

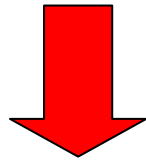
# An Empirical Study on Selective Sampling in Active Learning for Splog Detection

**Taichi Katayama<sup>1</sup>**  
**Takehito Utsuro<sup>1</sup>**  
**Yuuki Sato<sup>2</sup>**  
**Takayuki Yoshinaka<sup>3</sup>**  
**Yasuhide Kawada<sup>4</sup>**  
**Tomohiro Fukuhara<sup>5</sup>**

<sup>1</sup>University of Tsukuba, <sup>2</sup>Konami Corporation,  
<sup>3</sup>Tokyo Denki University,  
<sup>4</sup>Navix Co., Ltd., <sup>5</sup>University of Tokyo,

# Background

- Opinion Mining from Blogs



- Splogs are Serious Noise in Opinion Mining
  - e.g., larger scale statistics (2008 Mar.)
    - 40% of Japanese Blog Articles in BuzzPulse, nifty are Splogs, 2007 Oct. ~ 2008 Feb.
- Automatic Detection is highly Expected.

2007年08月08日

(^\_^)

エキスポランド 週内にも再開 ケミカルライトの液体で負傷 イチロー 松坂との対戦 鹿島? 鹿見島市、ロッセ受け入れ拒否 ハリウッドで1番もうかる俳優 鶴戸神宮の参道崩壊の恐れ 中越沖地震でペットもPTSDか

AAA DION EXILE IPO SKYPE YOU ナチュラルハイ ヒートアイランド現象 マキシマムザホル モンリアディゾン レンタカー 叶美香 株価 世界地図 川遊び 貸貸 堀北真希 郵便局 浴衣 鈴木保奈美 AKB48 EXILE IHI JR九州 JR東海 K-1 お盆 どんと晴れ ねぶた祭り はなまるマーケット アトピー 性皮膚炎 オーバードーズ クックパッド シュモクザハリースト マイクロソフト メガハウス ラッシュアワー 3リタリン レンタカー 為替 ちへ 華原朋美 叶美香 及川光博 原爆 厚生労働省 広島 江田五月 高速道 会保険庁 暑中見舞い 松たか子 松嶋菜々子 新垣結衣 森尾由美 万理生田斗真 台風情報 朝青龍 長澤まさみ 熱中症 布袋寅泰 本田昌毅 万理み 櫻井淳子 綾瀬はるか 華原朋美 叶美香 山口もえ 松たか子 松嶋菜々子 美 杉本エルザ 菅谷梨沙子 蒼井優 大後寿々花 大島優子 竹内結子 仲間由紀恵 長澤まさみ 浜崎あゆみ 万理沙ひとみ 優木まおみ 櫻井淳子 ジャッキー・チェン パク・ヨンハ 伊藤俊 加山雄三 及川光博 江田五月 高岡蒼甫 桜塚やっくん 山田涼介 小栗旬 小室哲哉 小沢一郎 真田広之 生田斗真 朝青龍 藤井裕久 内博貴 品川祐 布袋寅泰 本田昌毅 AKB48 EXILE HERO K-1 どんと晴れ はなまるマーケット らき☆すた ハリー・ポッター ビーチボーイズ ファーストキス ポケモン ラッシュアワー 3 花ざかりの君たちへ 関ジャニ∞ 金色の翼 桜蘭高校ホスト部 探偵学園Q 東方神起 篤姫 名探偵コナン ぼてふりん サークラヒマガーデン 仙台七夕まつり 楊枝橋コンテスト いえそば アイスクリーマー ホームベーカリー ホームメイド家電 豆乳メーカー 納豆メーカー FLASH PS3 Wii オンラインゲーム カラス キンブテン異 ゲームソフト フリーゲーム ミニゲーム 仮面ライダーカブト 暇つぶし 戦国無双2 無料ゲーム

posted by スイロ at 00:29 | 日記

2007年08月06日

(^\_^)

民主党初の参院議長に江田氏 TBSの不二家報道「重大な問題」タイ警察官 罰金はキティ風腕章 高砂親方が処分後初の面会 武蔵の不可解判定に怒り爆発 伊代、約17年ぶり山口生歌 利上げ 4割が景気腰折れ懸念

検索ボックス

検索語句

検索

<<2007年11月>>

日	月	火	水	木	金	土
---	---	---	---	---	---	---

25	26	27	28	29	30
----	----	----	----	----	----

RDF Site Summary  
RSS 2.0

keyword stuffed blog

FC東京: FC東京のウワサ - Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)

http://tr83657ytruw45.cocolog-nifty.com/blog/2009/04/post-6c42-8.html

よく見るページ Firefox を使ってみよう 最新ニュース HotMail の無料サービス Windows Media Windows リンクの変更

# FC東京

[<< FC東京のウワサ | トップページ](#)

2009年4月12日 (日)

## FC東京のウワサ


**粉骨砕身**  
 粉骨砕身 まあ、開始15分で2点もやっちゃあ、今の我が軍では勝ち  
 はない。FWの差がモロに出たゲームだった。高校出立での若造に  
 られるのは癪だけど、大迫は体の使い方が上手い。マルキもそう  
 けど、あれは優秀なFWとしての資質なんだろうね。...

**鹿島・大迫、リーグ戦初ゴールは決勝点!**  
 リーグ1部(旧J1)第5節で昨季王者・鹿島の大将ルーキーFW大迫勇  
 人がリーグ戦FC東京に先発出場し、決勝点となるリーグ戦初ゴー  
 ルを挙げた。【写真も見る】FC東京に40分、FW赤嶺のヘッドで1点  
 返し、後半も攻勢したが...

**藤原紀香のお宝ピキニ水素**  
 ... 東京ヴェルディファンの僕がFC東京主催ゲームに行くのは初めて  
 でしたが、お目当てがあつての観戦でした。ももいちご 桜前線  
 2009/04/05(Sun)01:17 今日は親族の花見に行って参りました 花  
 見って言っても別に桜の下で宴会とかではないですよ...

**びっくりした。**  
 ・J1 5節め。FC東京 -2鹿島 アウェイ新潟 それって感じ  
 に行ってるのに今日は... と去年みたい  
 ですけどね、軽く風邪ひいて... と去年みたい  
 なことになりそうで恐ろしいん

携帯URL



携帯にURL

2009年

		1	2	3	4	
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

バックナンバー

[2009年4月](#)  
[2009年3月](#)  
[2009年2月](#)  
[2009年1月](#)  
[2008年12月](#)

完了

Rumor of  
 "FC Tokyo"  
 (a football  
 team in  
 Japan)

Blog snippet  
 retrieved with  
 "FC Tokyo"

"FC Tokyo"

秘宝伝 - Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)

http://hihou-den.seesaa.net/index-6.html

よく見るページ Firefox を使ってみよう 最新ニュース HotMail の無料サービス Windows Media Windows リンクの変更

記事検索

検索

Y! ウェブ 記事

<< 2009年04月 >>

日	月	火	水	木	金	土	
				1	2	3	4
5	6	7	8	9	10	11	
12	13	14	15	16	17	18	
19	20	21	22	23	24	25	
26	27	28	29	30			

国内航空券の航空券  
ドットネット  
旅行会社・チケットショップによる、国内格安航空券の検索・比較サイト  
www.kokukuen.net

群馬県・川原湯温泉 高  
田屋旅館  
温泉、砂塩温泉風呂、北投石ラ  
ジウムミスチ岩盤浴を楽しむ  
温泉宿！  
www.takadayaryokan.com

旅・コンシェルジュ  
相談できる旅行会社誕生 - 海  
外旅行の航空券、ホテル選ぶを  
サポート  
www.the-sky.jp

<<前の10件 ... 4 5 6 7 8... 次の10件>>

2008年03月14日

### ルイ・ヴィトン キーケース

東京の建築ストリート:表参道パート1  
ルイ・ヴィトン表参道先程の日本橋区向島町にある、竣工したこちらの建築物は、ルイ・ヴィトン名古屋ビル、松屋ク店など、一連のルイ・ヴィトンの建築物を手がけている...

リチャード・プリンスがデザインしたルイ・ヴィトンの...  
リチャード・プリンスがデザインしたルイ・ヴィトンのパ...

チェ・ジウ チョウ・ユンファ ミュージシャン・ヨーと共に世界的な ...  
チェ・ジウは来たる14日、ア...規模で開催される<ルイ・ヴィトン カントン・ロードショップ>の  
オープニングイベントに出席する。チェ・ジウの所属事務所<オリブナイン>の関係者は「チェ  
ジウに対する高い関心で、今回のイベントに韓国をはじめ...」(続きを読む)

今日買ったルイ・ヴィトンの製造番号がおかしいです。  
今日買ったルイ・ヴィトンの製造番号がおかしいです。今日、札幌の大黒屋でルイ・ヴィトンのナイルを  
(118000円)中古品を購入しました。ナンバーをみると、AR1017と書いてありました。番号によると  
2007年11月に製造されたみたいですが、ありますか？新品同様で少し焼けてます。新品なら心  
配ありませんか？ スポンサーサイト (続きを読む)

キーケース で検索

質屋さんを利用したいと思っています。ルイ・ヴィトンの新品バッグ、財布、キーケース、...  
質屋さんを利用したいと思っています。ルイ・ヴィトンの新品商品、財布、キーケース等売りたいので  
すが、都内で買取価格が高いお店を教えてください。今後の参考のために買取強化商  
品数を教えて欲しいです。(続きを読む)

布の偽物を大量に出品している人がいます  
(続きを読む)

Blog snippet  
retrieved with  
"LOUIS VUITTON  
Key case"

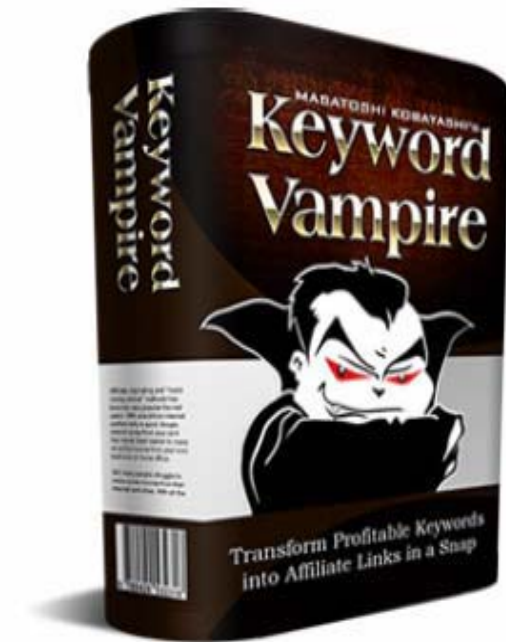
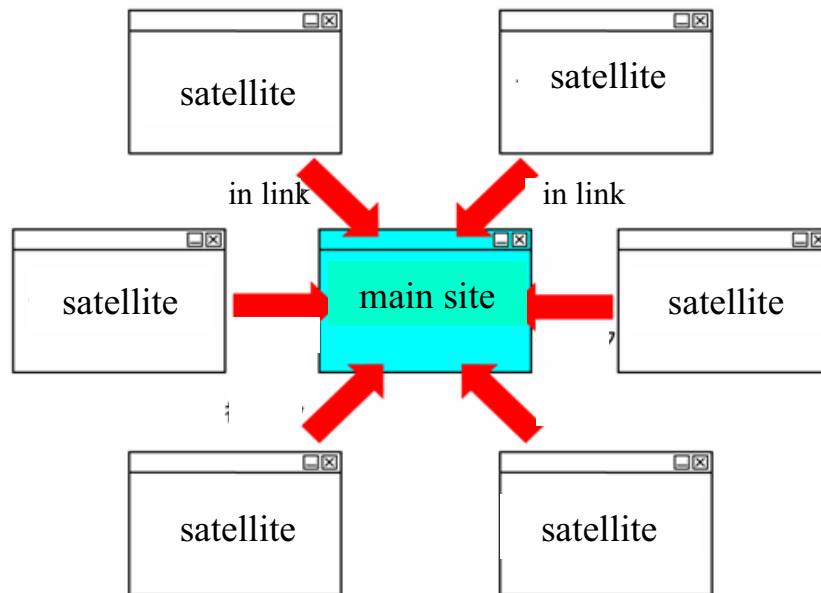
pop-up advertisement automatically  
inserted by the blog host system

http://www.seesaa.jp/hihou-den/seesaa-key-words/...

# \$50 Software Package for Massive Splog Creation

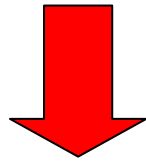
*Featuring*

- *SEO*
- *Affiliate Program*



# Background

- Opinion Mining from Blogs



- Splogs are Serious Noise in Opinion Mining
  - e.g., larger scale statistics (2008 Mar.)
    - 40% of Japanese Blog Articles in BuzzPulse, nifty are Splogs, 2007 Oct. ~ 2008 Feb.
- Automatic Detection is highly Expected.

# Previous studies on splog detection

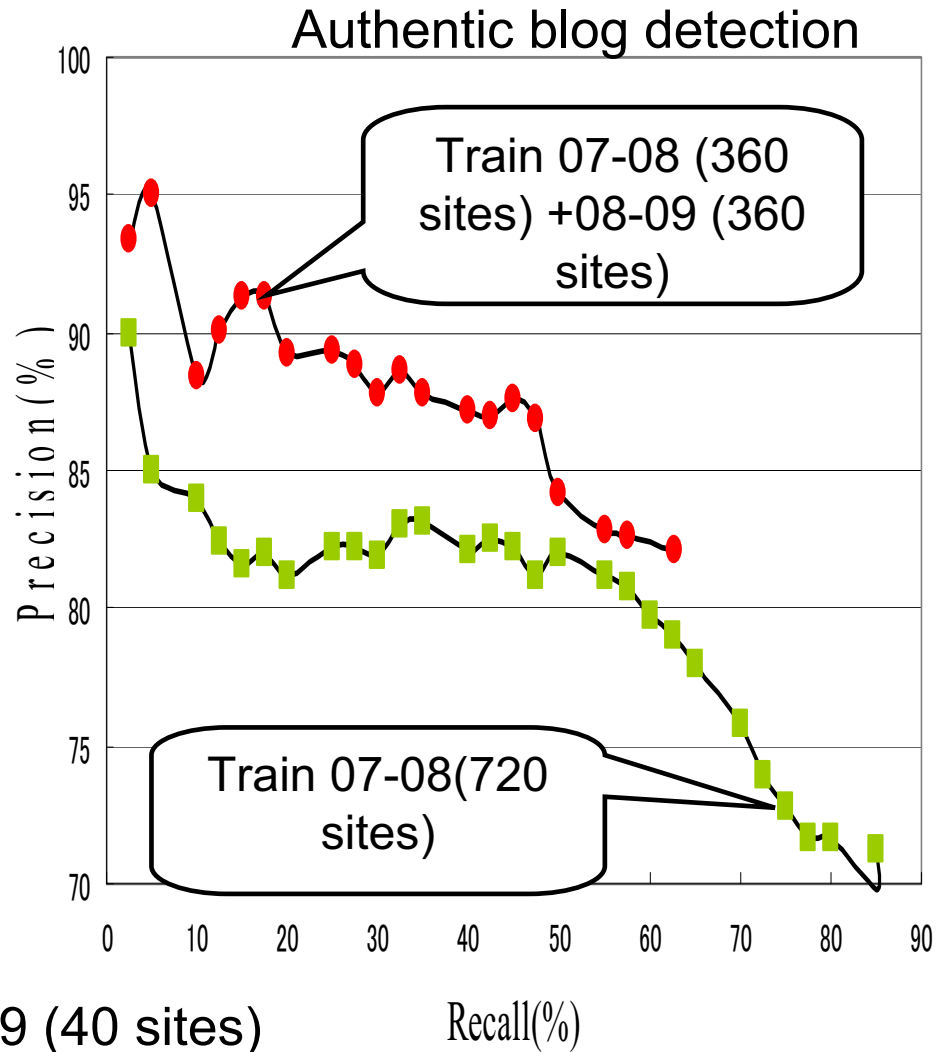
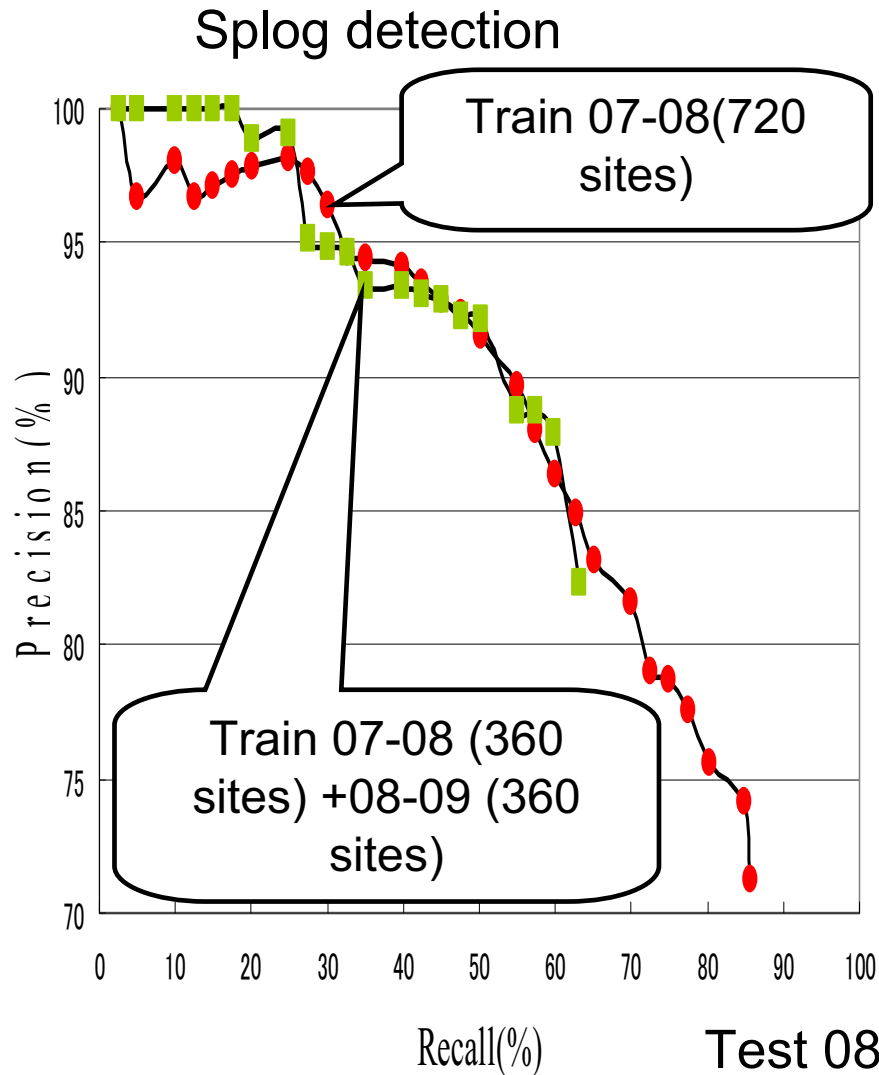
- [P.Kolari 2007]
  - Words
  - URLs
  - Anchor texts
  - Links
  - HTML meta tags
- [Y.-R.Lin 2007]
  - Temporal self similarities of
    - Posting time
    - Posting contents
    - Affiliated links
- [G.Mishne 2005]
  - Language models among the blog post , the comment ,and pages linked by the comments

# Evaluation with two data sets

## “Does splog change over time?”

1. Years 2007-2008 (720 sites)
2. Years 2008-2009 (720 sites)

# Recall/Precision curves with confidence measure



# Purpose of This Research (1)

- Needs for continuously updating splog/authentic blog data sets year by year



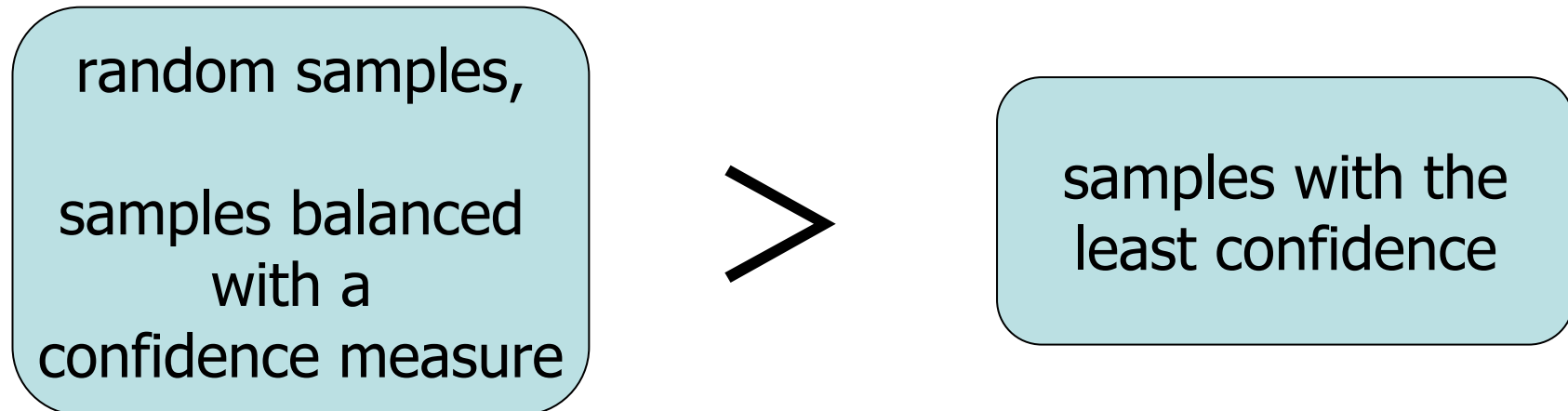
- How to reduce human supervision?



- May *active learning* framework work?

# Purpose of This Research (2)

- Optimal Strategies for Selective Sampling in Active Learning
- Guided by Certain Confidence Measure



# Outline

1. Definition of splog sites
2. Splog detection by Machine learning
  - SVM
  - Confidence Measure
  - Features
3. Active learning
4. Evaluation
5. Future works

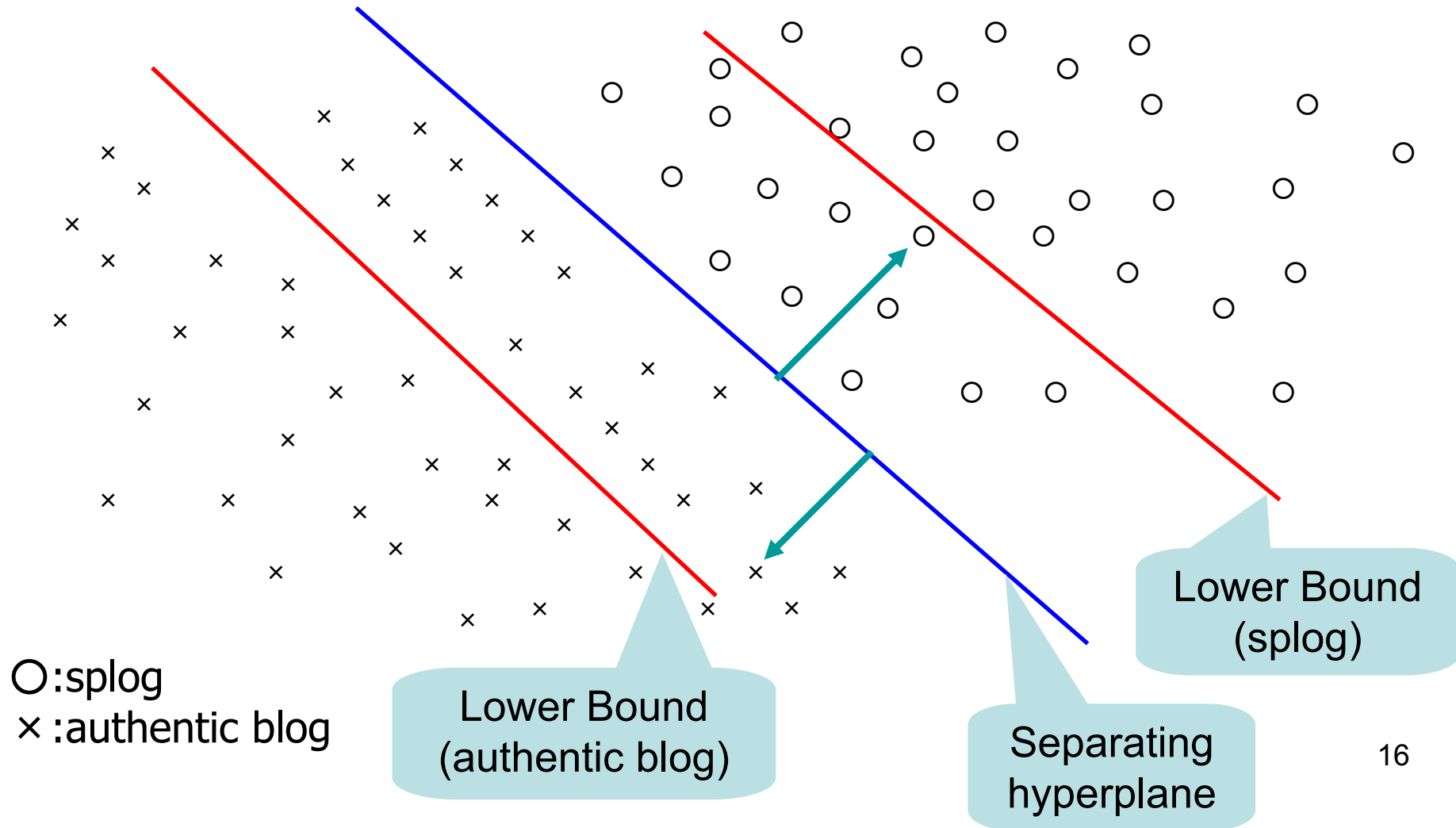
# Definition of splog sites

- If one of the followings holds for the given blog sites, then it is mostly splog
  - **originally written text is not included**
  - originally written text is included but many
    - “links top affiliated sites” or
    - ”advertisement articles” or
    - “articles with adult content”are included (judged individually by considering the contents of each blog)
- Otherwise, the given blog sites is an authentic blog

# Splog Detection by SVMs

- a tool
  - TinySVM
- the kernel function:
  - 2nd order  $>$  linear
- confidence measure
  - the distance from the separating hyperplane to each test instance

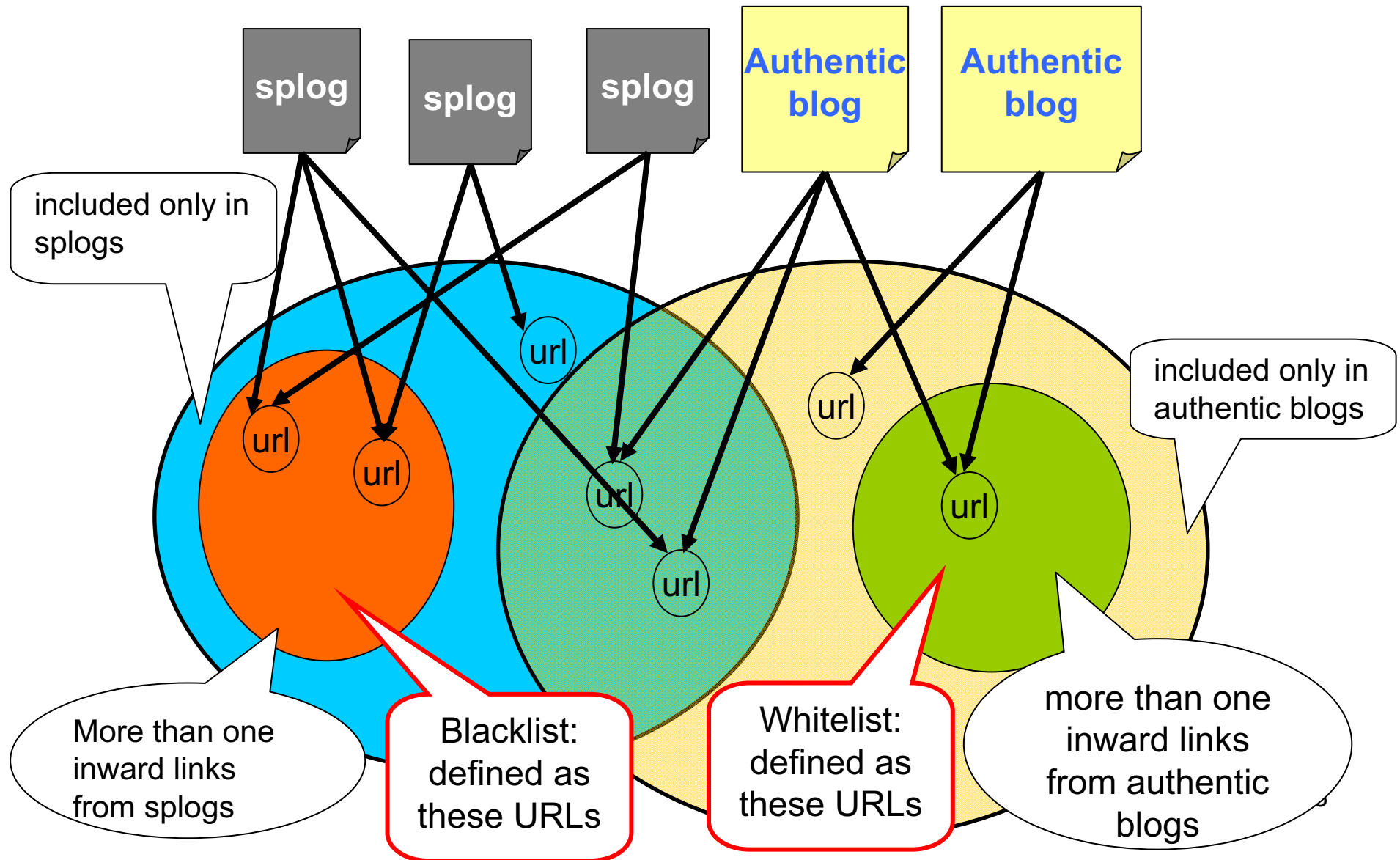
# A Confidence Measure



# Features for splog detection

1. Total frequency of URLs not linked from splogs
2. Co-occurrence between Noun Phrases and Splogs
  - Sum of  $\phi^2$  (splog, noun phrase  $w$ )
3. Noun Phrases in Anchor Texts and linked URLs
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to splog URLs and Blacklist URLs*
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to authentic blog URLs Whitelist URLs*<sup>7</sup>

# Feature1: URLs are not linked from splog



# Value of the Whitelist URLs feature

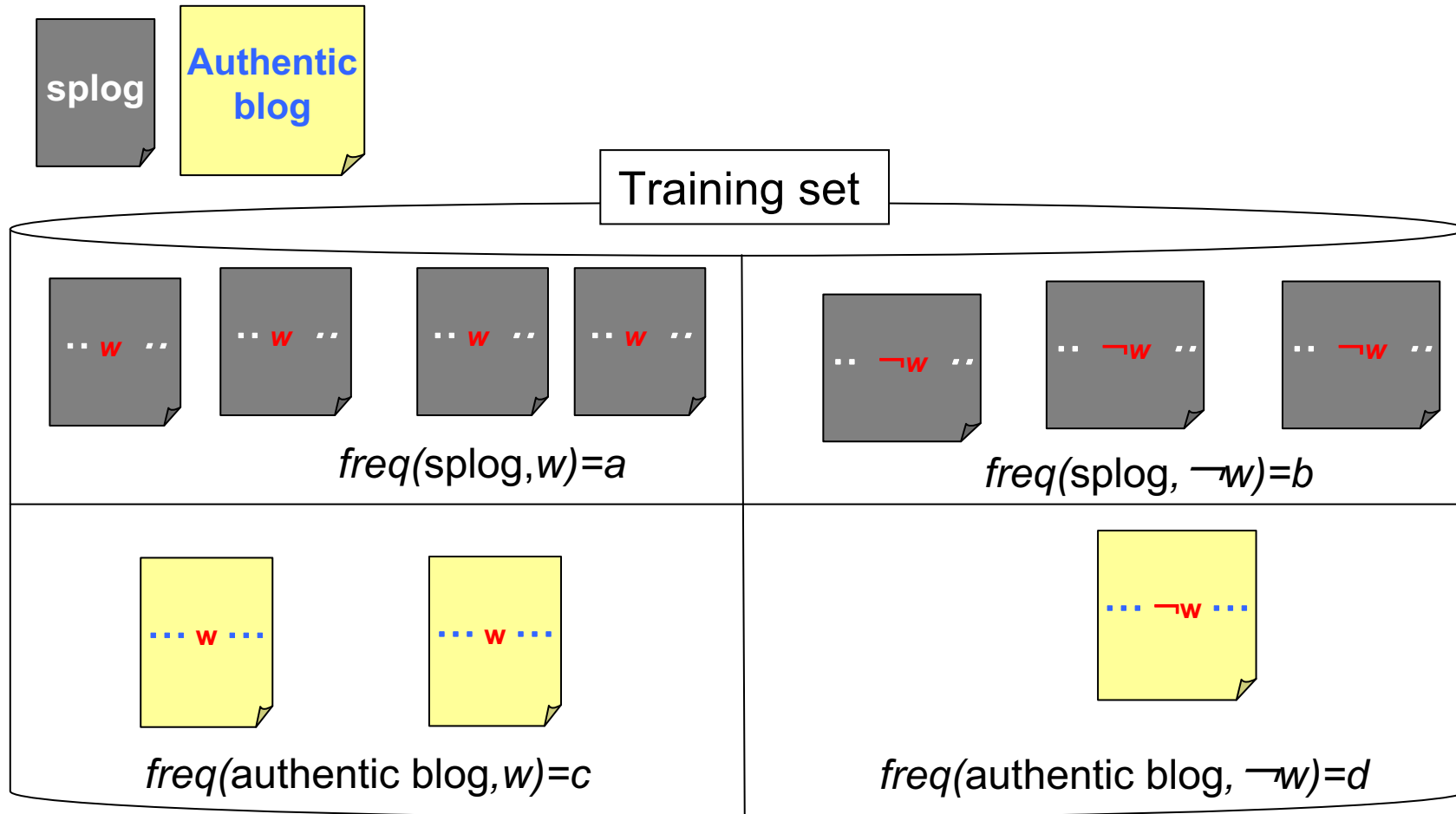
$$\log \sum_u \left( \begin{array}{l} \text{total frequency} \\ \text{of } u \text{ in the whole} \\ \text{training instances} \\ \text{of authentic blog} \\ \text{homepages} \end{array} \right) \times \left( \begin{array}{l} \text{total} \\ \text{frequency} \\ \text{of } u \text{ in} \\ \text{the test} \\ \text{instance} \end{array} \right)$$

$u$ : Whitelist URLs

# Features for splog detection

1. Total frequency of URLs not linked from splogs
2. Co-occurrence between Noun Phrases and Splogs
  - Sum of  $\phi^2$  (splog, noun phrase  $w$ )
3. Noun Phrases in Anchor Texts and linked URLs
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to splog URLs and Blacklist URLs*
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to authentic blog URLs Whitelist URLs<sup>0</sup>*

# Feature2: Noun Phrases



**w**: a noun phrase

# Value of the splog noun phrase feature

$$\phi^2(\text{splog}, w) = \frac{(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)}$$

$$\log \sum_w \phi^2(\text{splog}, w) \times \left( \begin{array}{l} \text{total frequency of } w \\ \text{in the test instance} \end{array} \right)$$

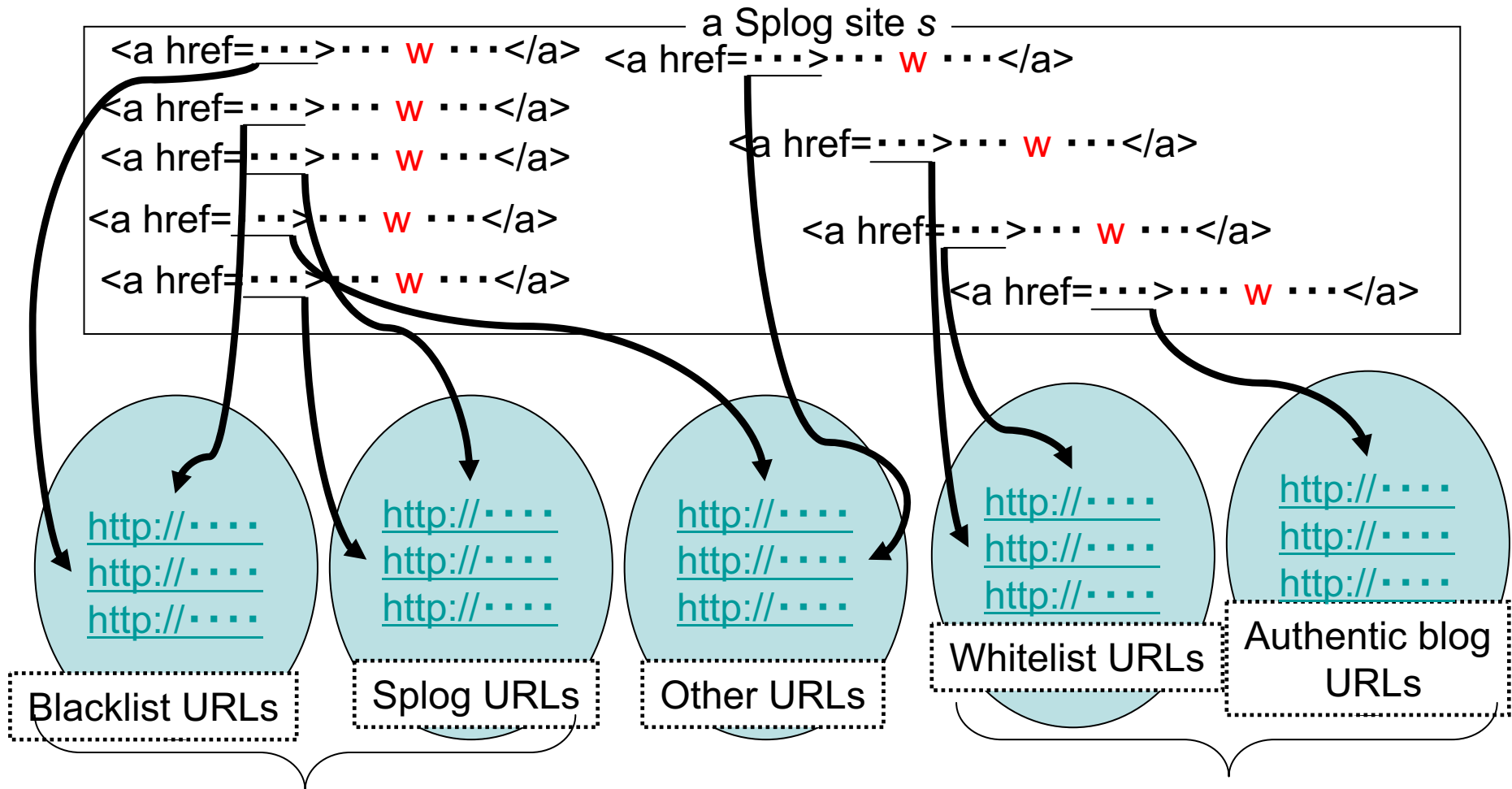
# Features for splog detection

1. Total frequency of URLs not linked from splogs
2. Co-occurrence between Noun Phrases and Splogs
  - Sum of  $\phi^2$  (splog, noun phrase  $w$ )
3. Noun Phrases in Anchor Texts and linked URLs
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to splog URLs and Blacklist URLs*
  - *Total frequency of anchor text noun phrases*
    - *in splogs*
    - *out-linked to authentic blog URLs Whitelist URLs*<sup>3</sup>

# Feature3:

## Noun Phrases in Anchor Texts and linked URLs

*w*: a noun phrase in Anchor text



$$AnchB(w,s)=freq\ of\ w$$

$$AnchW(w,s)=freq\ of\ w\ 24$$

# Noun Phrases in Anchor Texts and linked URLs: two features

the value of a feature named *anchor text noun phrase out-linked to Blacklist URLs* for a test instance blog homepage

$$\log \sum_w \left( \sum_s AncfB(w, s) \right) \times AncfB(w, t)$$

the value of a feature named *anchor text noun phrase out-linked to Whitelist URLs* for a test instance blog homepage

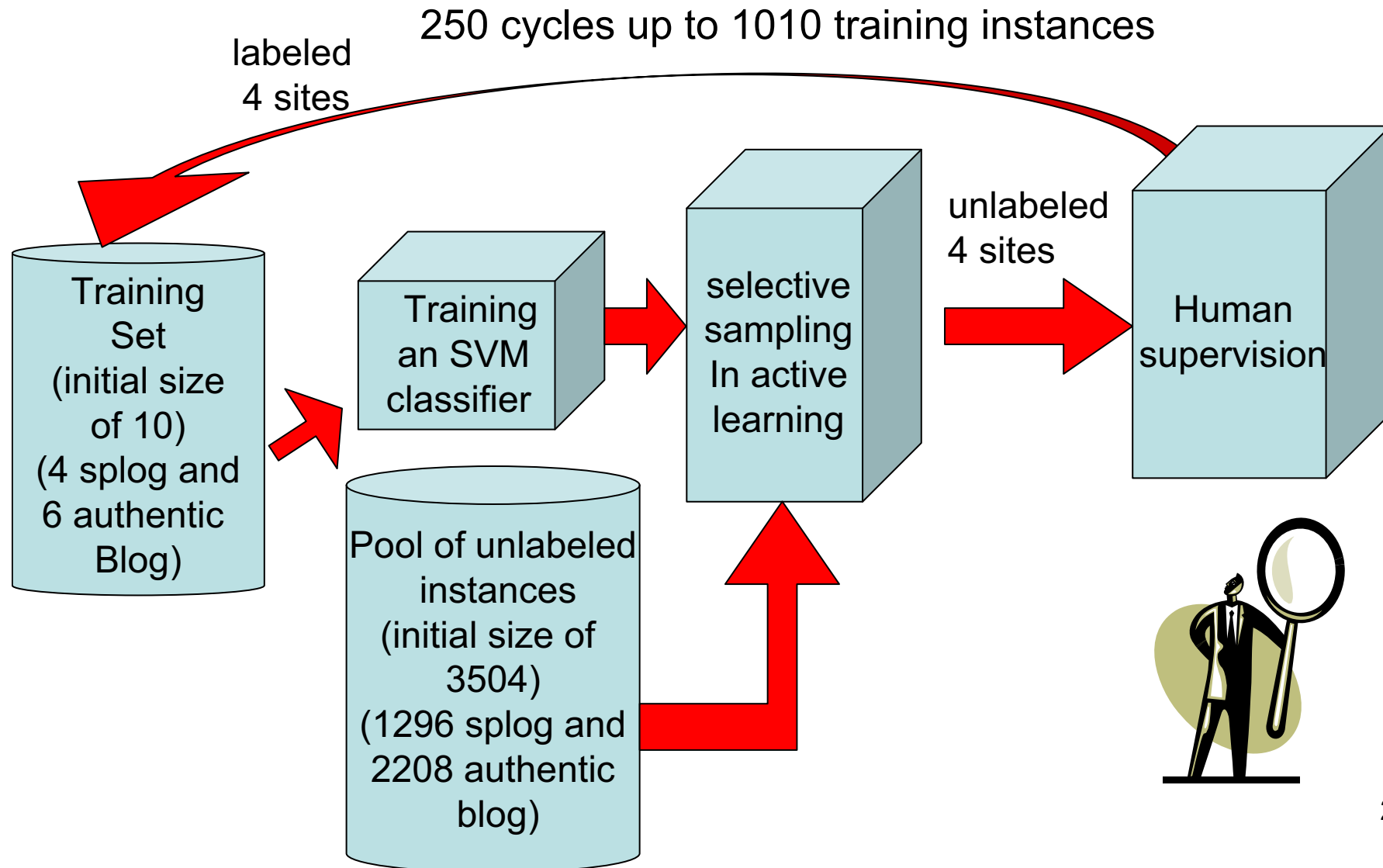
$$\log \sum_w \left( \sum_{\substack{\text{trainingsplog} \\ \text{homepages}}} AncfW(w, s) \right) \times AncfW(w, t)$$

*w*: noun phrase

*s*: a training splog homepage

*t*: a test instance blog homepage

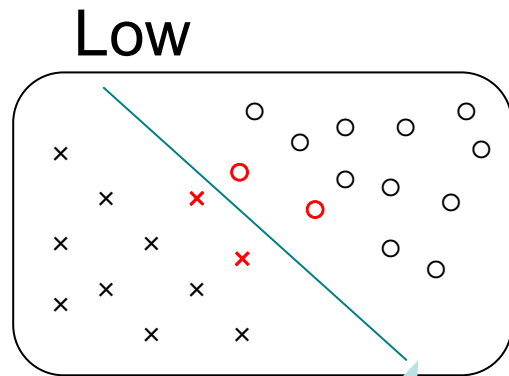
# Framework of Active learning



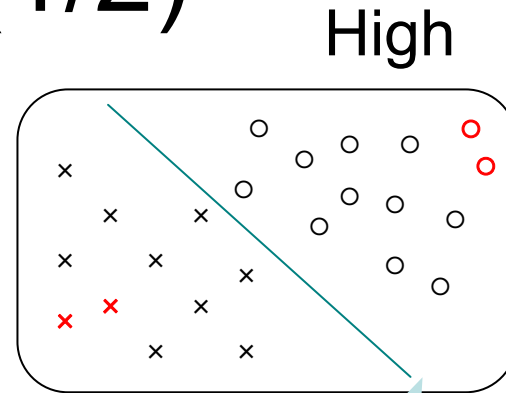
# Statistics of Splog/Authentic Blogs Data Set

Data Sets	# of splogs	# of authentic blogs	total
Years 2008-2009	1445	2459	3904

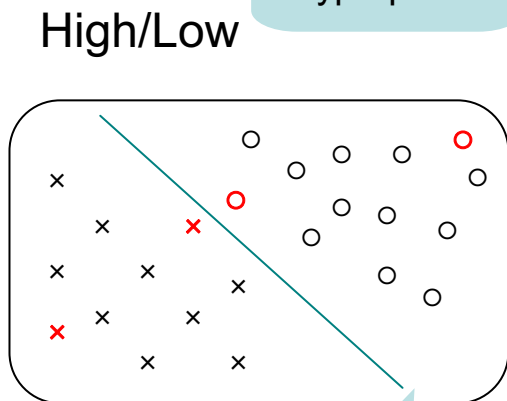
# Strategies of selective sampling(1/2)



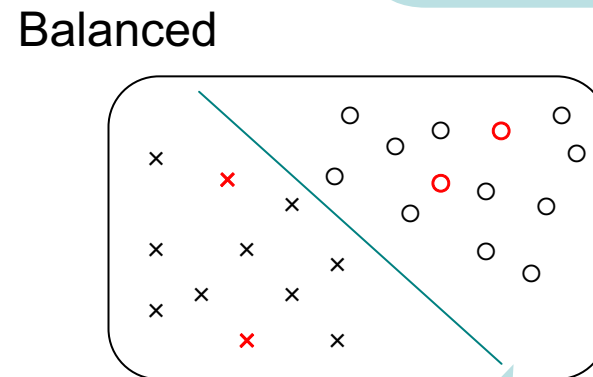
Separating hyperplane



Separating hyperplane



Separating hyperplane

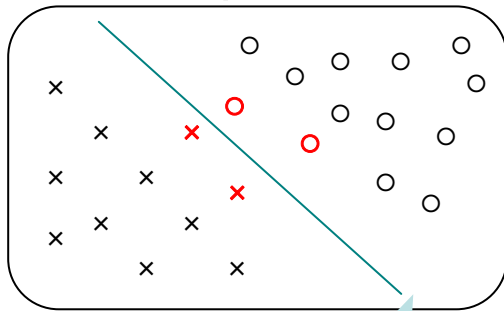


Separating hyperplane

O:splog  
×:authentic blog

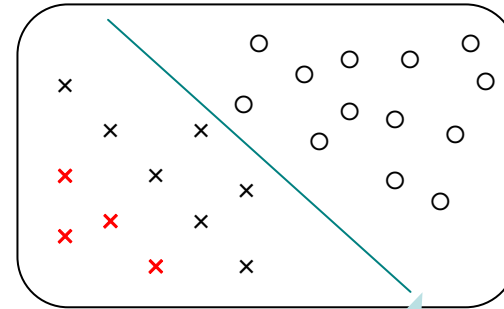
# Strategies of selective sampling(2/2)

Low-Sp/Au



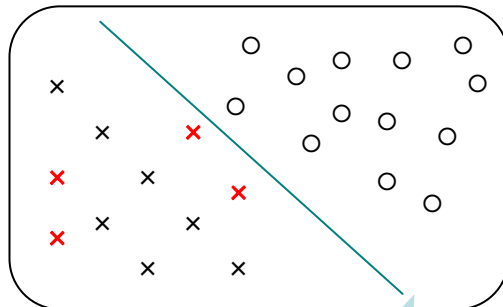
Separating hyperplane

High-Au



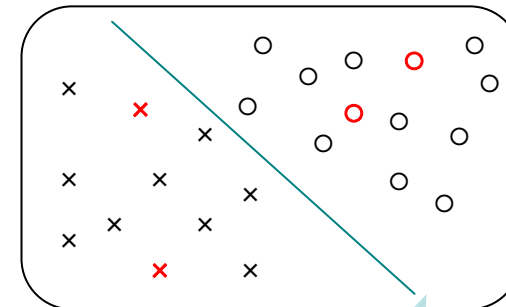
Separating hyperplane

High/Low-Au



Separating hyperplane

Balanced-Sp/Au



Separating hyperplane

O:splog  
×:authentic blog

# Outline

1. Definition of splog sites
2. Splog detection by Machine learning
  - SVM
  - Confidence Measure
  - Features
3. Active learning
4. Evaluation
5. Future works

# Measure for Performance evaluation after active learning cycles

- Recall/Precision

- Splog detection

$$\text{precision} = \frac{|\text{Ts}(\text{splog}) \cap \text{Ts}(\text{LBD}_s)|}{|\text{Ts}(\text{LBD}_s)|}$$

$$\text{recall} = \frac{|\text{Ts}(\text{splog}) \cap \text{Ts}(\text{LBD}_s)|}{|\text{Ts}(\text{splog})|}$$

- Authentic blog detection is considered in a similar fashion

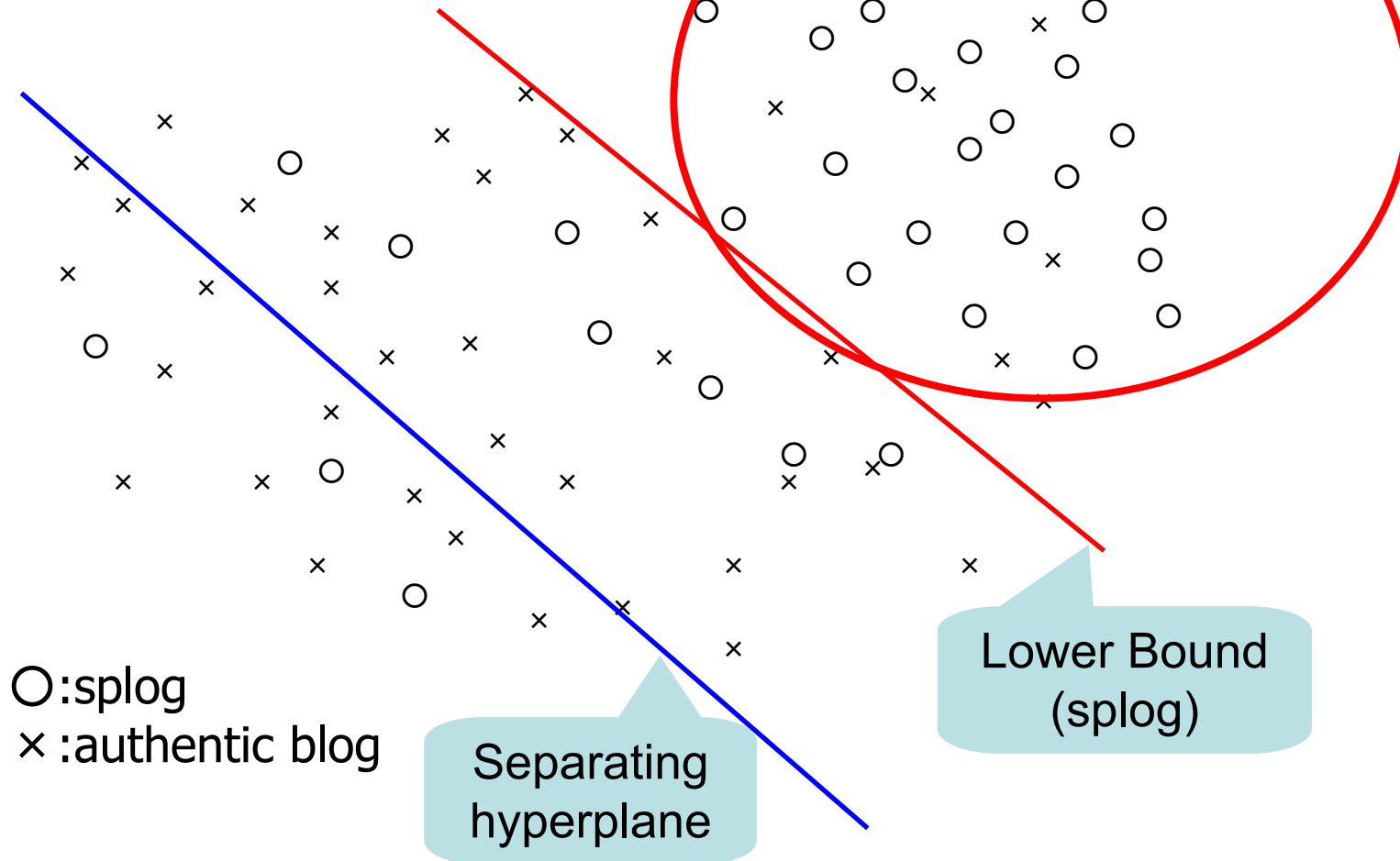
- “|  $Tr$  |= 3500”, “Random”

- “|  $Tr$  |= 3500” indicates a classifier trained with the whole 3504 instances in the pool
- “Random” indicates a classifier trained with randomly selected training instances

# Lower Bound of the Confidence Measure

$T_s(\text{splog})$ : the set of reference splog sites

$T_s(LBD_s)$



# Measure for Performance evaluation after active learning cycles

- Recall/Precision

- Splog detection

$$\text{precision} = \frac{|T_s(\text{splog}) \cap T_s(\text{LBD}_s)|}{|T_s(\text{LBD}_s)|}$$

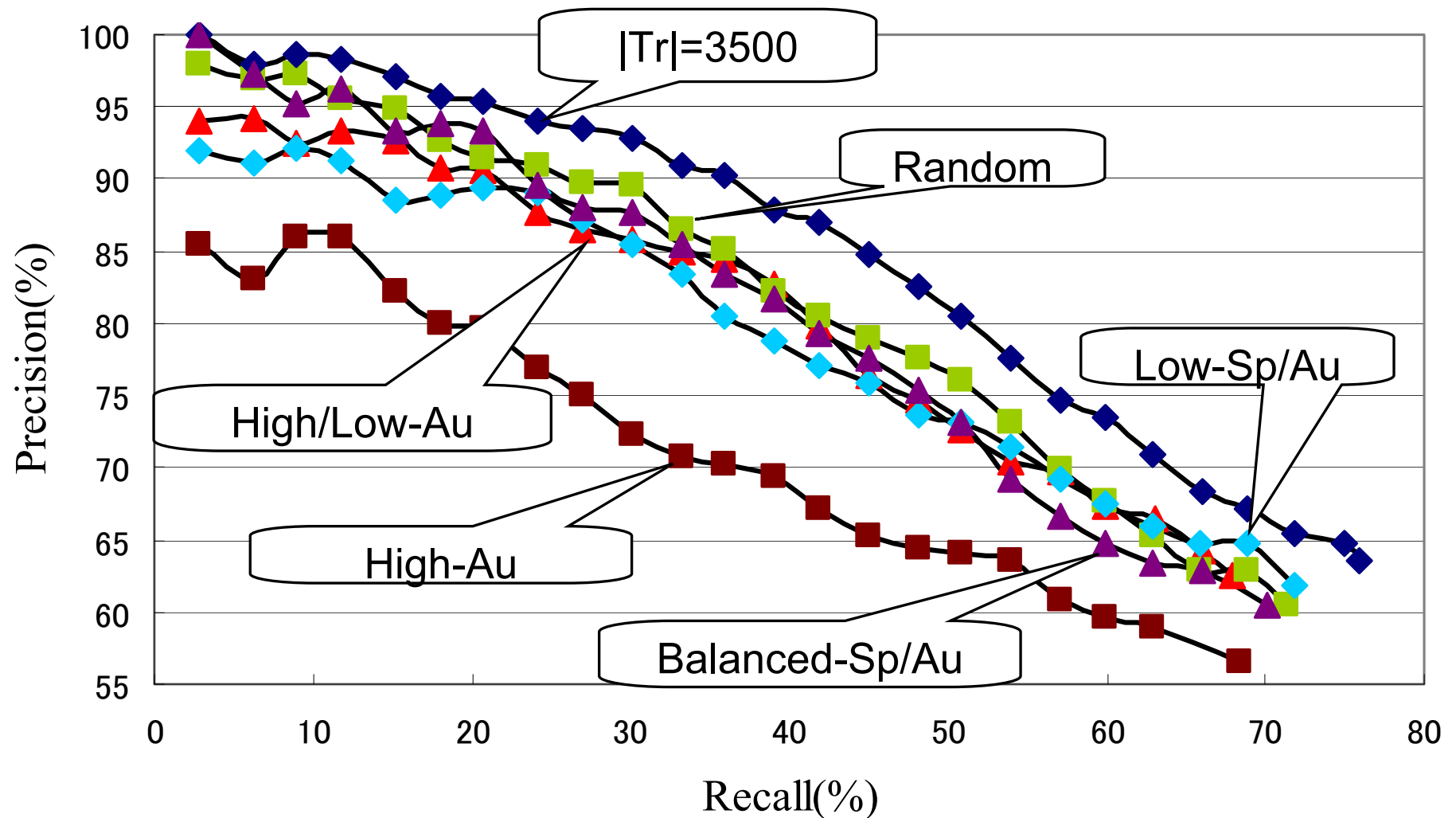
$$\text{recall} = \frac{|T_s(\text{splog}) \cap T_s(\text{LBD}_s)|}{|T_s(\text{splog})|}$$

- Authentic blog detection is considered in a similar fashion

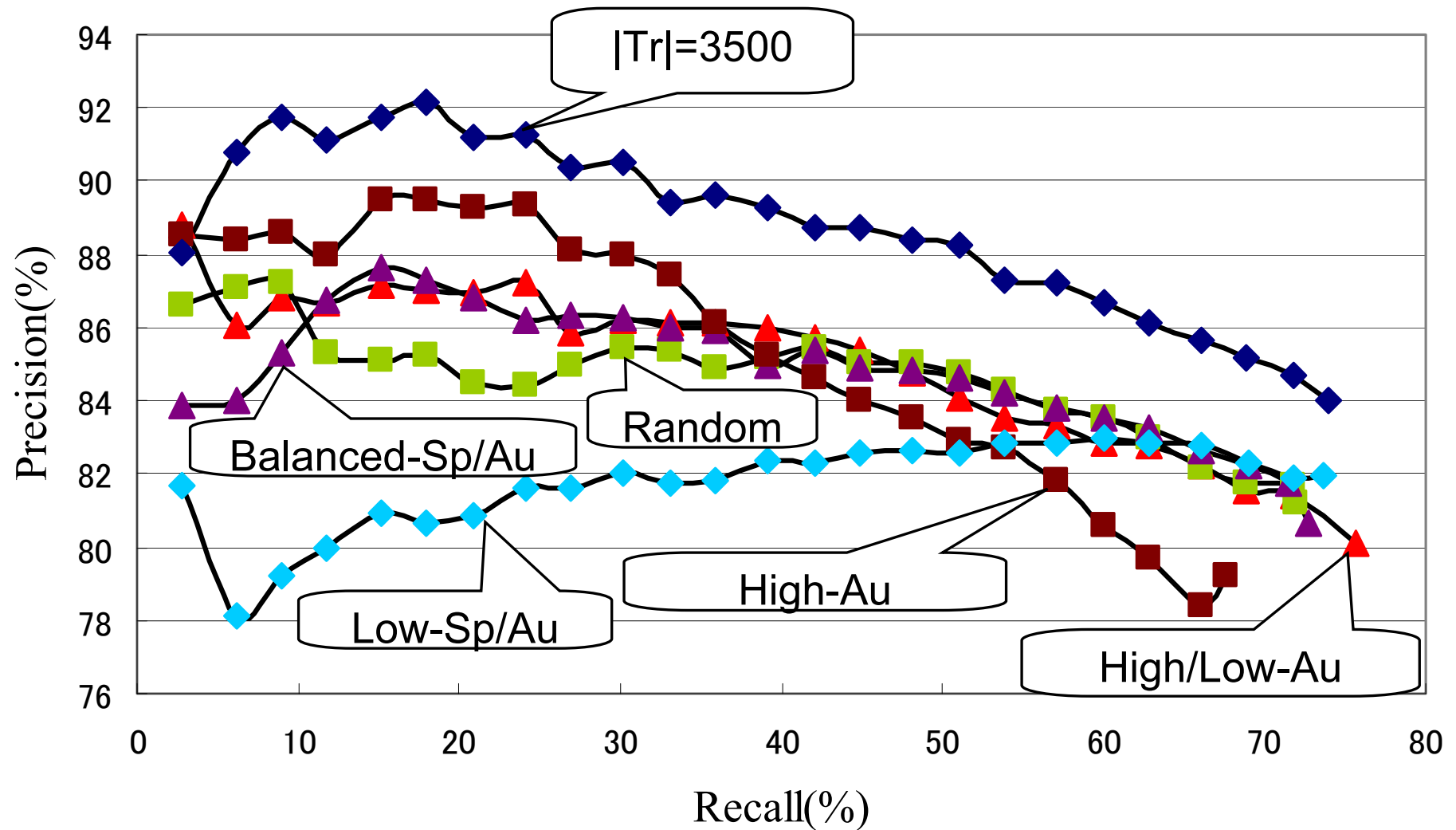
- “ $|Tr|=3500$ ”, “Random”

- “ $|Tr|=3500$ ” indicates a classifier trained with the whole 3504 instances in the pool
- “Random” indicates a classifier trained with randomly selected training instances

# Recall/precision curve of Splog detection

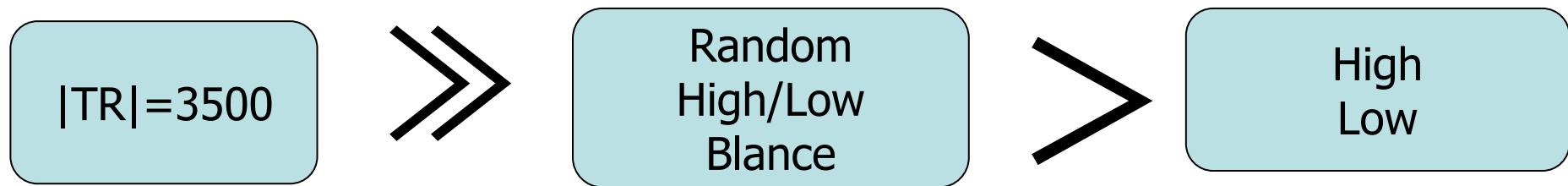


# Recall/precision curve of Authentic blog Detection

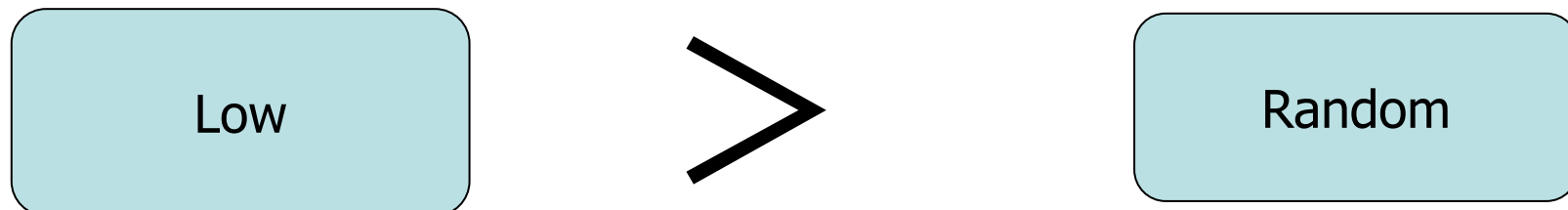


# Evaluation results: comparison of strategies for selective sampling

Splog/authentic blog detection

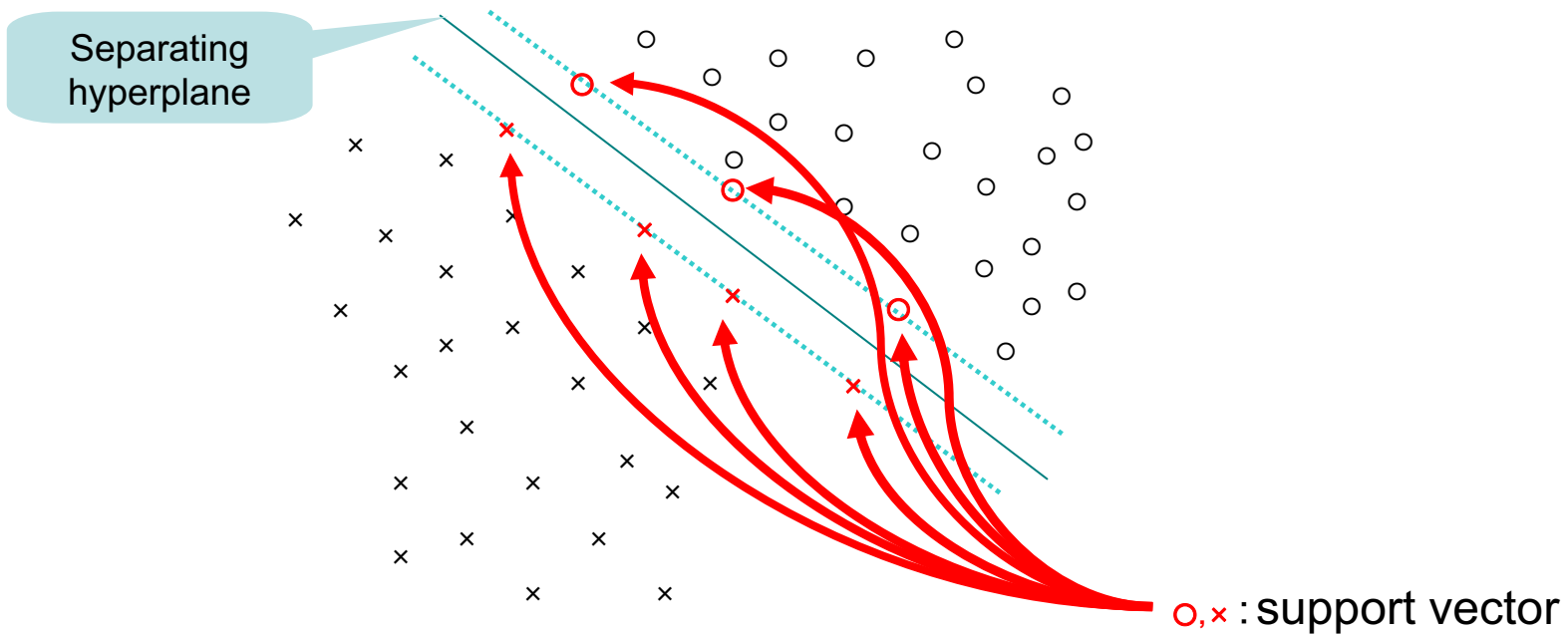


Previous studies of active learning for text classification tasks

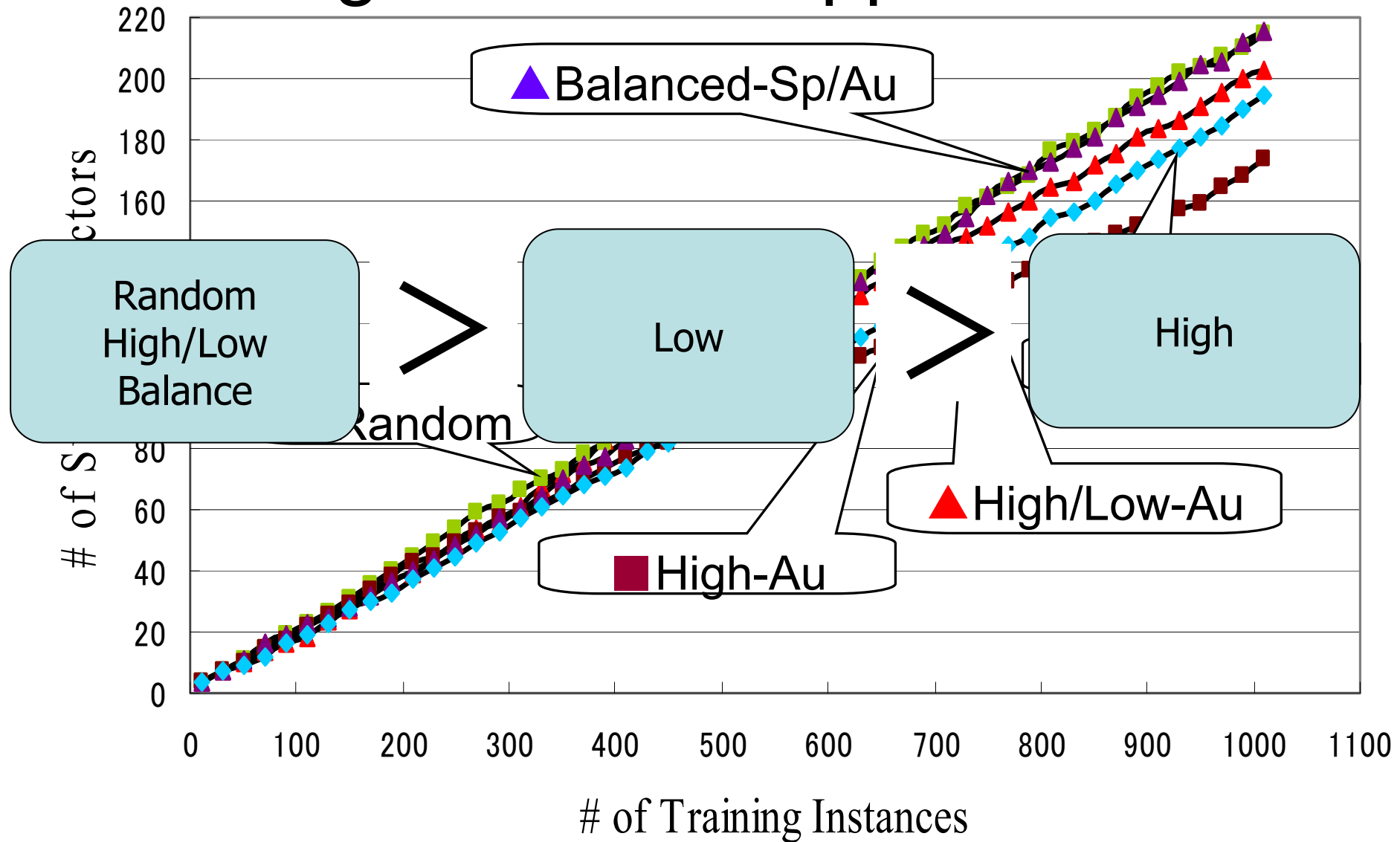


# Support Vectors

- only the support vectors have effect on deciding the position of the separating hyperplane
- the number of support vectors can be regarded as the complexity of the learning task

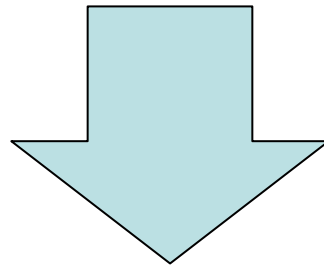


# Changes in # of Support Vectors



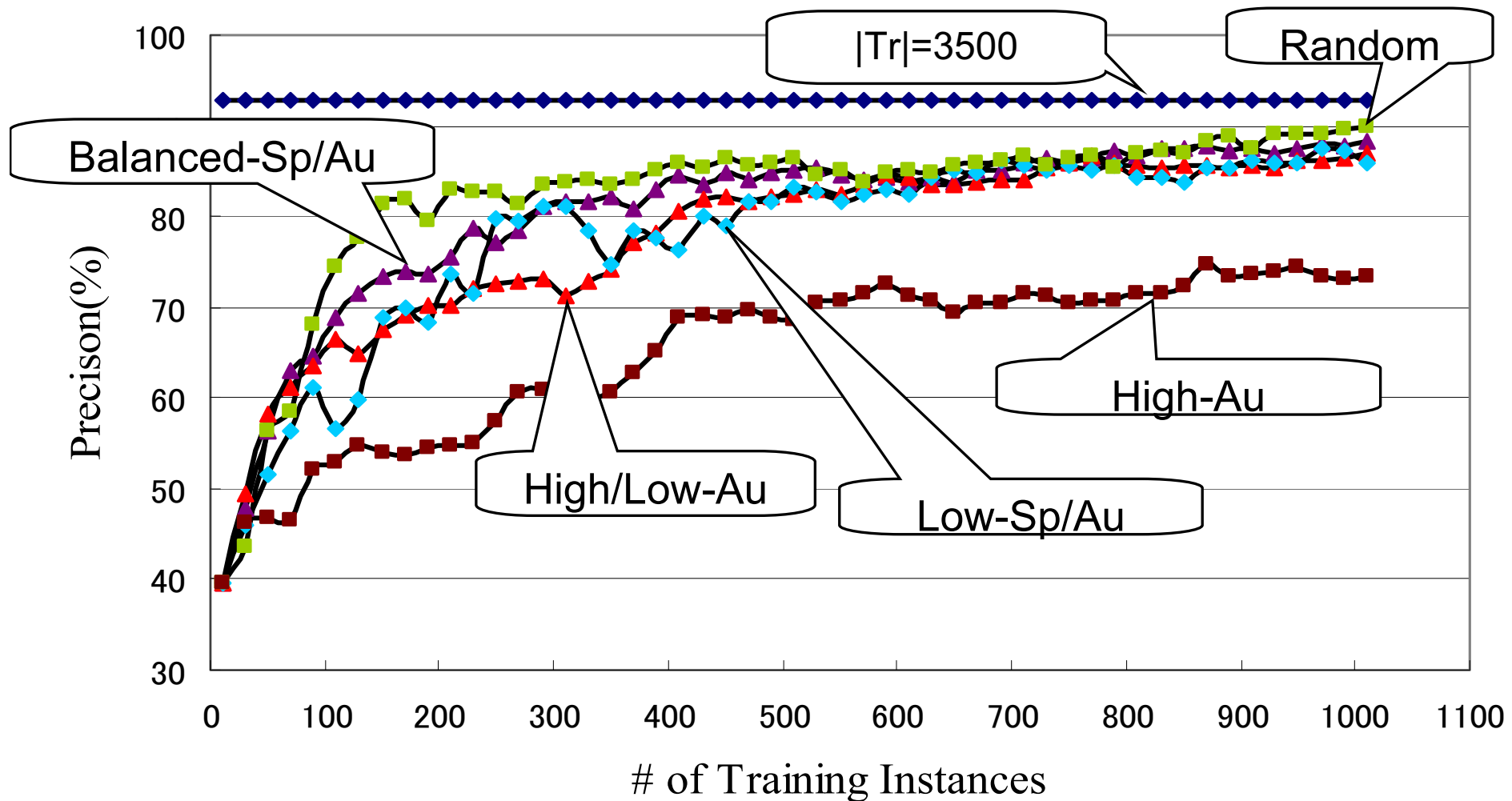
# Evaluation result: # of support vectors

- The number of support vectors linearly increases
- Performance of splog/authentic blog detection increase much more slowly
- About 20% of training instances are constantly selected as support vectors

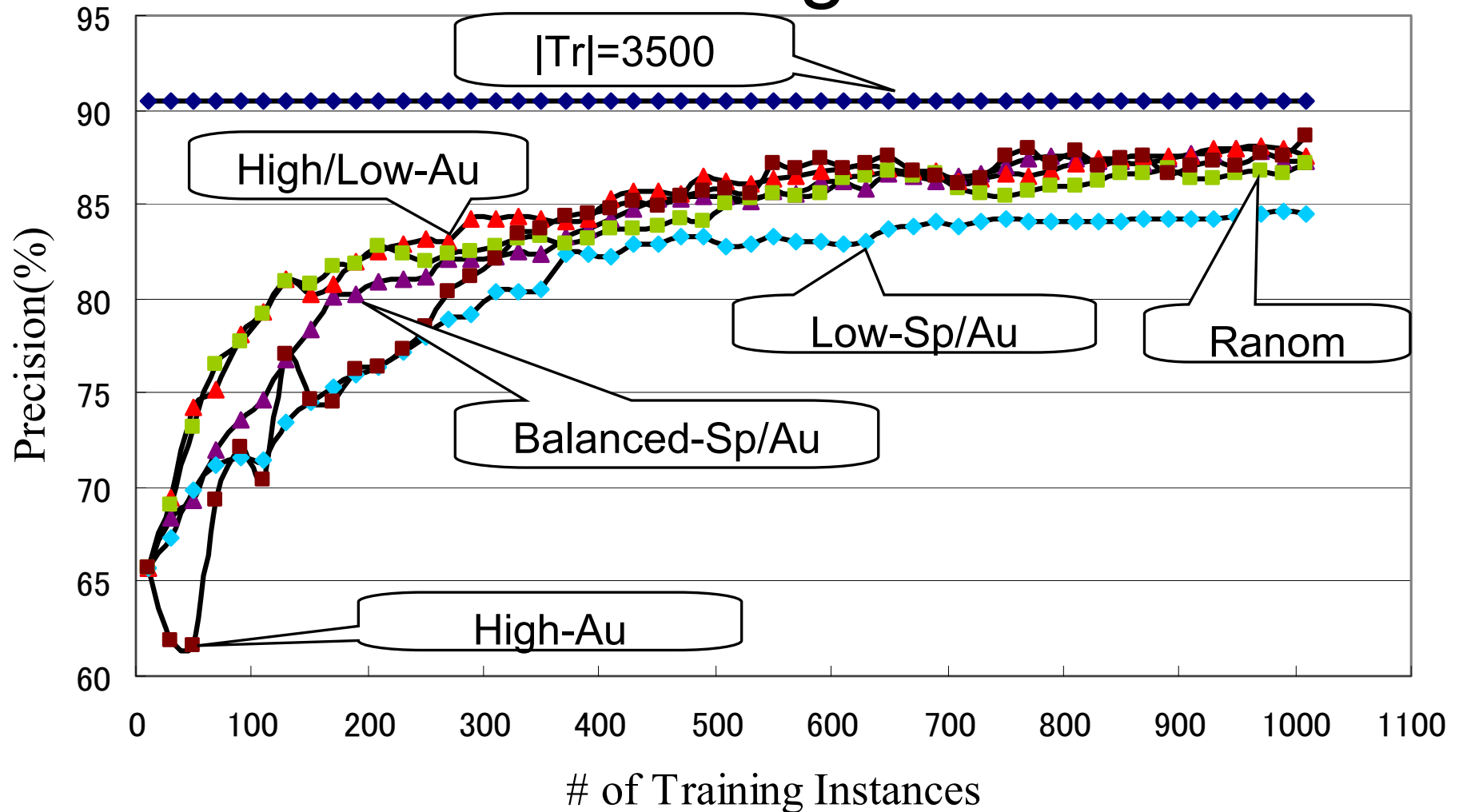


- In this task, more effective features should be added.

# Change in maximum precision with recall as 30 % of Splog Detection

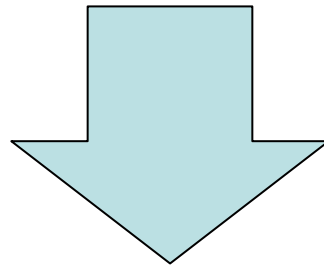


# Change in maximum precision with recall as 30 % of Authentic blog Detection



# Evaluation result: # of support vectors

- The number of support vectors linearly increases
- Performance of splog/authentic blog detection increase much more slowly
- About 20% of training instances are constantly selected as support vectors



- In this task, more effective features should be added.

# Future works

- Incorporating other features
  - Post time and intervals
  - Html structures
- Manual examination of support vectors

Thanks for your attention