マルウェア対策研究人材育成ワークショップ 2010 (MWS 2010)
October 19—21, セッション3F1 : D3M

実行ファイルに含まれる文字列の学習に基づくマルウェア検出方法

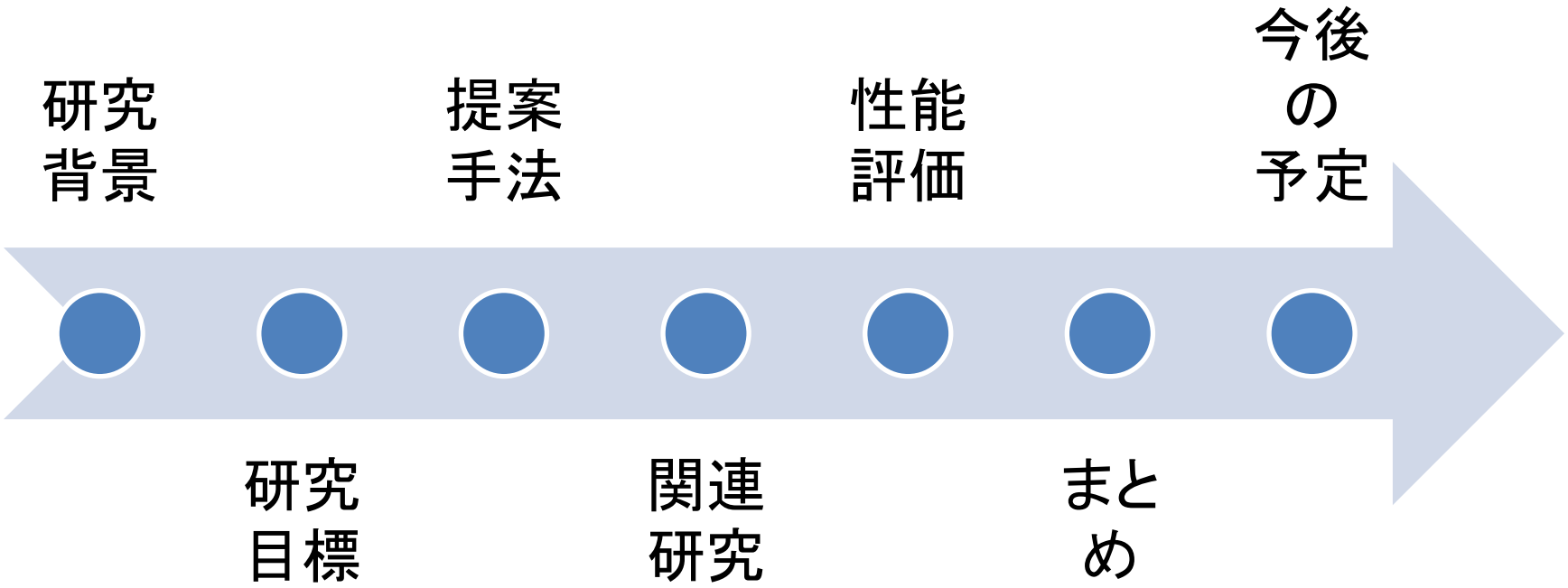
戸部 和洋[†], 森 達哉[†‡], 千葉 大紀[††], 下田 晃弘[†], 後藤 滋樹[†]

[†]早稲田大学 基幹理工学研究科 情報理工学専攻

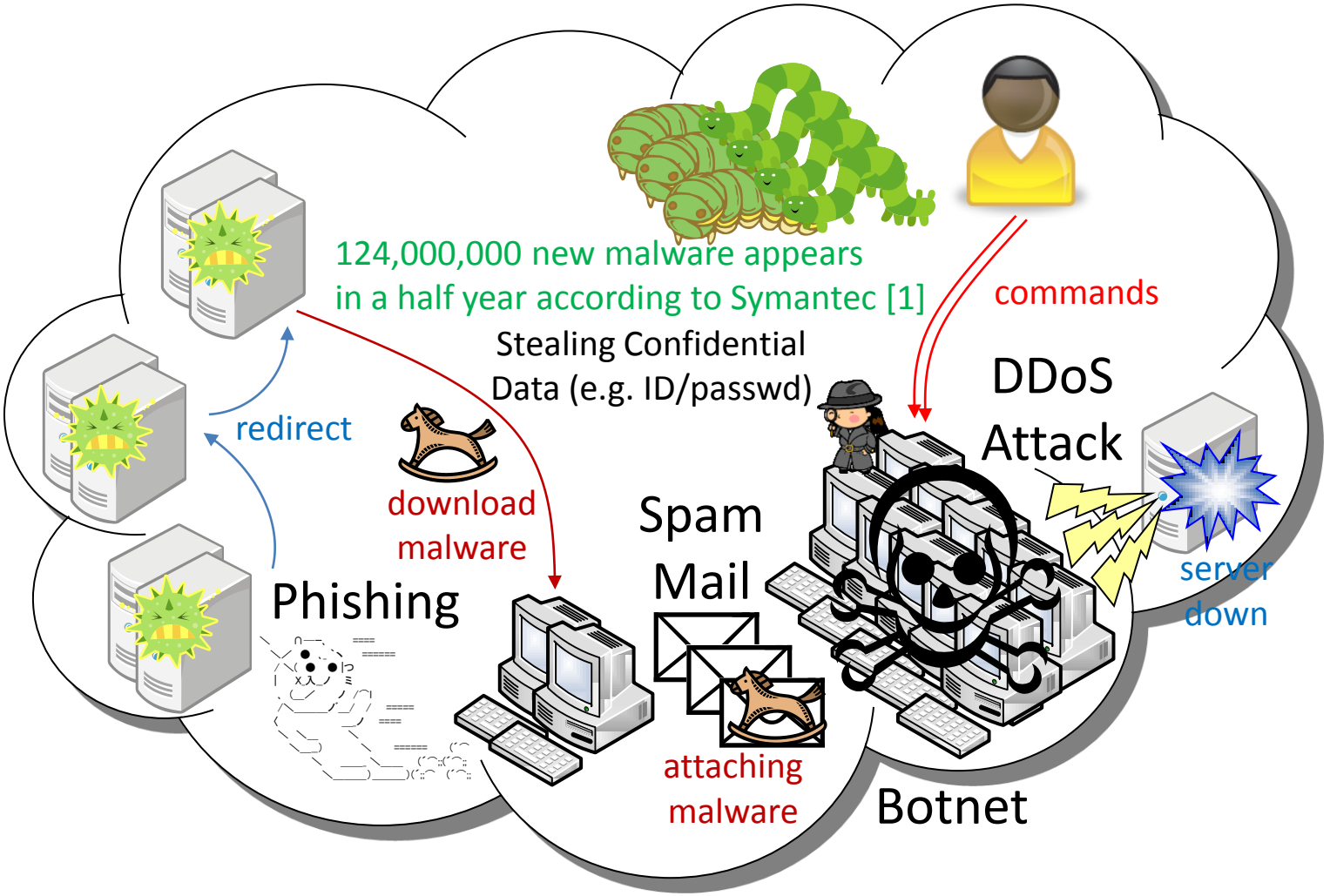
[‡]NTTサービスインテグレーション基盤研究所

[††]早稲田大学 基幹理工学部 情報理工学科

目次



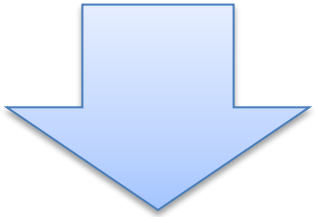
研究背景


124,000,000 new malware appears
in a half year according to Symantec [1]
commands
Stealing Confidential
Data (e.g. ID/passwd)
DDoS
Attack
redirect
download
malware
Spam
Mail
Phishing
server
down
attaching
malware
Botnet

研究背景 (cont.)

半年で1億2400万の新種マルウェアが出現 [1]

- 68万検体/日、2万8千検体/時、470検体/分
- 「ウイルス定義ファイル」の生成が間に合わない
  - 平均的なアナリスト：1時間/検体、優秀：5分/検体 [2]



従来のパターンマッチング方式は限界

- ふるまい検知（動的解析）では検知できない
  マルウェアが存在する… Anti-Anti-Virus機能

研究目標

解析に専門知識・技術を必要としない

e.g. ファイル構造、Win32 API、x86アーキテクチャ

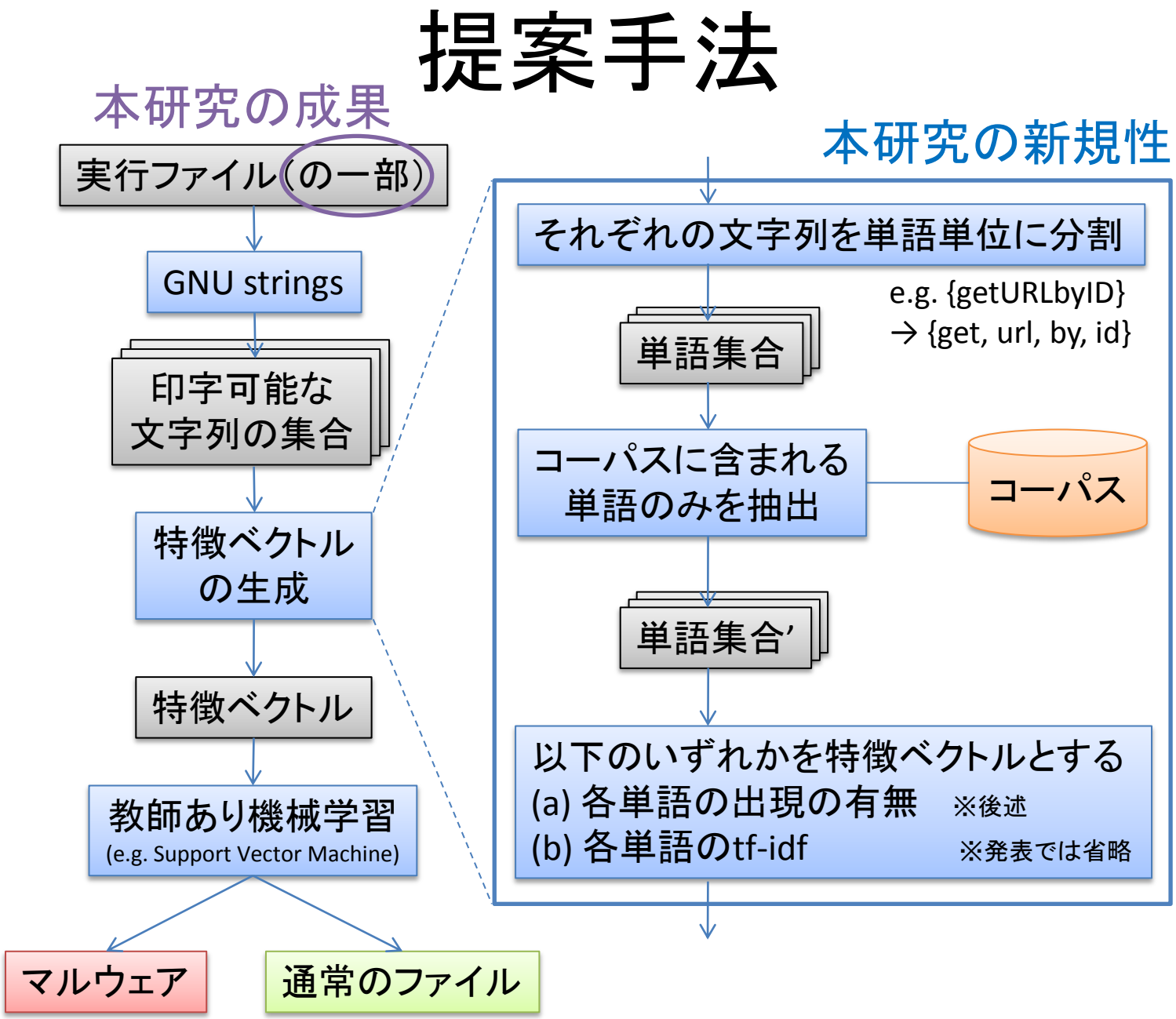
cf. ウイルス対策ソフトベンダの「職人」による解析

新種・亜種のマルウェアも高精度で検出可能

cf. 従来のパターンマッチング方式

高速・軽量でネットワークレイヤで適応可能

提案手法


本研究の成果
実行ファイル（の一部）
GNU strings
印字可能な
文字列の集合
特徴ベクトル
の生成
特徴ベクトル
教師あり機械学習
(e.g. Support Vector Machine)
マルウェア
通常のファイル
本研究の新規性
それぞれの文字列を単語単位に分割
e.g. {getURLbyID}
→ {get, url, by, id}
単語集合
コーパスに含まれる
単語のみを抽出
コーパス
単語集合'
以下のいずれかを特徴ベクトルとする
(a) 各単語の出現の有無 ※後述
(b) 各単語のtf-idf ※発表では省略

関連研究

interpretable

実行ファイルに含まれる可読な文字列を分析 [3]

- アンサンブル学習 (Bagging + Support Vector Machine)
- 実行ファイル全体を用いる ☹

  cf. 我々の研究では、ファイルの一部（先頭*n*バイト）の情報のみ ☺

バイト列を固定長に区切り、エントロピーを計算 [4]

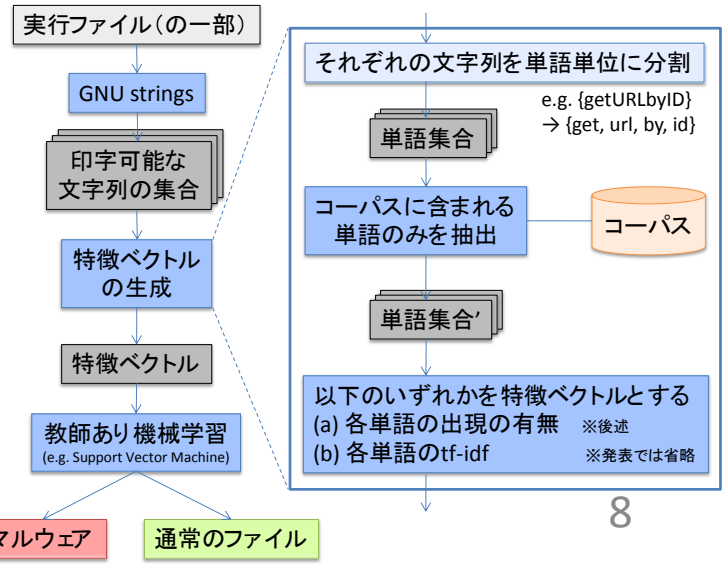
- 暗号化やパッキングされたファイル（≒マルウェア）は
  エントロピーが大きい ∵ゼロパディングされた領域→減少

ファイル構造を特徴とするパターン認識 [5]

- コード・データセクションの名前、書き込み・実行可能
  セクションの数などを特徴とし、決定木やk-NNなどで

性能評価

- 実行ファイルの収集
- コーパス
- 特徴ベクトル
- 各手法の精度比較
- 各手法のコスト比較

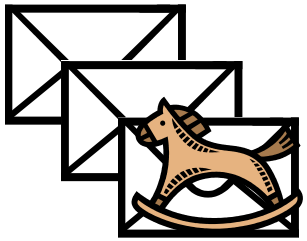


性能評価::実行ファイルの収集

マルウェア検体：1449個

- MWS DATAset 2010 [6]
  - CCC DATAset
  - D3M DATAset
- Nepenthes [7]

※マルウェア検体1449個と通常の実行ファイル1511個は、それぞれすべてハッシュ値が異なる



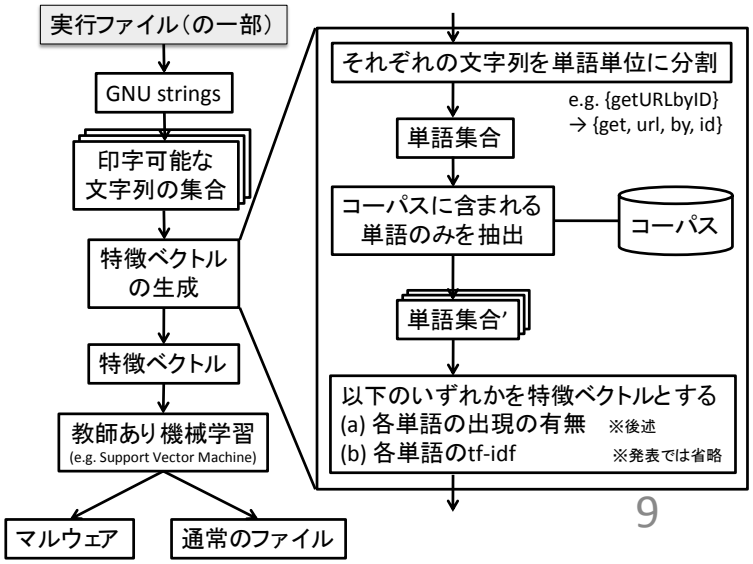
スパムメールに添付されていた実行ファイル

通常の実行ファイル：1511個

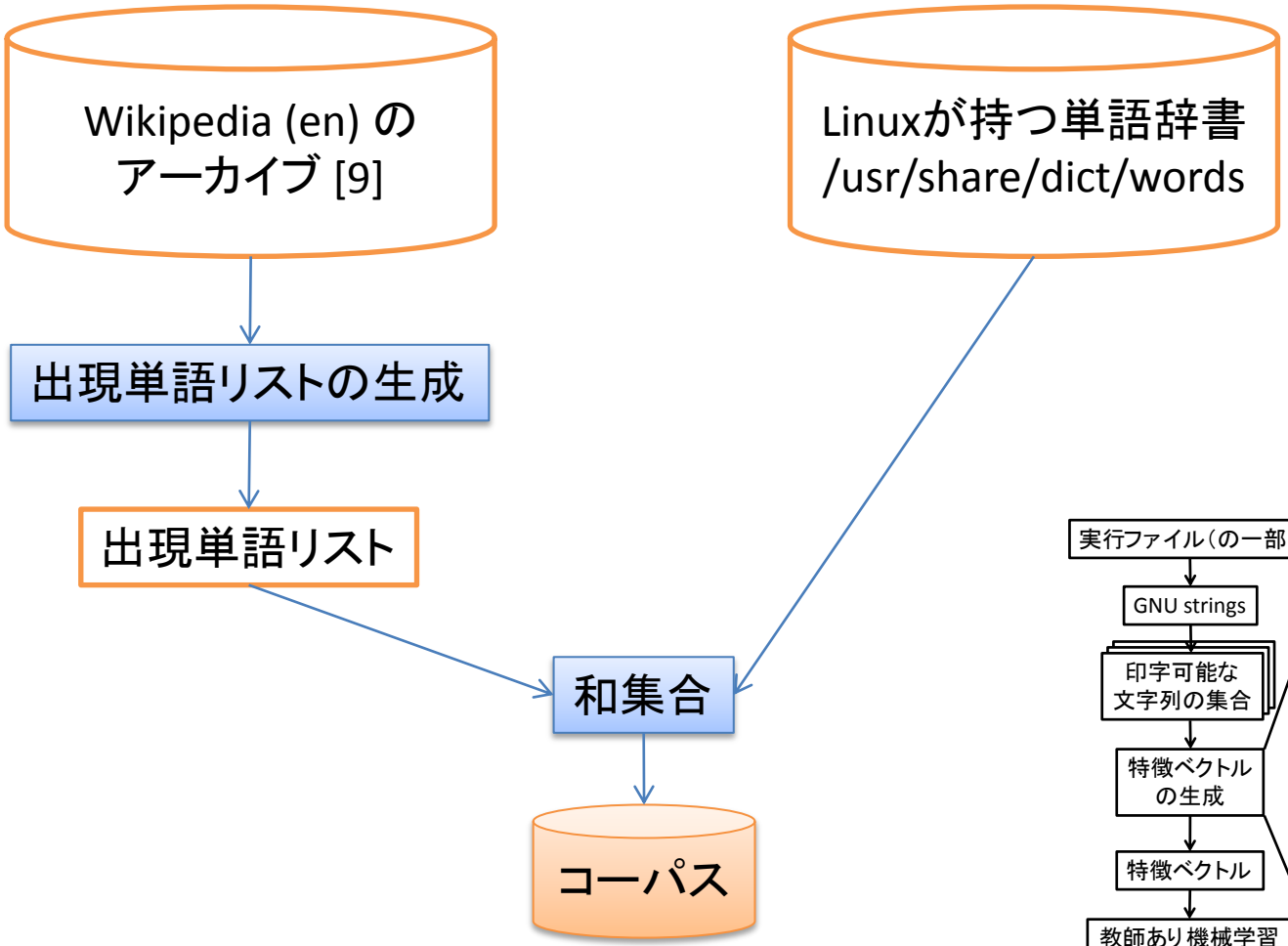
初期状態のWindows XP SP3に存在した実行ファイル

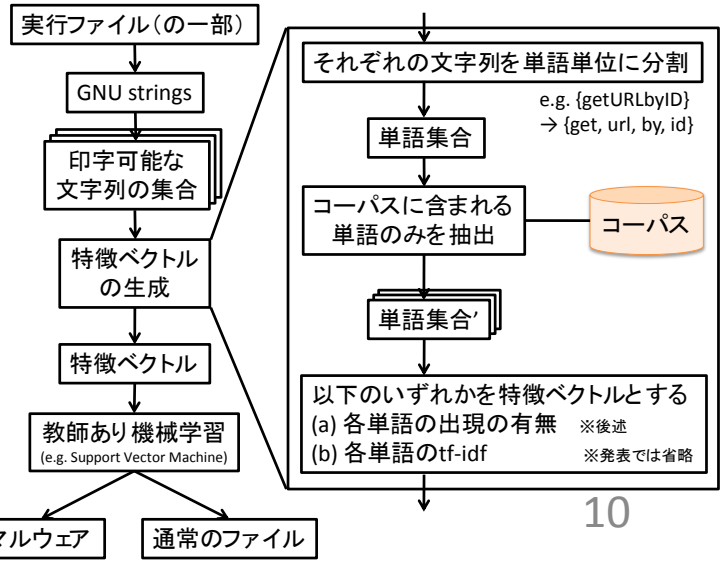
Vector

ソフトウェアライブラリVector [8] のWindows向けソフトウェア



提案手法::コーパス


Wikipedia (en) の
アーカイブ [9]
Linuxが持つ単語辞書
/usr/share/dict/words
出現単語リストの生成
出現単語リスト
和集合
コーパス


実行ファイル（の一部）
GNU strings
印字可能な
文字列の集合
特徴ベクトル
の生成
特徴ベクトル
教師あり機械学習
(e.g. Support Vector Machine)
マルウェア
通常のファイル
それぞれの文字列を単語単位に分割
e.g. {getURLbyID}
→ {get, url, by, id}
単語集合
コーパスに含まれる
単語のみを抽出
コーパス
単語集合'
以下のいずれかを特徴ベクトルとする
(a) 各単語の出現の有無 ※後述
(b) 各単語のtf-idf ※発表では省略

性能評価::特徴ベクトル

1. naïve_str (bin)：提案手法（簡易版）

- GNU stringsで抽出した文字列をそのまま使用
- 文字列の出現の有無 {0, 1} を特徴ベクトルの値とする

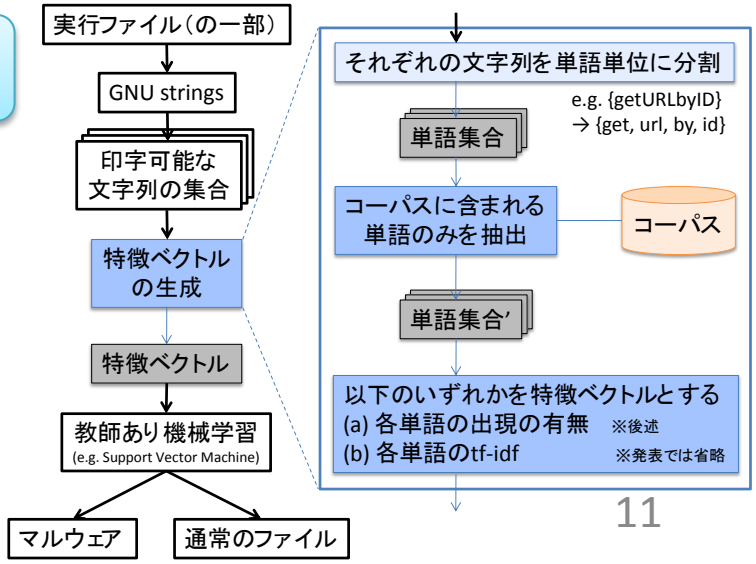
2. dict_words (bin)：提案手法

- 複合語を分割して、コーパスに含まれる単語のみを使用

e.g. {{ResetKeySecurity, {L$^L^]}} => {{reset, key, security}, {}}

3, 4. entropy_2d/4d：既存手法（改）

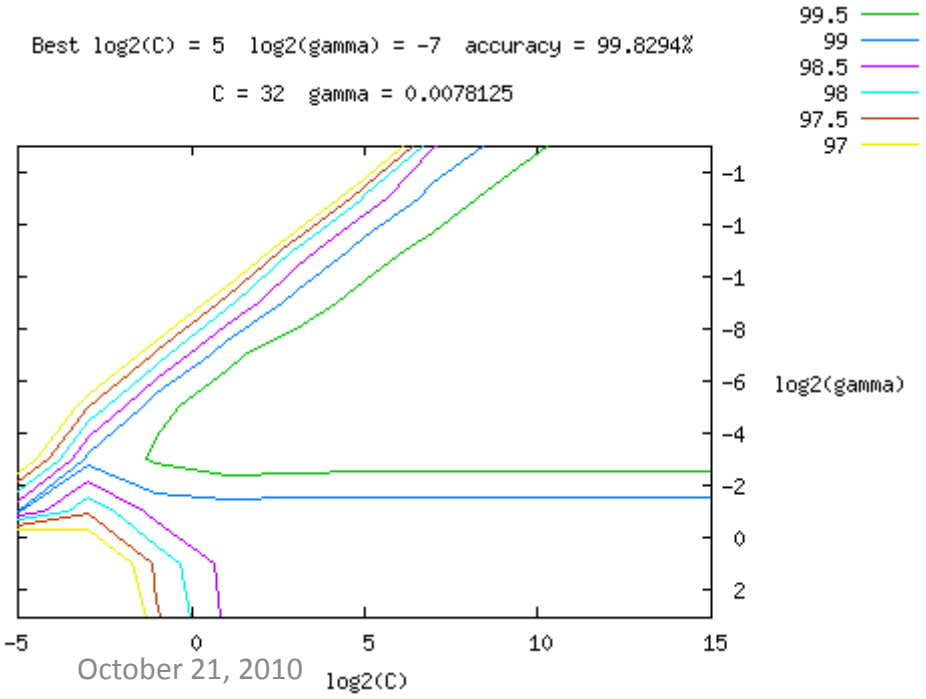
- 256バイトごとに計算した
  エントロピーの代表値を使用
  - max, (min), mean, (SD)
- 既存手法：99.99% CIで線引き
- 本研究：SVMで線引き



性能評価::精度評価方法

チューニング [10]

- 特徴量の線形スケーリング
- 最適パラメータのグリッド探索
- 5-fold Cross Validation (CV)

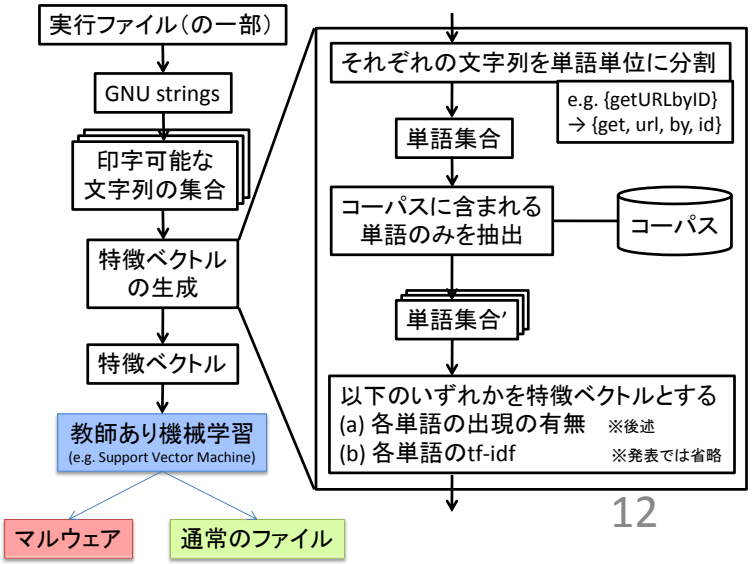


Support Vector Machine (SVM)

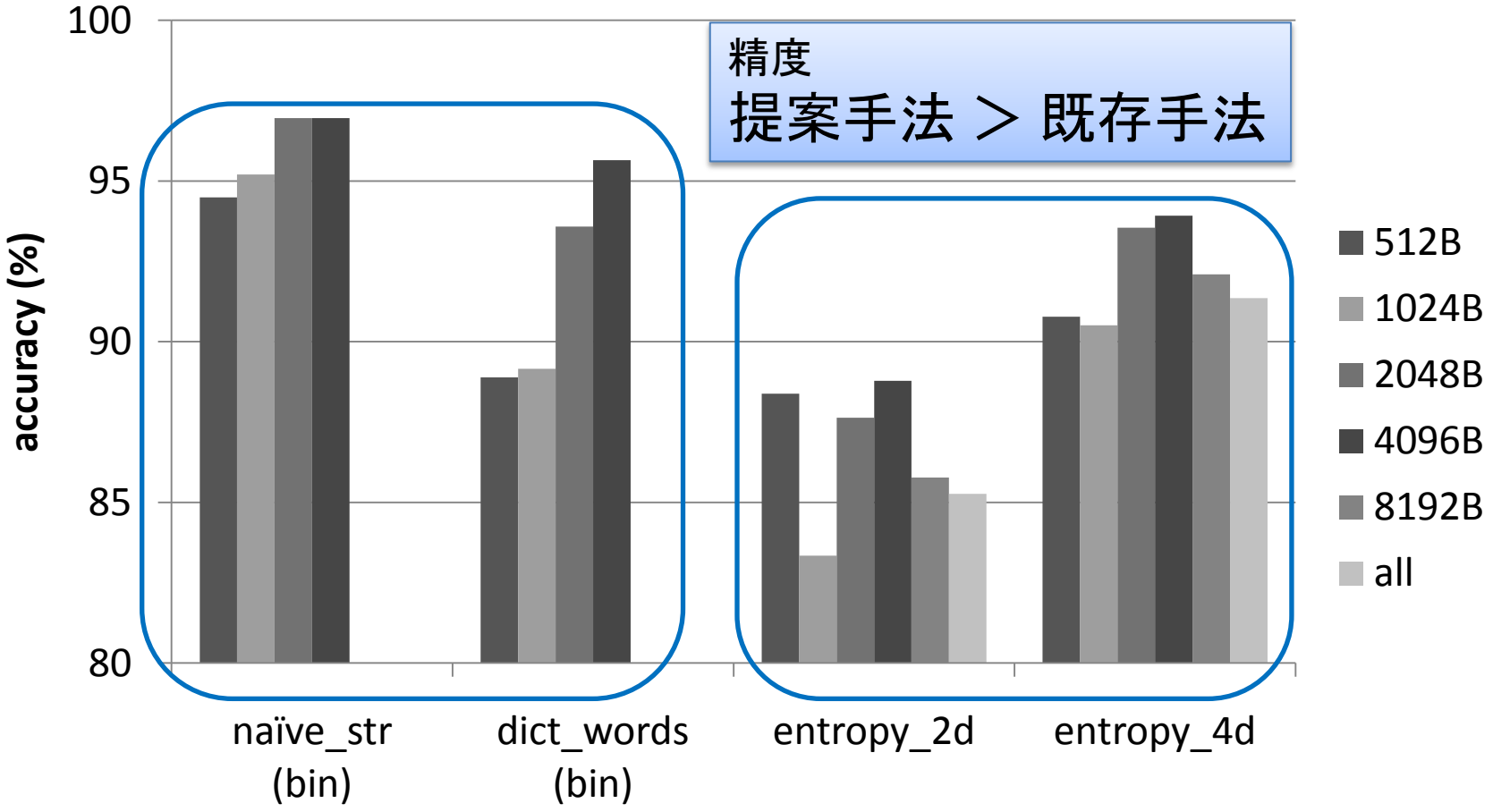
- 教師あり機械学習のひとつ
- ライブラリ：LIBSVM [11]

  /** @param *C* */ C-SVM

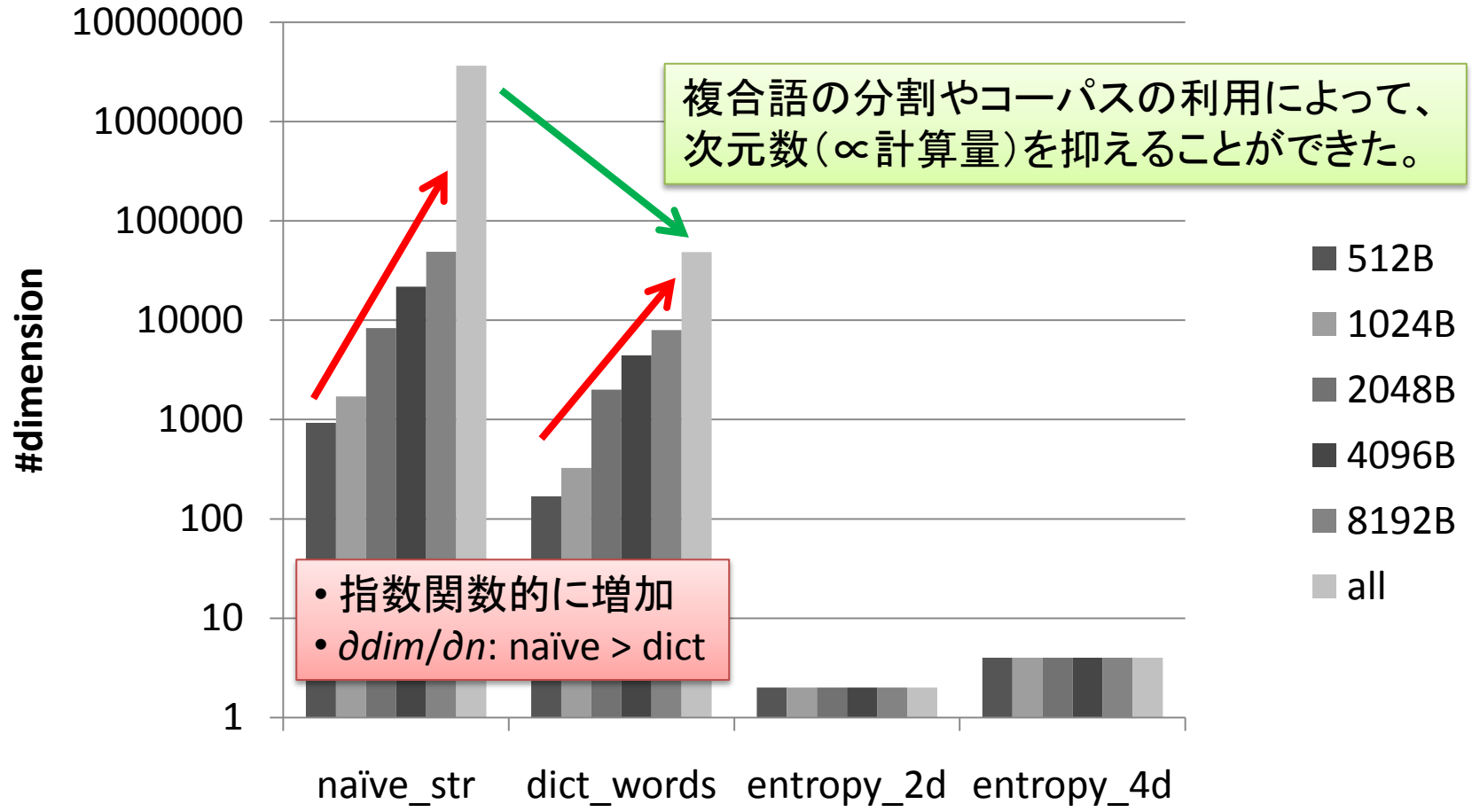
  /** @param *γ* */ RBF kernel



性能評価::精度比較
提案手法による精度の向上



性能評価::コスト比較
提案手法による次元数の削減



まとめ

- 解析に専門知識・技術を必要としない
  - 実行ファイルに含まれる文字列から特徴ベクトルを自動生成し、機械学習 (SVM) による識別を行う
- 新種・亜種のマルウェアも高精度で検出可能
  - 精度: 95%以上（> 既存手法の精度） by 5-fold CV
- 高速・軽量でネットワークレイヤで適応可能
  - ファイルの先頭512B（< 1パケット）で検出可能
  - 複合語の分割とコーパスの利用による次元数削減

今後の展望

特徴ベクトル生成方法の洗練

- N-gram, Stemming, etc

⇒ 計算時間の短縮、精度の向上

機械学習アルゴリズムの変更

- LIBSVM → LIBLINEAR [12]
- 精度はあまり落ちないまま、計算時間を大幅に短縮できることが分かっている

ご清聴ありがとうございました

参考文献

[1] 新種ウイルスが半年で1億2400万件、「従来の対策では不十分」- ニュース：Itpro
http://itpro.nikkeibp.co.jp/article/NEWS/20100902/351743/

[2] 3つの要素を組み合わせたマカフィーのセキュリティ基盤、GTI ー ＠IT
http://www.atmarkit.co.jp/news/201009/17/mcafee.html

[3] Y. Ye, L. Chen, D. Wang, T. Li, and Q. Jiang, "SBMDS: an interpretable string based malware detection system," Journal in Computer Virology, Vol. 5, No. 4. pp. 283–293. 2009.

[4] R. Lyda and J. Hamrock, "Using Entropy Analysis to Find Encrypted and Packed Malware," Security & Privacy, IEEE, Volume 5, Issue 2, 2007 pp. 40–45.

[5] R. Perdisci, A. Lanzi, and W. Lee, “Classification of packed executables for accurate computer virus detection,” Pattern Recognition Letters, Volume 29, Issue 14, 2008, pp. 1941–1946.

[6] 畑田充弘, 中津留勇, 秋山満昭, 三輪信介. マルウェア 対策のための研究用データセット ～MWS 2010 Datasets～. マルウェア対策研究人材育成ワークショップ2010.

参考文献 (cont.)

[7] P. Baecher, M. Koetter, T. Holz, M. Dornseif, and F. Freiling. “The Nepenthes Platform: An Efficient Approach to Collect Malware,” In Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID), pp. 165–184. Springer. 2006.

[8] Vector：ソフトライブラリ＆ PC ショップ- 国内最大級のフリーソフトダウンロードサイト http://www.vector.co.jp/

[9] Wikimedia Downloads, http://download.wikimedia.org/

[10] C. W. Hsu, C. C. Chang, C. J. Lin, "A practical guide to support vector classification," http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf

[11] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

[12] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear