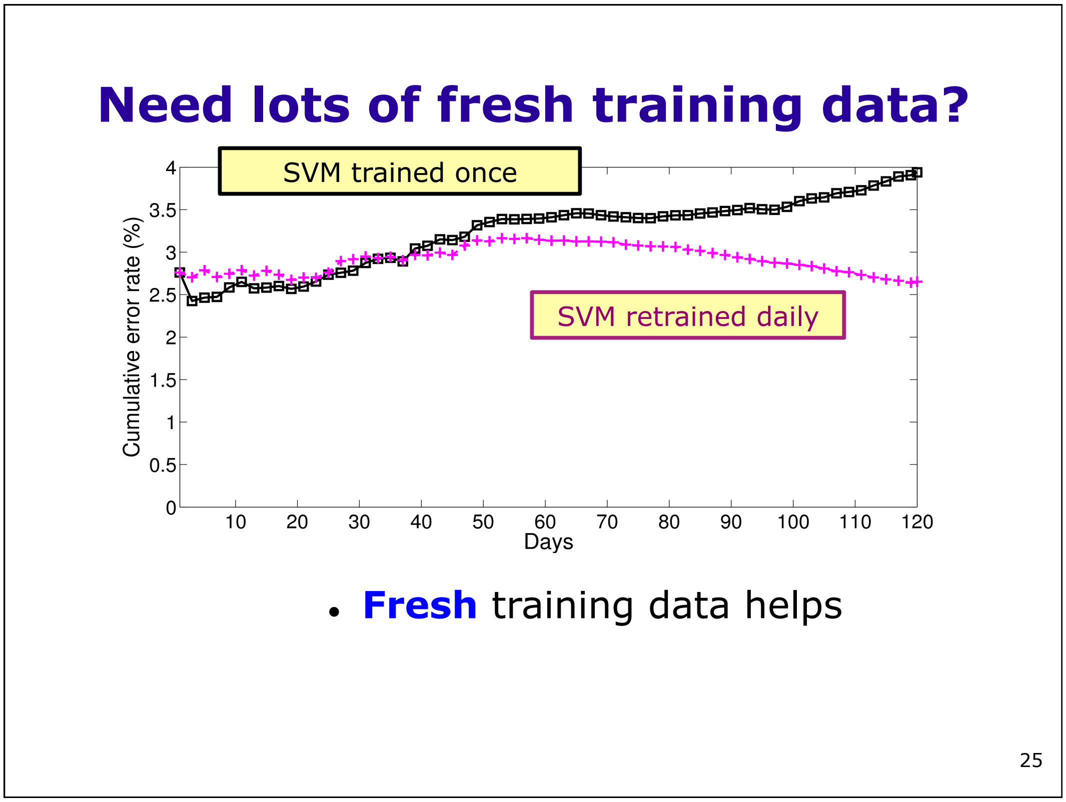
- Batch/offline learning
  - SVM, logistic regression, decision trees, etc
  - Multiple passes over data
  - No incremental updates
  - Potentially high memory and processing overhead
- Online learning
  - Perceptron-style algorithms
  - Single pass over data
  - Incremental updates
  - Low memory and processing overheard

Online learning addresses scale and non-stationarity



Evaluations

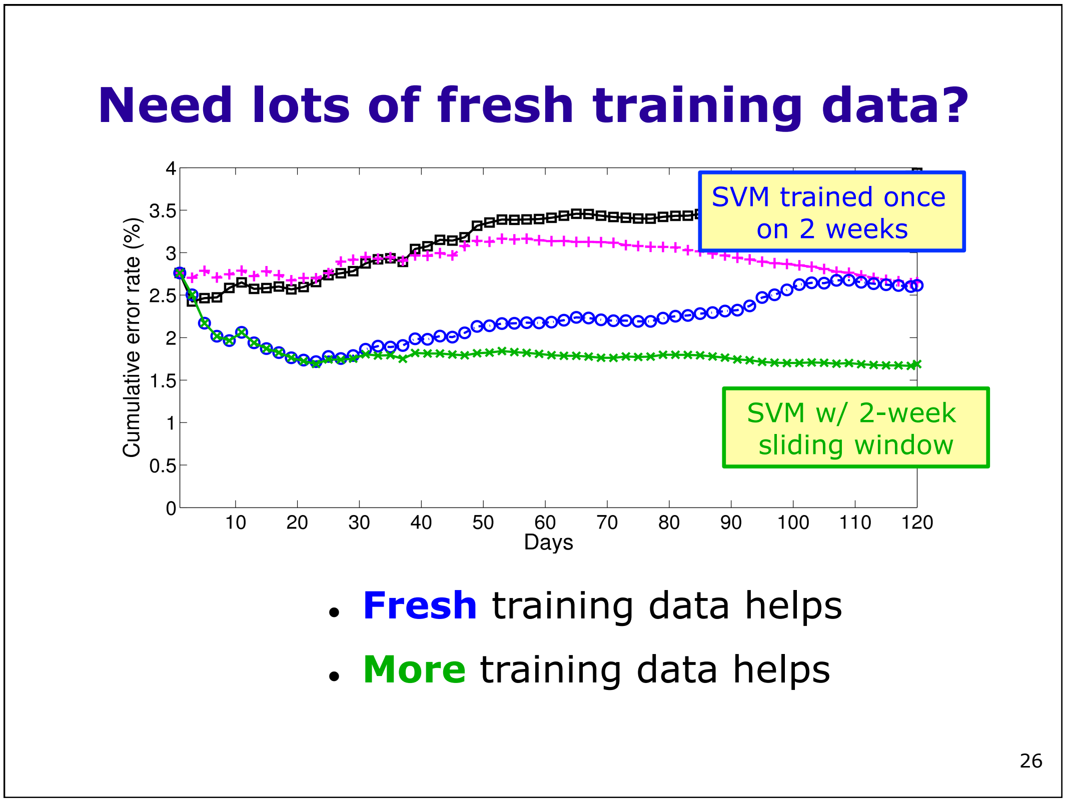
- Online learning for URL reputation
  - Need for large, fresh training sets
  - Comparing online algorithms
  - Continuous retraining
  - Growing feature vector



Need lots of fresh training data?

- **Fresh** training data helps

25



Need lots of fresh training data?

- **Fresh** training data helps
- **More** training data helps

26

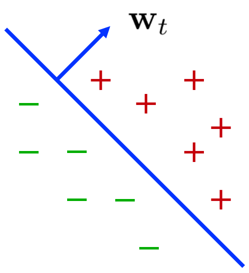
Which online algorithm?

- Perceptron
- Stochastic Gradient Descent for Logistic Regression
- Confidence-Weighted Learning

27

Perceptron

[Rosenblatt, 1958]



- Convergence result:

Number of mistakes ≤ $\frac{R^2}{\gamma^2}$

radius

margin

- Update on each mistake:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \mathbf{x_t}$$

28

Logistic Regression with SGD

[Bottou, 1998]

- Log likelihood:

$$L_t(\mathbf{w}) = \log \sigma(y_t(\mathbf{w} \cdot \mathbf{x}_t))$$

- For every example:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \gamma \frac{\partial L_t}{\partial \mathbf{w}} = \mathbf{w}_t + \boxed{\gamma \Delta_t \mathbf{x}_t}$$

where

$$\boxed{\Delta_t = \tfrac{y_t+1}{2} - \sigma(\mathbf{w_t} \cdot \mathbf{x_t})}$$

Proportional

Confidence-Weighted Learning

[Dredze et al., 2008] [Crammer et al., 2009]

- Maintain Gaussian distribution over weight vector:

$$\mathbf{w}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

- Constrained problem:

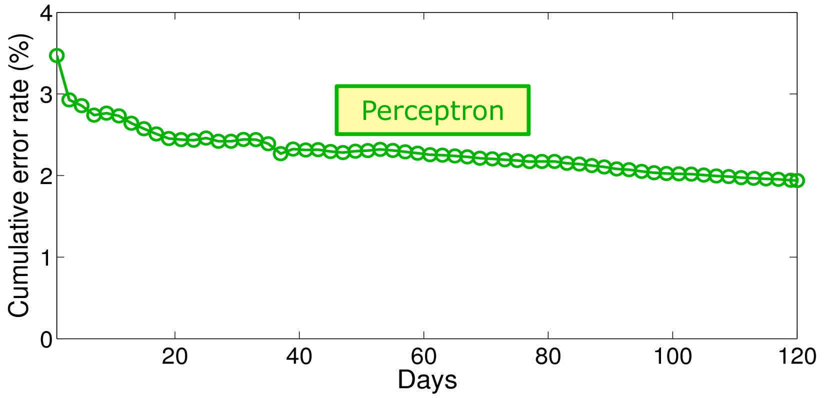
$$(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}) \leftarrow \operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \| \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t))$$
$$\text{s.t. } \Pr[y_t(\mathbf{w} \cdot \mathbf{x}_t) \geq 0] \geq \eta$$

- Closed-form update:

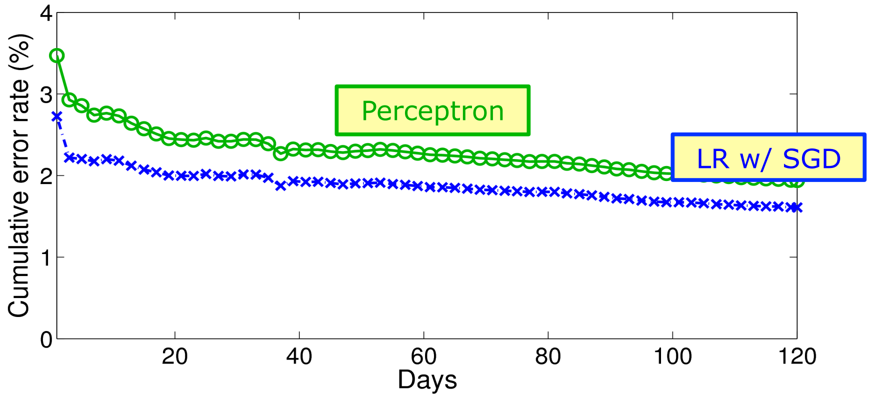
$$\begin{aligned} \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \alpha_t y_t \boxed{\boldsymbol{\Sigma}_t \mathbf{x}_t} \\ \boldsymbol{\Sigma}_{t+1} &\leftarrow \boldsymbol{\Sigma}_t - \beta_t \boldsymbol{\Sigma}_t \mathbf{x}_t \mathbf{x}_t^{\top} \boldsymbol{\Sigma}_t \end{aligned}$$

Treat features differently

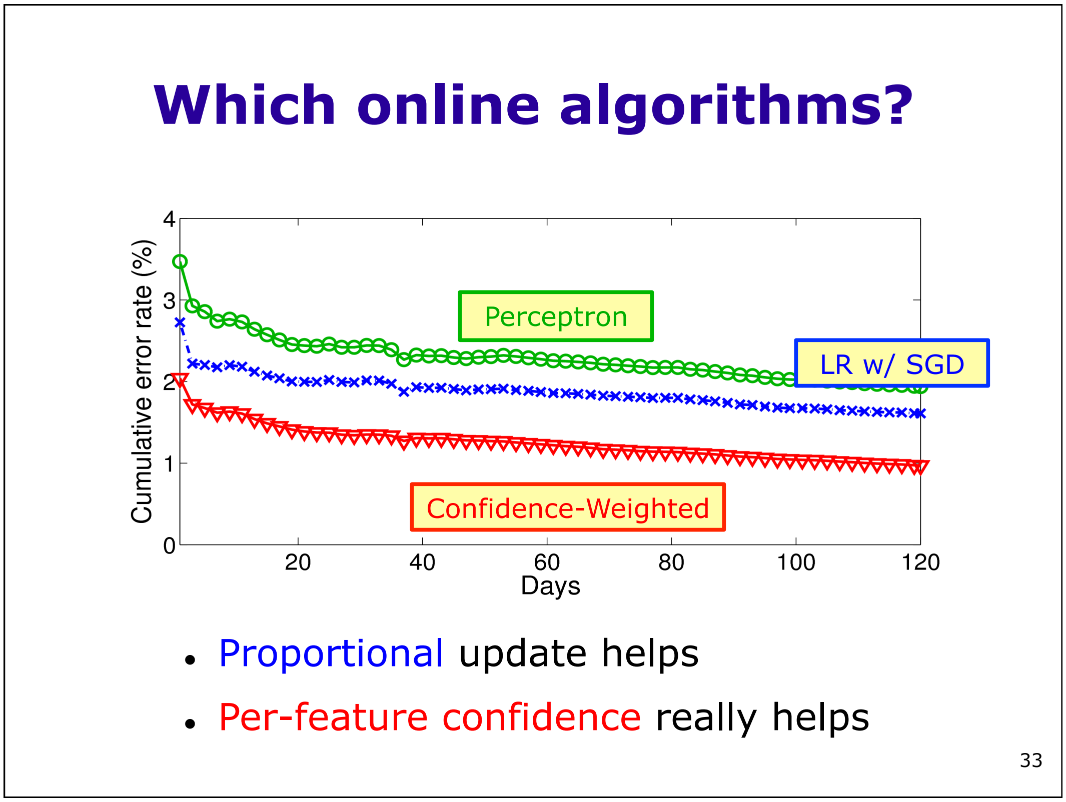
Which online algorithms?



Which online algorithms?



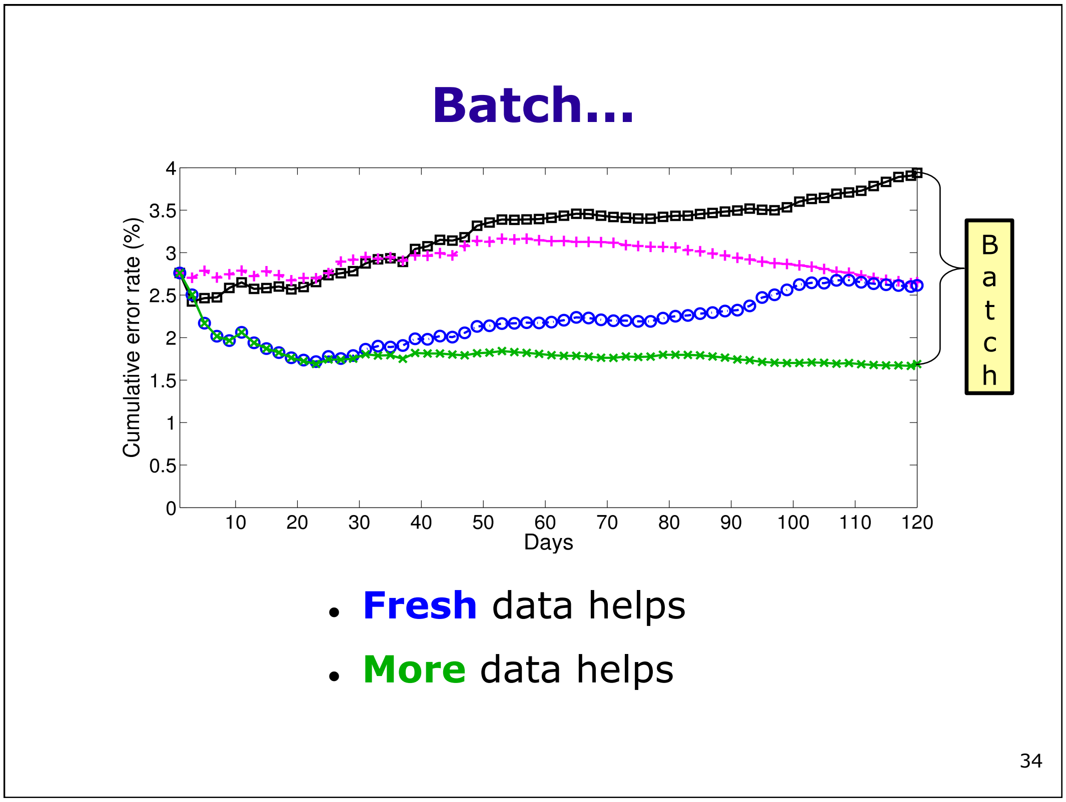
- Proportional update helps



Which online algorithms?

- Proportional update helps
- Per-feature confidence really helps

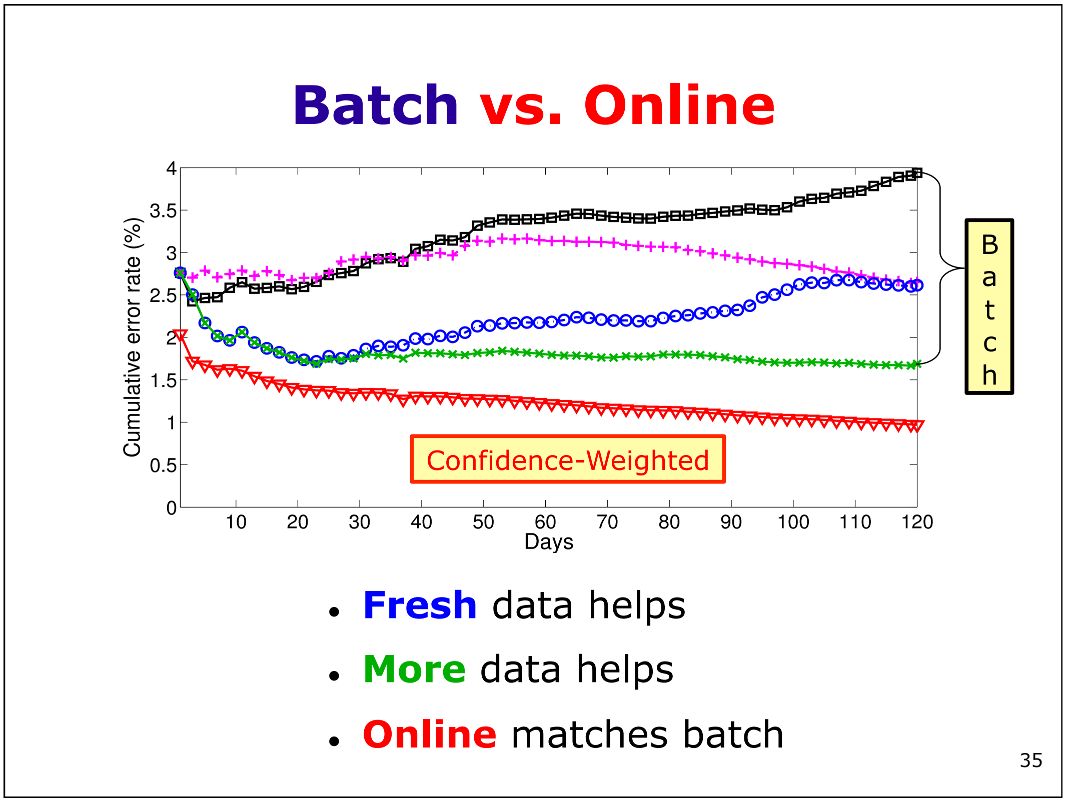
33



Batch…

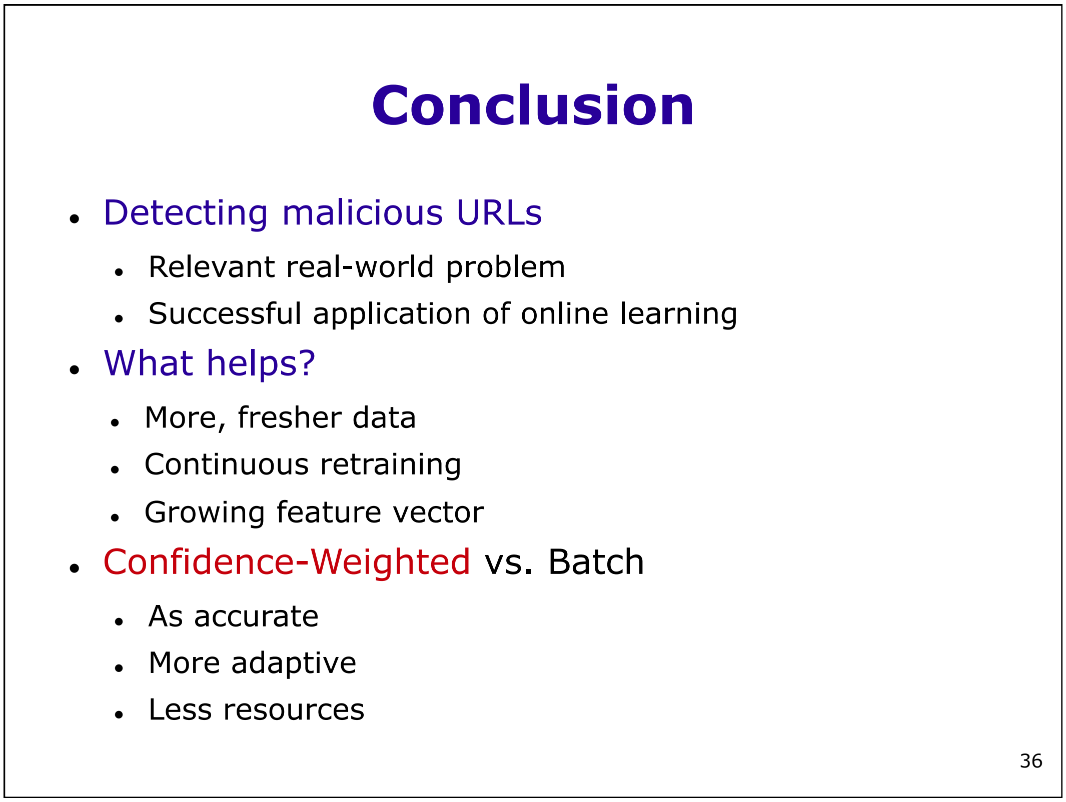
- **Fresh** data helps
- **More** data helps

34



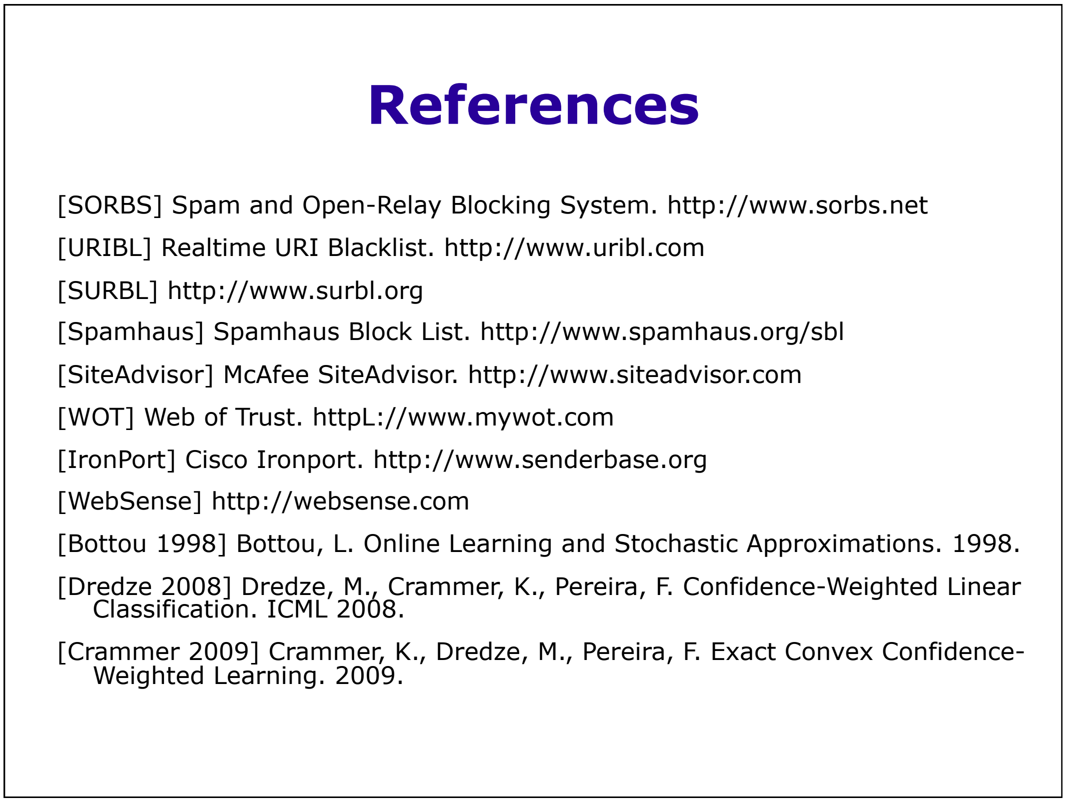
Batch vs. Online

- **Fresh** data helps
- **More** data helps
- **Online** matches batch



Conclusion

- Detecting malicious URLs
  - Relevant real-world problem
  - Successful application of online learning
- What helps?
  - More, fresher data
  - Continuous retraining
  - Growing feature vector
- Confidence-Weighted vs. Batch
  - As accurate
  - More adaptive
  - Less resources



References

[SORBS] Spam and Open-Relay Blocking System. http://www.sorbs.net

[URIBL] Realtime URI Blacklist. http://www.uribl.com

[SURBL] http://www.surbl.org

[Spamhaus] Spamhaus Block List. http://www.spamhaus.org/sbl

[SiteAdvisor] McAfee SiteAdvisor. http://www.siteadvisor.com

[WOT] Web of Trust. httpL://www.mywot.com

[IronPort] Cisco Ironport. http://www.senderbase.org

[WebSense] http://websense.com

[Bottou 1998] Bottou, L. Online Learning and Stochastic Approximations. 1998.

[Dredze 2008] Dredze, M., Crammer, K., Pereira, F. Confidence-Weighted Linear Classification. ICML 2008.

[Crammer 2009] Crammer, K., Dredze, M., Pereira, F. Exact Convex Confidence-Weighted Learning. 2009.