Text Classification for Data Loss Prevention

<u>Michael Hart</u>, Pratyusa Manadhata, Rob Johnson

Symantec Research Laboratory

Data Loss has become a serious problem

- Data breaches cost an average of $6.6 million to an organization
- Almost 500 million records with personal info have been leaked since 2005
- Recent targets: CIA, RSA, Senate.gov
- WikiLeaks
  - 250K State Department cables
  - > 70K Afghanistan War reports
  - And have a “contingency plan”
- Hacker gangs
  - Lulz Sec and Anonymous

Latest Incidents   twitter/ DataLossDB

RECORDS	DATE	ORGANIZATIONS
19,799	2011-07-20	Swedish Medical Center
0	2011-07-20	Policía Nacional de Colombia (Colombia National Police)
0	2011-07-19	Mountain Mike's Pizza
0	2011-07-18	REWE Group
4,827	2011-07-18	Unknown Organization, JL Audio, Inc.
2,021	2011-07-18	Beth Israel Deaconess Medical Center
340	2011-07-17	Unknown Organization, Federal Emergency Management Agency, Williams Chevrolet Inc. Customers
188	2011-07-17	Haartman Hospital
25	2011-07-16	Kirklees Council
0	2011-07-16	Meath Council

Largest Incidents

RECORDS	DATE	ORGANIZATIONS
130,000,000	2009-01-20	Heartland Payment Systems, Tower Federal Credit Union, Beverly National Bank
94,000,000	2007-01-17	TJX Companies Inc.
90,000,000	1984-06-01	TRW, Sears Roebuck
77,000,000	2011-04-26	Sony Corporation
40,000,000	2005-06-19	CardSystems, Visa, MasterCard, American Express
32,000,000	2009-12-14	RockYou Inc.
26,500,000	2006-05-22	U.S. Department of Veterans Affairs
25,000,000	2007-11-20	HM Revenue and Customs, TNT
24,600,000	2011-05-02	Sony Online Entertainment, Sony Corporation
17,000,000	2008-10-06	T-Mobile, Deutsche Telekom

From datalossdb.org

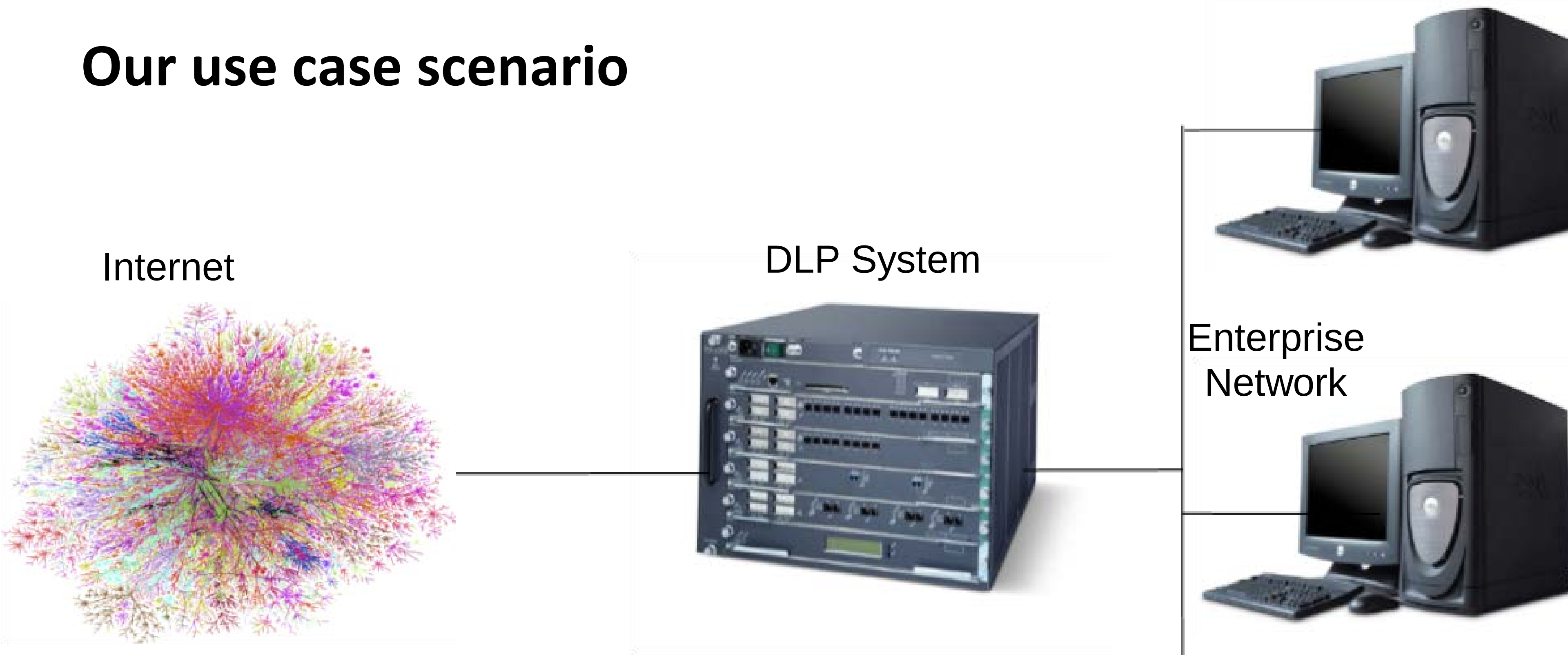
Protecting Data

Type	Description	DLP goal
Data-at-rest	Information stored on enterprise devices such as document management systems, email servers and file servers.	Scan data, identify unsecured confidential information and report.
Data-in-motion	Enterprise data contained in outbound network traffic such as emails, instant messages and web traffic.	Block transmission of sensitive data.
Data-in-use	Data currently used at the end point such as Outlook, http, https, print and file to USB.	Prevent unauthorized usage of data (e.g. copying to a thumb drive).

Why Machine Learning for Data Loss Prevention?

- Need for more effective approaches to stop data breaches
- Downsides to current approaches
  - Impossible to describe all CI entirely in rule based formats
  - Potentially large number of documents that constantly evolve
  - Requires allowing IT staff access to sensitive materials
- Text classification
  - Long history of research and many different techniques
  - Handles unstructured data
  - Requires minimal supervised interaction
- **Goal**: automatically learn what is secret

Our use case scenario



- Build message classifier for outgoing messages
- Train on examples of private and public messages
- Use classifier to detect outgoing messages with private data
- Block or log outgoing messages with private data

Performance Metrics

- Achieve a high recall on confidential (*secret*) documents
- Maintain a low false positive rate on both:
  - Company media (*public*) documents
  - Non-enterprise (*NE*) documents
- Constraints
  - Scale well
  - Require no metadata
  - Be agnostic to message type

Baseline Approach

- Simply train a standard classifier on confidential and public documents
- Employed a search for classifiers with WEKA
  - Best classifier: SVM with a linear kernel
  - Best features: Unigrams with binary weights

Potential issues with enterprise training data

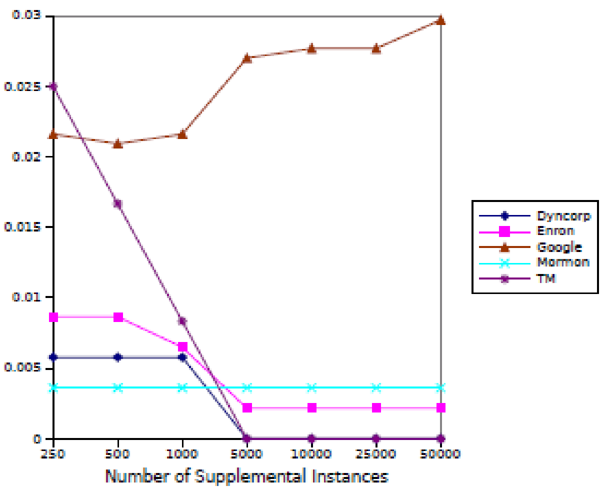
- Suffer from high FP on *NE* documents
- Can weight common words strongly towards *secret*
  - Example words: `Policy, Police, Procedure, 1, Afghanistan`
  - Feature behavior for *public* documents absent in training set
- 40% of classifiers were biased towards the *secret* class
  - Performed poorly for instances inadequately represented in vector space
- **Underlying problem**
  - Can the organization even describe what is not *secret*?

Addressing inadequate training data (Step 1)

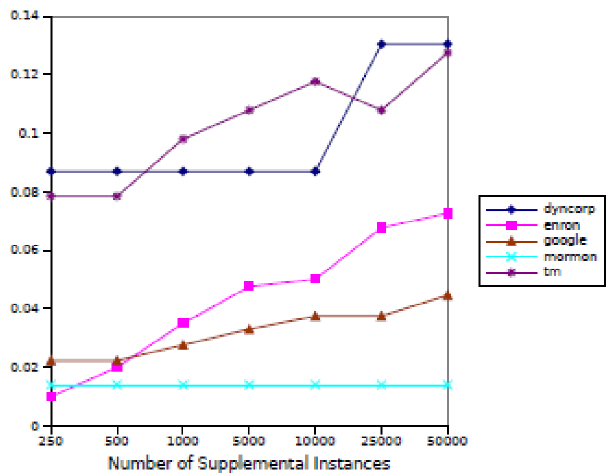
- Better learn what is secret by supplementing
  - Add 10K supplemental instances from Wikipedia to the training set
  - Key point: do **not** expand feature set
  - Gives classifier more representative training set
    - Better learn which features correlate with secret
- Adjust the classifier
  - Adding more instances increases false negative rate
  - Adjust decision plane within 10% of the closest TN
- Call this classifier SA_{private}

Effect of supplementation

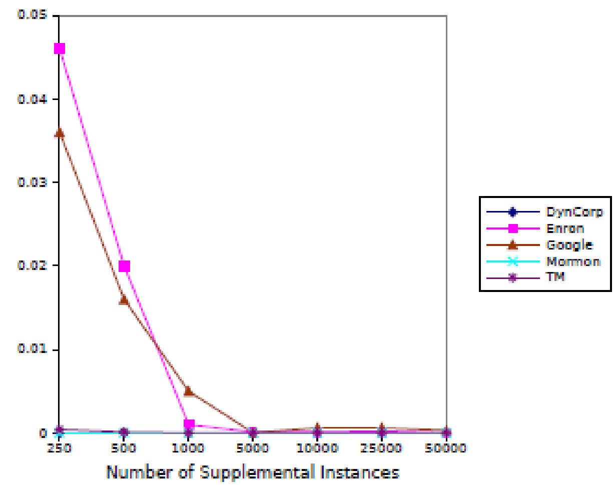
Public False Positive Rate



False Negative Rate

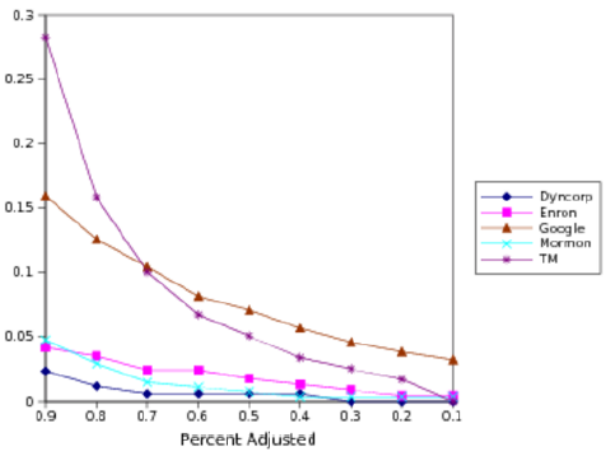


NE False Positive Rate

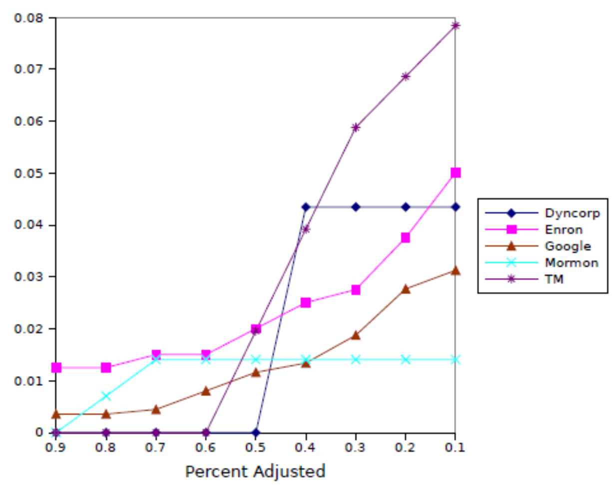


Effect of adjustment

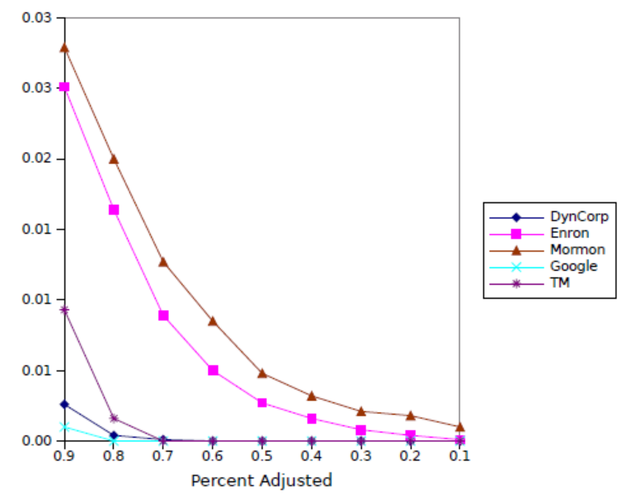
Public False Positive Rate



False Negative Rate



NE False Positive Rate



Correcting for mistakes (Step 2)

- In some cases
  - Still observe FPs by *NE*s
  - Increased FP rate on *public*
  - Classifier more sensitive to knowledge domain than *secret*
- Train a second classifier with new features
  - Eliminate *NE* false positives by measuring the topical relatedness of documents
  - Address *public* false positives by learning what *public* means using an SA_{public} classifier
  - Change the classification decision from *secret*, ¬*secret* to *secret, public* and *NE*

Targeting *NE* false positives

- Related documents should share similar language
  - Measure amount of new vocabulary contained in document
- Introduce new attribute: $xtra.info_{class}$ where
  - class in {*secret,public*}
  - Percentage of words in document that exist in the document, but not in any document labeled class
  - Only consider words with a document frequency less than 0.5%
- **Hypothesis**
  - A document in class should have a lower $xtra.info_{class}$

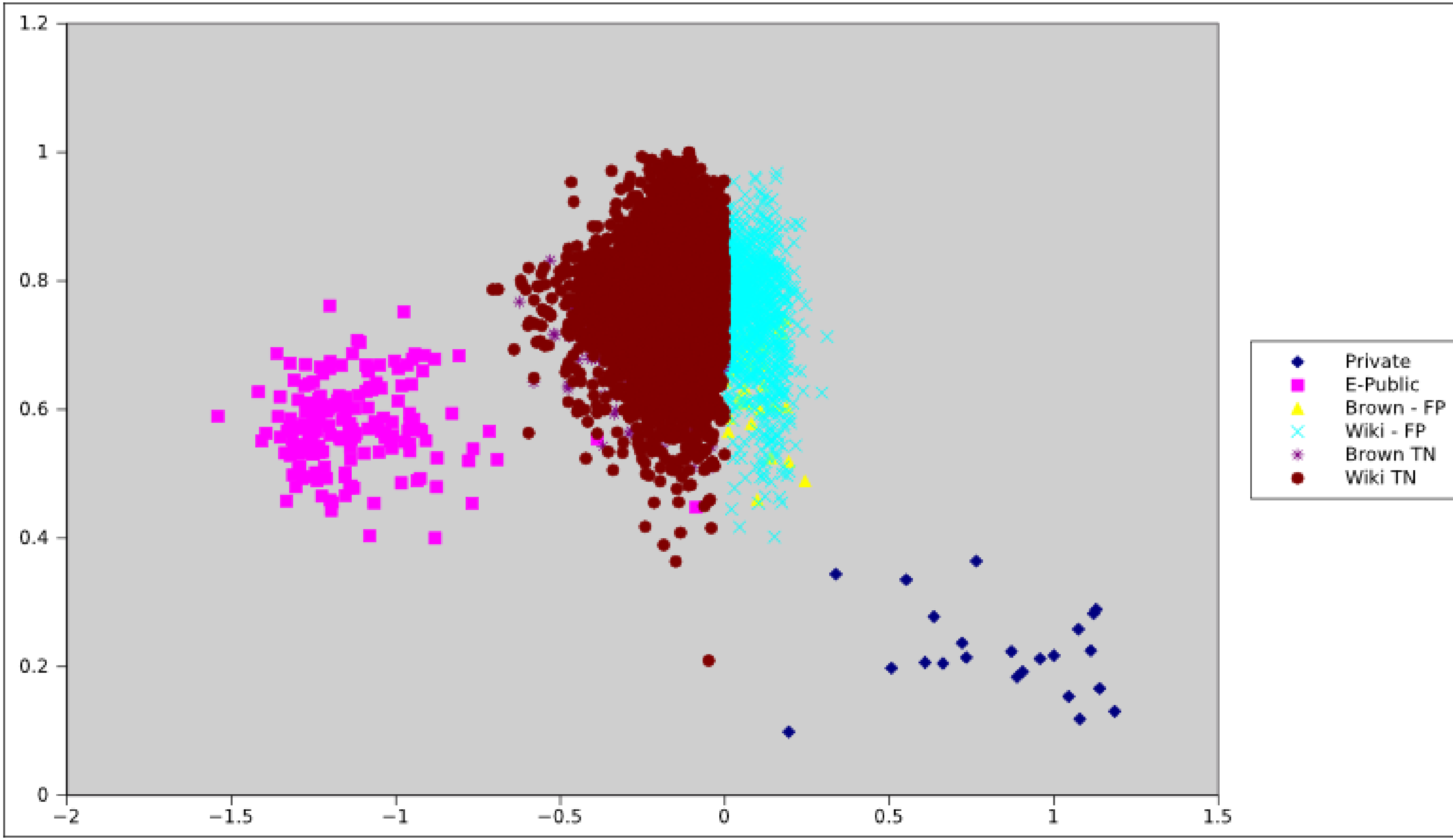
xtra.info$_{secret}$ comparison for *NE* and *secret*

xtra.info$_{secret}$	Dyncorp	Enron	Google	Mormon	TM
secret	0.54 (0.10)	0.83 (0.09)	0.70 (0.15)	0.49 (0.15)	0.66 (0.11)
NE	0.96 (0.03)	0.99 (0.02)	0.98 (0.04)	0.95 (0.08)	0.99 (0.02)

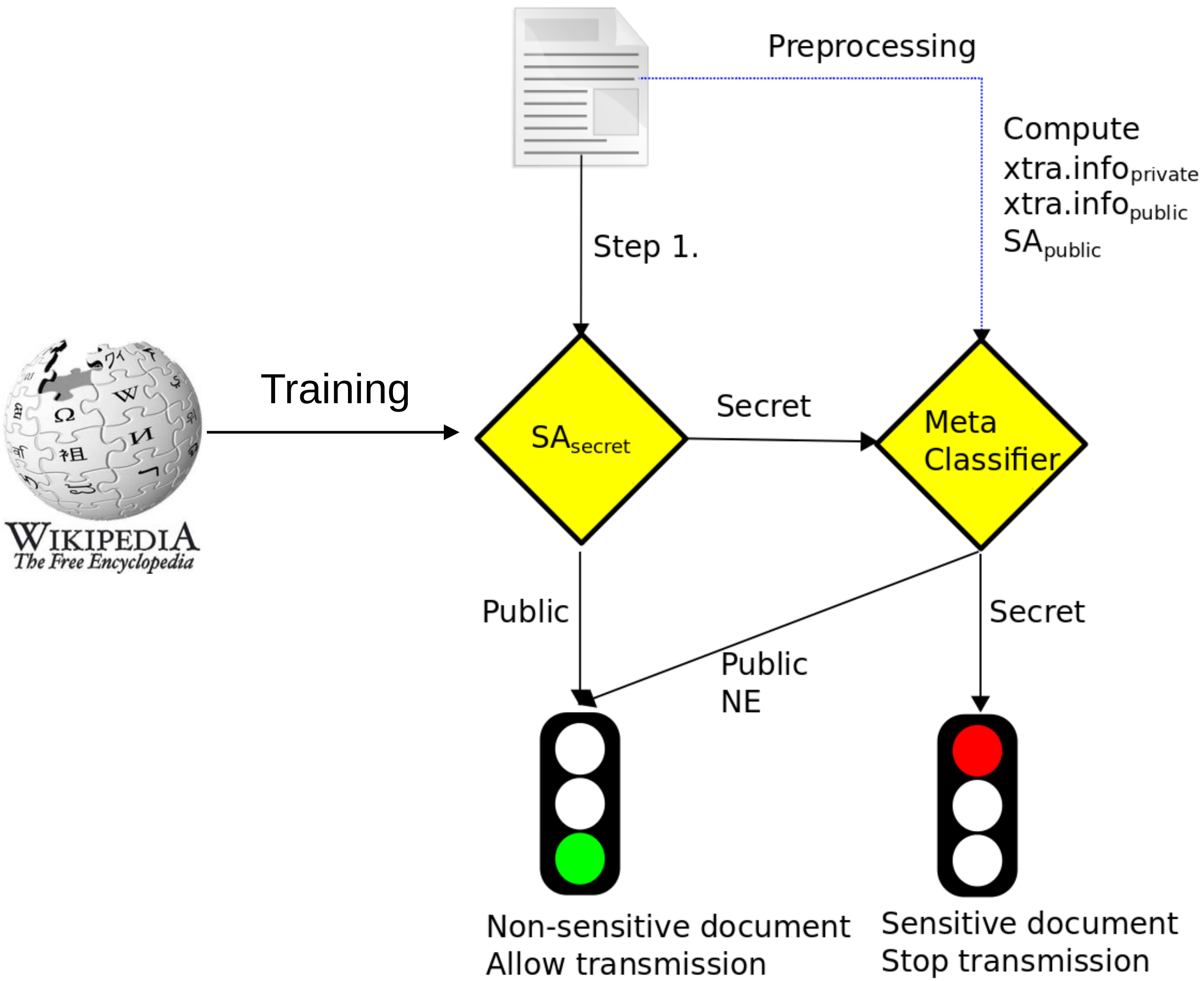
Why three classes?

- *NE* and *public* false positives are topically dissimilar
- Grouping together increases the variance in
  - xtra.info attributes
  - SA_{public} classifier
- Change the decision to *NE, secret, public*
- Increase separability between *secret* and *NE* for $\text{xtra-info}_{private}$ attribute
- Observe decrease in mislabeling of *public* documents as *secret*

SVM output + xtra.info$_{private}$ for Dyncorp



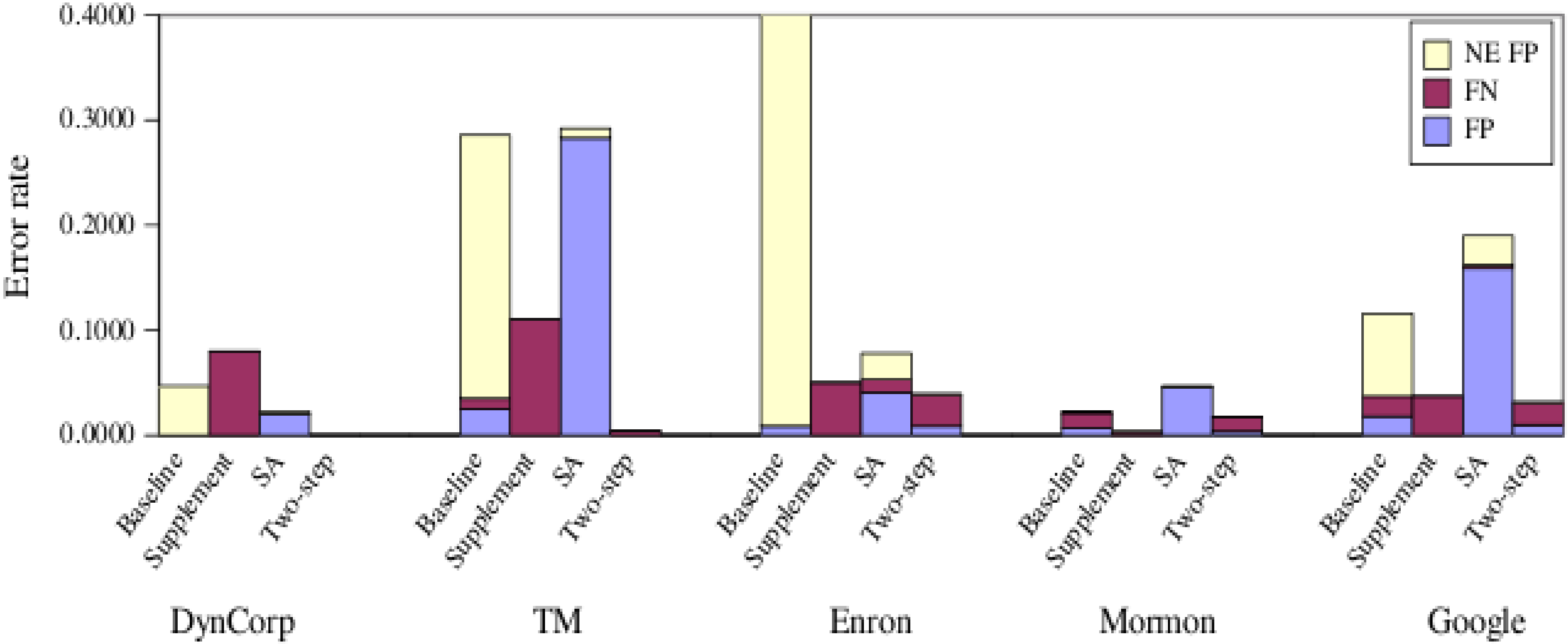
Review: DLP pipeline



Corpora for DLP

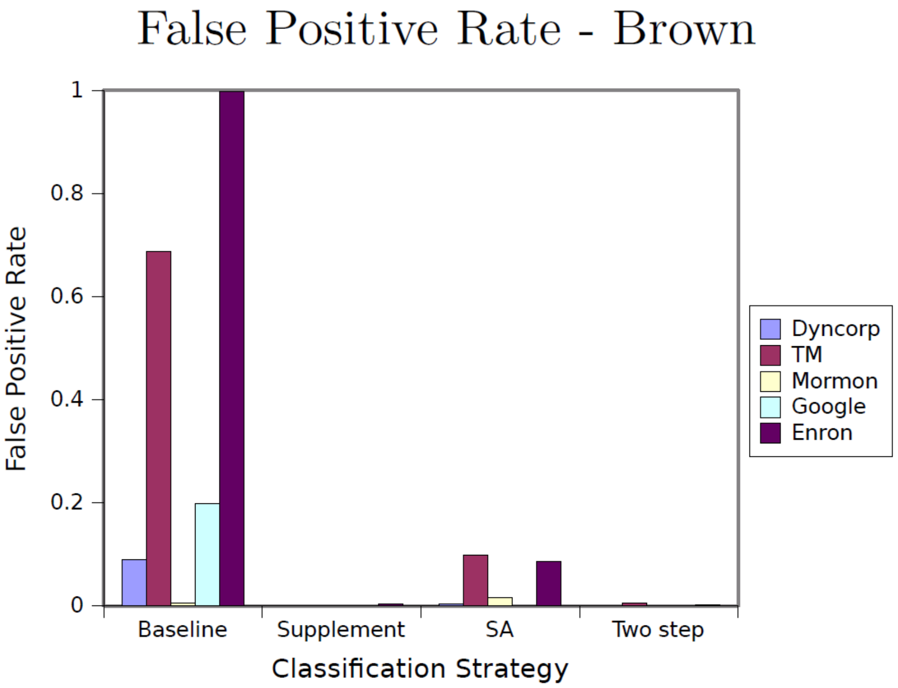
Dataset	Source of Sensitive Documents	Source of Public Documents	Description
Dyncorp	WikiLeaks	www.dyncorp.com	23 private documents leaked from the military contractor Dyncorp
TM	WikiLeaks	www.alltm.org, www.tmscotland.org	102 documents from high ranking officials in the Transcendental Meditation movement
Mormon	WikiLeaks	www.lds.org	Private Mormon handbook split into 1000 word chunks
Enron	Enron Email dataset	Enron's former website via the Wayback Machine	399 emails labeled by Hearst et al. as business-related
Google	Google Product blogs	Google PR Blogs	Label product-related posts as private and public relations posts as public
Wikipedia		English Wikipedia	10K randomly selected articles for false positive detection
Brown Corpus		Press releases, reviews and books	500 texts selected to represent modern American English
Reuters-21758		Reuters News Service	10788 news items published by the news service

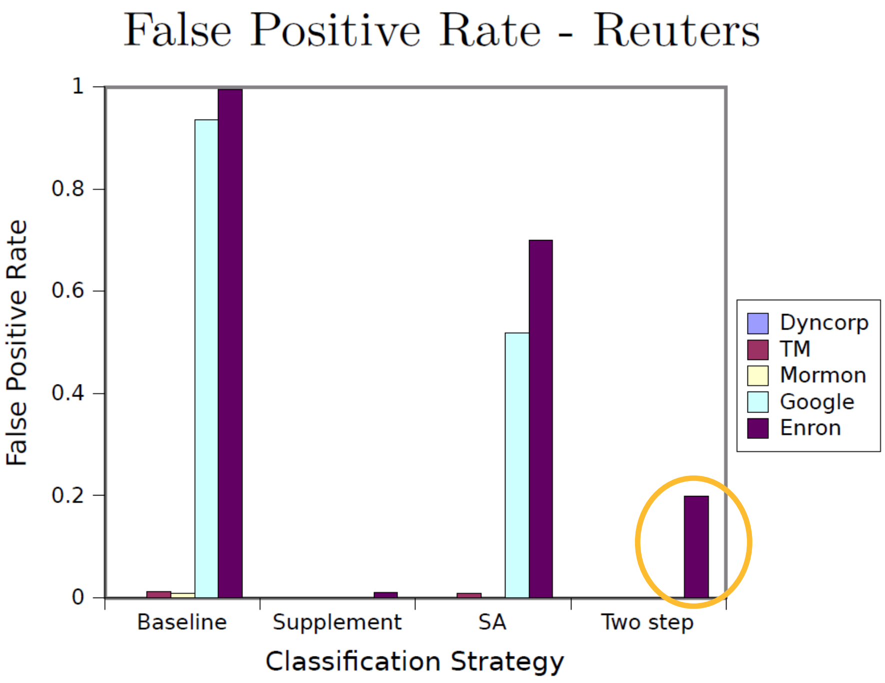
Results: Error rates



FN Rate	FP public	FP NE	FDR
1.6%	0.46%	~0.0%	0.47%

False positive rate on other *NE* corpora





Outstanding research questions

- Given a set of documents, how well will it work in deployment?
  - If it performs poorly, can I fix it?
- What about sensitive documents that are not classified well?
  - Likely scenario: new project initial documents
- What if I am given a large number of diverse documents?
- Intra-organizational DLP?
- What about this document is confidential?
  - Can we highlight, redact, sanitize?
- Sensitivity score?
- What can I do for my Smartphone?

Conclusion

- An algorithm to train text classifiers for DLP
  - Enhance the text classification process to prevent data loss
  - Add supplemental examples to better understand what is *secret*
  - First step filters out majority of FPs generated by non-enterprise documents
  - Employs a second classifier with contextual information
- Approach motivated by understanding and modeling the data
  - Confidential documents do contain publically known entities
  - Are the salient features
  - But can cause false positives
  - It is the relationship between these entities that must be protected

Thank you!

Michael Hart

Michael_Hart@symantec.com