TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, Jiawei Han

University of Illinois at Urbana-Champaign

U.S. Army Research Laboratory

czhang82@illinois.edu

The Emergence of Geo-Tagged Social Media

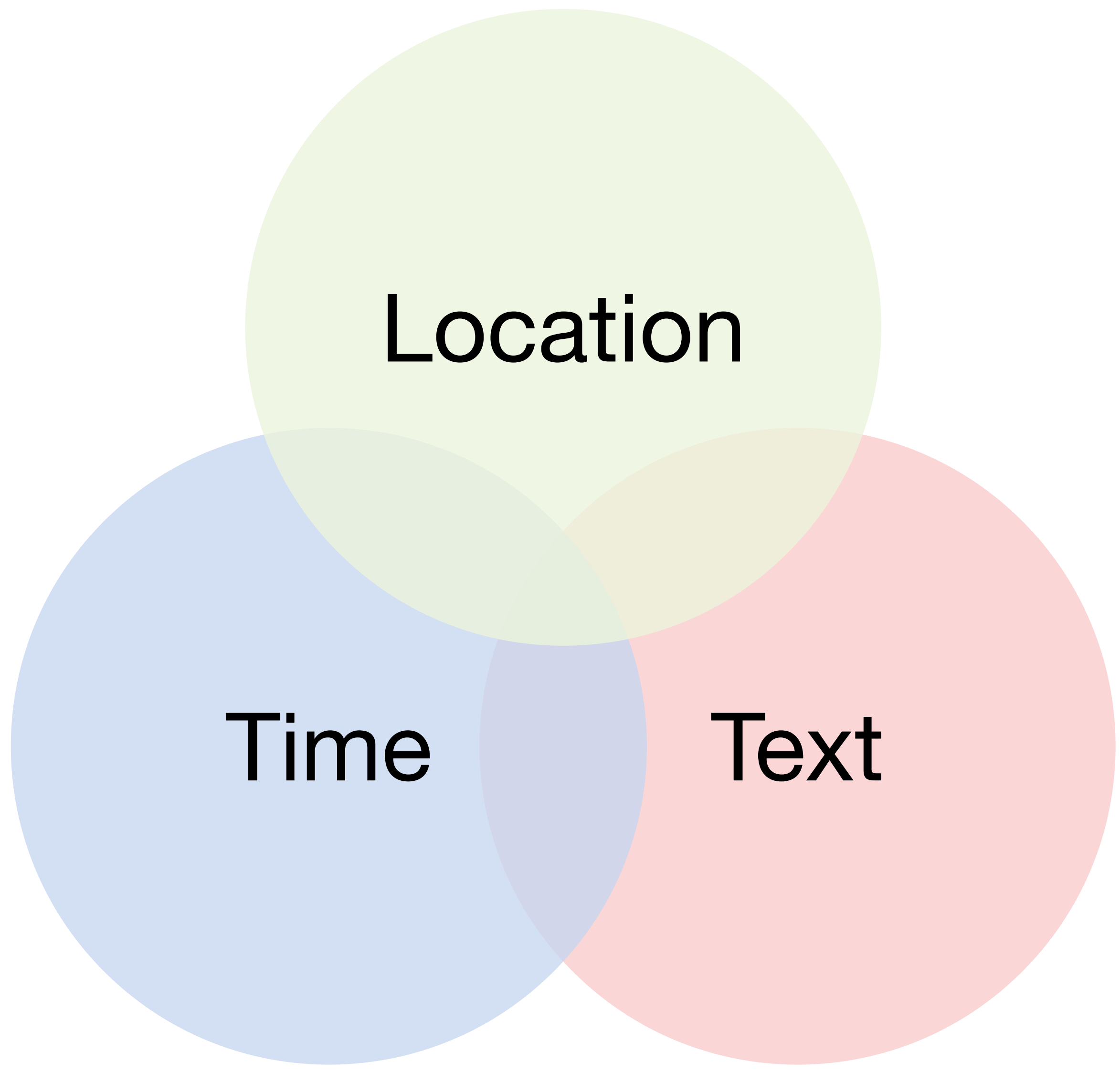
Everyone and everything is getting increasingly connected.

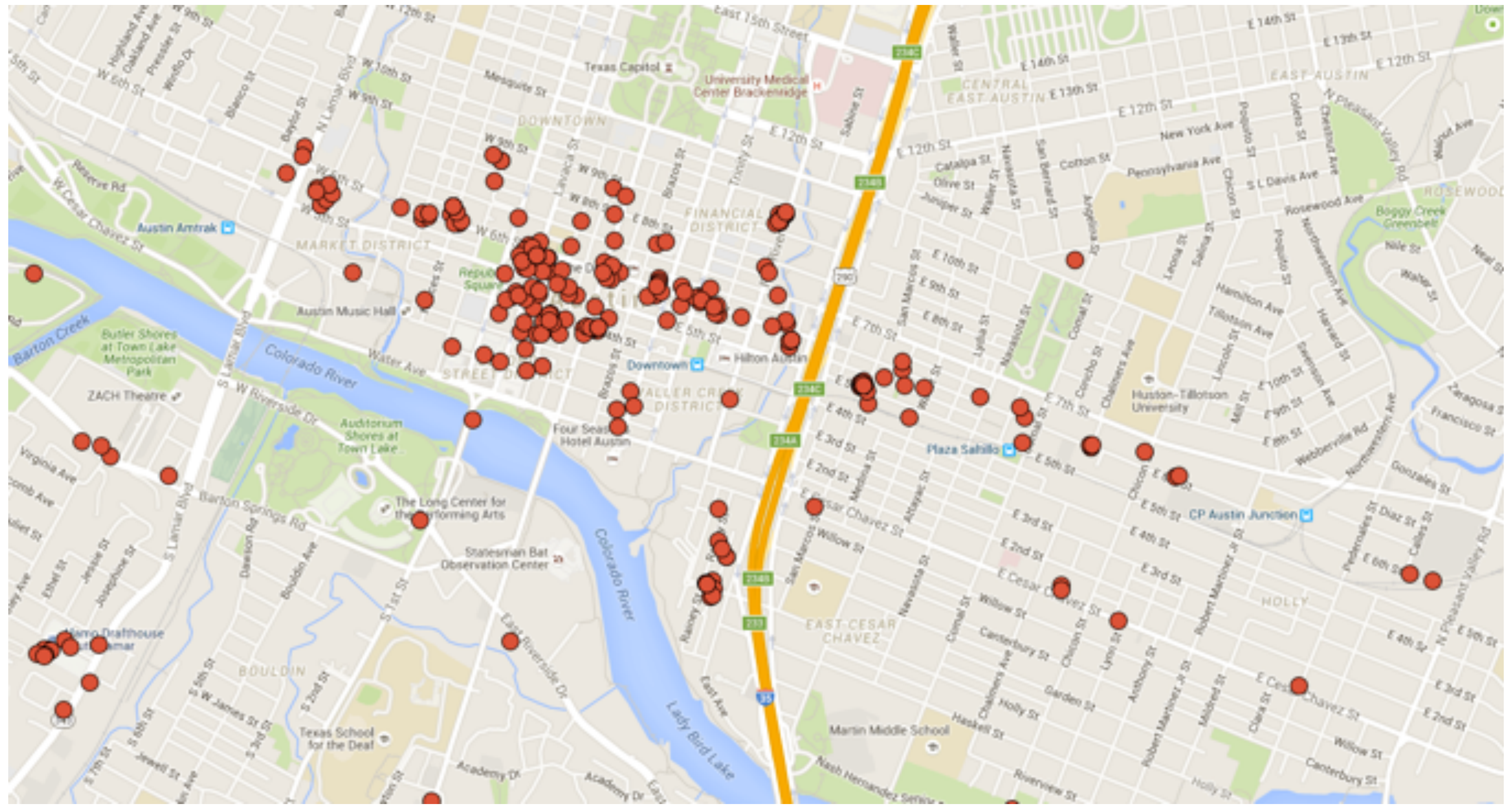
Meal

Travel

Concert

Geo-Tagged Social Media: A Result of Human Sensing





Local Event Detection from Social Media Streams

A local event is an unusual activity bursted in a local area and a specific duration while engaging a considerable number of participants

Examples:

Concert

Football Game

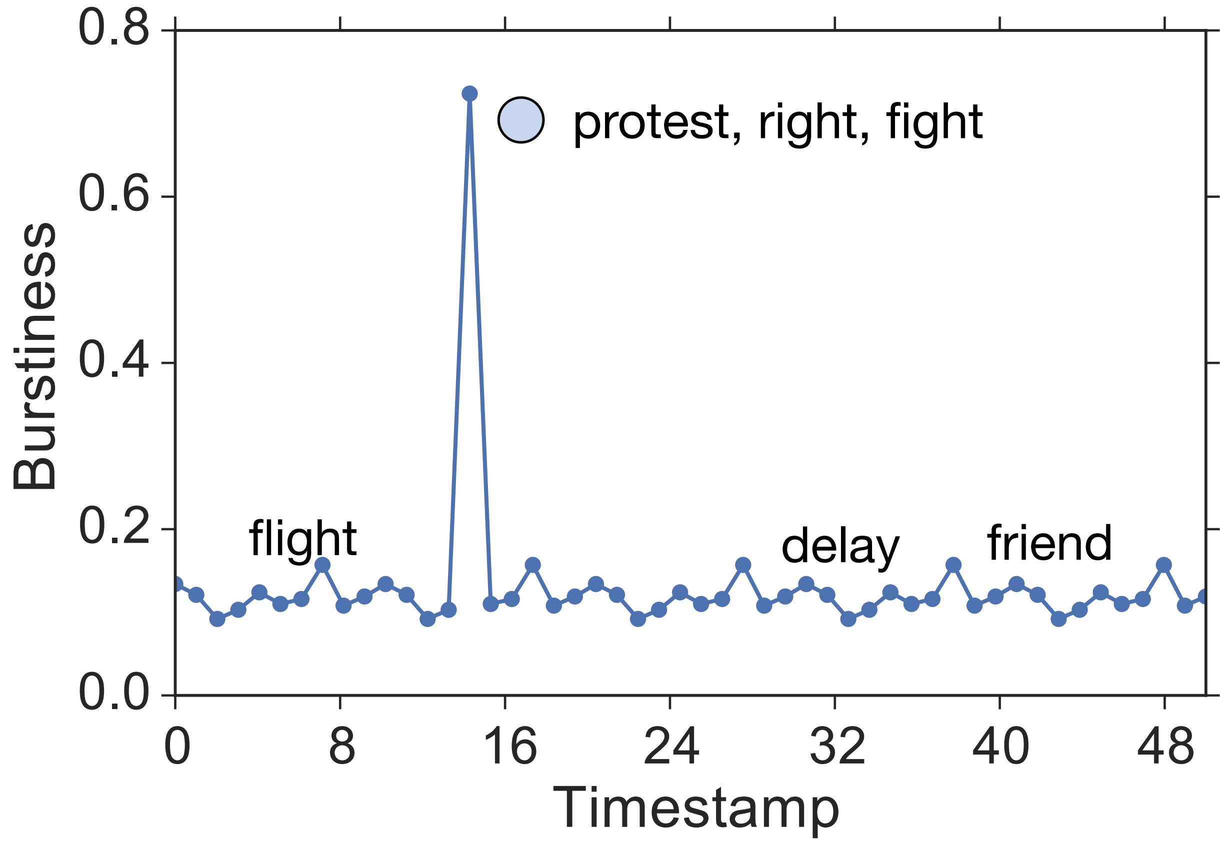
Protest

Problem: Online Local Event Detection from Social Media

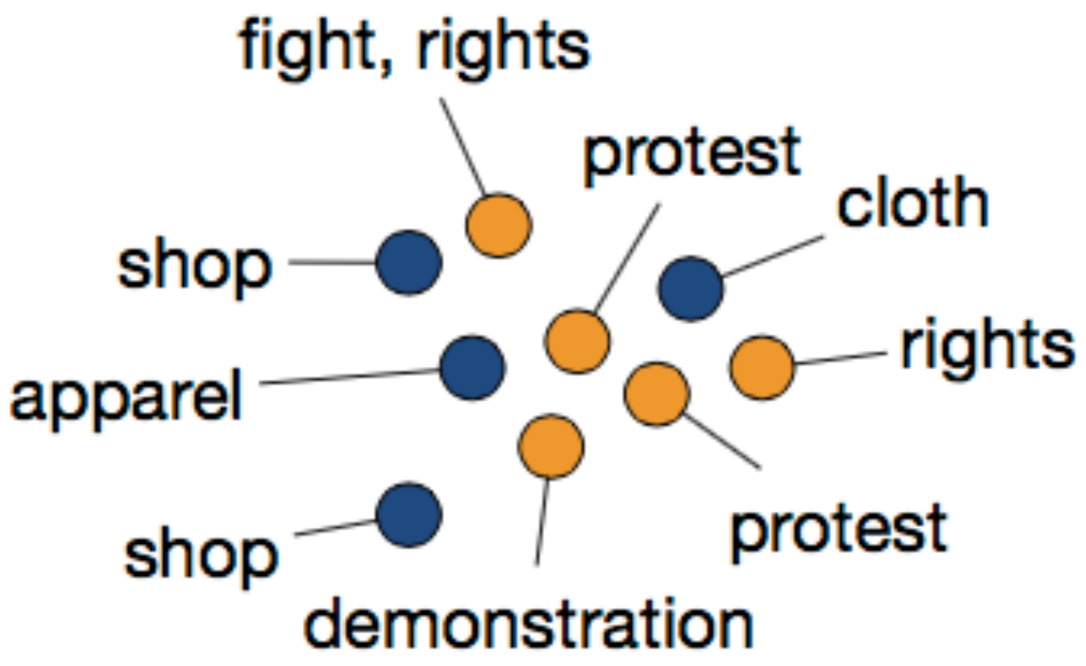
- Input:
  - A continuous geo-tagged social media stream D
  - A query time window Q
- Task:
  - Find all the local events falling in Q
  - Update the event list as Q shifts continuously

Previous Methods

- Feature-based methods
  - Find bursty keywords (e.g., Wavelet transform)
  - Cluster bursty keywords into candidate events
  - Select top-K bursty candidates



- Document-based methods
  - Cluster tweets into geo-topic clusters
  - Measure the bustiness of different clusters
  - Identify top-K bursty clusters



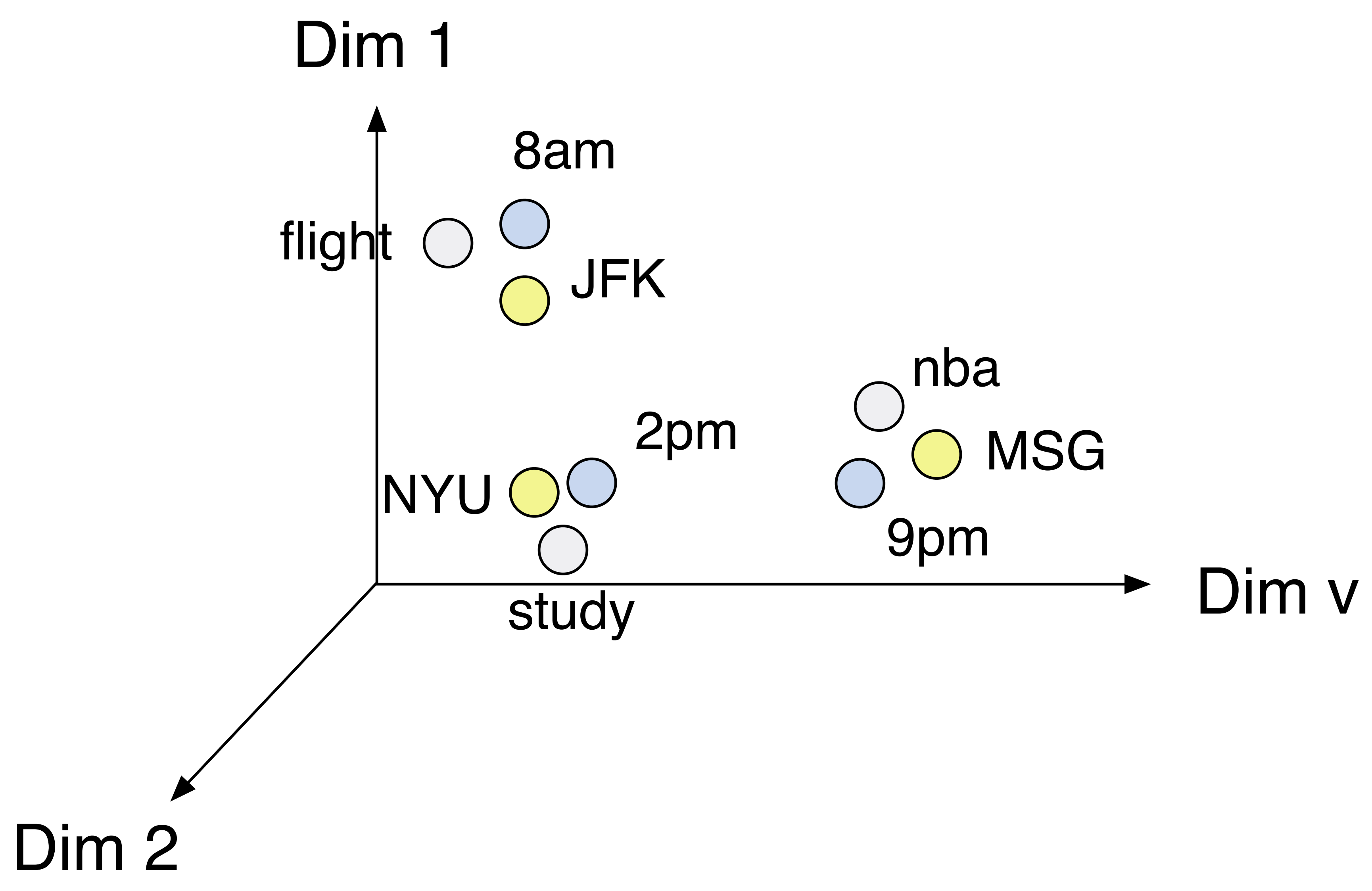
[1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. PVLDB, 6(12):1326–1329, 2013.
[2] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In SIGIR, pages 513–522, 2016.

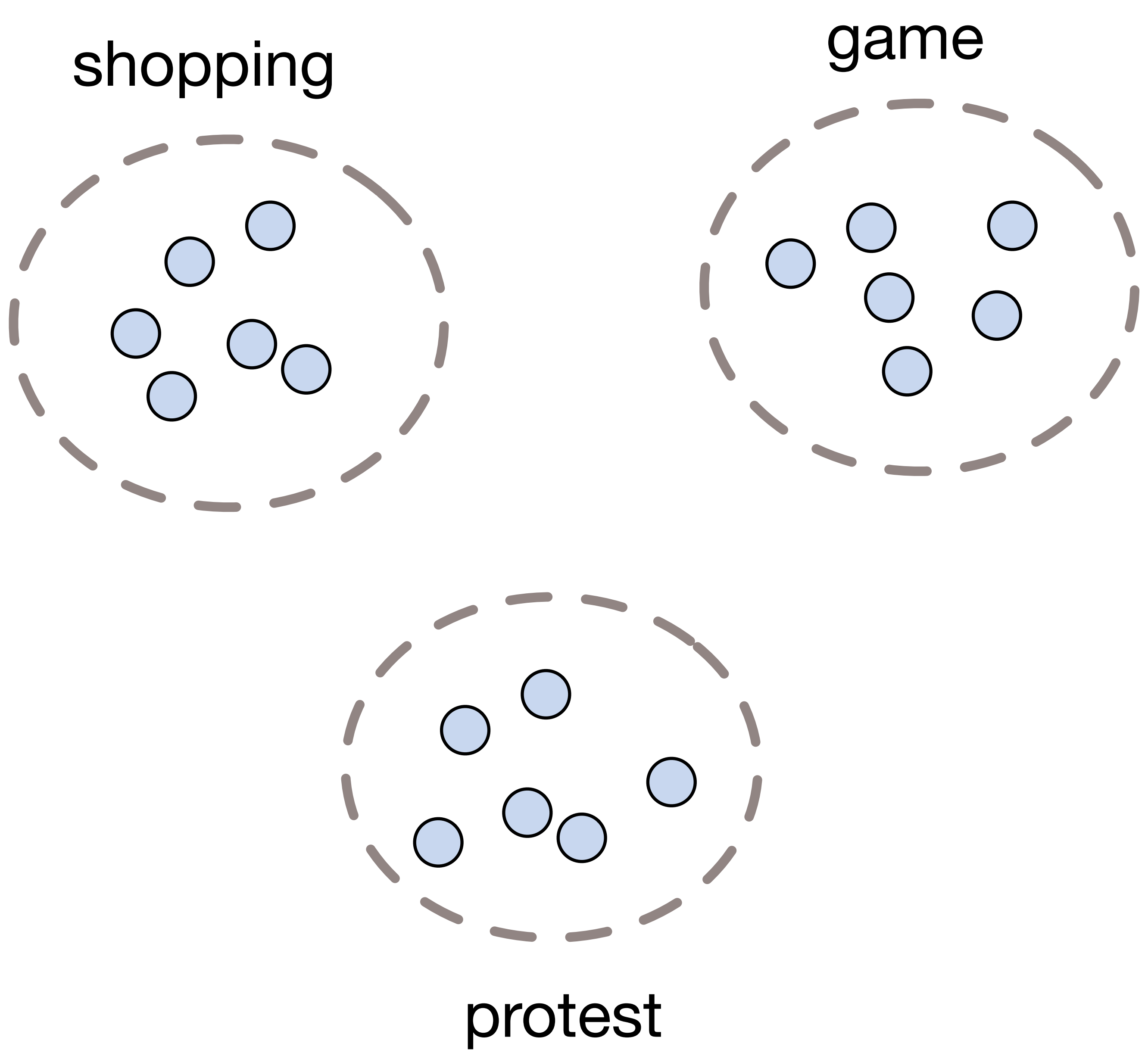
Two Drawbacks of Existing Methods

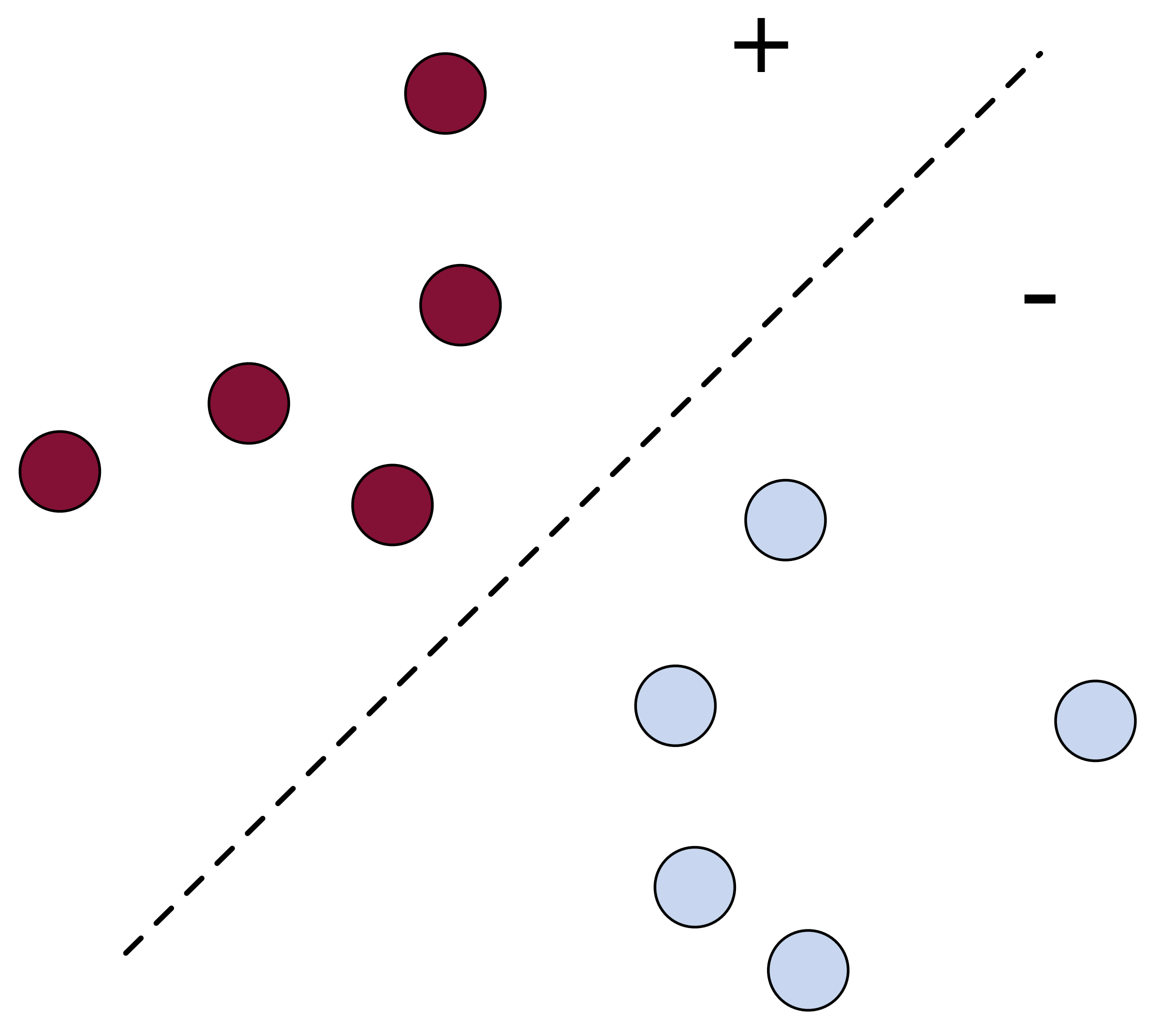
- Short-text semantics should be captured in a better way
  - Previous studies use bag-of-words for computing text similarity
  - They suffer from short text sparsity and do not capture word correlation well, e.g., “kobe” and “kb”
- Rigid top-K selection hurts the performance of local event detection
  - It relies on manually designed ranking functions for measuring bustiness
  - It is not flexible enough: some query windows can have more than K events, while others many have no events at all.

TrioVecEvent: Embedding-Based Local Event Detection

1. Generate multimodal embedding of location, time, text
2. Use embeddings to find geo-topic clusters as candidate events
3. Extract features for the candidates and train a classifier to identify true events







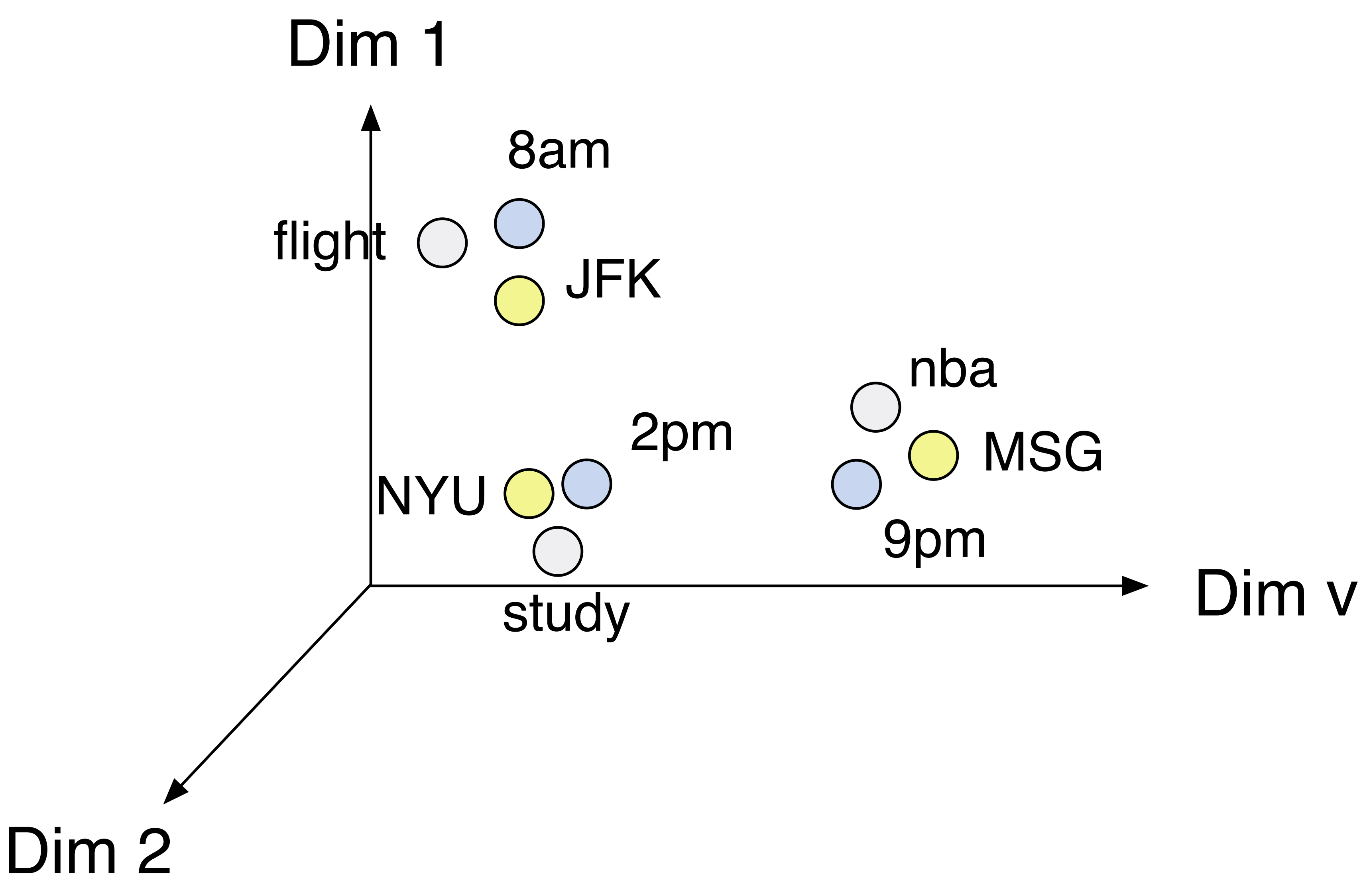
Multimodal Embedding: Extending the CBOW Model

Goal: embed all the location, time, and keywords into the same latent space

Objective function: for each attribute, predict it based on the rest attributes in the record:

$$p(i|d_{-i}) = \exp(s(i, d_{-i})) / \sum_{j \in X} \exp(s(j, d_{-i}))$$

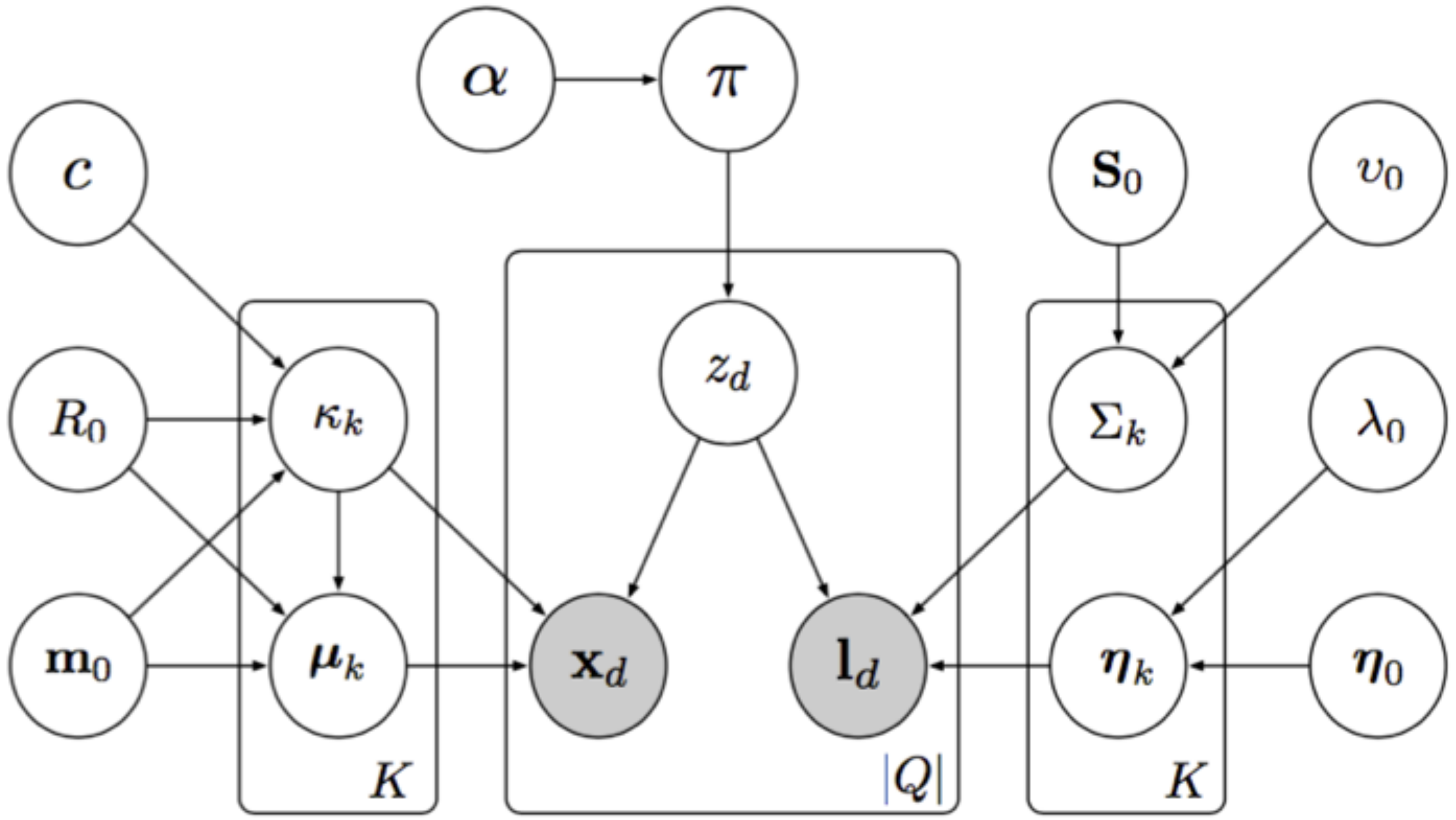
$$s(i, d_{-i}) = \mathbf{v}_i^{\mathrm{T}} \sum_{j \in d_{-i}} \mathbf{v}_j / |d_{-i}|.$$



[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119, 2013.

Candidate Generation: A Bayesian Mixture Model

- Each tweet *d* is described by: (1) a two-dimensional location **l**; and (2) a text embedding **x**
- Goal: find all the geo-topic clusters in the query window
    - The tweets in the same cluster are geographically close and semantically similar



$\pi \sim \textbf{Dirichlet}(.|\alpha)$

$\{\boldsymbol{\eta}_k, \Sigma_k\} \sim \textbf{NIW}(.|\boldsymbol{\eta}_0, \lambda_0, \mathbf{S}_0, v_0) \quad k = 1, 2, \ldots, K$

$\{\boldsymbol{\mu}_k, \kappa_k\} \sim \Phi(.|\mathbf{m}_0, R_0, c) \quad k = 1, 2, \ldots, K$

$z_d \sim \textbf{Categorical}(.|\pi) \quad d \in Q$

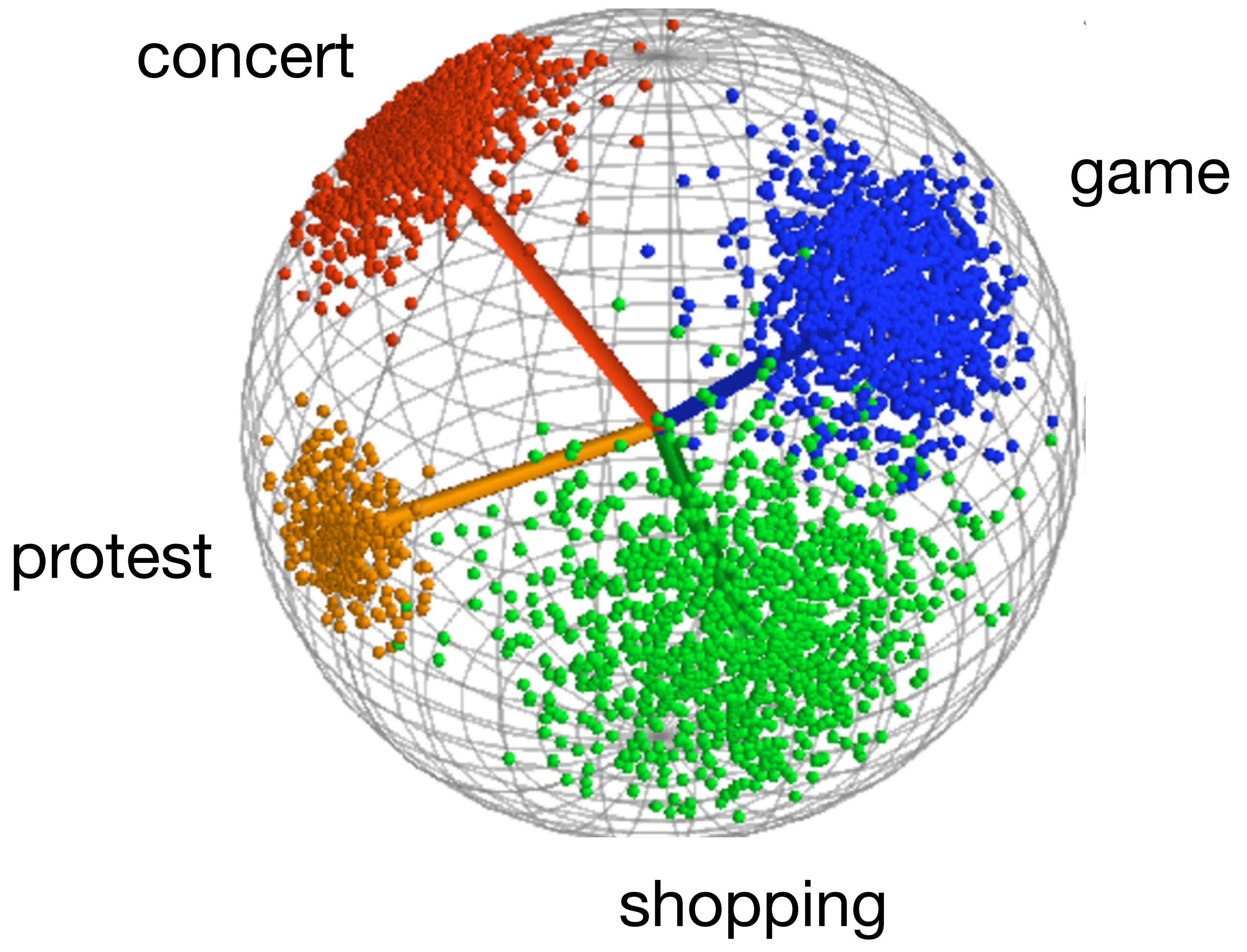
$\boxed{\mathbf{l}_d \sim \mathcal{N}(.|\boldsymbol{\eta}_{z_d}, \Sigma_{z_d}) \quad d \in Q}$ location: Gaussian

$\boxed{\mathbf{x}_d \sim \textbf{vMF}(.|\boldsymbol{\mu}_{z_d}, \kappa_{z_d}) \quad d \in Q}$

text embedding: Von-Mises Fisher distribution

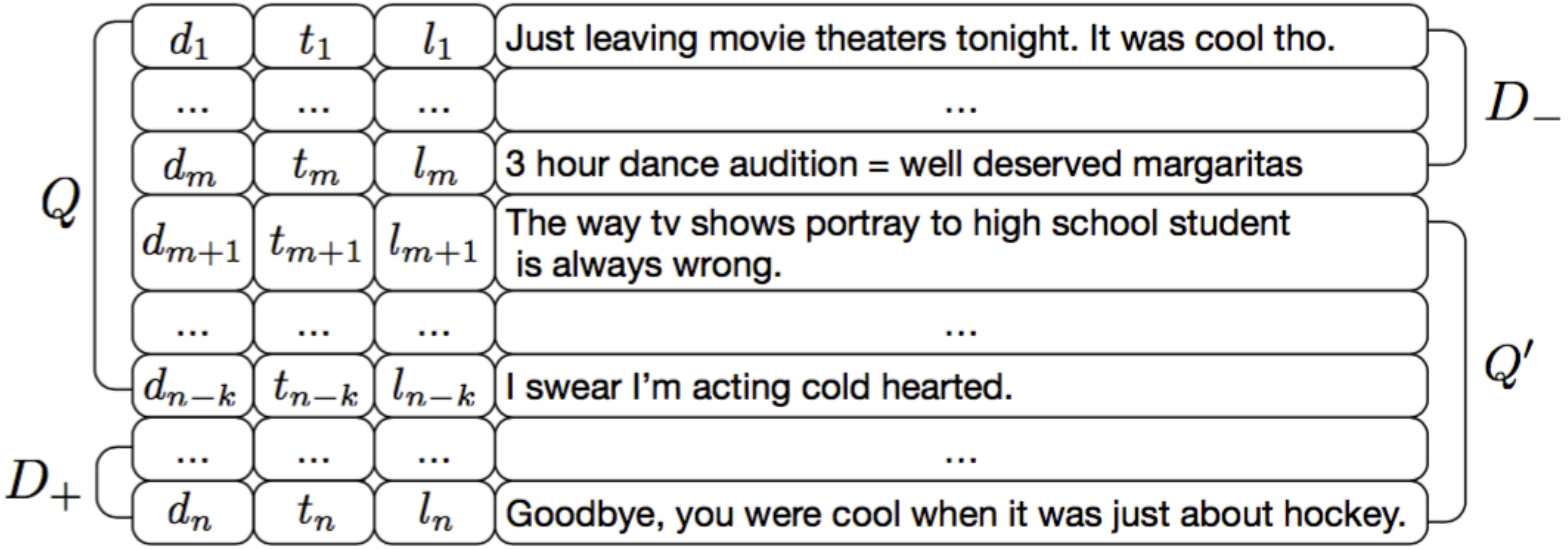
Why Using vMF to Model Text Embedding

- vMF has two major parameters:
  - (1) the mean direction of the cluster on the unit sphere;
  - (2) the concentration of the points in the cluster
- The cosine similarity naturally leads to the choice of using vMF for finding clusters
  - The mean direction acts as a semantic focus on the sphere
  - It produces relevant embeddings around it



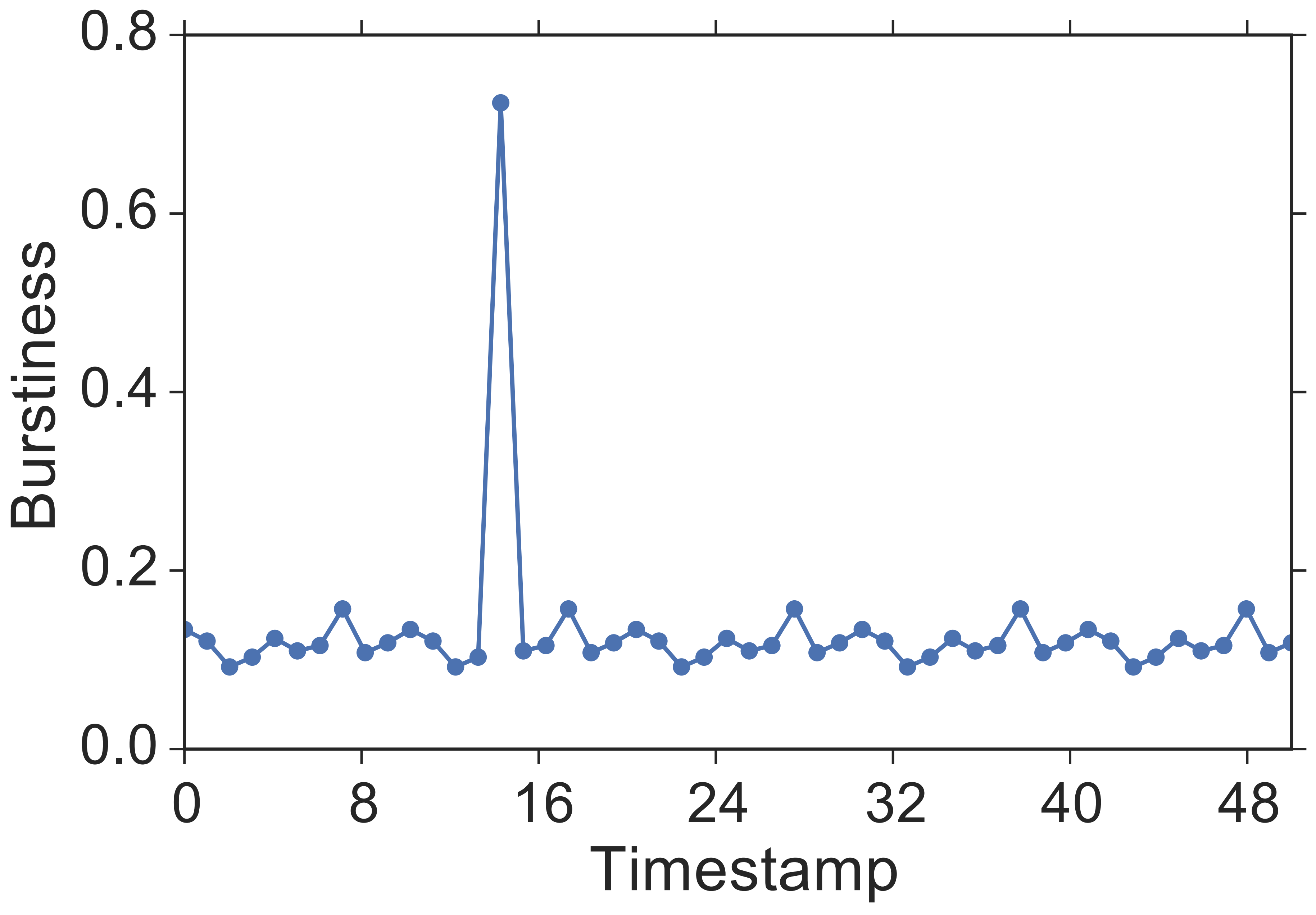
Online Updating

- When the query window shifts from Q to Q'
  - Remove the outdated tweets
  - Sample the cluster memberships for the newly arrived tweets
  - Periodically re-compute the clustering results from scratch to avoid error accumulation



Candidate Event Filtering

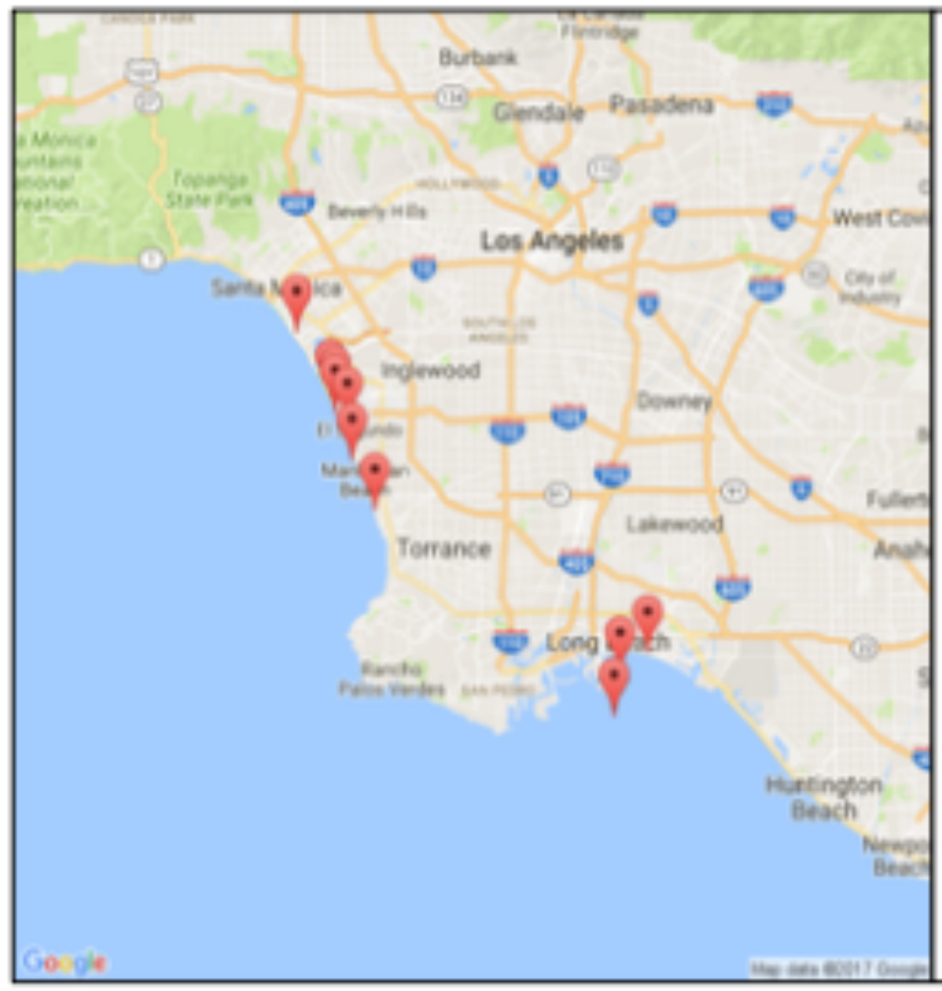
- Goal: determine whether each geo-topic cluster is a local event
- Cast the candidate filtering as a binary classification problem
  - Label the candidates in 100 windows, and use logistic regression to select true local events
- Features
  - Spatial and temporal unusualness
  - Spatial and temporal concentration
  - Semantic concentration
  - Bustiness



Experimental Settings

- Data
  - Geo-tagged tweets in Los Angeles during 2014.08 — 2014.11
  - Geo-tagged tweets in New York City during 2014.08 — 2014.11
  - Preprocessing: extract entities and noun phrases from raw tweet text
- Compared methods
  - 1. EvenTweet (PVLDB’13); 2. GeoBurst (SIGIR’16); 3. GeoBurst+ (TIST’17)
- Evaluation
  - Randomly generate 100 query windows
  - Detect events in those query windows and evaluate the results using CrowdFlower

Example Multimodal Embeddings



lax	lakers	dodgers	beachlife
international	kobe	ladders	sand
losangeles	bryant	dogerstadium	boardwalk
united	bulls	itfdb	ocean
people	cavs	letsgododgers	wave
tsa	kevin	game	beachday
sfo	knicks	dodgergame	pacificocean
food	clipper	play	santamonica
flight	lebron	losdoyers	pier
travel	cp3	win	wave
"33.942, -118.409"	"nba"	"baseball"	"beach"

"beach"

(a) Examples on LA (the second query is the location of the LAX Airport).

jfk	knicks	mlb	rockaway
airport	melo	yankees	beachday
international	lebron	mets	howard_beach
johnfkennedy	durant	yanks	brighton
burger	basketball	inning	longbeach
terminal	kobe	yankee	coney
john	cavs	ballpark	atlantic
kennedy	theknicks	pitch	island
sfo	game	jeter	boardwalk
flight	lakers	game	long
"40.641, -73.778"	"nba"	"baseball"	"beach"

"beach"

(b) Examples on NY (the second query is the location of the JFK Airport).

Example Local Events

- Standing for **justice**! @ LAPD Headquarters http://t.co/YxNUAloQcE
- At the LAPD **protest** downtown **#EzellFord #MikeBrown** http://t.co/kWphv6dXOr
- Hands Up. Don't **Shoot**. @ Los Angeles City Hall
- Black, Brown, poor white, ALL **oppressed** people **unite**. #ftp #lapd **#ferguson** #lapd **#mikebrown #ezellford** http://t.co/szf3mJRJwV
- Finished **marching** now **gathered** back at LAPD police as organizers speak some truth **#EzellFord #MikeBrown #ferguson** http://t.co/M33n9IMOzC

(a) LA local event I: a protest rally at the LAPD Headquarter.

- Thanks for making my Teenage Dreams come true @arjanwrites!! AHHH @**KATYPERRY**!! (at @**STAPLESCenter** for **Katy Perry**) https://t.co/TVEaghr1Tt
- **Katy perry** with my favorite. http://t.co/FpfPYAQNBR
- @MahoganyLOX are you at the **Katy perry concert**?
- One of the beeeeest **concerts** in history!
- My two minutes of fame was me and my friends picture getting put on the TV screens at the **Katy Perry concert**.

(b) LA local event II: Katy Perry's concert at the Staples Center.

Example Local Events

- **Hoboken Fall Arts & Music Festival** with bae @alli_holmes93 @ Washington St. Hoboken
- On Washington Street. (at **Hoboken Music And Arts Festival**) https://t.co/YbLSdZhLZV
- Sweeeeet. Bonavita **Guitars**, at the **Hoboken festival**. http://t.co/2Cw1Qz4UGo
- I'm at **Hoboken Music And Arts Festival** in Hoboken, NJ https://t.co/i4bSM3mrjb
- It's a **festy music** day.

(a) NY local event I: the Hoboken Music and Arts Festival in Hoboken, NJ.

- **Knicks game** w literally a person. http://t.co/hxVYidpCzs
- **Knicks game** with my main man.
- It has been one of my dream to watch **NBA game**!! Let's go! http://t.co/GRJRvFw6vd
- Watching @nyknicks at @TheGarden for for the first time! Go **Knicks**! #nyk4troops
- I was outside of **msg** today pretending I liked the **Knicks**. It's that bad.

(b) NY local event II: The Knicks' basketball game at the Madison Square Garden.

Effectiveness Comparison

P: precision
R: pseudo recall
F1: pseudo F1 score

Method	LA			NY		
	P	R	F1	P	R	F1
EvenTweet	0.132	0.212	0.163	0.108	0.196	0.139
GeoBurst	0.282	0.451	0.347	0.212	0.384	0.273
GeoBurst+	0.368	0.483	0.418	0.351	0.465	0.401
TrioVecEvent	**0.804**	**0.612**	**0.695**	**0.765**	**0.602**	**0.674**

classification is better than top-k for candidate filtering

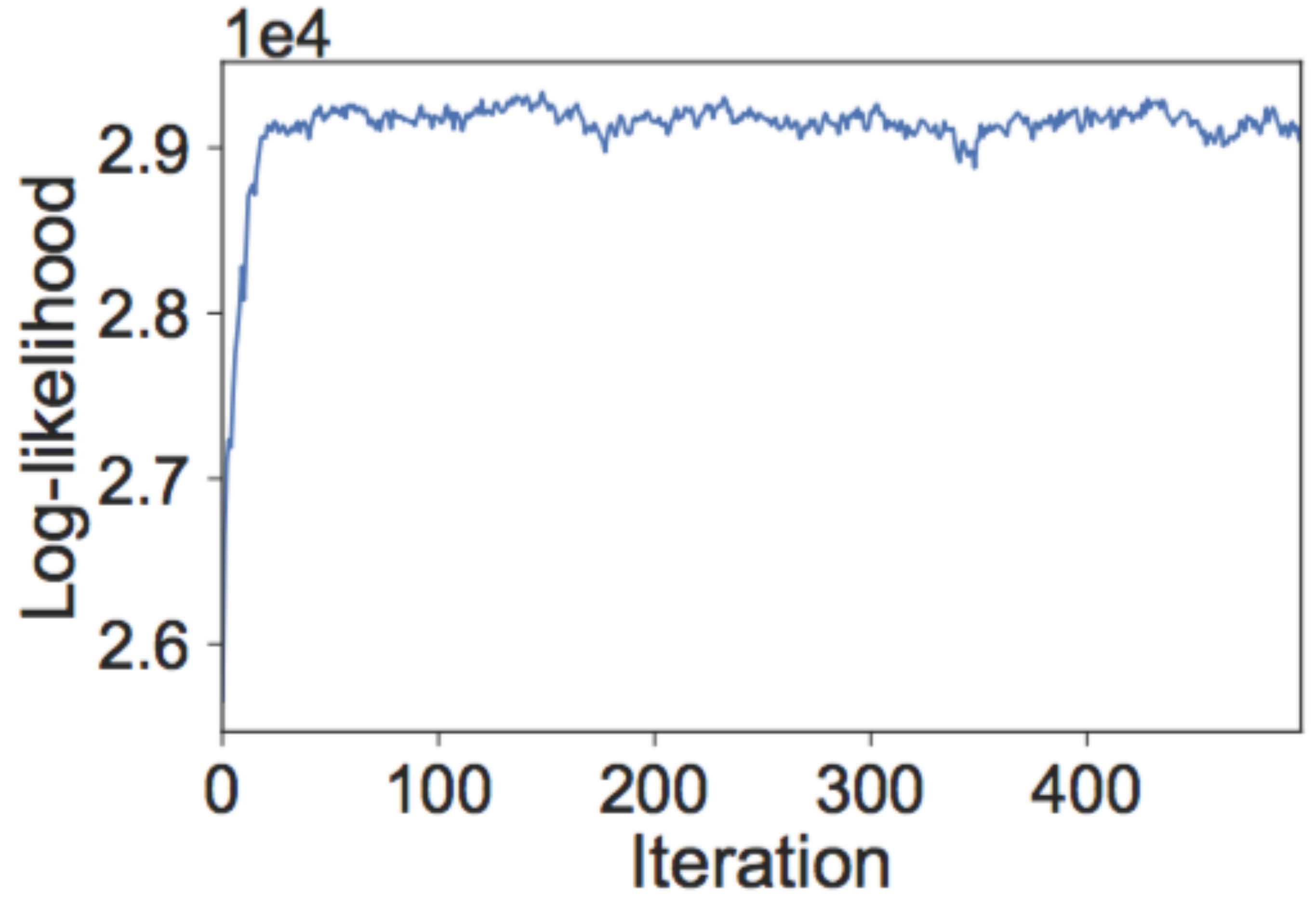
Multimodal embeddings help generate quality geo-topic clusters and extract discriminative features.

EvenTweet: H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. PVLDB, 6(12):1326–1329, 2013.

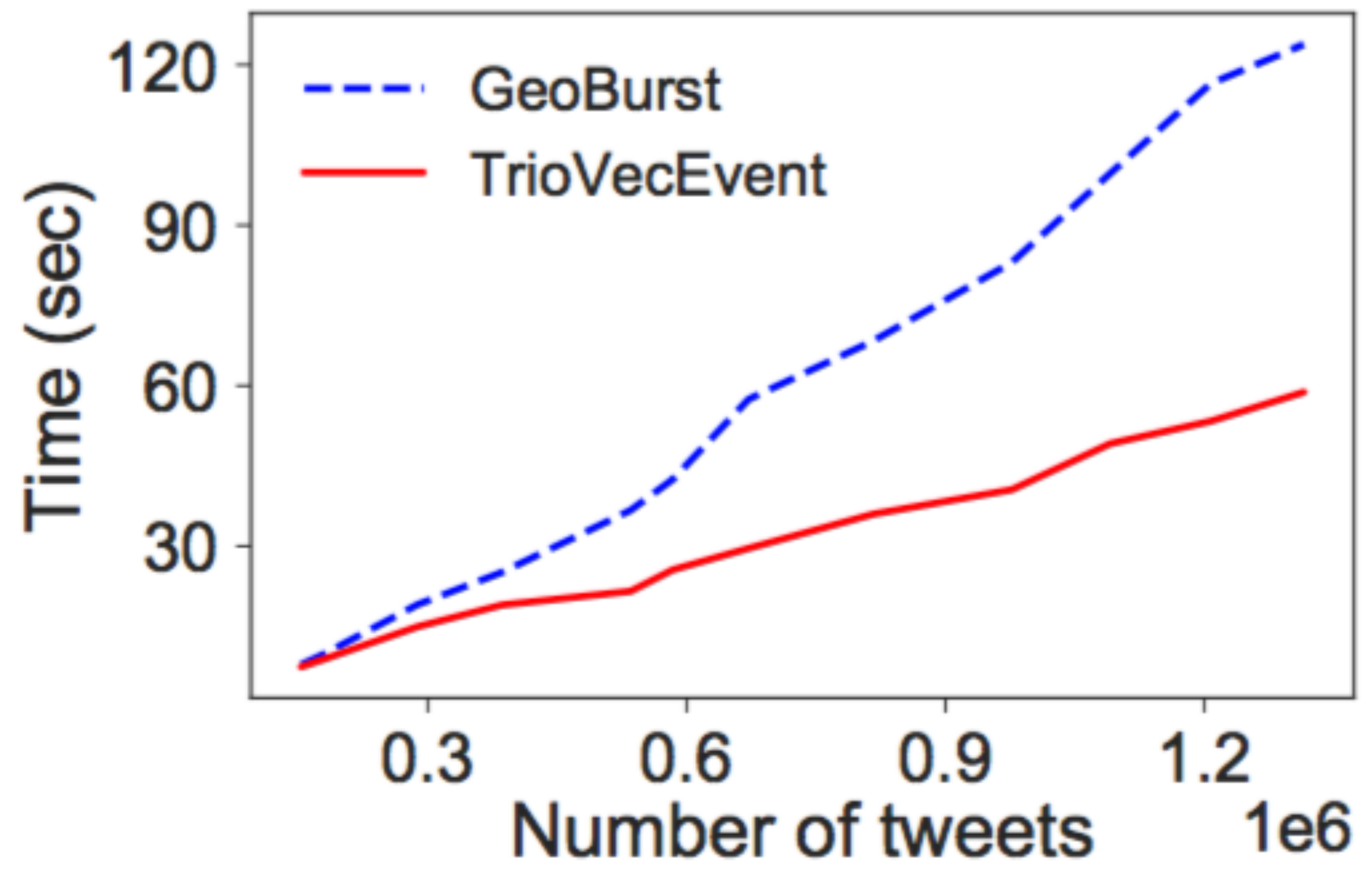
GeoBurst: C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In SIGIR, pages 513–522, 2016.

Efficiency

The clustering module converges within 10 iterations



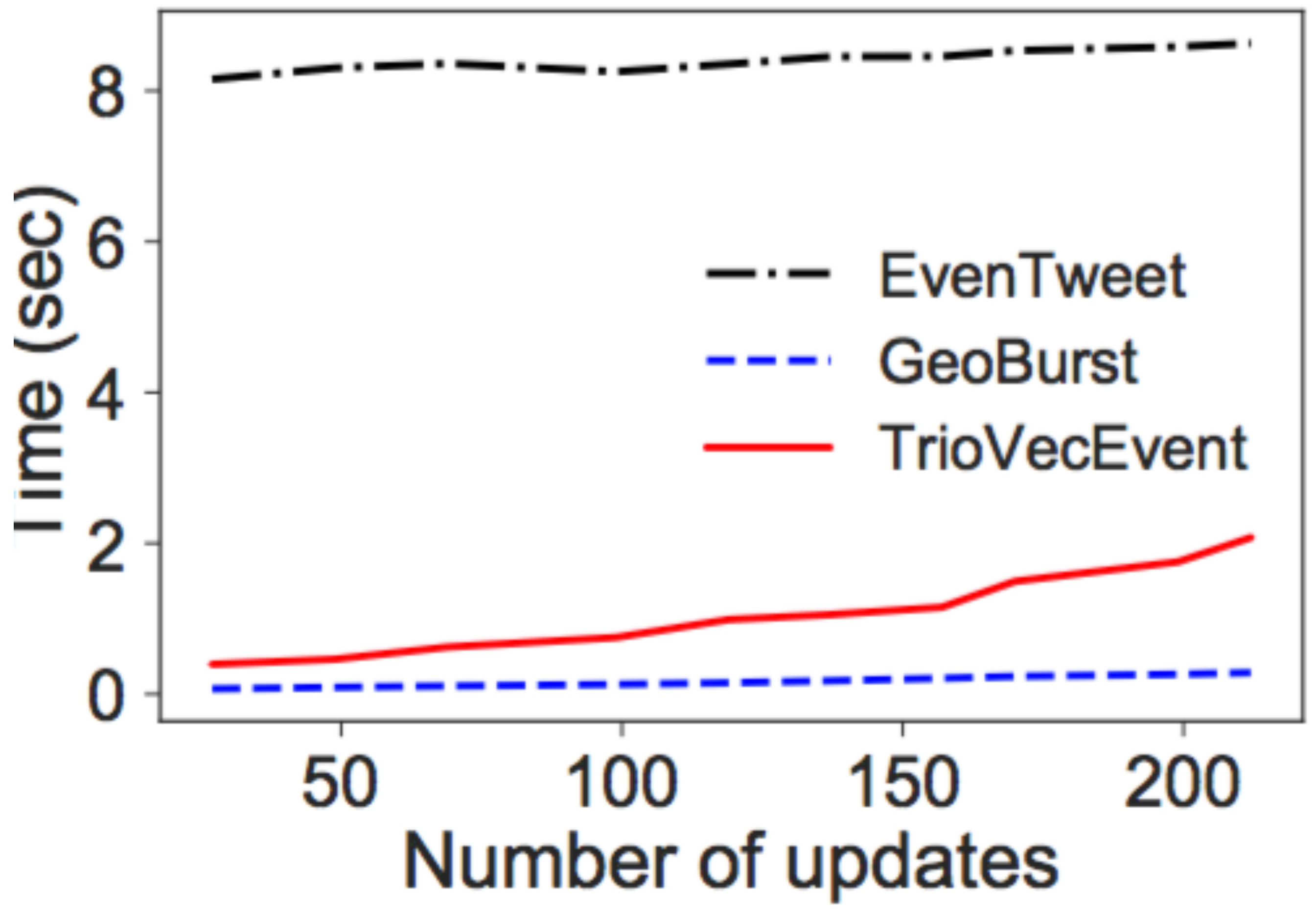
(a) Geo-topic clustering convergence.



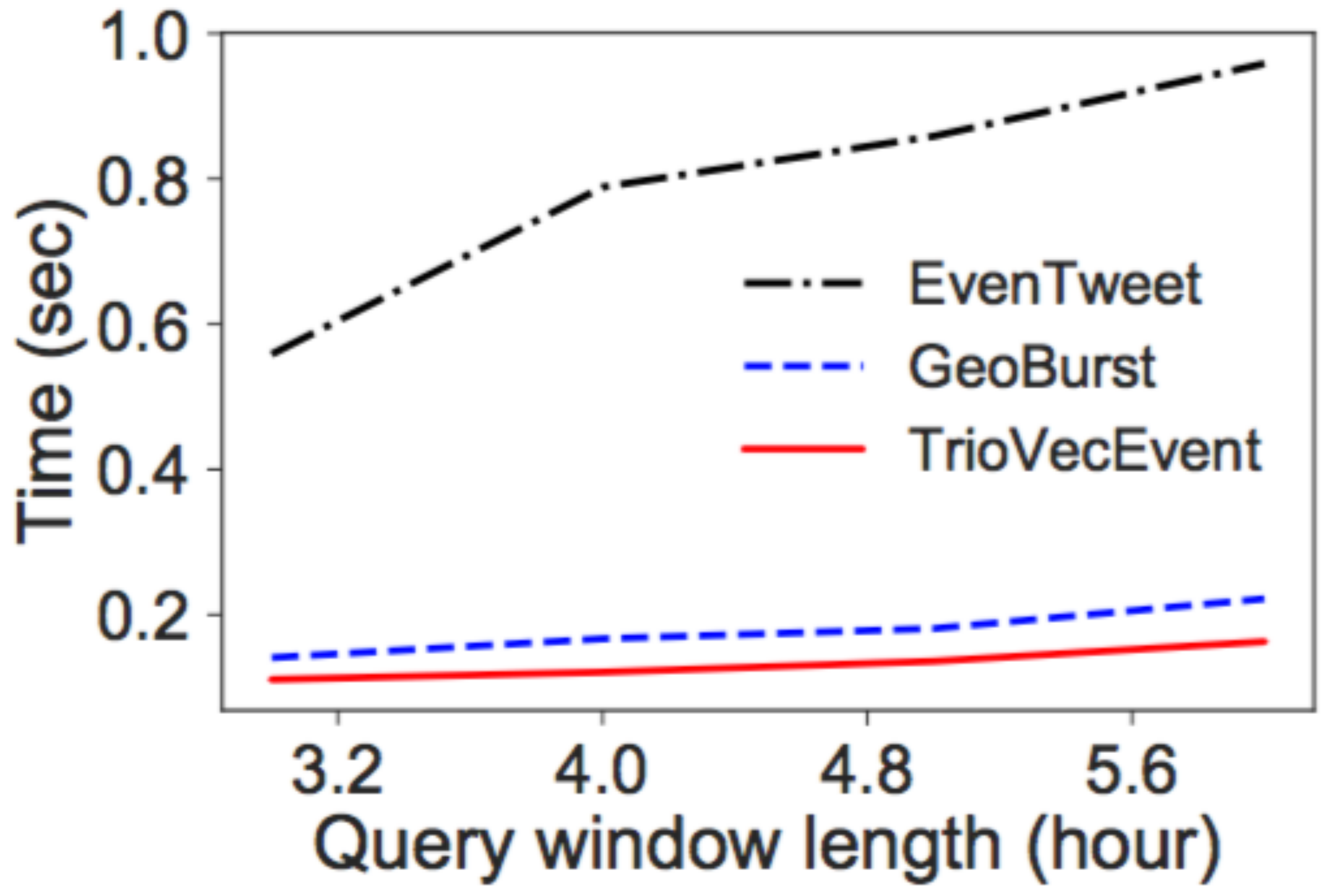
(b) Summarization throughput.

The embedding module takes less time than previous stream summarization-based methods

The time cost of the clustering step is slower than previous methods.



(c) Online clustering time.



(d) Candidate filtering time.

The candidate filtering step based on classification is fast.

Summary

- We proposed a local event detector based on multimodal embedding and discriminative candidate filtering.
- The multimodal embeddings serve well for event detection:
  - Generate very coherent geo-topic clusters as candidate events
  - Induce discriminative features for training accurate classifiers
- Future work:
  - Fine-grained local event detection (detect events with different types)
  - Mining the evolving patterns for local events: before, during, after
  - Structured event summarization

Thanks!