Information Cascades: Topic Detection & Tracking

Social Computing

Department of Computer Science
University of Massachusetts, Lowell

Hadi Amiri
hadi@cs.uml.edu

Lecture Topics

- Topic Detection
- Topic Tracking
- Early Prediction

Matrix Factorization for Topic Detection

Topic Detection

Tweets

Computer technology: 2-Tone L.E.D. to Simplify Screens

Stock Market: A Better Deal For Investors Isn't Simple. Large Sale 03/02

The Shape of Cinema, Transformed At the Click of a Mouse. Movie production.

The three big Internet portals begin to distinguish among themselves as shopping malls

Topic Detection

Topics

Computer	**0.02**
Technology	**0.03**
System	**0.04**
Internet	**0.01**
...	

Sale	**0.02**
Product	**0.03**
Market	**0.02**
Consumer	**0.04**
...	

Film	**0.05**
Movie	**0.04**
Theater	**0.02**
Production	**0.04**
...	

Tweets

Computer technology: 2-Tone L.E.D. to Simplify Screens

Stock Market: A Better Deal For Investors Isn't Simple. Large Sale 03/02

The Shape of Cinema, Transformed At the Click of a Mouse. Movie production.

The three big Internet portals begin to distinguish among themselves as shopping malls

A **topic** is a distribution over words

Topic Detection

TF/IDF

Topics		Tweets	Assignments

Computer 0.02
Technology 0.03
System 0.04
Internet 0.01
...

Sale 0.02
Product 0.03
Market 0.02
Consumer 0.04
...

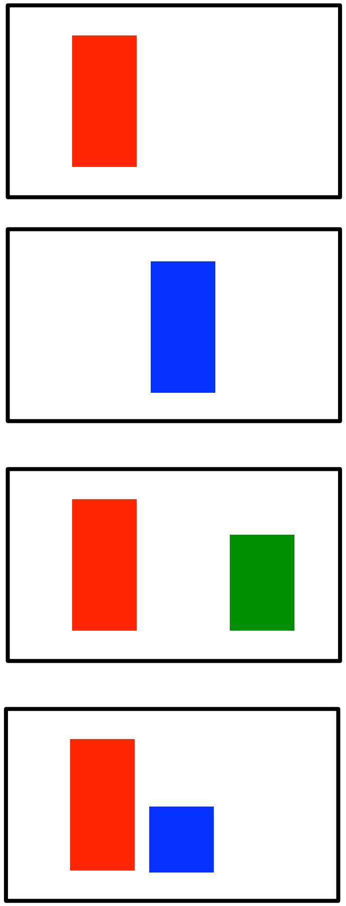
Film 0.05
Movie 0.04
Theater 0.02
Production 0.04
...

Computer technology: 2-Tone L.E.D. to Simplify Screens

Stock Market: A Better Deal For Investors Isn't Simple. Large Sale 03/02

The Shape of Cinema, Transformed At the Click of a Mouse. Movie production.

The three big Internet portals begin to distinguish among themselves as shopping malls



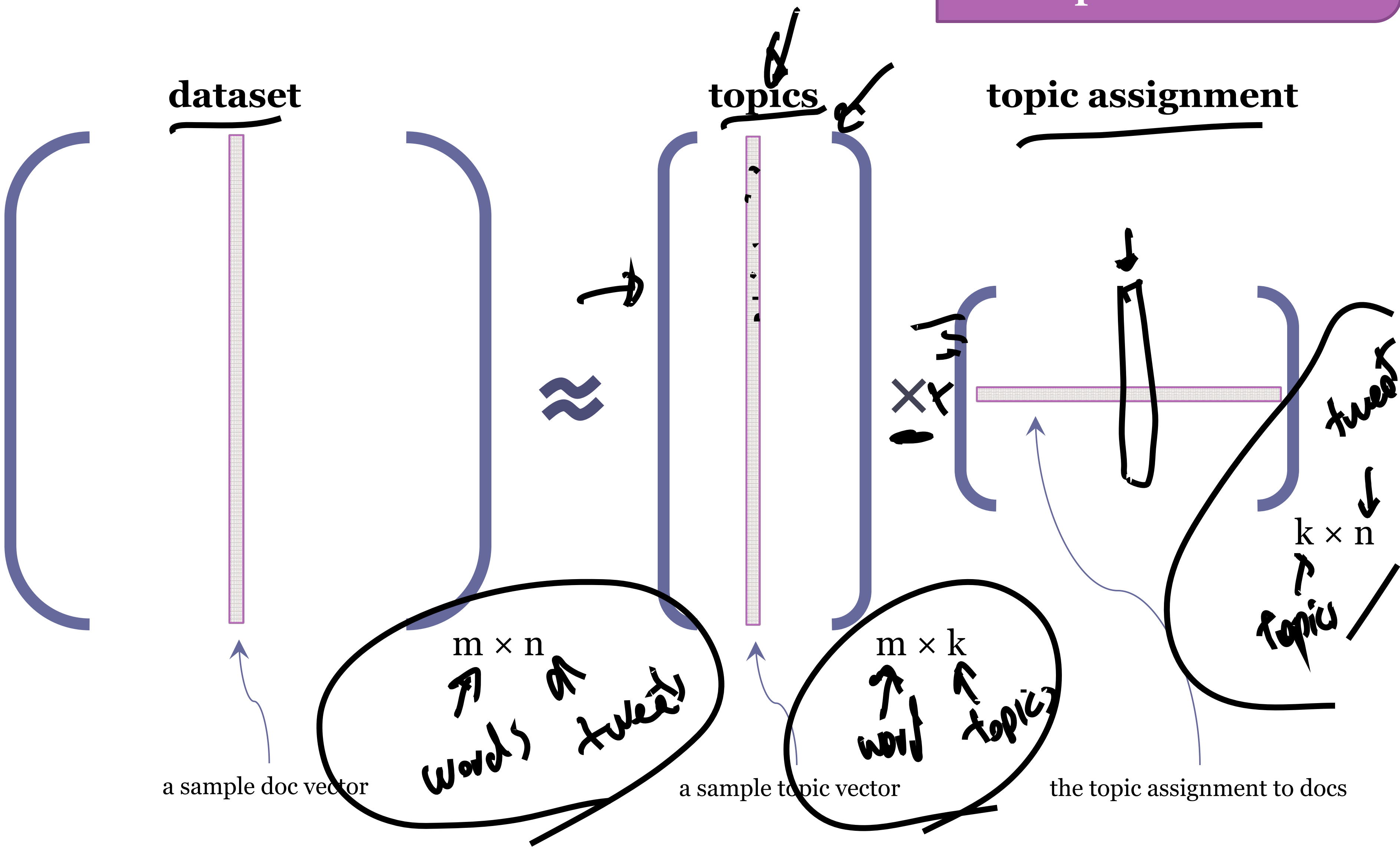
A **topic** is a distribution over words
A **tweet** is a mixture of topics / distribution over topics

Topic Detection

- Different learning techniques
- Matrix factorization methods
  - LU decomposition
  - Singular Value Decomposition(SVD)
  - Probabilistic Matrix Factorization(PMF)
  - (Online) Non-negative Matrix Factorization(NMF)
  - Etc.

Topic Detection - NMF

m: # **terms** in the dataset
n: # **docs** in the dataset
k: # **topics** in the dataset



Topic Detection - NMF

m: # **terms** in the dataset
n: # **docs** in the dataset
k: # **topics** in the dataset

$$\underset{m \times n}{\overset{\textbf{dataset}}{\begin{pmatrix} & & \\ & S & \\ & & \end{pmatrix}}} \approx \underset{m \times k}{\overset{\textbf{topics}}{\begin{pmatrix} \\ D \\ \\ \end{pmatrix}}} \times \underset{k \times n}{\overset{\textbf{topic assignment}}{\begin{pmatrix} & X & \end{pmatrix}}}$$

Topic Detection - NMF

m: # **terms** in the dataset
n: # **docs** in the dataset
k: # **topics** in the dataset

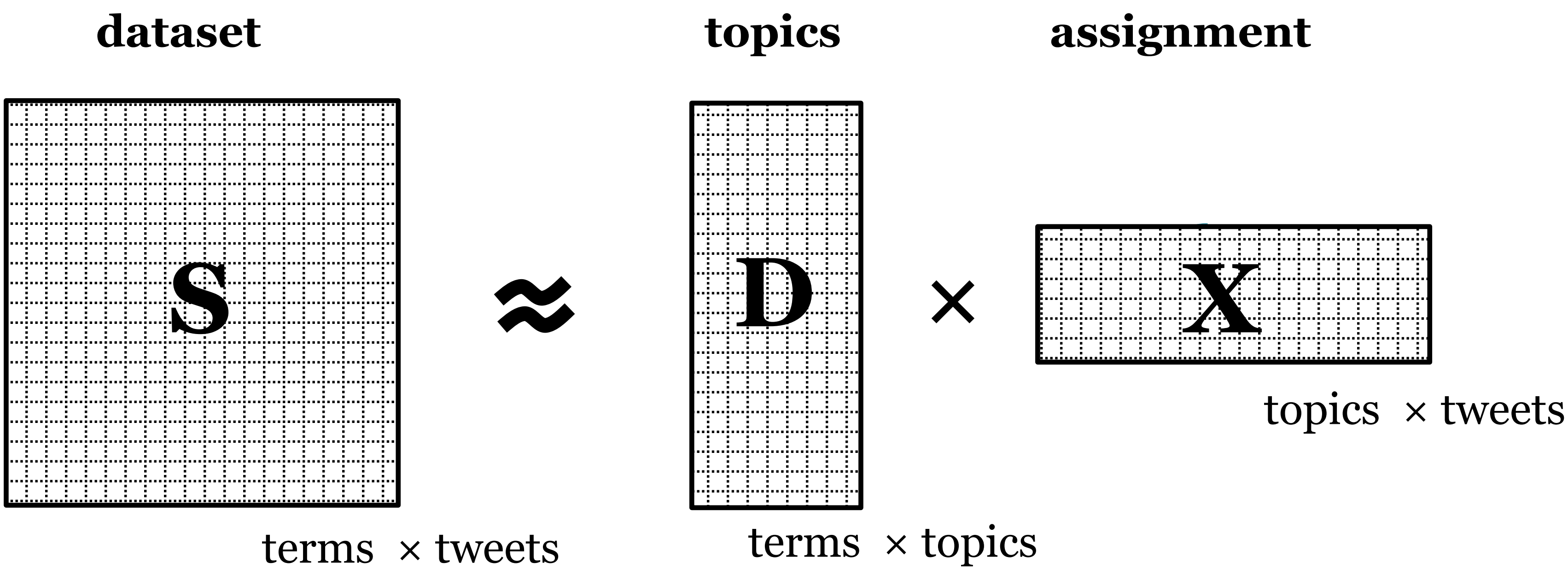
dataset		topics		topic assignment
$\mathbf{S}_{m \times n}$	$\approx$	$\mathbf{D}_{m \times k}$	$\times$	$\mathbf{X}_{k \times n}$

α

$$(\mathbf{D}, \mathbf{X}) = \mathbf{arg\,min}_{\,\mathbf{D},\mathbf{X}} \left\| \mathbf{S} - \mathbf{D}\mathbf{X} \right\|_F^2 + \lambda \left\| \mathbf{X} \right\|_1$$

s.t. $\mathbf{X} \geq \mathbf{0}$, $\mathbf{D} \geq \mathbf{0}$, $\|\boldsymbol{d_i}\| = \mathbf{1}$ *for* $i = \{1,\ldots k\}$

Topic Detection - NMF



$$(\mathbf{D}, \mathbf{X}) = \mathbf{arg\,min}_{\mathbf{D},\mathbf{X}} \left\| \mathbf{S} - \mathbf{D}\mathbf{X} \right\|_F^2 + \lambda \left\| \mathbf{X} \right\|_1$$

$$\text{s.t. } \mathbf{X} \geq \mathbf{0}, \mathbf{D} \geq \mathbf{0}, \|\boldsymbol{d}_i\| = \mathbf{1} \textit{ for } i = \{1,\ldots k\}$$

Source: Paatero and Tapper, 1994; Lee and Seung, 1999; Chih-Jen Lin, 2007

Topic Detection - NMF

$$(\mathbf{D}, \mathbf{X}) = \mathbf{arg\,min}_{\mathbf{D},\mathbf{X}} \left\| \mathbf{S} - \mathbf{D}\mathbf{X} \right\|_F^2 + \lambda \left\| \mathbf{X} \right\|_1$$

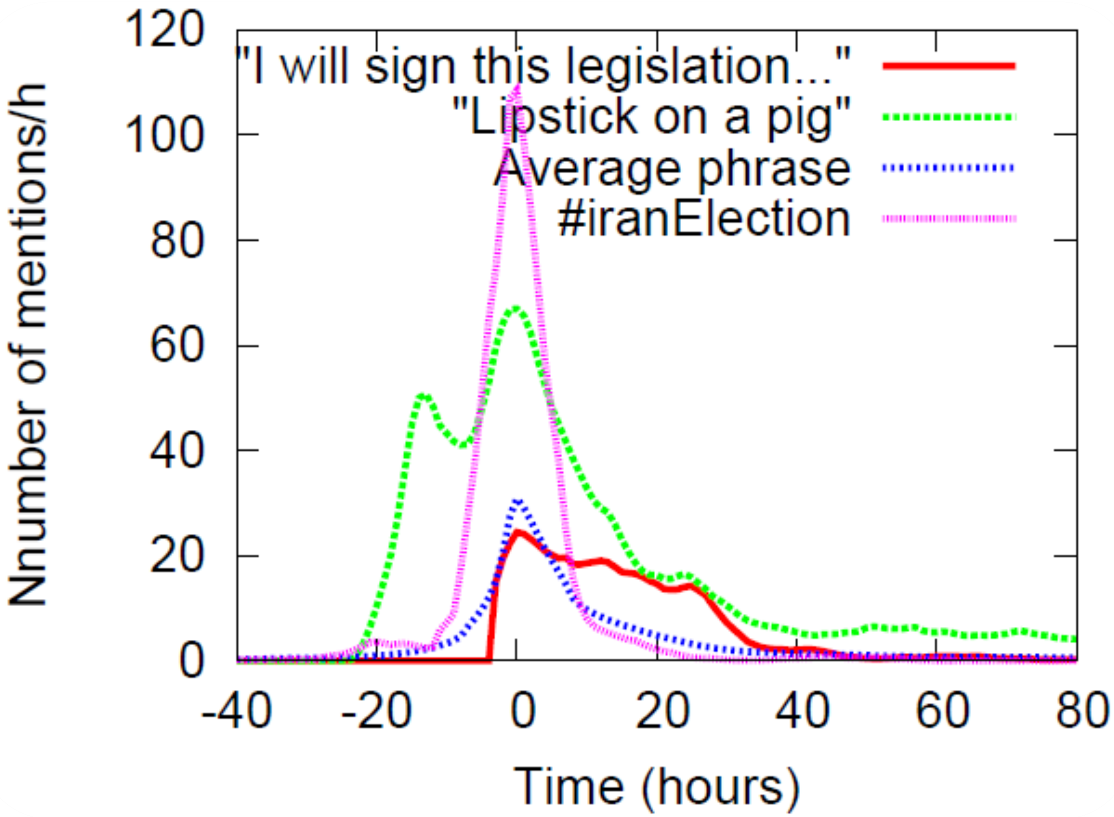
s.t. $\mathbf{X} \geq \mathbf{0}$, $\mathbf{D} \geq \mathbf{0}$, $\|\boldsymbol{d}_i\| = \mathbf{1}$ *for* $i = \{1,\ldots k\}$

- Non-convex optimization problem.
  - many local optimum.
- But, if one of the variables, either **D** or **X**, is known, optimization wrt the other will be convex.
  - Solution:
    - Iteratively optimize the objective function
      - Alternatively optimize wrt **D** and **X** while holding the other fixed!

Topic Tracking

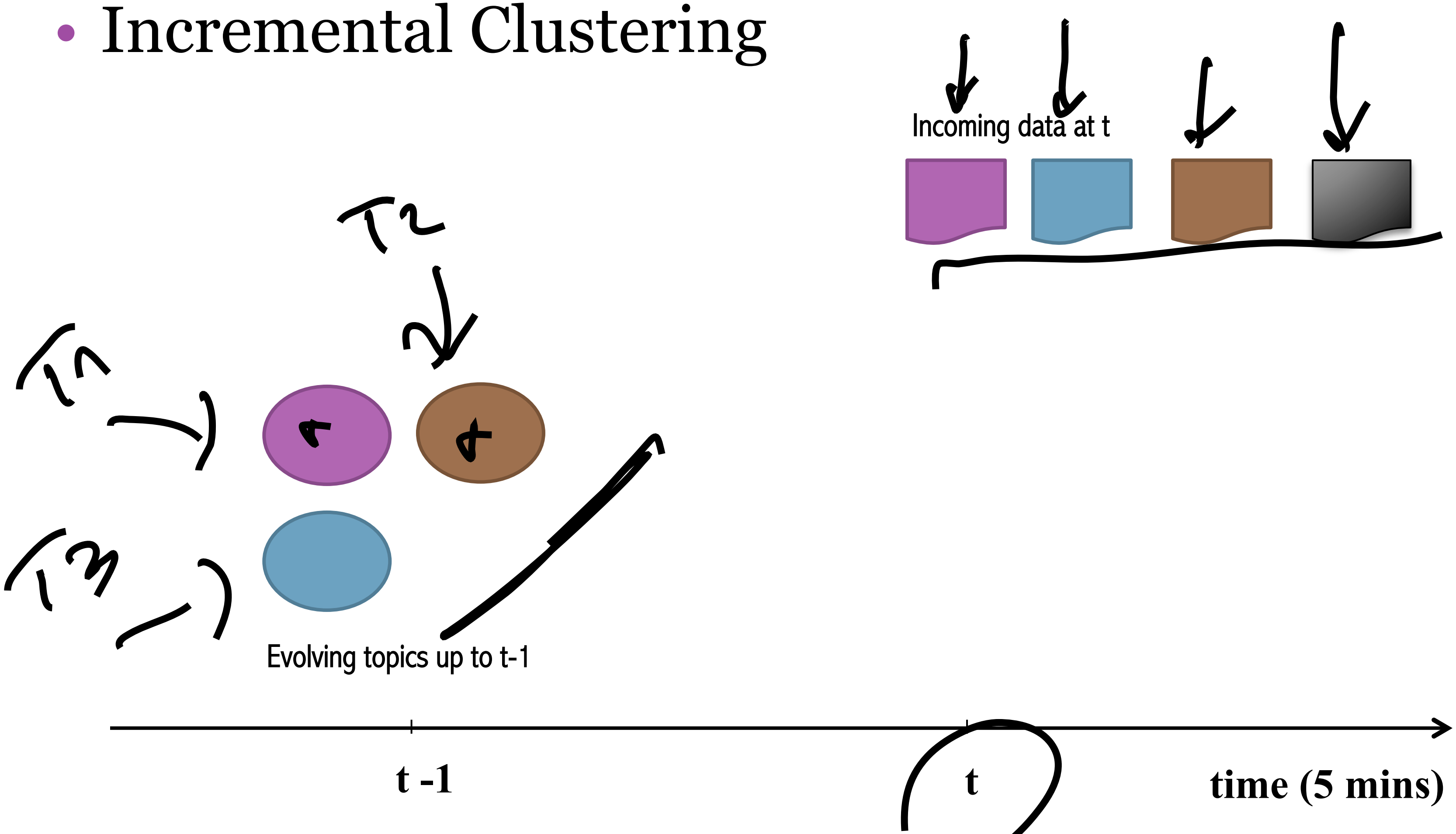
Topic Tracking

- Smooth evolution of topics through time



Topic Tracking

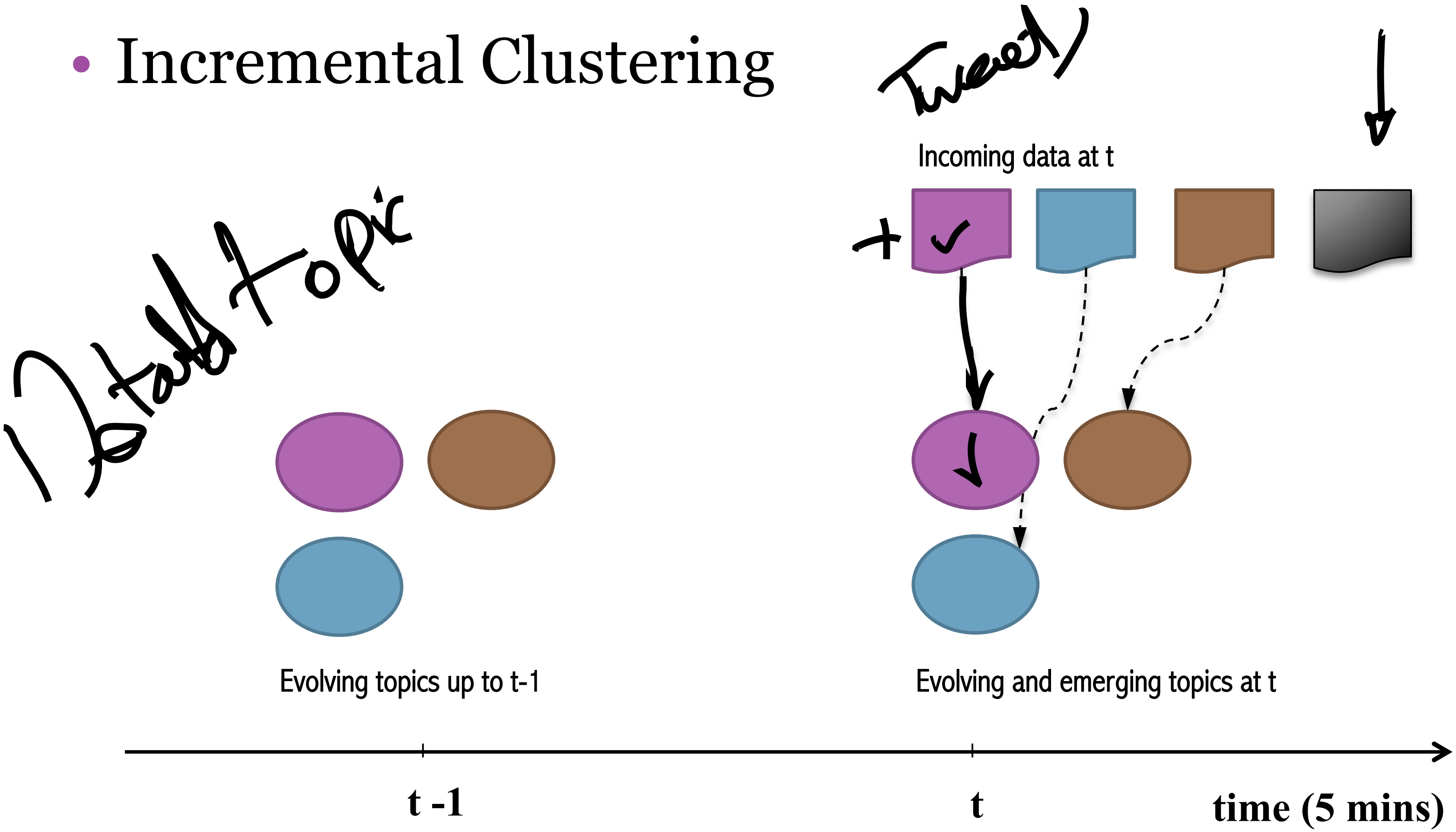
- Incremental Clustering



Evolving topic: a previously identified topic.
Emerging topic: new topics

Topic Tracking

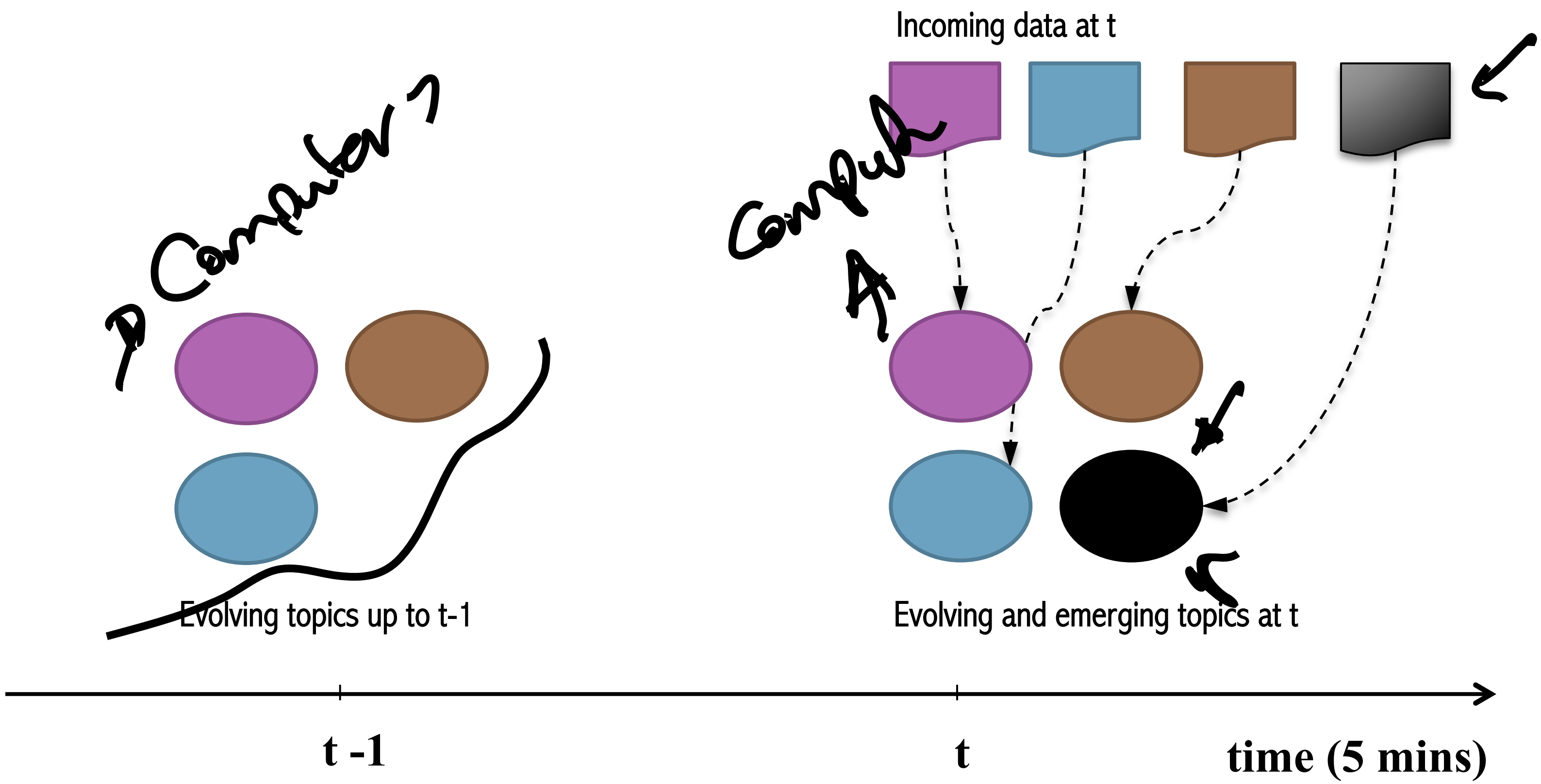
- Incremental Clustering



Evolving topic: a previously identified topic.
Emerging topic: new topics

Topic Tracking

- Incremental Clustering



Evolving topic: a previously identified topic.
Emerging topic: new topics

Topic Tracking

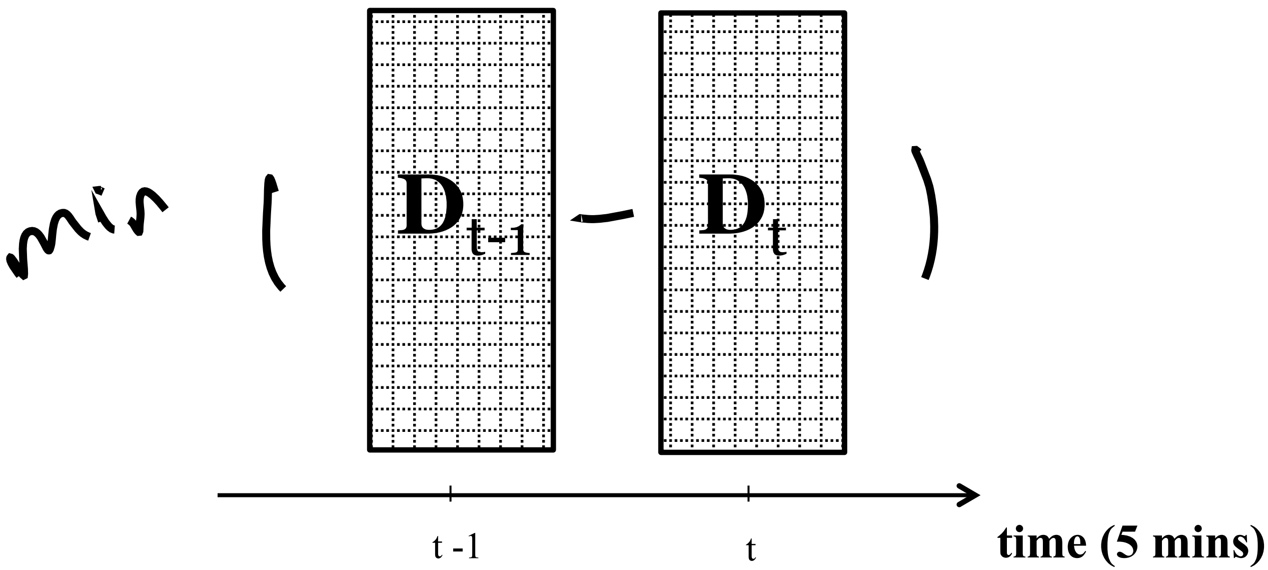
- Incremental Clustering for Topic Discovery

  - Compute similarity btw each incoming tweet and each cluster center.
  - If the maximum similarity value is greater than τ , assign the tweet to the cluster and update cluster center.
  - Otherwise, generate a new cluster and cluster center.
  - *faster* approach: Minhash or LSH

Input: tweet sets D, topic cluster set C, cluster cente
set $Center$, and threshold τ.
Output: update topic clusters C, and update cluster centers $Center$.
Process:
if $C = \emptyset$ **then**
    random select N tweets from D and add into C and $Center$.
end if
initialize max, tmp_C, tmp_{center}.
for $d_i \in D$ **do**
    for $center_j \in Center$ **do**
        compute Cosine Similarity sim between $center_j$ and d_i.
        if $sim > max$ **then**
            $max = sim$, $tmp_C = C_j$, $tmp_{center} = center_j$.
        end if
    end for
    if $max > \tau$ **then**
        distribute d_i to cluster tmp_C, and update tmp_{center}.
    else
        new cluster and centroid and add to C and $Center$.
    end if
end for
return C and $Center$.

Topic Tracking

- **Key Idea**: Temporal Coherence, smooth evolution

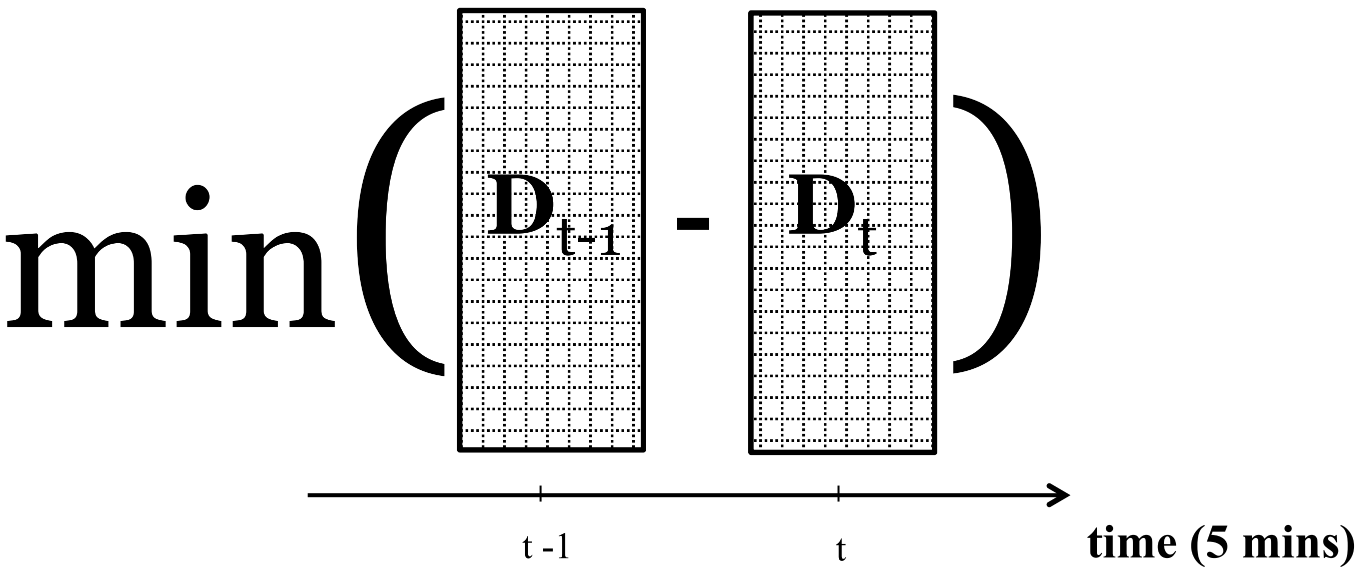


D at *t* to be a smooth evolution of **D** at *t-1*

No dramatic change in distribution over words for the same **evolving** topic in consecutive time stamps.
The nature of the topic remains the same.

Topic Tracking

- **Key Idea**: Temporal Coherence, smooth evolution



D at *t* to be a smooth evolution of **D** at *t-1*

No dramatic change in distribution over words for the same **evolving** topic in consecutive time stamps.
The nature of the topic remains the same.

Topic Tracking

$$\mathcal{L}(\mathbf{D}) = \| \mathbf{S} - \mathbf{D}\mathbf{X} \|_F^2 + \lambda \| \mathbf{X} \|_1 + \mu \| \mathbf{D} - \mathbf{D}^{t-1} \|_F^2$$

$$\mathcal{H}[\mathcal{L}(\mathbf{D})] = \mathbf{X}\mathbf{X}^T + 2\mu \mathbf{I}_k \qquad \mathbf{D}_{i+1} = P\left[\mathbf{D}_i - \alpha_i \nabla_{\mathbf{D}} \mathcal{L}(\mathbf{D})_{[\mathbf{D}_i, \mathbf{X}]}\right]$$

Algorithm 5.2. Computing $\mathbf{D}^t$ and $\mathbf{X}^t$ at time t, see TL in Figure 4

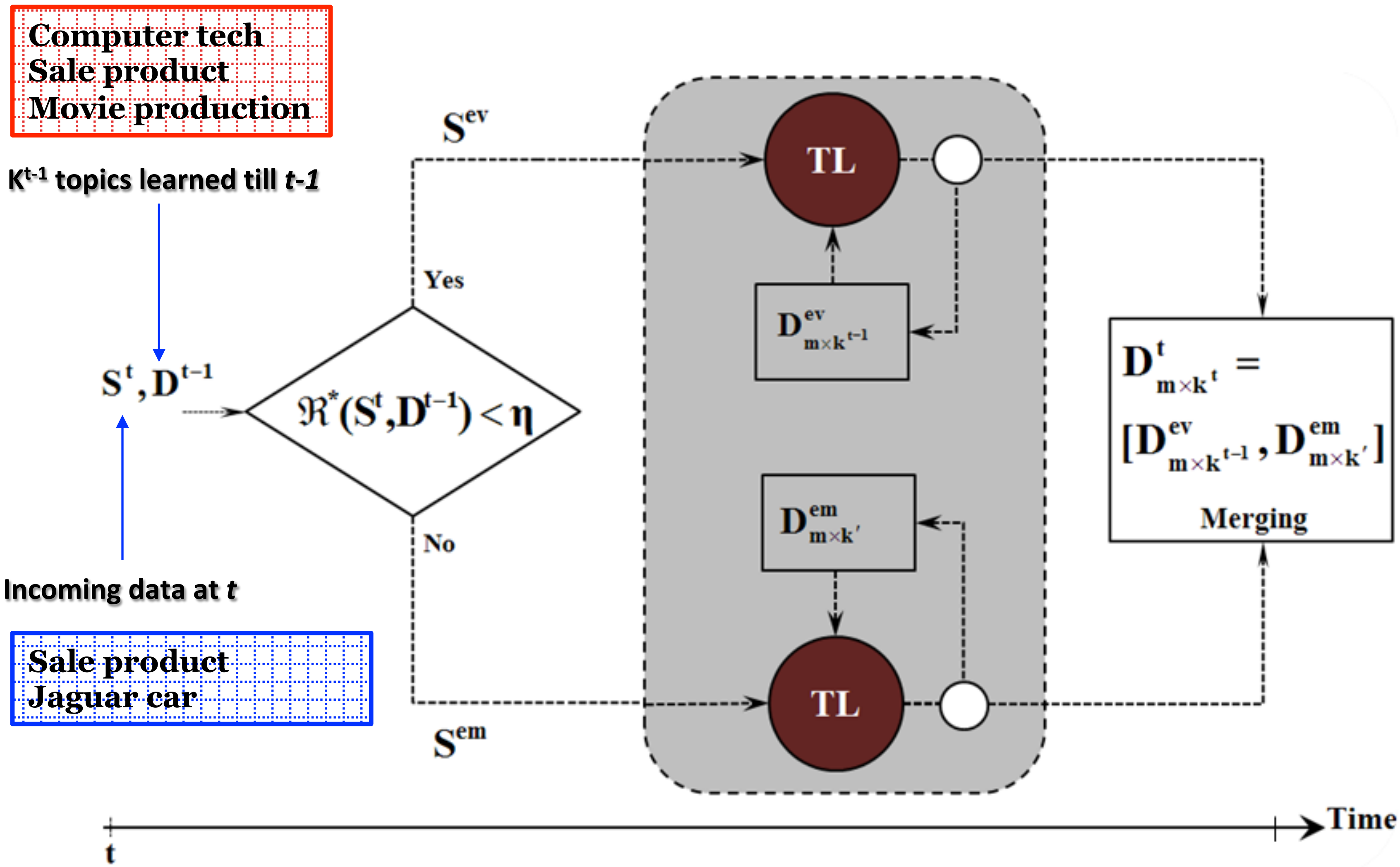
Input: $\mathbf{S}^t$, $\mathbf{D}^{t-1}$, itr: number of iterations

Output: $\mathbf{D}^t$, $\mathbf{X}^t$

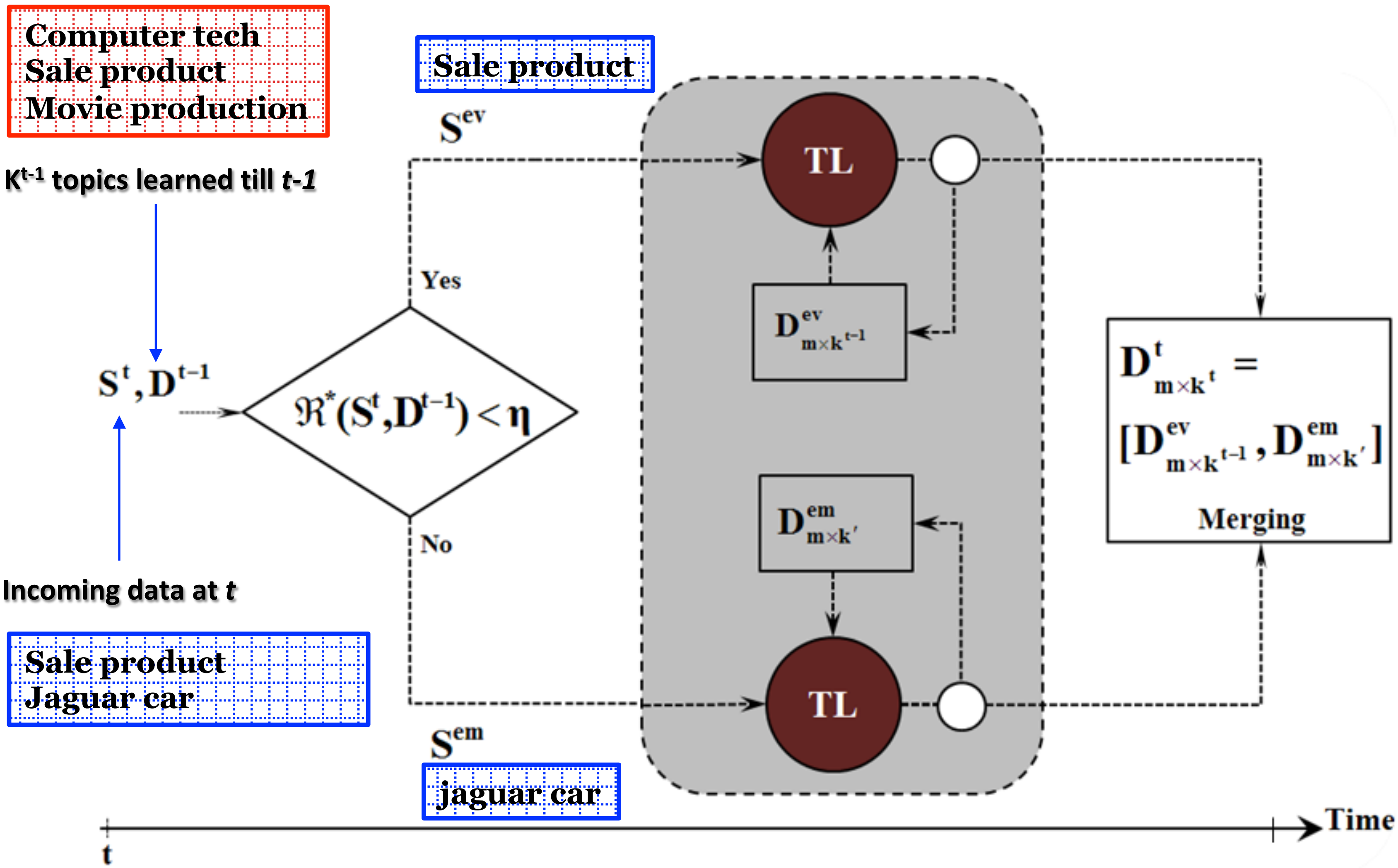
1. Compute $\mathbf{X}^t$ using $\mathbf{S}^t$ and $\mathbf{D}^{t-1}$
2. $\mathbf{D}_0^t = \mathbf{D}^{t-1}$
3. for i=1 : itr do
4. compute $\nabla_{\mathbf{D}} \mathcal{L}(\mathbf{D}_{\mathbf{i-1}}^{\mathbf{t}})$
5. $\mathbf{U} = \nabla_{\mathbf{D}} \mathcal{L}(\mathbf{D}_{\mathbf{i-1}}^{\mathbf{t}}) diag^{-1}\left(\mathcal{H}[\mathcal{L}(\mathbf{D})]_{[\mathbf{X}^{\mathbf{t}}]}\right) + \mathbf{D}_{\mathbf{i-1}}^{\mathbf{t}}$
6. $\mathbf{D}_{\mathbf{i}}^{\mathbf{t}} = max(\mathbf{0}, \mathbf{U})$
7. end for

[1] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro: *Online Learning for Matrix Factorization and Sparse Coding*. Journal of Machine Learning Research 11: 19-60 (2010)

Temporal Coherence

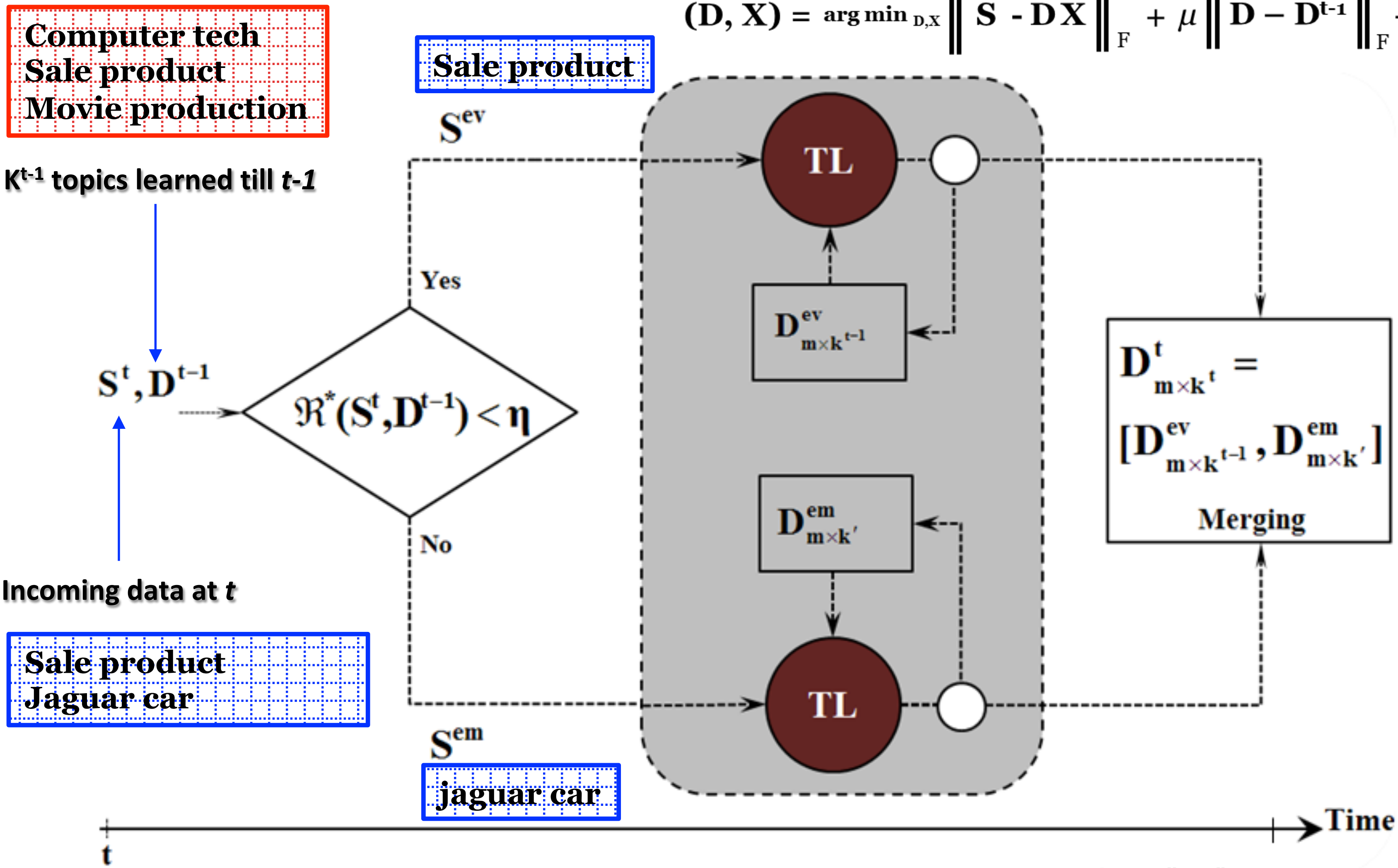


Temporal Coherence



Temporal Coherence

$$(D, X) = \arg\min_{D,X} \left\| S - DX \right\|_F^2 + \mu \left\| D - D^{t-1} \right\|_F^2 + \lambda \left\| X \right\|_1$$



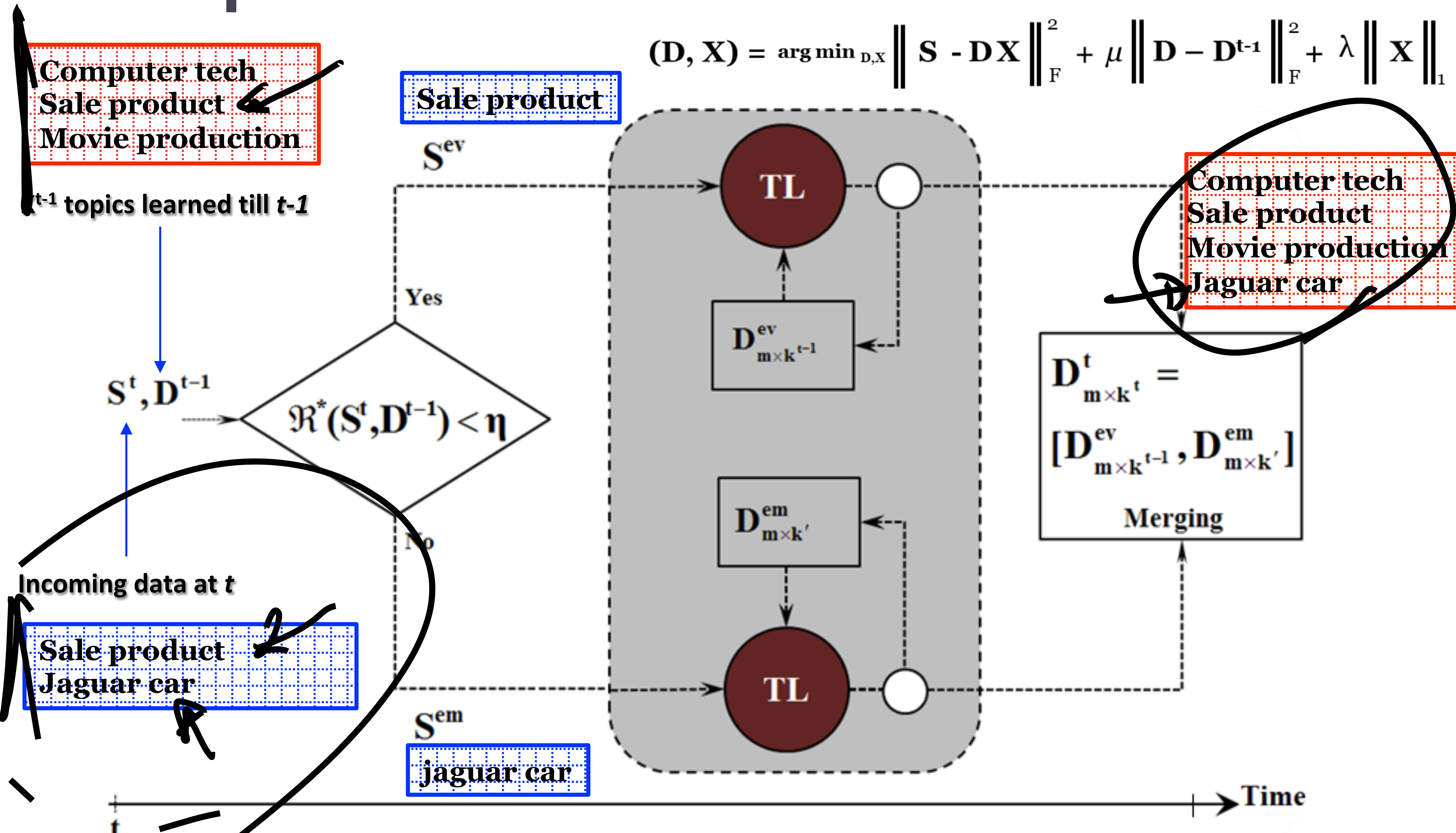
$$(D, X) = \arg\min_{D,X} \left\| S - DX \right\|_F^2 + \lambda \left\| X \right\|_1$$

Emerging Topics

Temporal Coherence

Evolving Topics

$$(D, X) = \arg\min_{D,X} \left\| S - DX \right\|_F^2 + \mu \left\| D - D^{t-1} \right\|_F^2 + \lambda \left\| X \right\|_1$$



$$(D, X) = \arg\min_{D,X} \left\| S - DX \right\|_F^2 + \lambda \left\| X \right\|_1$$

Emerging Topics

Topic Tracking- Cnt.

- Temporal Coherence constraint for topic learning:
  - $\mathbf{D}^{ev}$ to be a smooth evolution of $\mathbf{D}^{t\text{-}1}$

$$(\mathbf{D}, \mathbf{X}) = \arg\min_{\mathbf{D},\mathbf{X}} \left\| \mathbf{S} - \mathbf{D}\mathbf{X} \right\|_F^2 + \lambda \left\| \mathbf{D} - \mathbf{D}^{\mathbf{t-1}} \right\|_F^2 + \lambda \left\| \mathbf{X} \right\|_1$$

s.t. $\mathbf{X} \geq \mathbf{0}$, $\mathbf{D} \geq \mathbf{0}$, $\|\boldsymbol{d_i}\| = \mathbf{1}$ *for* $i = \{1,\ldots k\}$

- Can be solved efficiently
  - Space: O(n*m), given that m >> k
  - Running time: O(n)

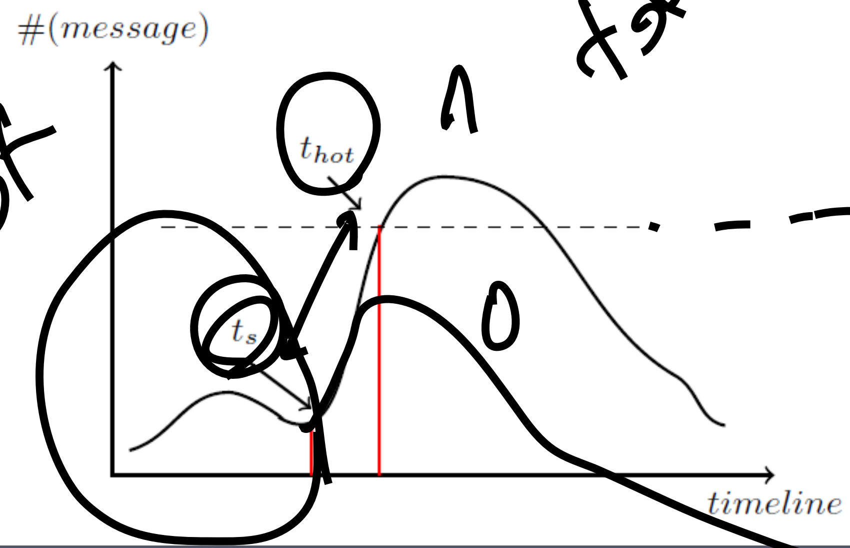
Early Detection of Emerging Topics

Chen, Yan, et al. "Emerging topic detection for organizations from microblogs." SIGIR 2013.

Early Detection of Topics

- Evolution of a hot topic
  - t_s topic detection time
  - t_{hot} the time by which topic becomes major.
    - tweets number exceeds a threshold.
- We aim to predict if an already-detected topic will be major in the near future!

Hot a topic



Early Detection

View 1: rate indicators

- Rate of increase in #users
- Rate of increase in #tweets
- Rate of increase in #re-tweets

View 2: overlap indicators

- Overlap btw users posted about topic and influential users
- Overlap btw topic keywords and top influential keywords

Co-training (Co): Two SVM classifiers trained on the above two orthogonal views of features

Ensemble Learner (En): Ensemble of three classifiers (Decision Tree, SVM, and Naive Bayesian) vote for each unlabeled topic.

Early Detection

- User authority / user influence against the topic
- Tweet authority / derived from topical user auth.

- f_1 is the rate of increase of user number,

$$f_1 = \frac{|U^t|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |U^x|}. \qquad (6)$$

- f_2 is the rate of increase of tweets number,

$$f_2 = \frac{|Tw^t|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |Tw^x|}. \qquad (7)$$

- f_3 is the rate of increase of re-tweets number,

$$f_3 = \frac{|Rtw^t|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |Rtw^x|}. \qquad (8)$$

- f_4 is the overlap between org keyusers and top N influential topic users,

$$f_4 = \frac{\#(ku_{tp} \cap ku)}{\#ku_{tp}}. \qquad (9)$$

- f_5 is the overlap between org keywords and top N influential topic keywords, and

$$f_5 = \frac{\#(kw_{tp} \cap kw)}{\#kw_{tp}}. \qquad (10)$$

- f_6 represents the rate of increase of influence of the accumulated weight of tweets,

$$f_6 = \frac{|A^t|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |A^x|}, \qquad (11)$$

where $A = \frac{\sum_{tw \in Tw_{tp}} auth_{tp}(tw)}{|Tw_{tp}|}$.

Evaluation

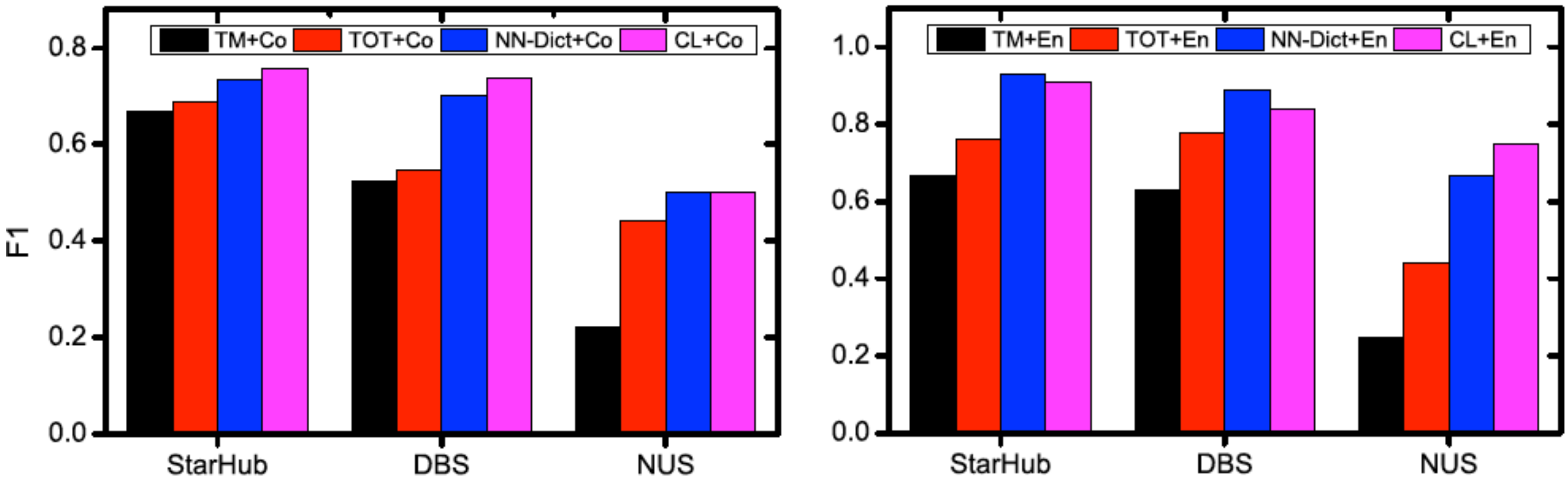


Figure 5: Performance of emerging topic Detection when $T_L = t_{hot}$

CL: Incremental clustering
Co: Co-training
En: Ensemble Learner

Evaluation

Table 2: Performance of emerging topic detection when $T_L = t_{hot}$

Methods	Organization	recall	precision	F_1
CL+En	*StarHub*	**0.93**	**0.87**	**0.90**
CL+TSVM		0.86	0.75	0.80
CL+Semi-NB		0.86	0.71	0.77
CL+En	*DBS*	**0.89**	**0.80**	**0.84**
CL+TSVM		**0.89**	0.73	0.80
CL+Semi-NB		**0.89**	0.67	0.70
CL+En	*NUS*	**1.00**	**0.60**	**0.75**
CL+TSVM		**1.00**	0.50	0.67
CL+Semi-NB		**1.00**	0.42	0.73

CL: Incremental clustering
Co: Co-training
En: Ensemble Learner

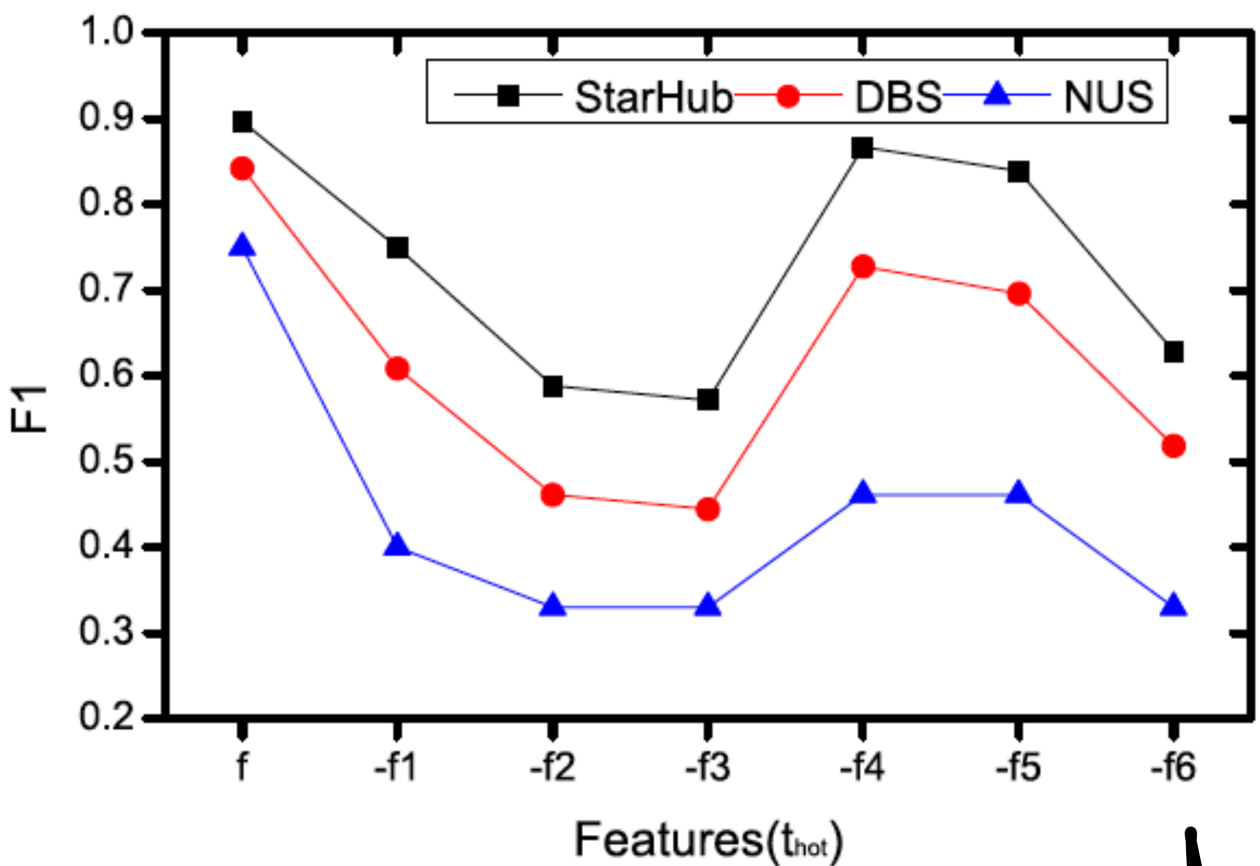
Evaluation

Table 3: Performance of emerging topic detection when $T_L = t_{mid}$

Methods	Organization	recall	precision	F_1
CL+En	*StarHub*	**0.71**	**0.83**	**0.77**
CL+TSVM		**0.71**	0.71	0.71
CL+Semi-NB		**0.71**	0.67	0.69
CL+En	*DBS*	**0.78**	**0.78**	**0.78**
CL+TSVM		**0.78**	0.70	0.74
CL+Semi-NB		**0.78**	0.64	0.70
CL+En	*NUS*	**0.67**	**0.50**	**0.57**
CL+TSVM		**0.67**	0.40	0.50
CL+Semi-NB		**0.67**	0.40	0.50

CL: Incremental clustering
Co: Co-training
En: Ensemble Learner

Evaluation



f2: rate of increase in #tweets
f3: rate of increase in #re-tweets,
f6: overall accumulated influence of tweets

Summary

- Effective NMF model with temporal coherence constraint
  - Improves topic tracking in streaming data.
- Effective framework for early prediction of emerging topics.
  - Rate and overlap features

Reading

- Emerging topic detection for organizations from microblogs. Chen, Y., et al. SIGIR'13.