Leibniz Universität Hannover

Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme
Fachgebiet Wissensbasierte Systeme (KBS)

Seminar: Topics in Data Mining

Summer Semester 2016

STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream

Vaibhav Kasturia

M.Sc. ITIS

27th June 2016

Overview

- Introduction
- Event as a cluster of hashtags
- Data warehousing structure
- Hashtag clustering
- Event ranking
- Experimental study
- Conclusion

Introduction

- Twitter : Most famous microblogging website
- Over a billion tweets posted per week
- Each tweet: 140 characters
- Tweet composition:
  - Text
  - Link to author
  - Hashtags (words preceded by '#')
  - Usertags
  - Timestamp
  - Links to external resources
- Geo-tagging of tweets

Sample Tweet Record

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Event as a Cluster of Hashtags

- Twitter Users: Post updates on ongoing events
- Detect events in real time
- Analyze long-term events
- Event as a cluster of hashtags
- Syrian Uprising of 2012[1]
  - #Bashar
  - #Assad
  - #AssadCrime
  - #GenocideInSyria
  - #RiseDamascus

Data Warehousing Structure

- STREAMCUBE: Data Cube Structure[3] extension
- Explore events along
  - Time dimension
  - Space dimension
- Data arrangement in real time
- Rank events in real time using
  - Popularity
  - Burstiness
  - Localness

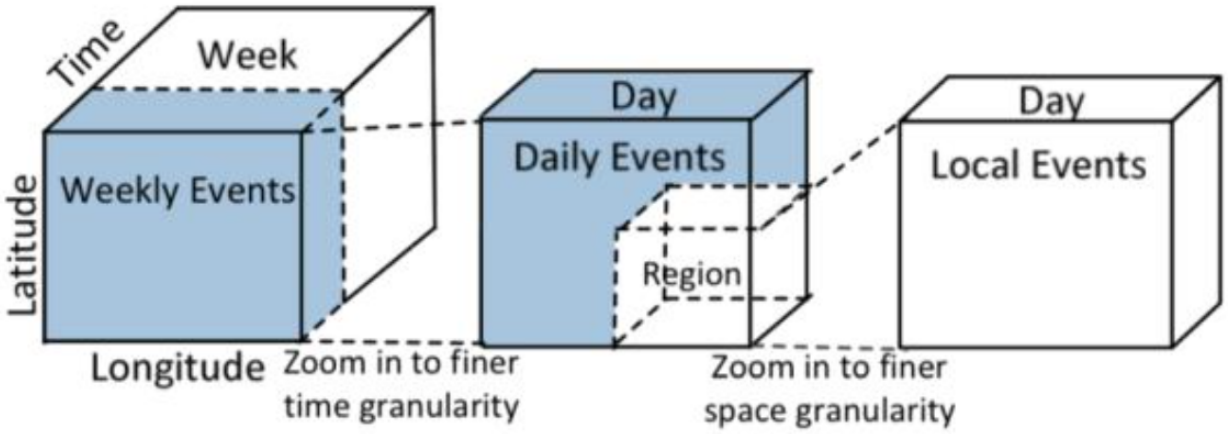


Fig. 1. Zoomable Event Cube[2]

Data Warehousing Structure

- Event Cube structure: Spatio-temporal aggregation
- Space Hierarchy
  - Entire global space
  - Quad Tree like hierarchy
  - Country, city and district
- Time Hierarchy
  - Coarsest granularity: 24 hrs
  - Finest granularity: 6 hrs
- Space Hierarchy embed in Time Hierarchy
- Current 6 hr frame: Main Memory
- Older frames: Disk Storage

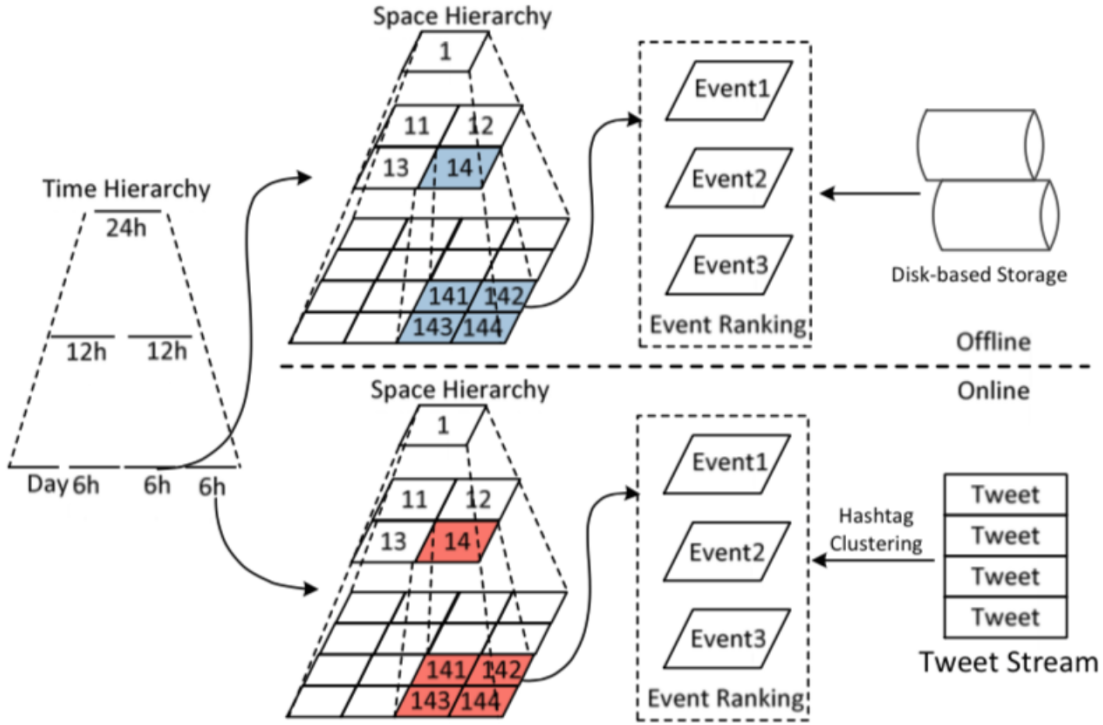


Fig. 2. Framework[2]

Data Warehousing Structure

- Connecting Space and Time Hierarchy
- Creating New Cubes
  - Assigning tweet to lowest hierarchy
  - Mapping to global snapshot in current time frame
  - Diff. tweets, diff. regions: Parallel Mapping
- Spatial Merge
  - All sub-regions combined into a region
  - Current as well as previous cube
- Temporal Merge
  - Matching same regions of current & previous 6 hr frame
  - Current 12 hr frame merged with previous 12 hr frame

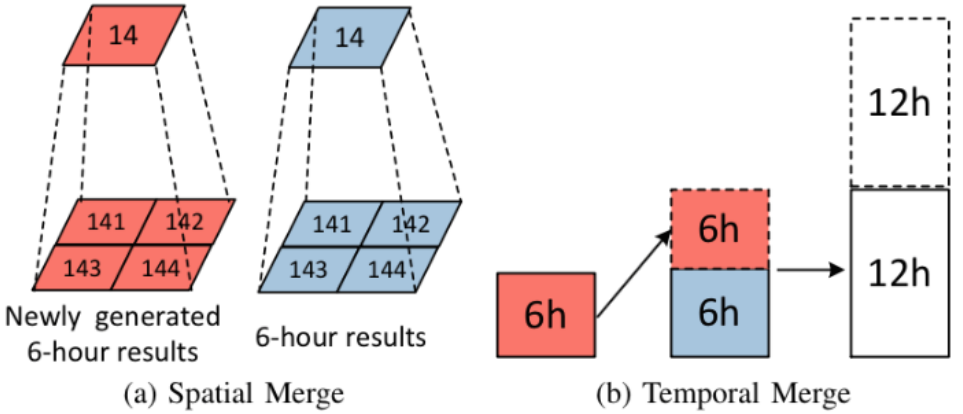


Fig. 3. Spatio-temporal aggregation[2]

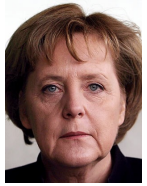
Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

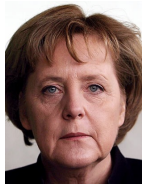
#merkel



Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

#merkel

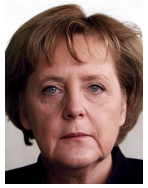


#hannovermesse

Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

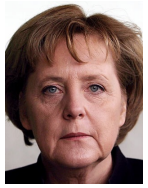
#merkel #hannovermesse



Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

#merkel #hannovermesse



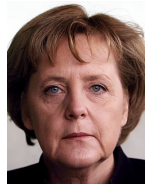
Merkel's Visit to Hannover Messe

Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

#brexit

#merkel #hannovermesse



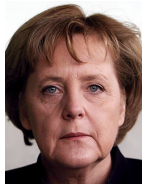
Merkel's Visit to Hannover Messe

Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

#brexit

#merkel



#hannovermesse

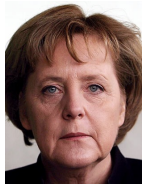
Merkel's Visit to Hannover Messe

Hashtag Clustering

- **Major Challenge**: Hashtags are not Static Data Points

#brexit

#merkel



Angela Merkel on UK's exit from EU

#hannovermesse

Merkel's Visit to Hannover Messe

Hashtag Clustering

Other Challenges:

- Avoiding Iterative Computation
  - Single-Pass Algorithm
  - Little Cost for merging events
- Clear understanding of
  - Localness
  - Burstiness
  - Popularity
- Users interested in
  - Breaking News
  - Past Year's Events
  - Local News
  - Global News

Breaking News Vs. Past Year's Events

Global News Vs. Local News

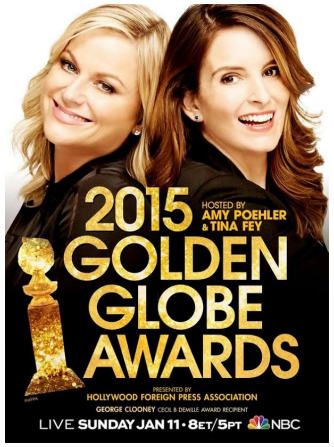
Hashtag Clustering

Hashtag Representation and Similarity:

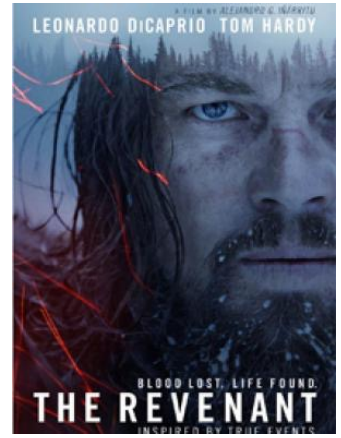
- Hashtag as a Bag of Words
  - Collect all tweets containing hashtag h
  - Identify words in tweets important for hashtag
  - Hashtag h as Normalized Weighted Vector

    $h_{word} = (w_1, w_2, \ldots, w_{|W|})$ & $||h_{word}|| = 1$

- #GoldenGlobes
  - Movies: The Revenant, The Martian
  - Actors: Leonardo DiCaprio, Matt Damon
  - Actresses: Brie Larson, Jennifer Lawrence



Golden Globes 2015

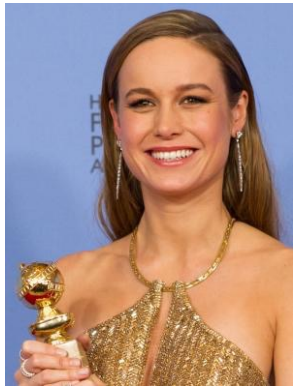


The Revenant

The Martian


Matt Damon


Leonardo DiCaprio


Brie Larson

Jennifer Lawrence

Hashtag Clustering

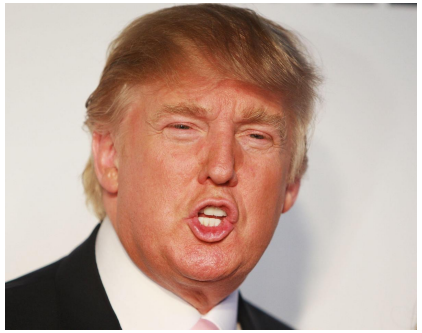
Hashtag Representation and Similarity:

- Hashtag as a Bag of Hashtags
  - Hashtag describing an event co-occur with each other
  - Hashtag set H and hashtag h

    $h_{tag} = (h_1, h_2, \ldots, h_{|H|})$ & $||h_{tag}|| = 1$

- US Presidential Elections 2016
  - #Trump2016
  - #Hillary2016
  - #Sanders
  - #USElections2016

US Presidential Elections 2016



Donald Trump

Bernie Sanders

Hillary Clinton

Hashtag Clustering

Hashtag Representation and Similarity: What to use ?

- Use Combination of both
  - Hashtag as bag of words
  - Hashtag as bag of hashtags
- Given two hashtags h_1 and h_2, their similarity can be calculated as follows
  - $\alpha + \beta = 1$
  - $\beta = 0.7$
  - System gives more weight to hashtags than extracted words

$$
\begin{aligned}
sim(\mathbf{h}_1, \mathbf{h}_2) &= \alpha \cdot cos(\mathbf{h}^1_{word}, \mathbf{h}^2_{word}) + \beta \cdot cos(\mathbf{h}^1_{tag}, \mathbf{h}^2_{tag}) \\
&= \alpha \frac{\mathbf{h}^1_{word}, \mathbf{h}^2_{word}}{||\mathbf{h}^1_{word}|| \cdot ||\mathbf{h}^2_{word}||} + \beta \frac{\mathbf{h}^1_{tag}, \mathbf{h}^2_{tag}}{||\mathbf{h}^1_{tag}|| \cdot ||\mathbf{h}^2_{tag}||} \\
&= \alpha \sum_{i=1}^{|W|} w^1_i w^2_i + \beta \sum_{i=1}^{|H|} h^1_i h^2_i \\
&= (\alpha^{\frac{1}{2}} \mathbf{h}^1_{word}, \beta^{\frac{1}{2}} \mathbf{h}^1_{tag}) \cdot (\alpha^{\frac{1}{2}} \mathbf{h}^2_{word}, \beta^{\frac{1}{2}} \mathbf{h}^2_{tag})
\end{aligned}
$$

Hashtag Clustering

Hashtag Representation and Similarity:

- What does the system do?
  - Normalized word weighted vector for hashtag **h**
  - Normalized hashtag weighted vector for hashtag **h**
  - Re-normalization using combination of both for finding similarity between hashtags
  - Hashtag can be finally represented as the vector:

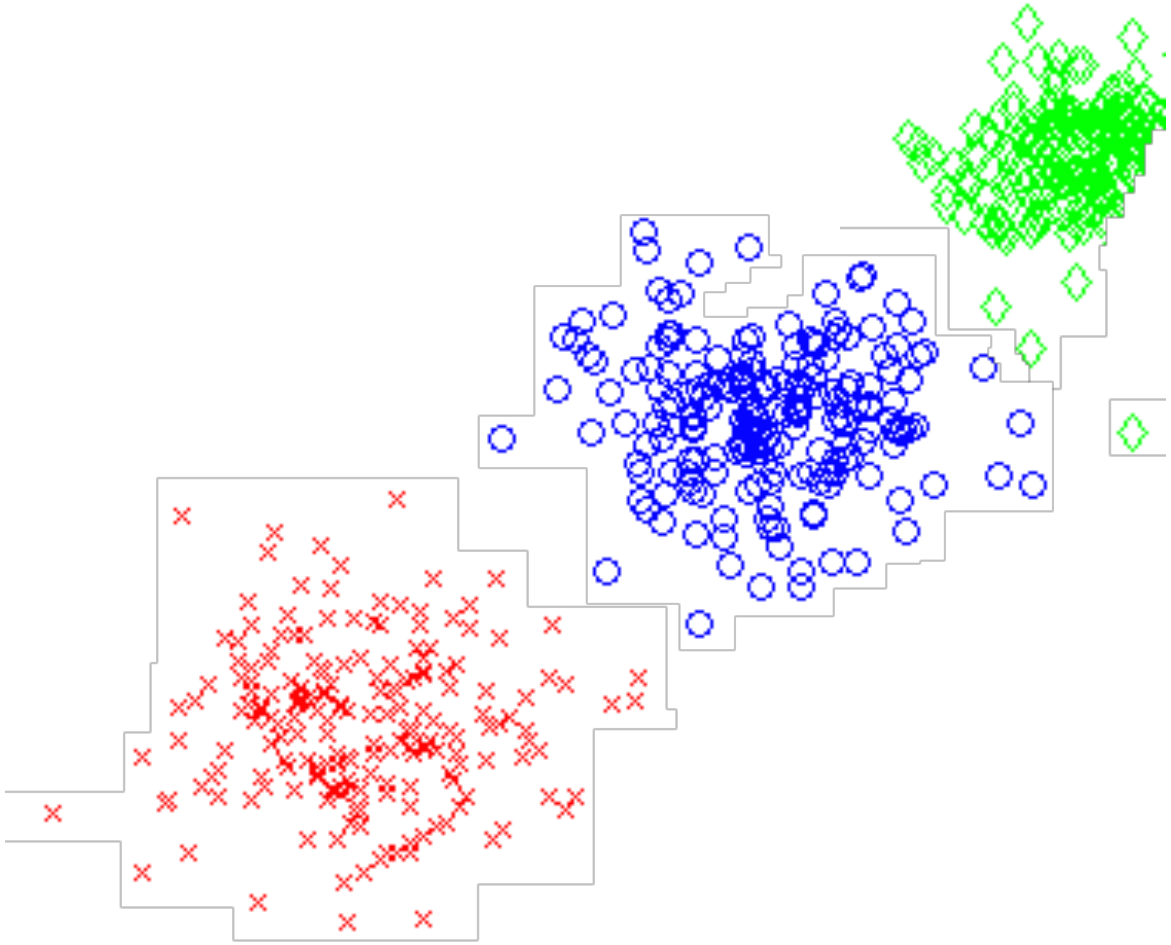
$$\mathbf{h} = (\alpha^{\frac{1}{2}}\mathbf{h}_{word}, \beta^{\frac{1}{2}}\mathbf{h}_{tag})$$

- Event Representation
  - Event: Group of Hashtags
  - Can be represented in the same way as hashtags, i.e.,

$$\mathbf{e} = (\alpha^{\frac{1}{2}}\mathbf{e}_{word}, \beta^{\frac{1}{2}}\mathbf{e}_{tag})$$

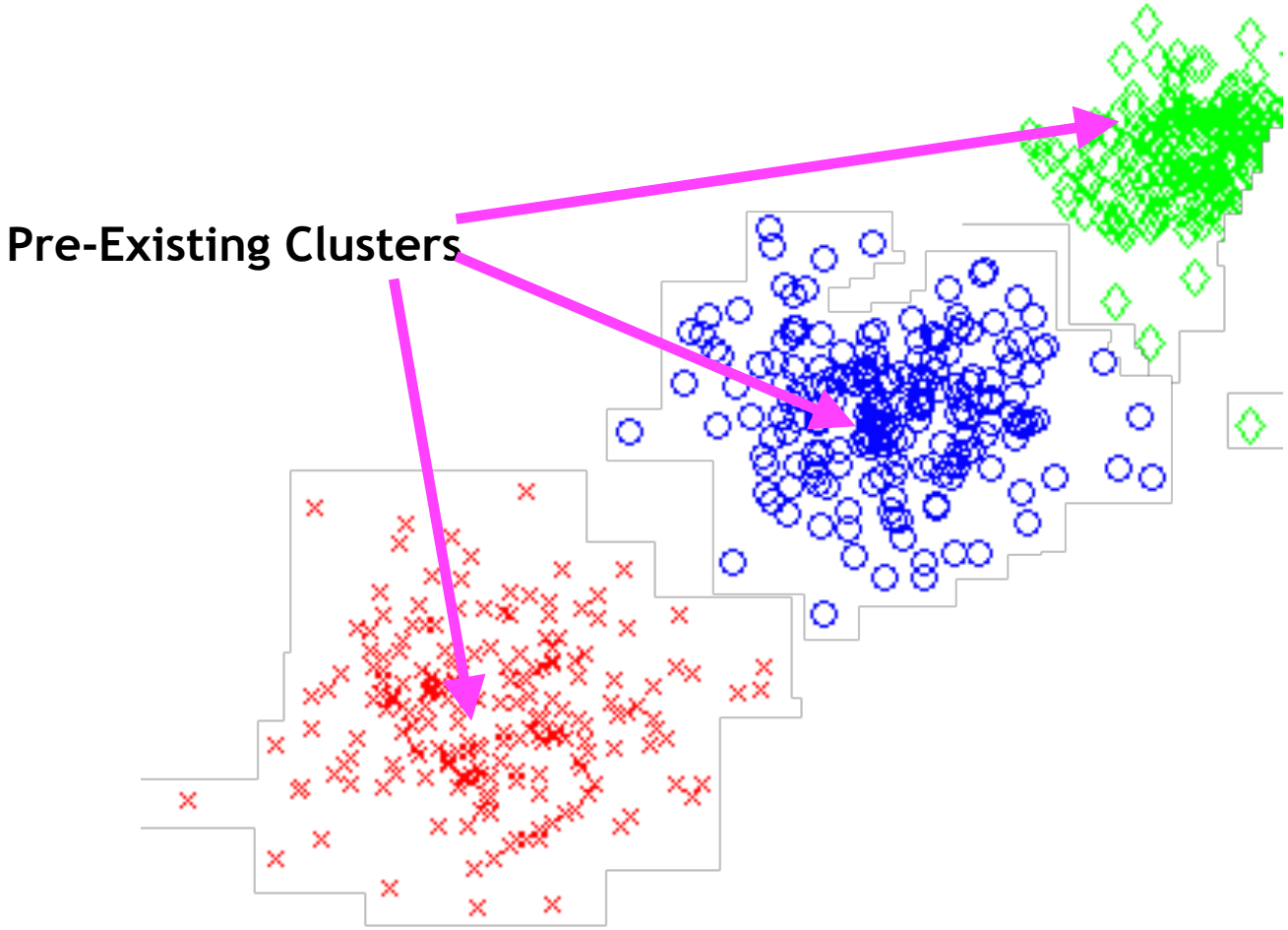
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



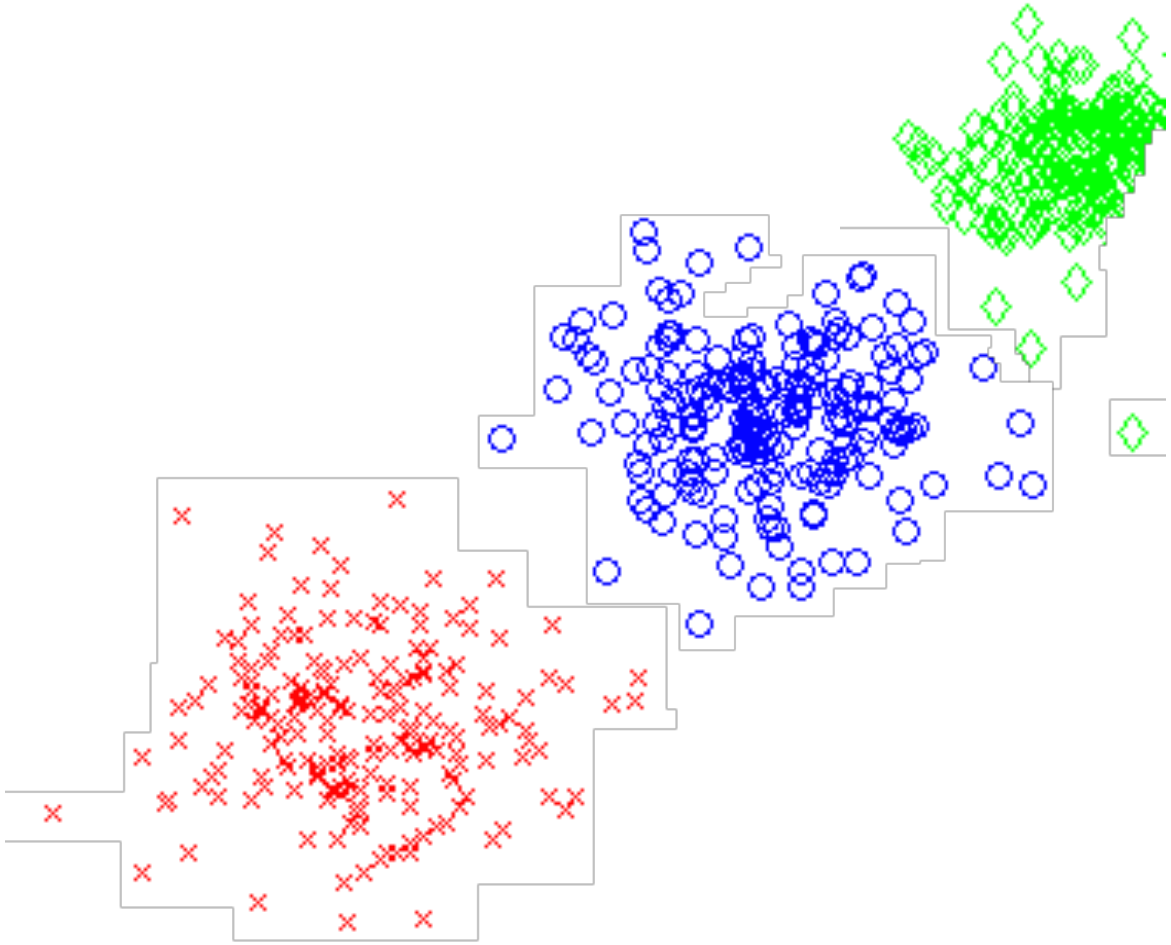
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



Hashtag Clustering

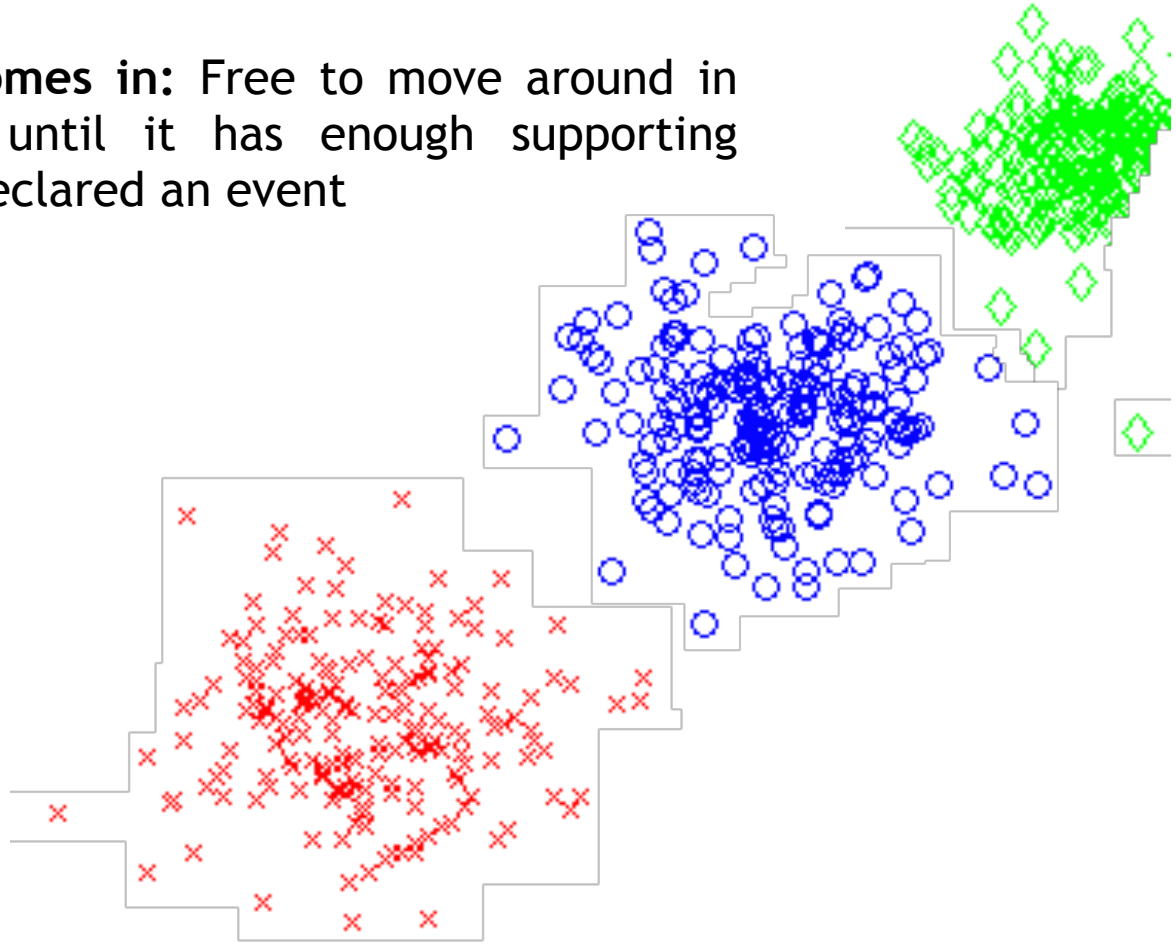
Part Of Vector Space (22-24. April 2016)



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

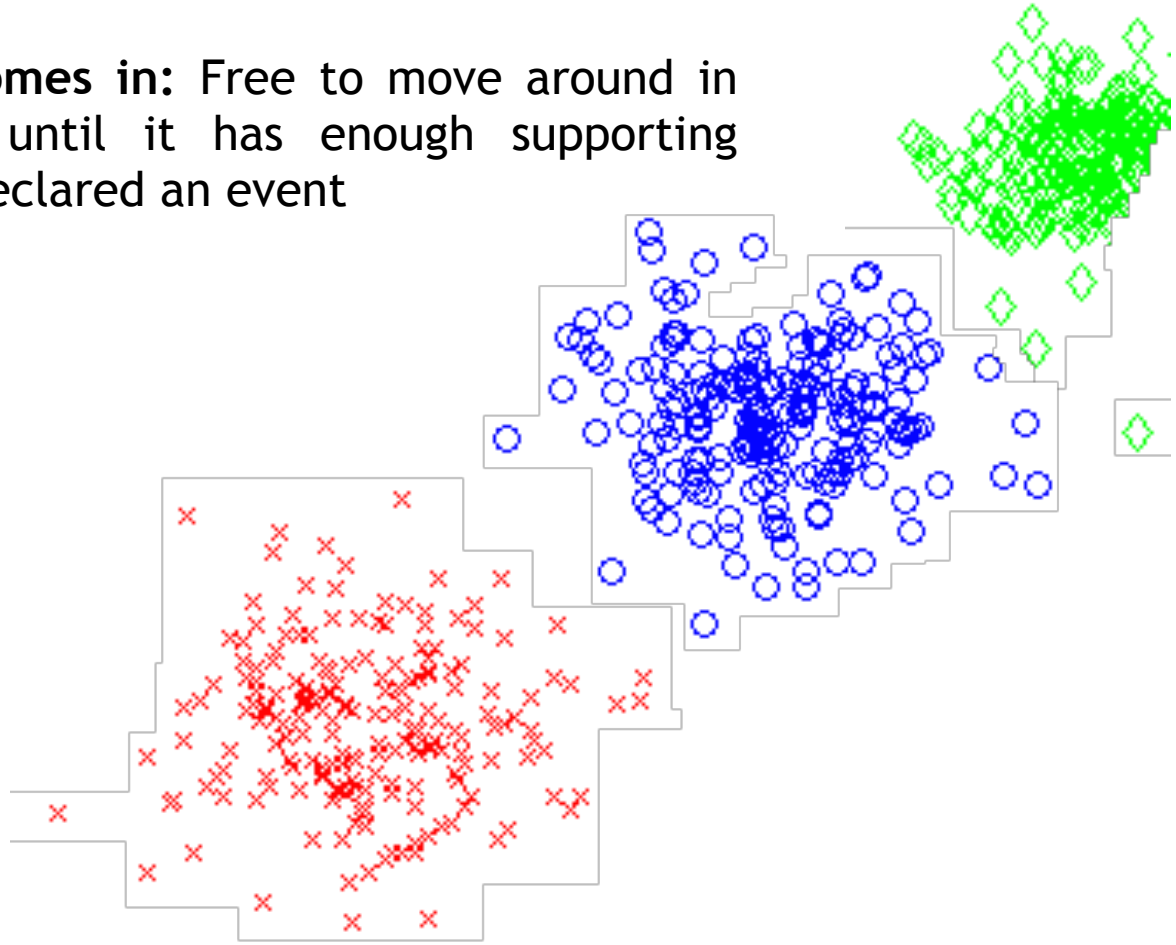
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

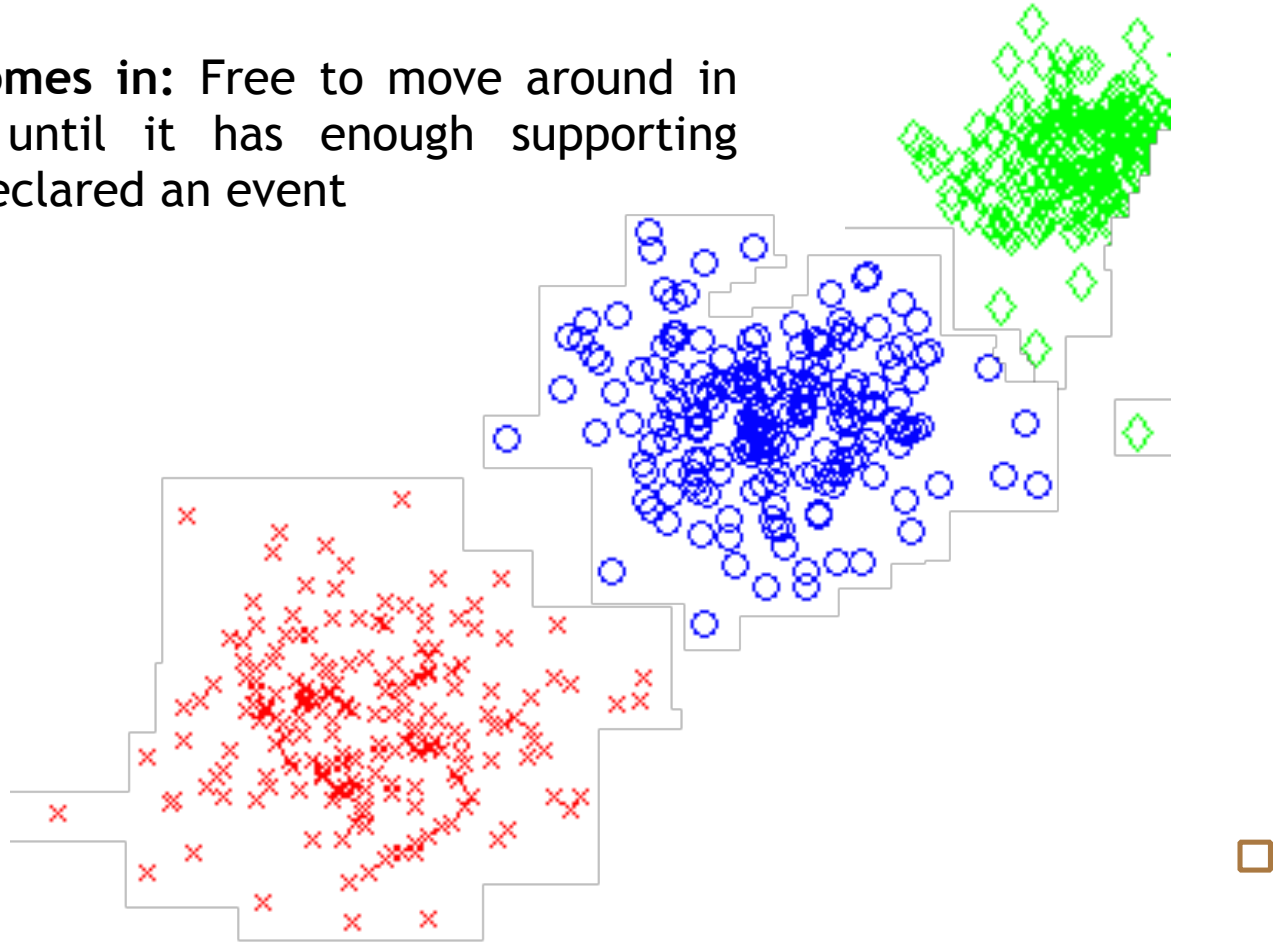
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

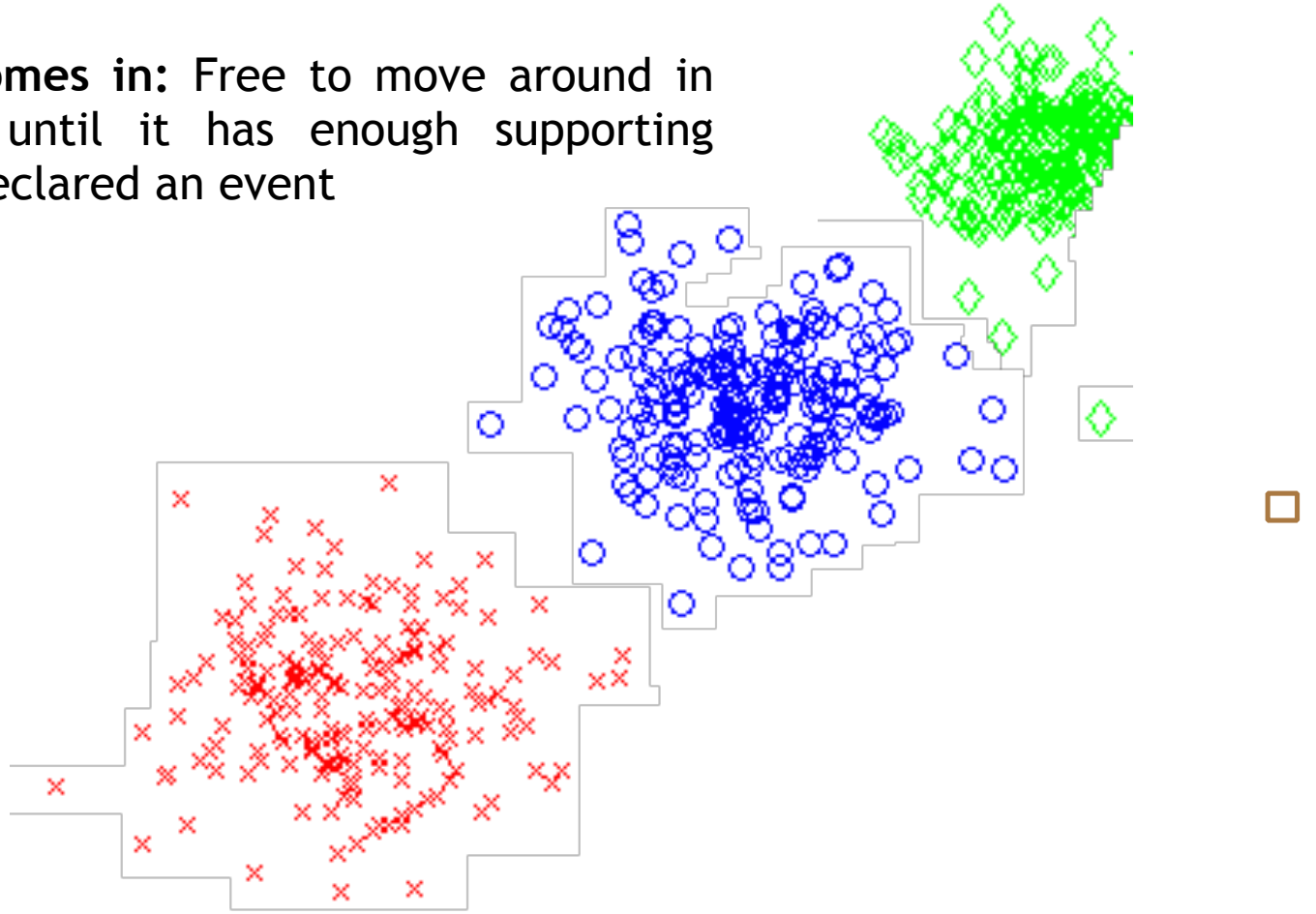
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

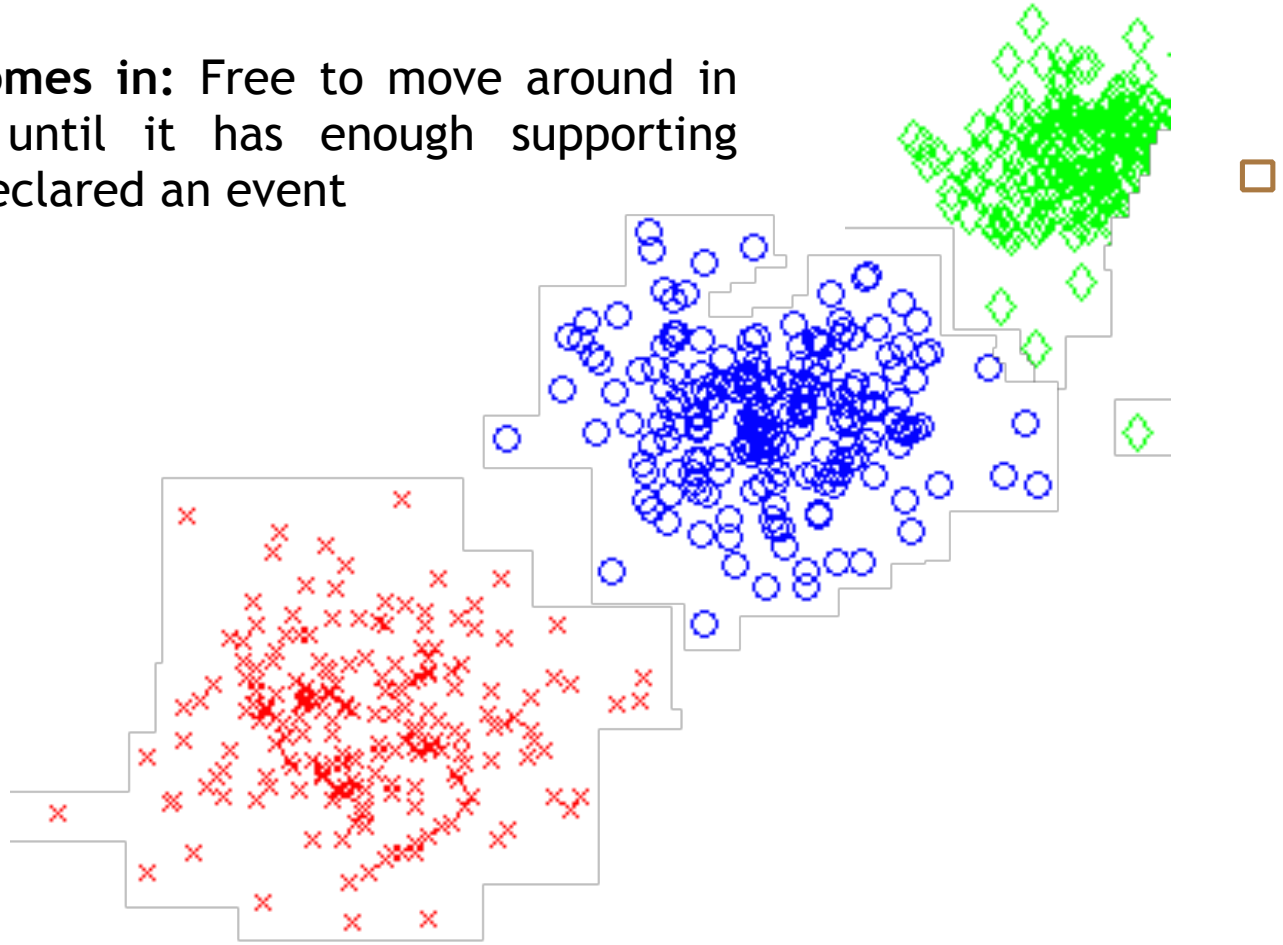
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

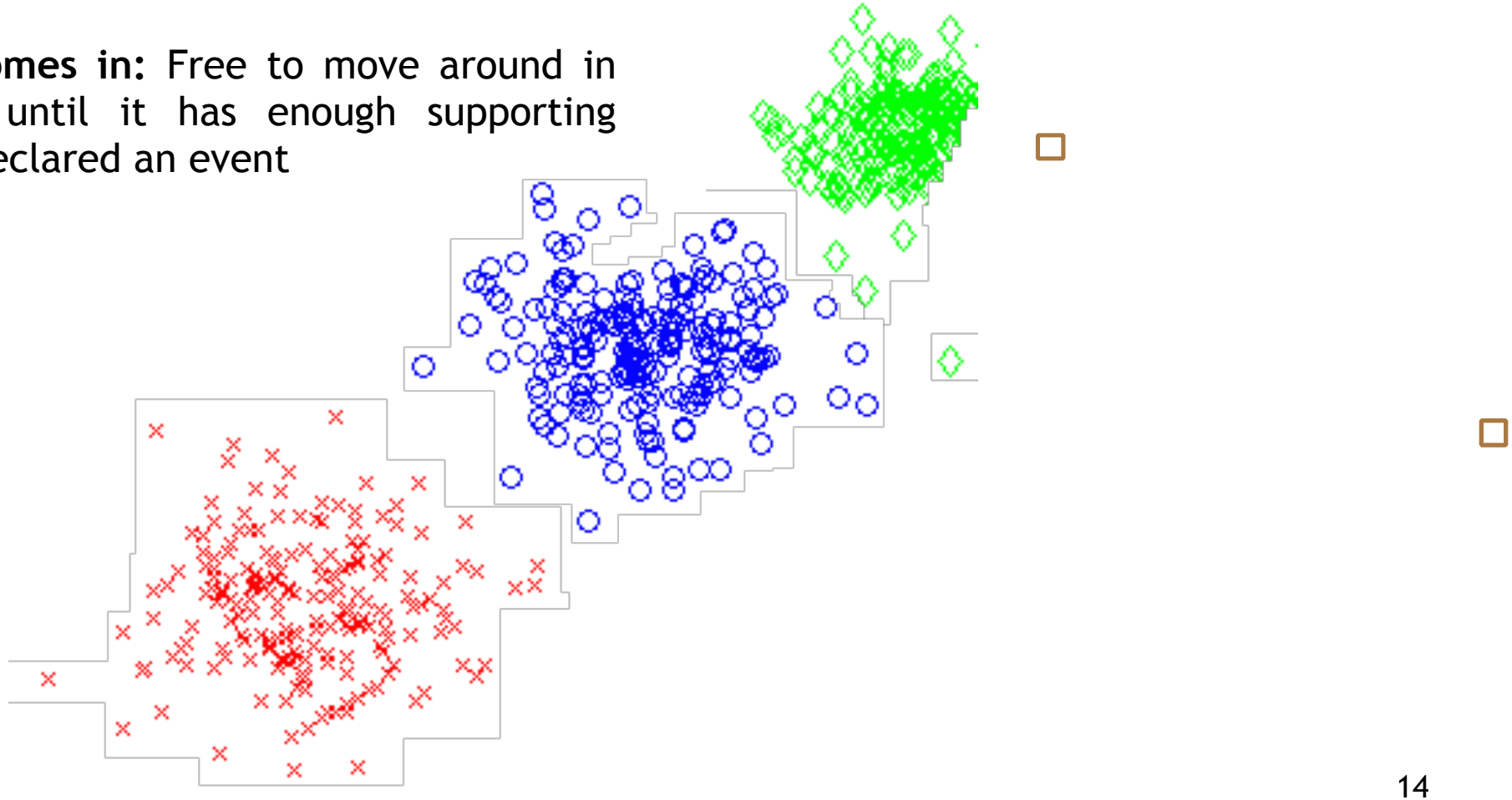
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

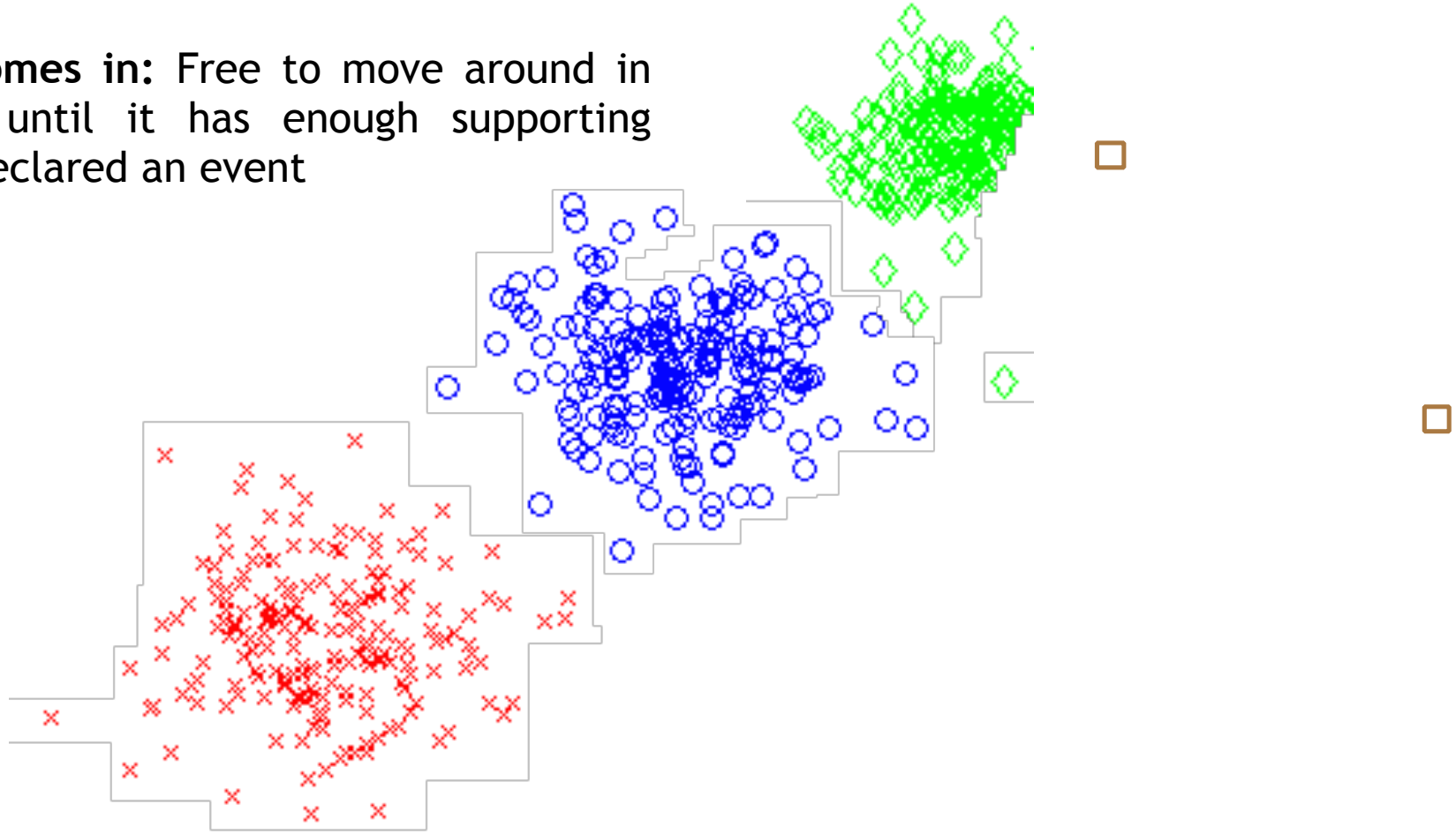
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

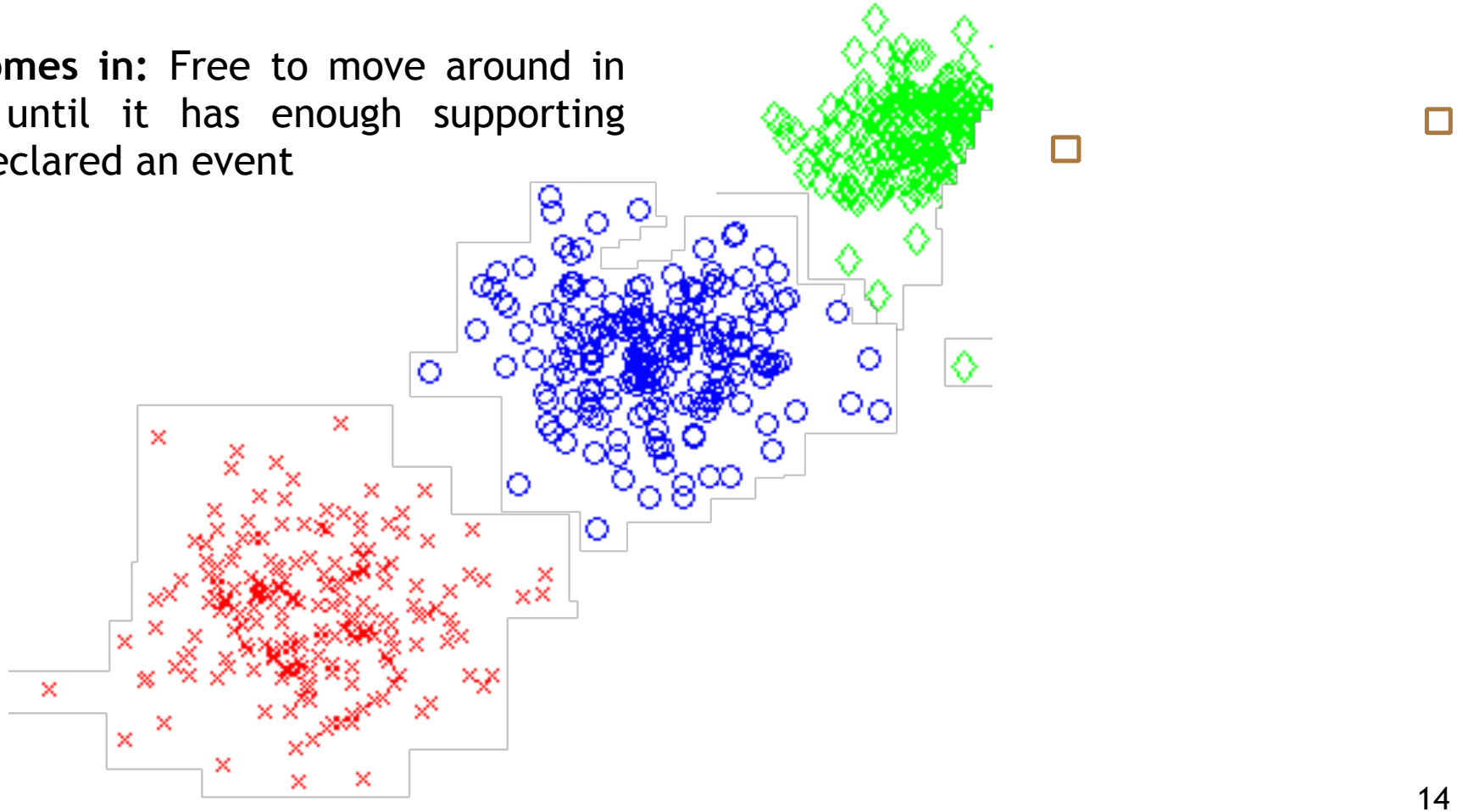
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

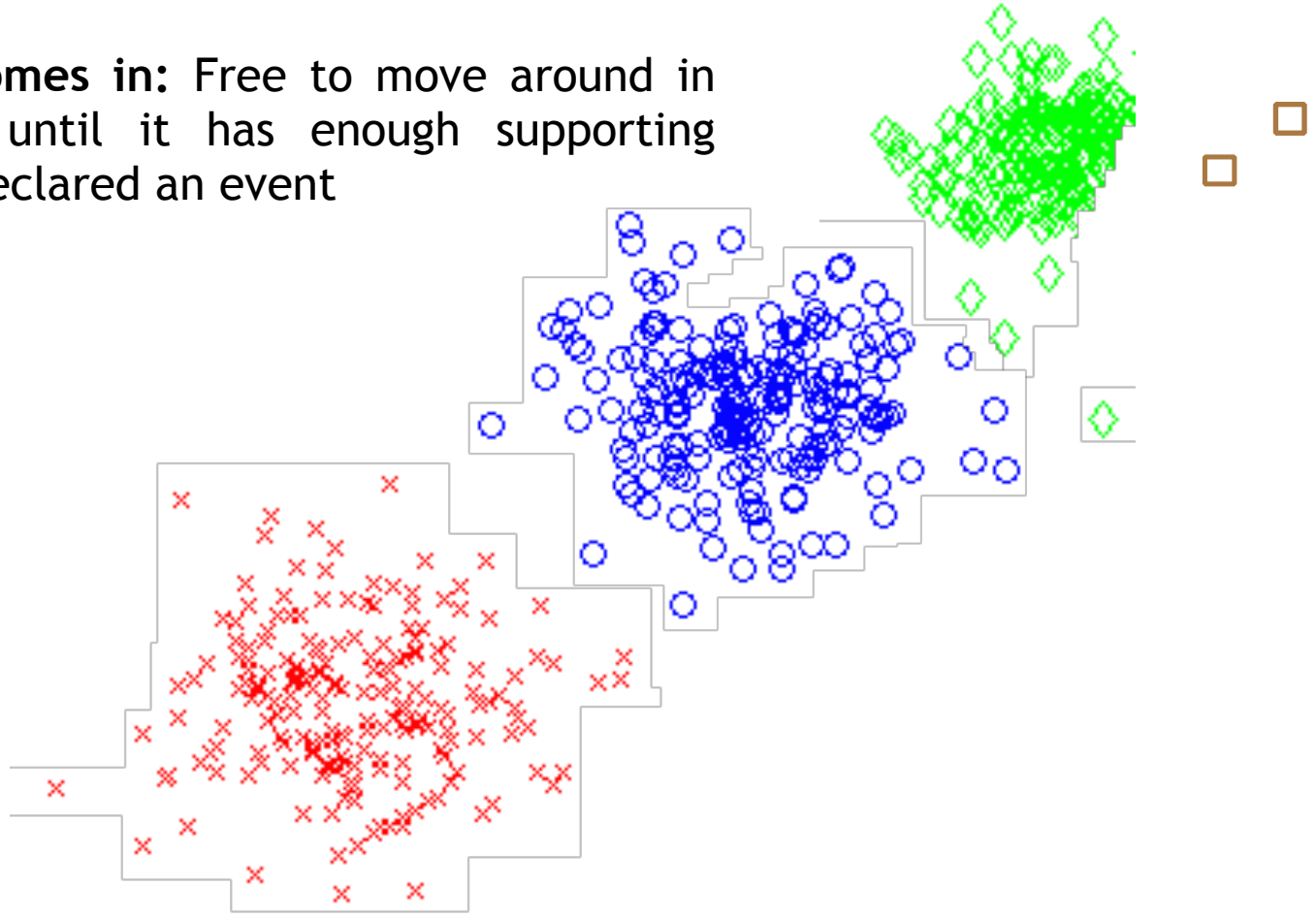
New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



Hashtag Clustering

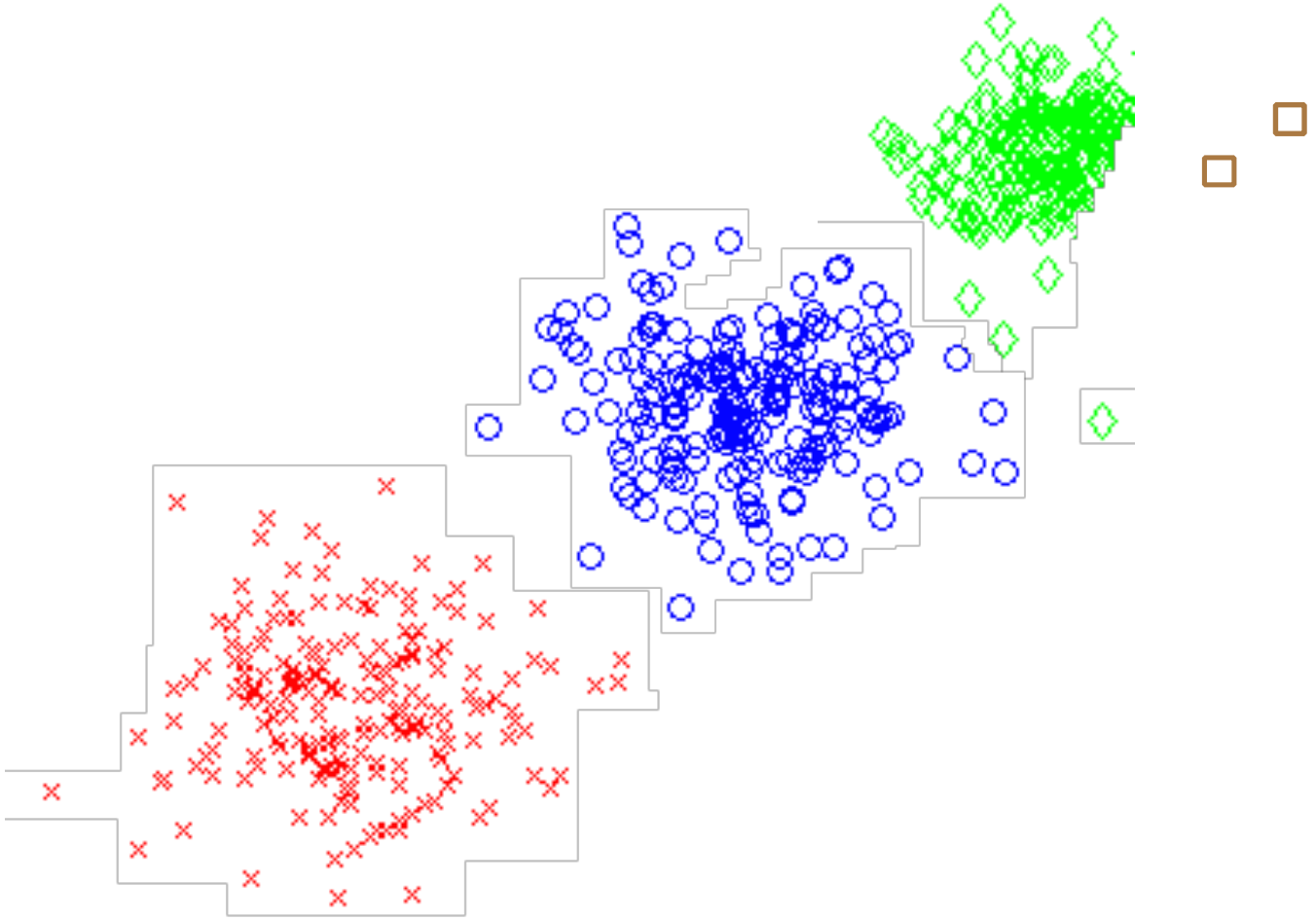
Part Of Vector Space (22-24. April 2016)

New Tweet comes in: Free to move around in Vector Space until it has enough supporting tweets to be declared an event



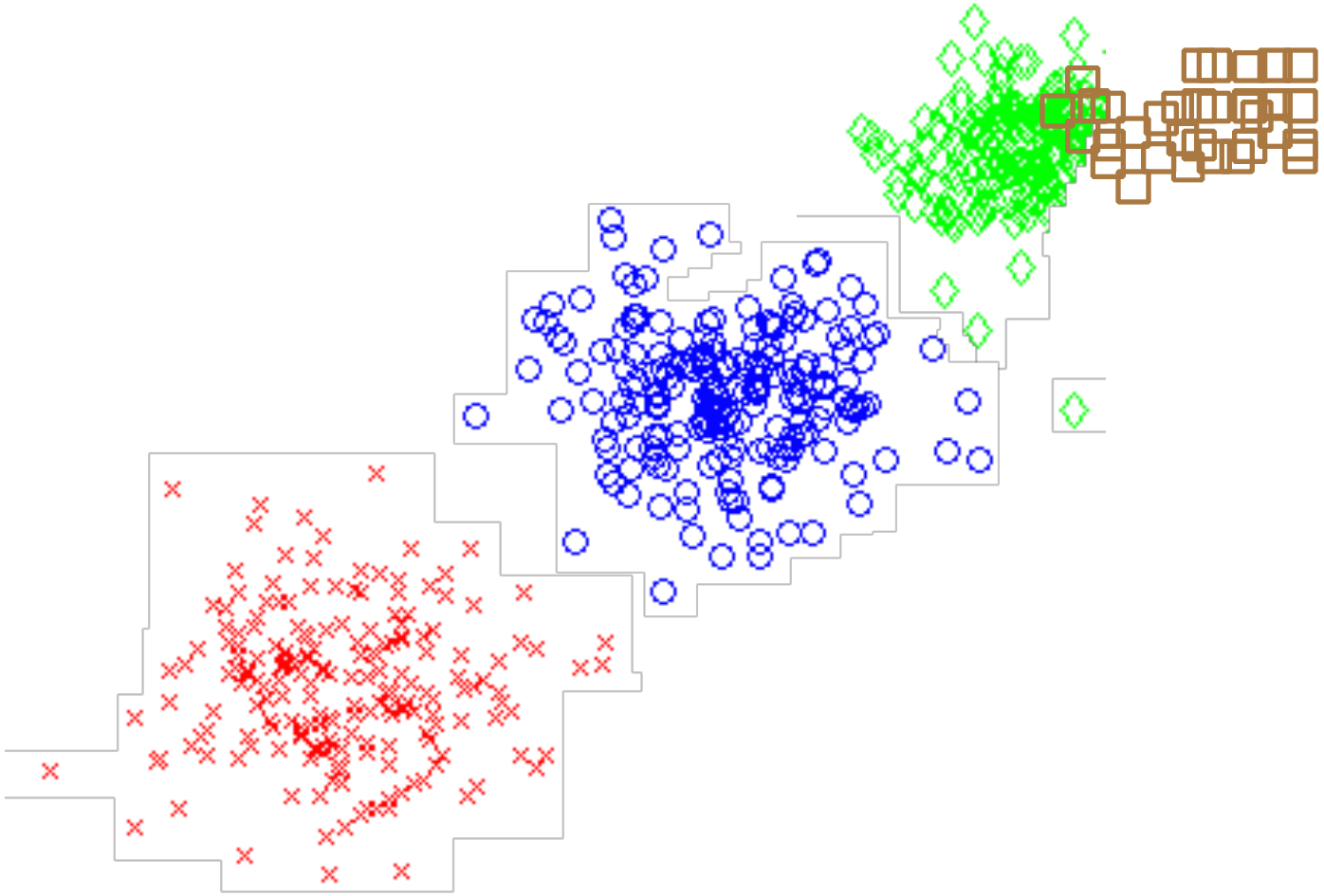
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



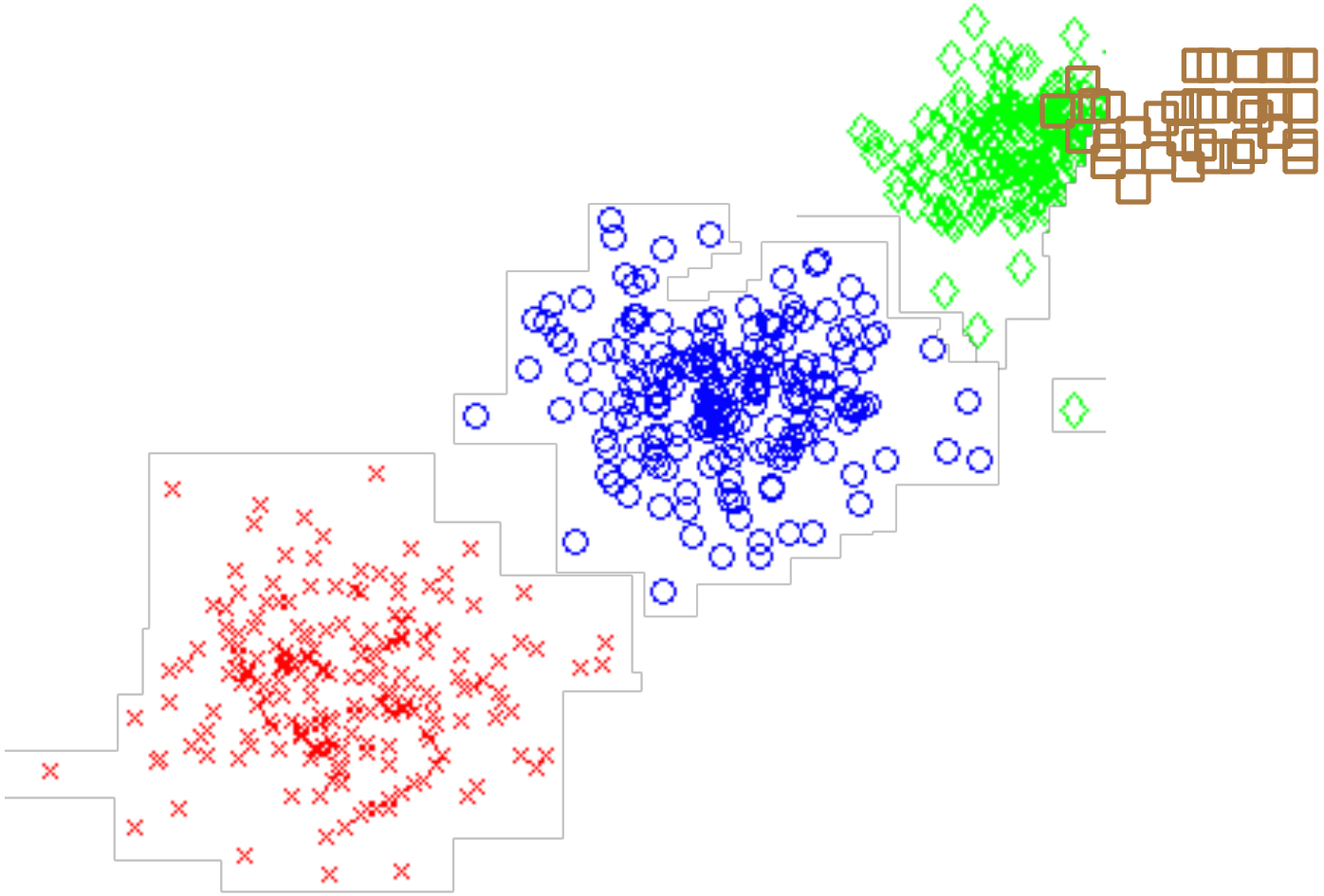
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



Hashtag Clustering

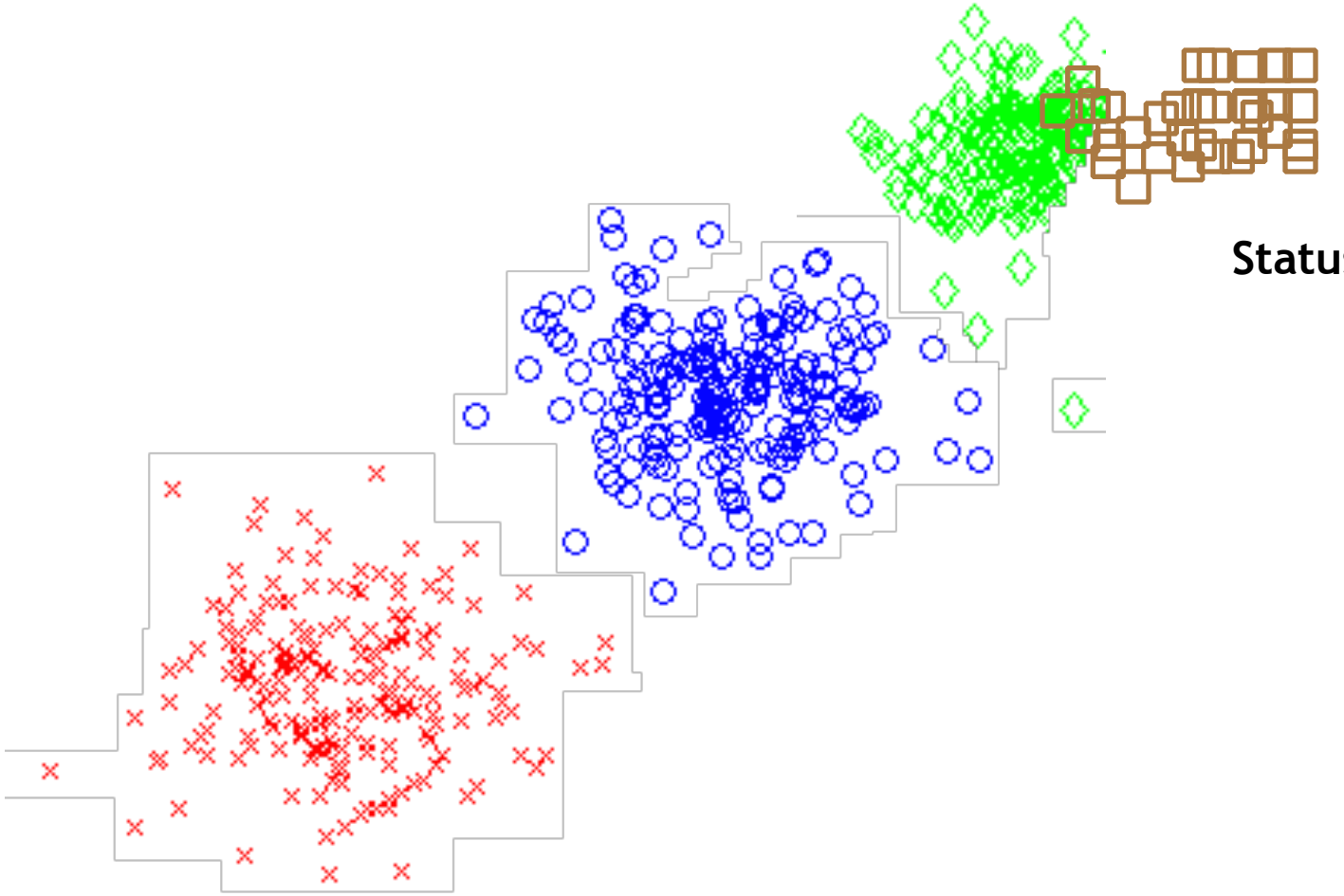
Part Of Vector Space (22-24. April 2016)



Event > 30 Tweets

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

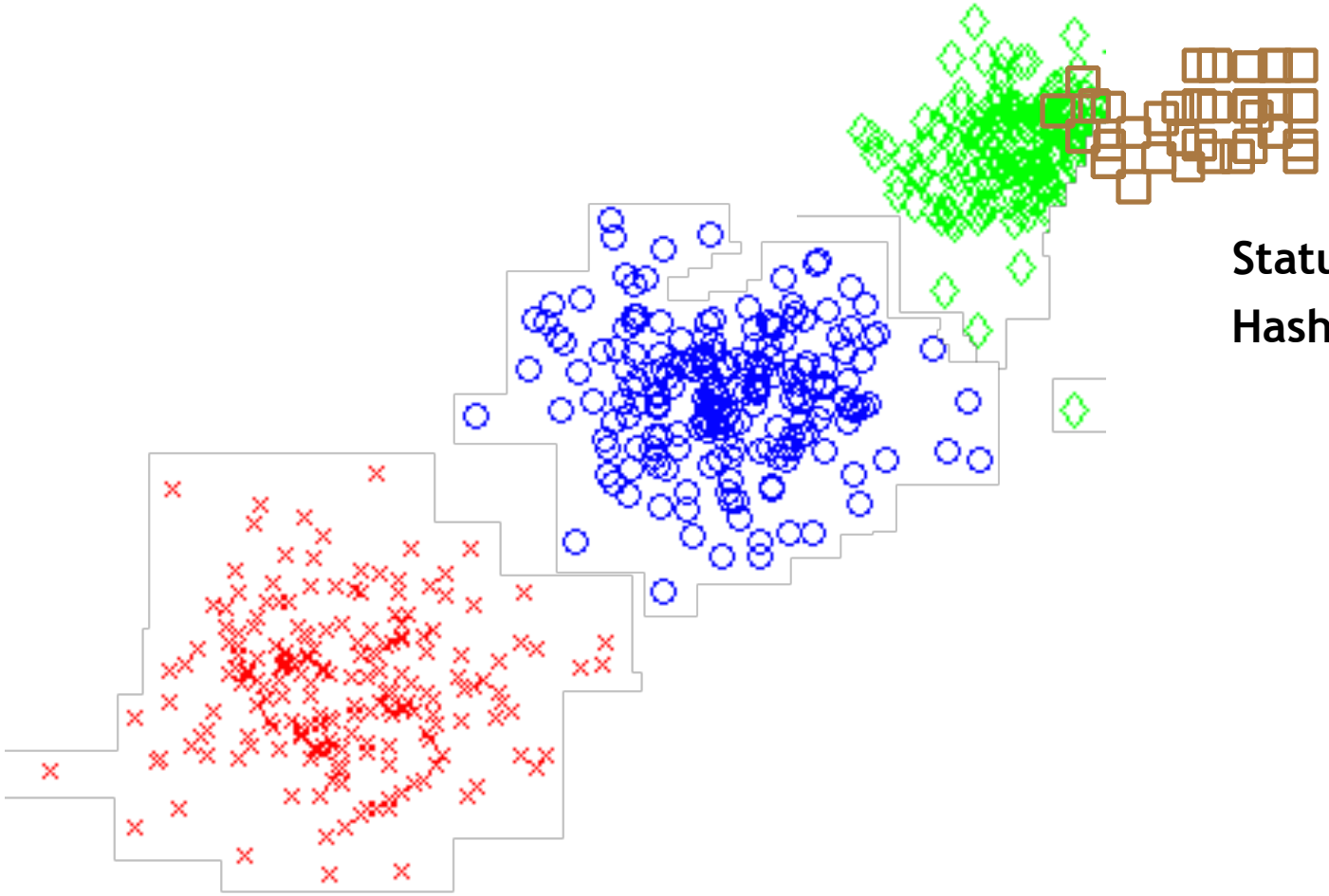


Event > 30 Tweets

Status Change: Inactive to Active

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



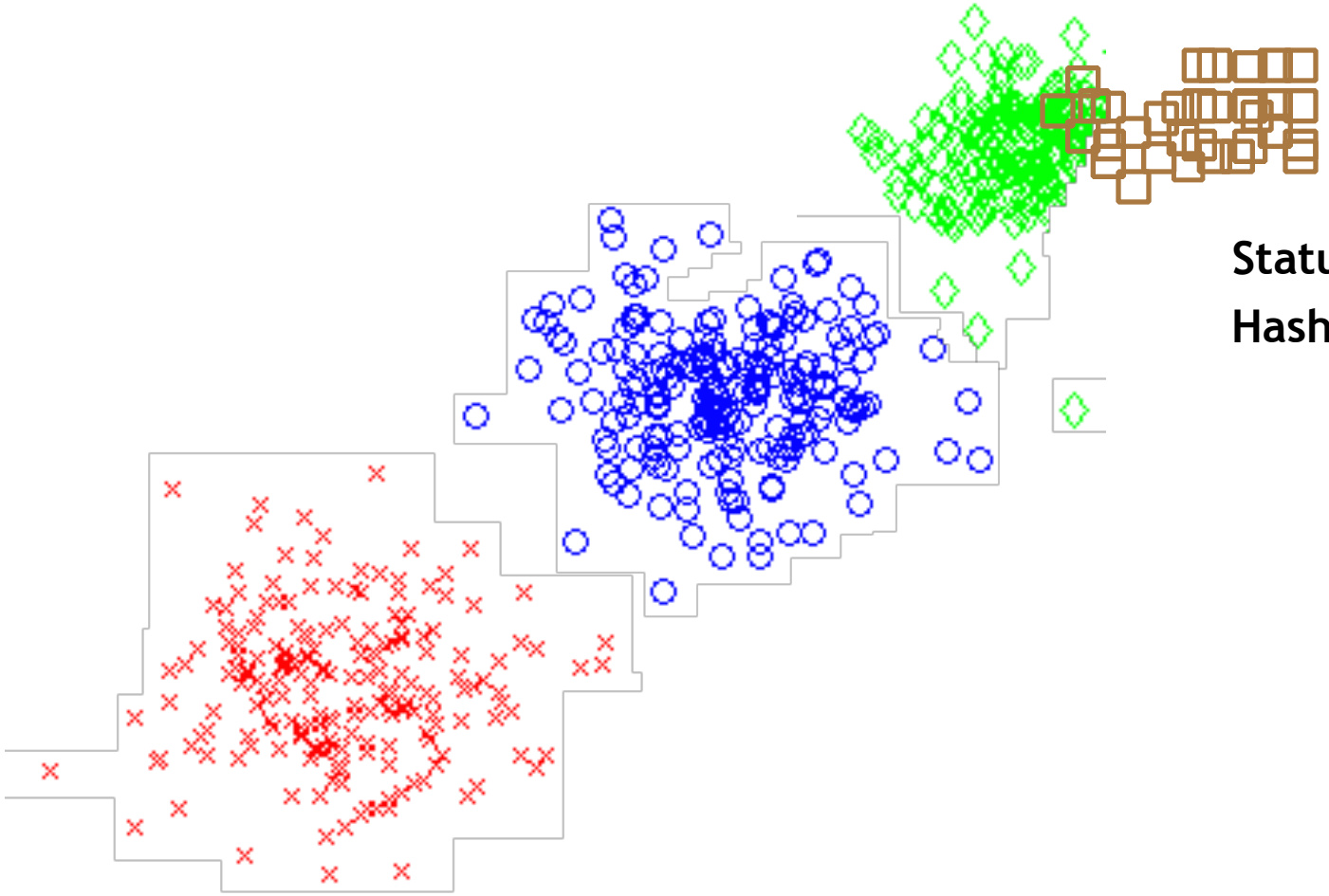
Event > 30 Tweets

Status Change: Inactive to Active

Hashtags: Dynamic to Static

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

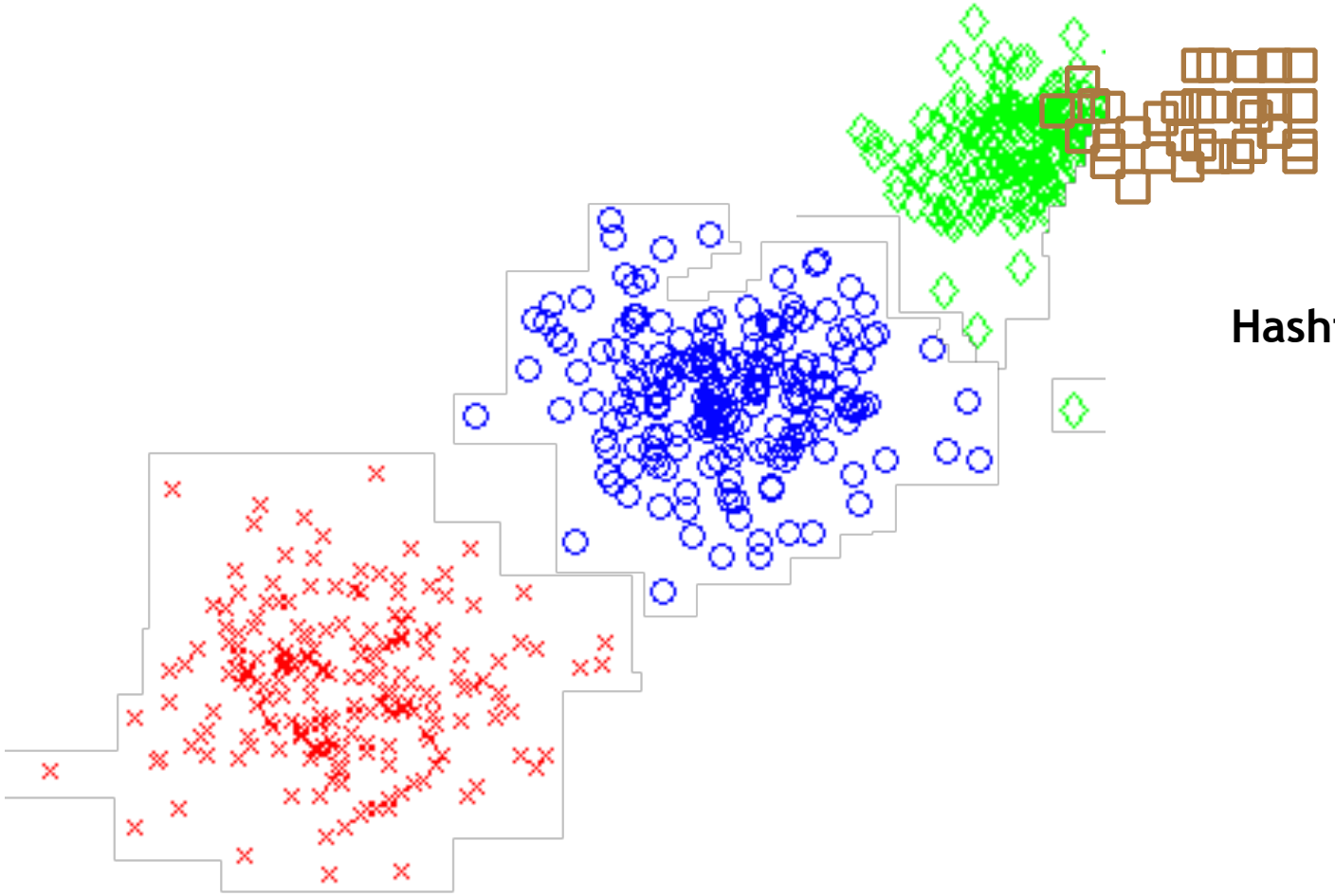


Status Change: Inactive to Active

Hashtags: Dynamic to Static

Hashtag Clustering

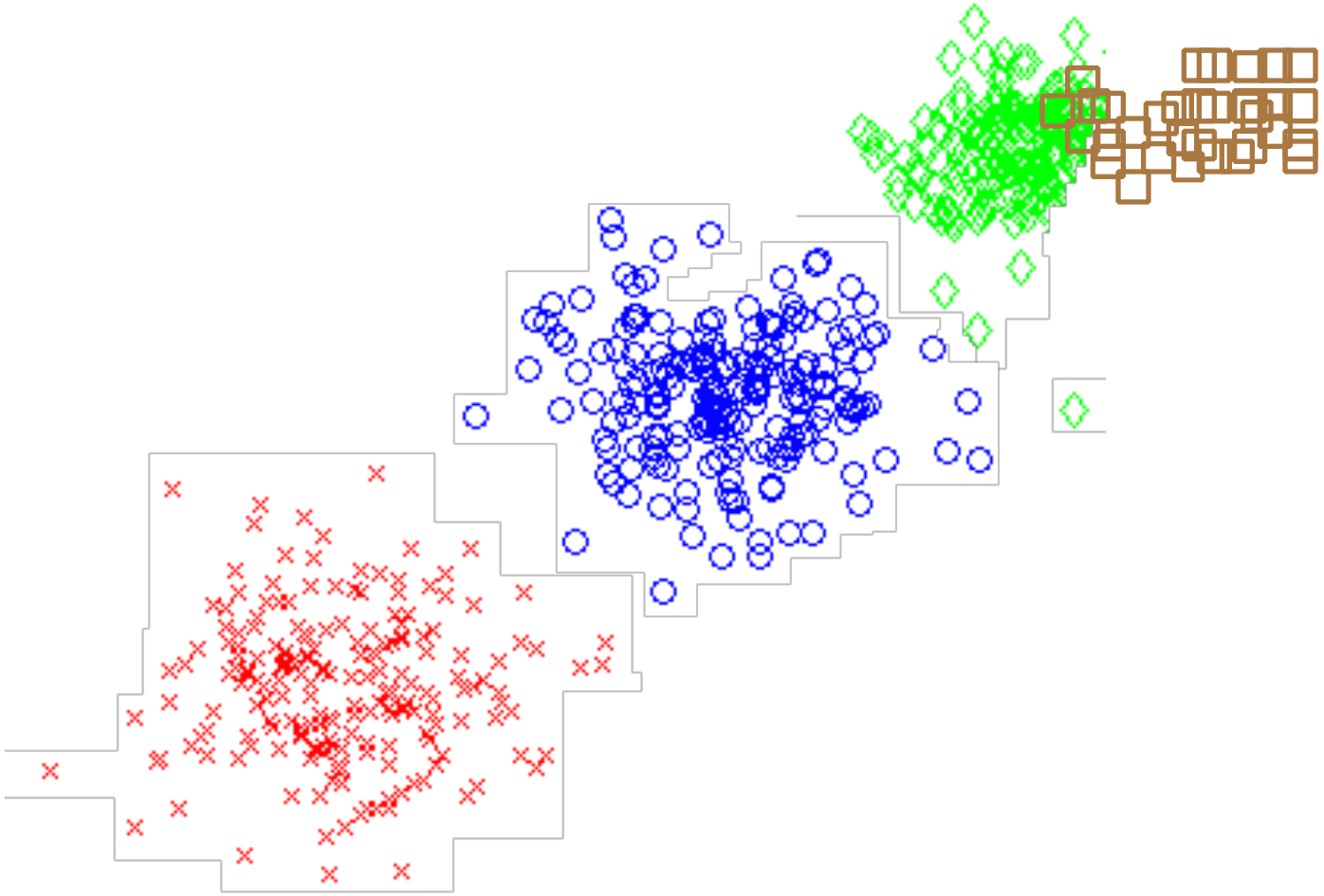
Part Of Vector Space (22-24. April 2016)



Hashtags: Dynamic to Static

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

St. George Day's Celebration (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Leave EU Campaign Launch (22. April)

St. George Day's Celebration (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Leave EU Campaign Launch (22. April)

St. George Day's Celebration (23. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Leave EU Campaign Launch (22. April)

Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

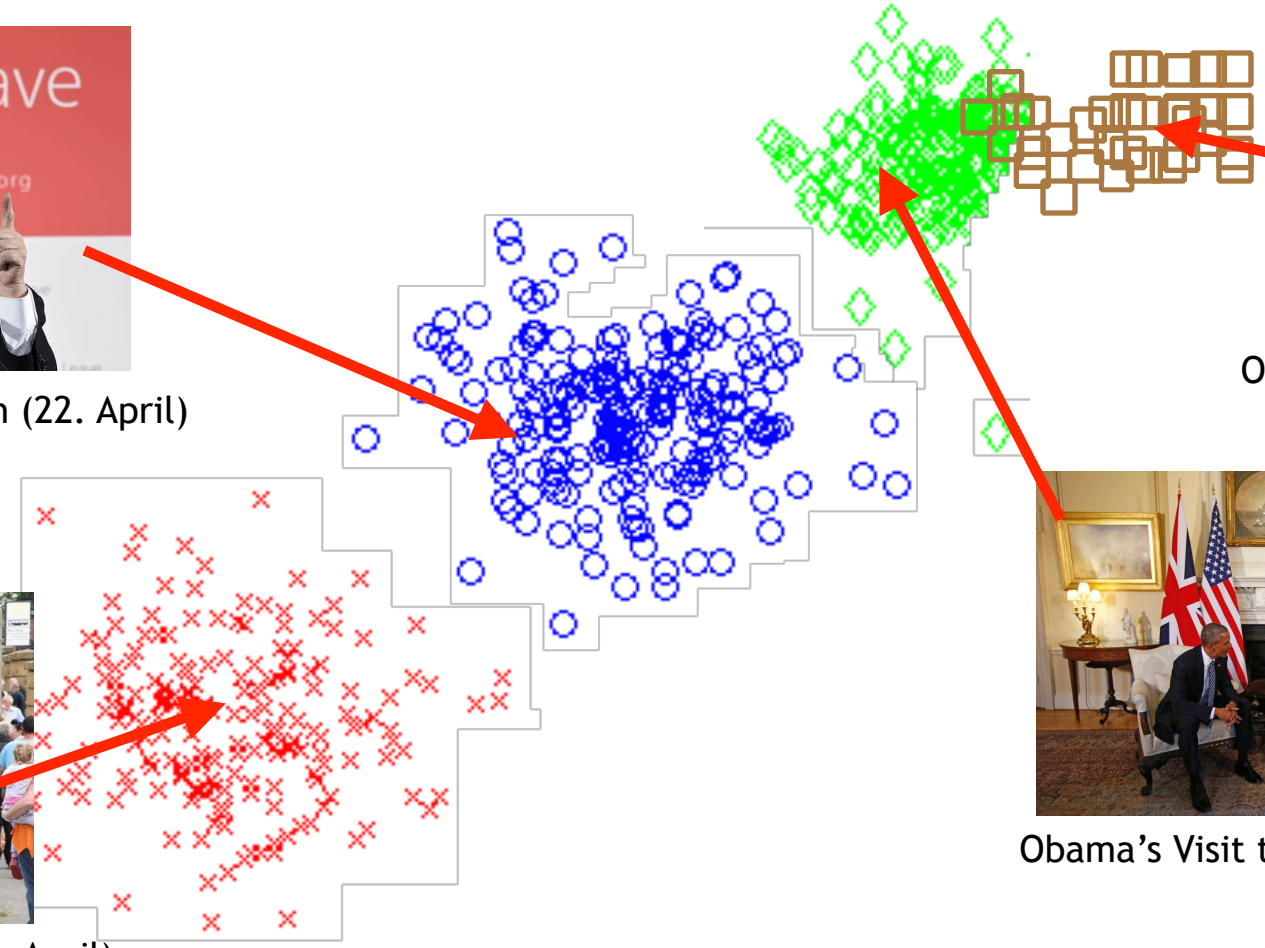
St. George Day's Celebration (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Leave EU Campaign Launch (22. April)

St. George Day's Celebration (23. April)



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Leave EU Campaign Launch (22. April)

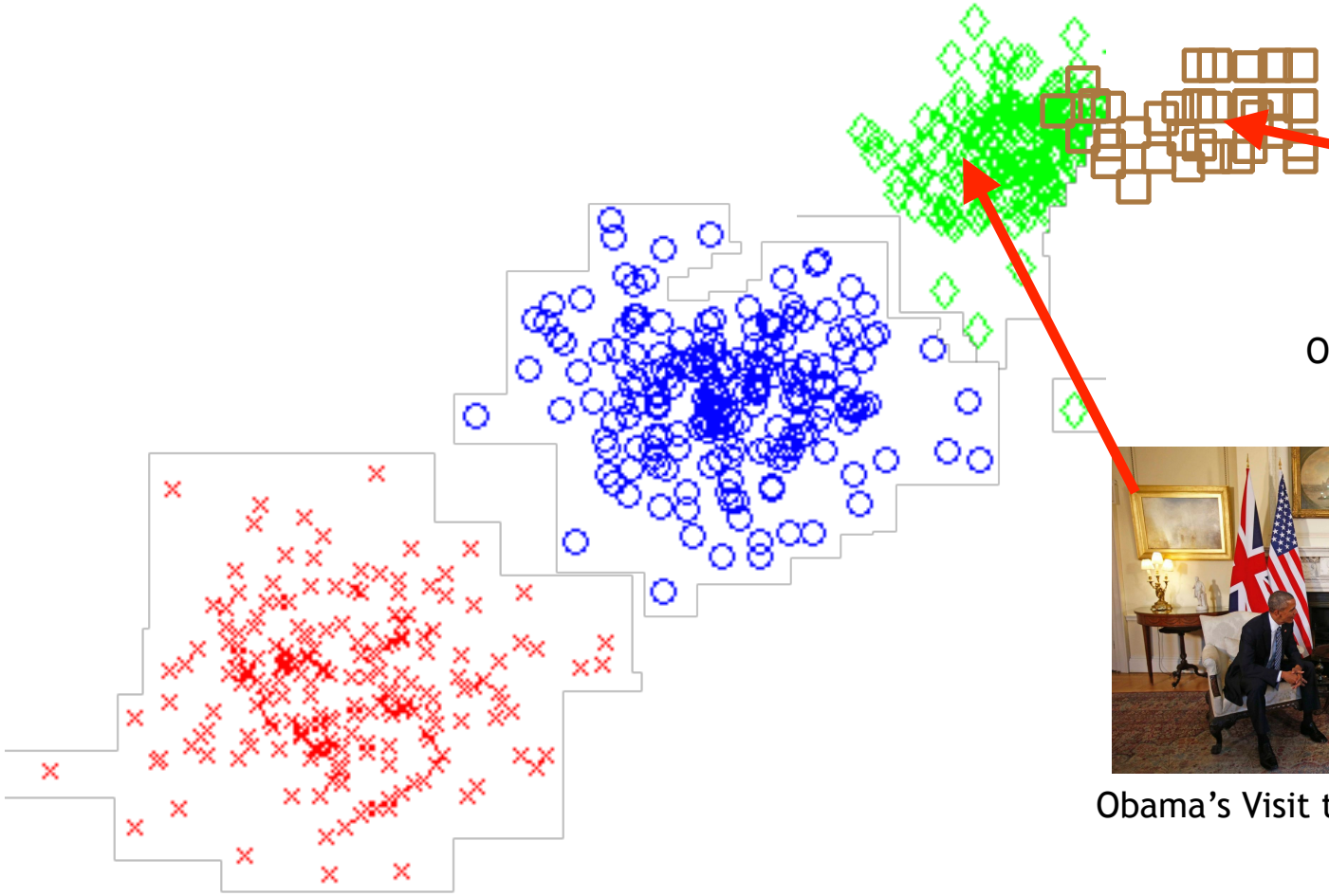
Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)



Nearest Neighbour Search

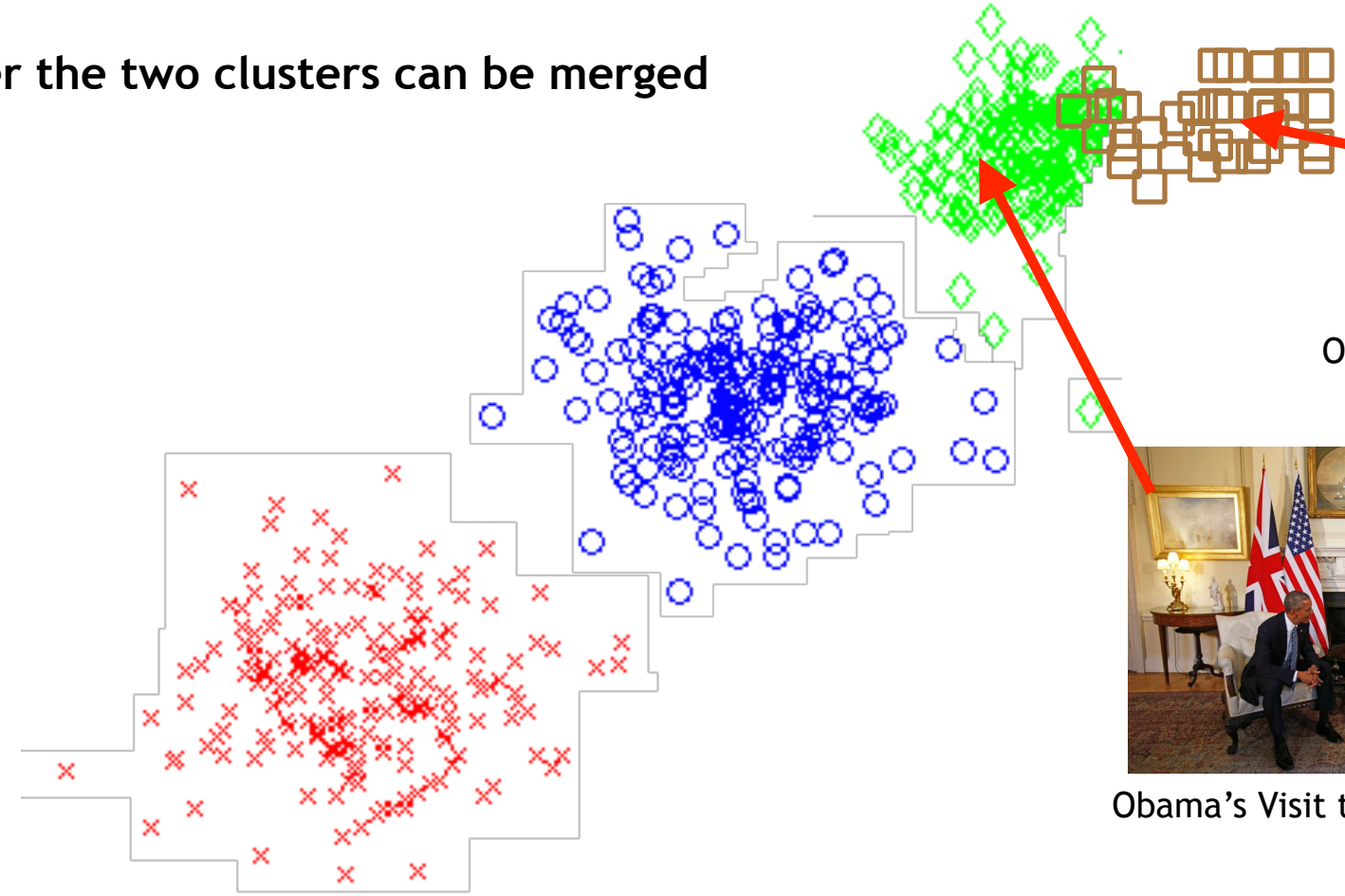
Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

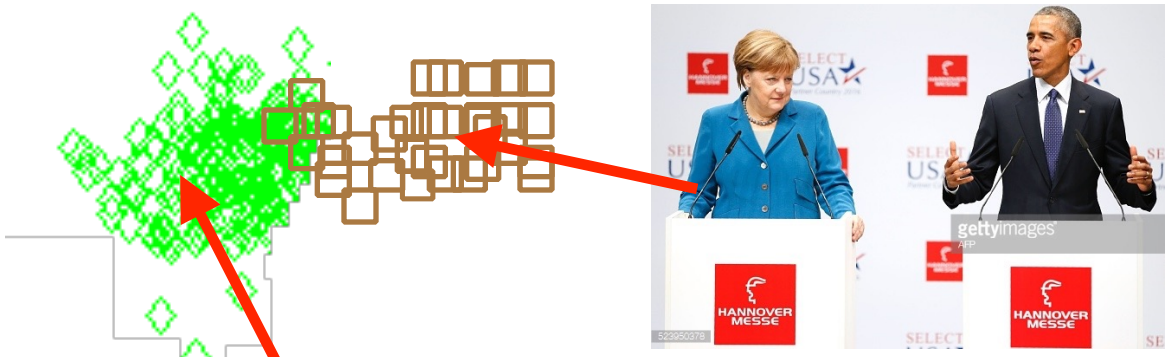
Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

Nearest Neighbour Search



Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

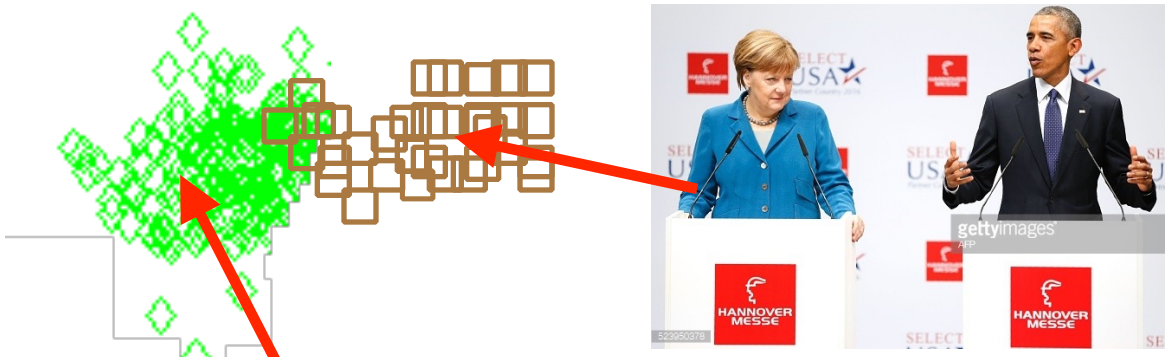
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common

Nearest Neighbour Search



Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

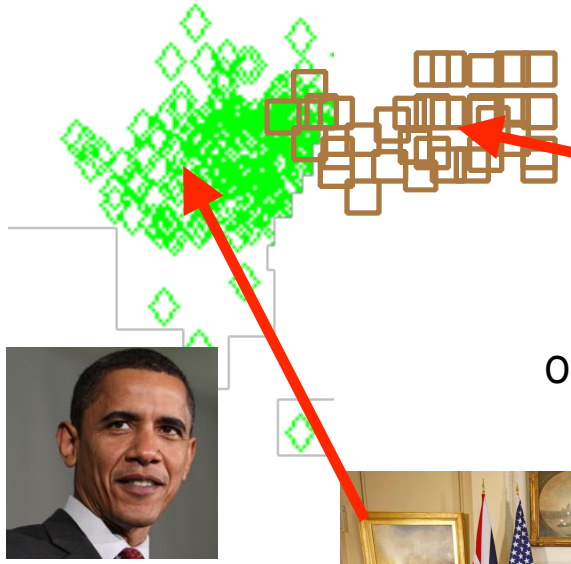
Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common

Nearest Neighbour Search



#obama

Obama's Visit to Hannover Messe (24. April)

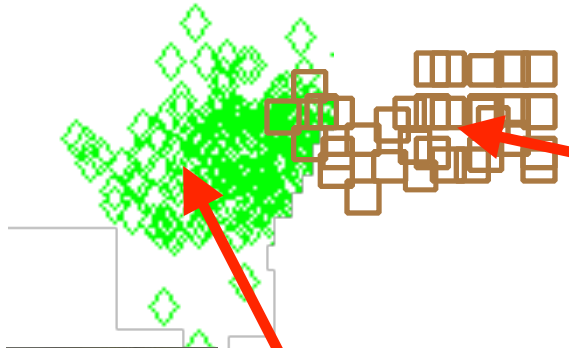
Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

#foreignvisit

#obama

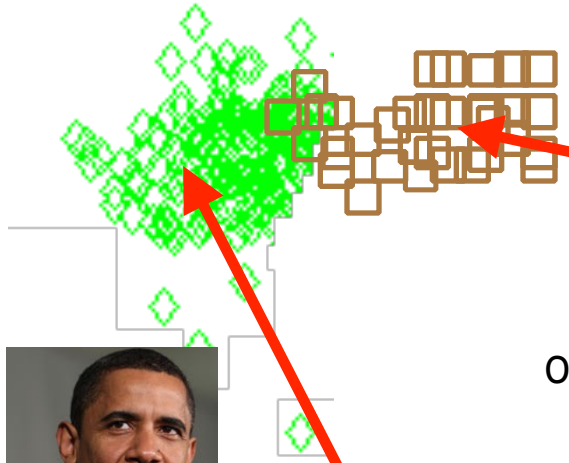
Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

#foreignvisit

#obama

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

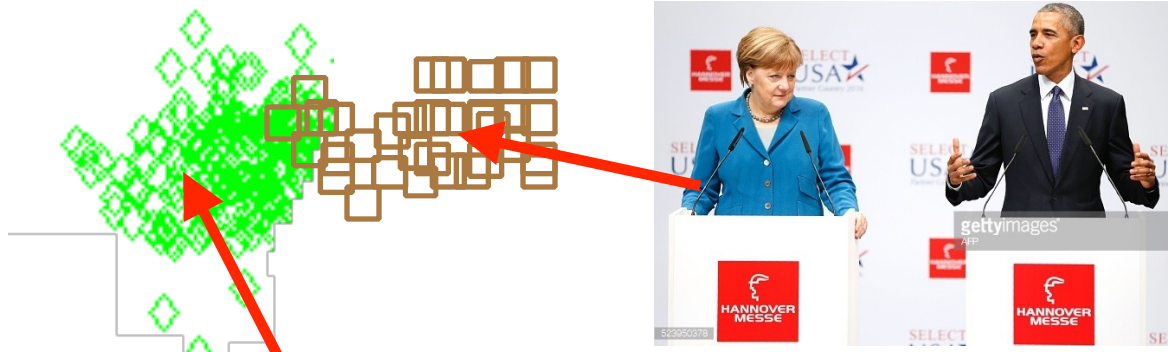
Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon

#foreignvisit

#obama

Nearest Neighbour Search



Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon

#foreignvisit

#obama

#brexit

Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon

#foreignvisit

#obama

#brexit

#hannovermesse

Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

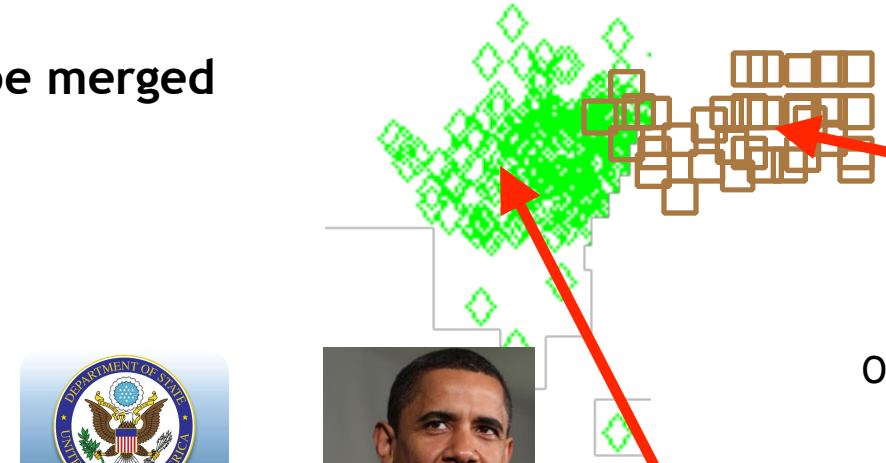
Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon
- If uncommon hashtags are within absorbance threshold: Merge Clusters



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

#obama

Obama's Visit to UK (23. April)

#foreignvisit

#brexit

#hannovermesse

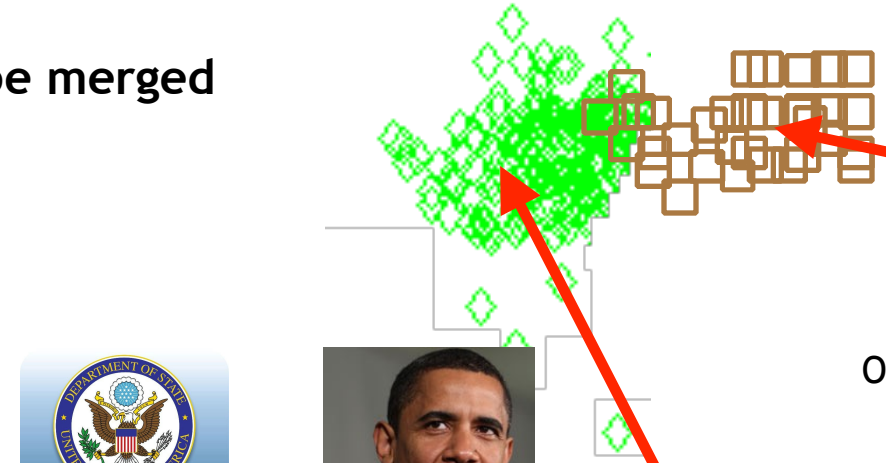
HANNOVER MESSE

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon
- If uncommon hashtags are within absorbance threshold: Merge Clusters
- Else: Keep clusters separate



Nearest Neighbour Search

Obama's Visit to Hannover Messe (24. April)

#obama

#foreignvisit

#brexit

#hannovermesse

HANNOVER MESSE

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon
- If uncommon hashtags are within absorbance threshold: Merge Clusters
- Else: Keep clusters separate

Separate Clusters ⇒ Separate Events

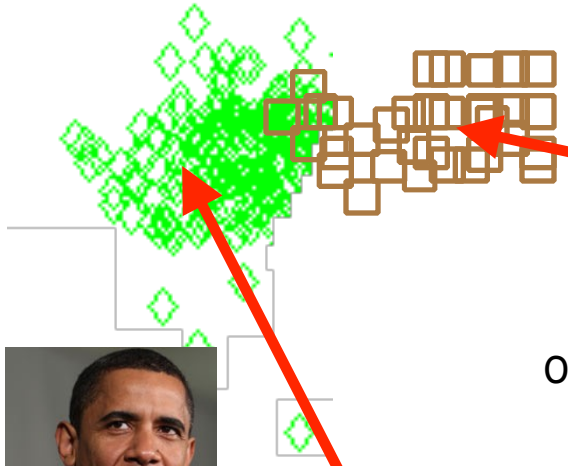
#foreignvisit

#obama

#brexit

#hannovermesse

Nearest Neighbour Search



Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Hashtag Clustering

Part Of Vector Space (22-24. April 2016)

Check whether the two clusters can be merged

- Check which hashtags are common
- Common hashtags get absorbed
- Check which hashtags are uncommon
- If uncommon hashtags are within absorbance threshold: Merge Clusters
- Else: Keep clusters separate

Separate Clusters ⇒ Separate Events

Same Cluster ⇒ Same Event

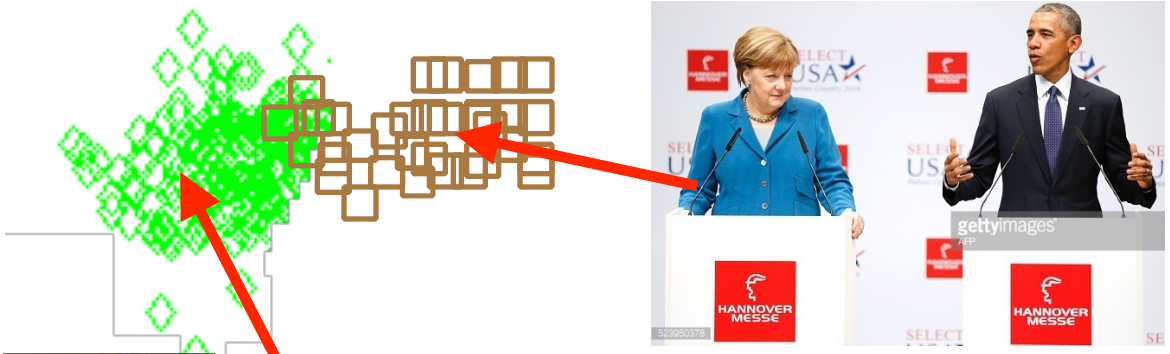
#foreignvisit

#obama

#brexit

#hannovermesse

Nearest Neighbour Search



Obama's Visit to Hannover Messe (24. April)

Obama's Visit to UK (23. April)

Event Ranking

Ranking Factors

- Burstiness

  - Burst Events: Events exhibiting high popularity during very short period of time

Event Ranking

Ranking Factors

- Burstiness

  - Burst Events: Events exhibiting high popularity during very short period of time

Kröpcke, Hannover

Event Ranking

Ranking Factors

- Burstiness

  - Burst Events: Events exhibiting high popularity during very short period of time

Event Ranking

Ranking Factors

- Burstiness

  - Burst Events: Events exhibiting high popularity during very short period of time

Shopping in Kröpcke

Demonstrations in Kröpcke

Event Ranking

Ranking Factors

- Burstiness

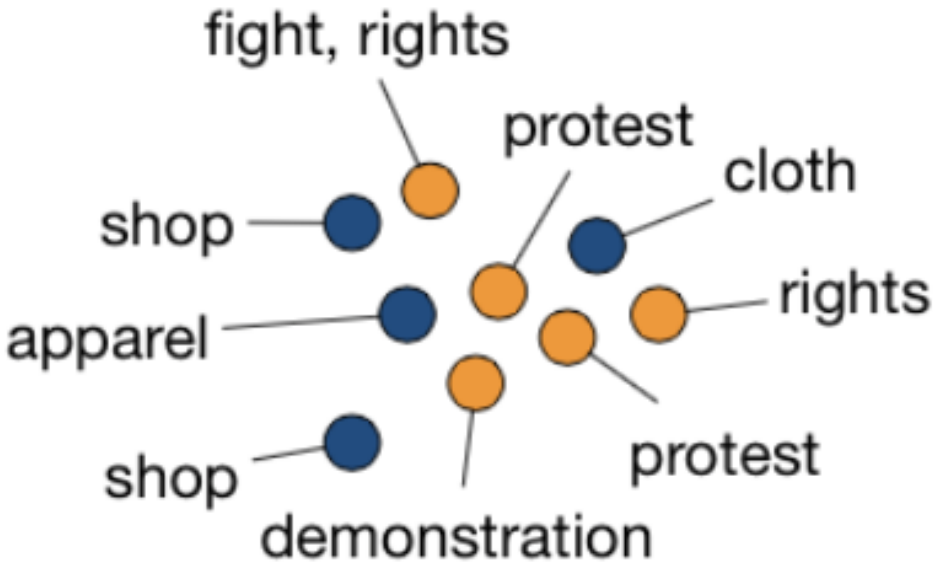
  - Burst Events: Events exhibiting high popularity during very short period of time

Event Ranking

Ranking Factors

- Burstiness

  - Burst Events: Events exhibiting high popularity during very short period of time



Tweets around Kröpcke, Hannover

Event Ranking

Ranking Factors

- Burstiness

  - Model historical popularity as Gaussian Distribution

  - p_t = popularity in time frame t, μ = mean, σ = standard deviation

$$burstiness(e) = \frac{p_t - \mu}{\sigma}$$

- Localness

  - Similar to Burstiness: Region instead of time

$$localness(e) = \frac{p_r - \mu}{\sigma}$$

- Popularity

  - freq(e) = frequency of event e, N= Total number of tweets

$$popularity(e) = \frac{freq(e)}{N}$$

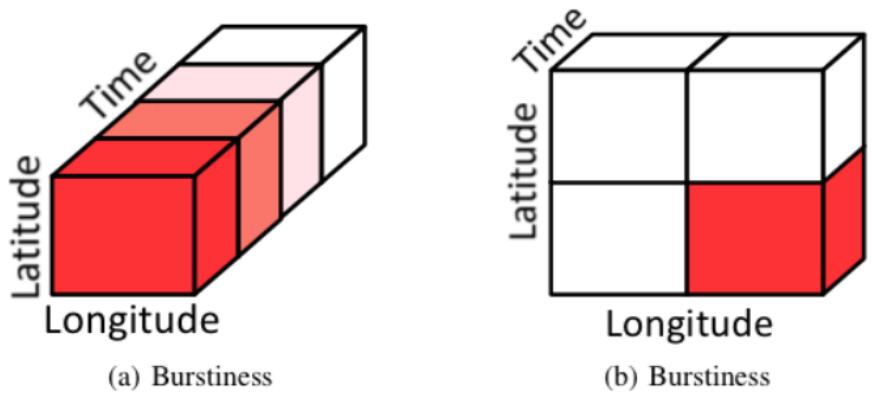


Fig. 4. Localness[2]

Event Ranking

Ranking Score

- Total Ranking Score

$$score(e) = \sum_{i=1}^{k} w_i score(h_i)$$
$$= \alpha \sum_{i=1}^{k} w_i pop(h_k) + \beta \sum_{i=1}^{k} w_i burst(h_k) + \gamma \sum_{i=1}^{k} w_i local(h_k)$$
$$= \alpha \cdot pop(e) + \beta \cdot burst(e) + \gamma \cdot local(e)$$

- Values of α, β, and γ varied according to the interest of the users

- For local events, higher value of γ when compared to α and β

- Popular events of an year: higher value of α when compared to β and γ

Experimental Study

Baselines

- TwitterMonitor[4]
  - Classify tweets on basis of important keywords
  - Finds events by clustering keywords with high burstiness
- SCMA[5]
  - Batch clustering of keywords : Offline mode
  - Incremental clustering of each new tweet that comes in on basis of keywords
- SUMBLR[6]
  - Batch clustering of tweets instead of keywords
  - Incremental clustering of each new tweet that comes in

Experimental Study

Qualitative Evaluation

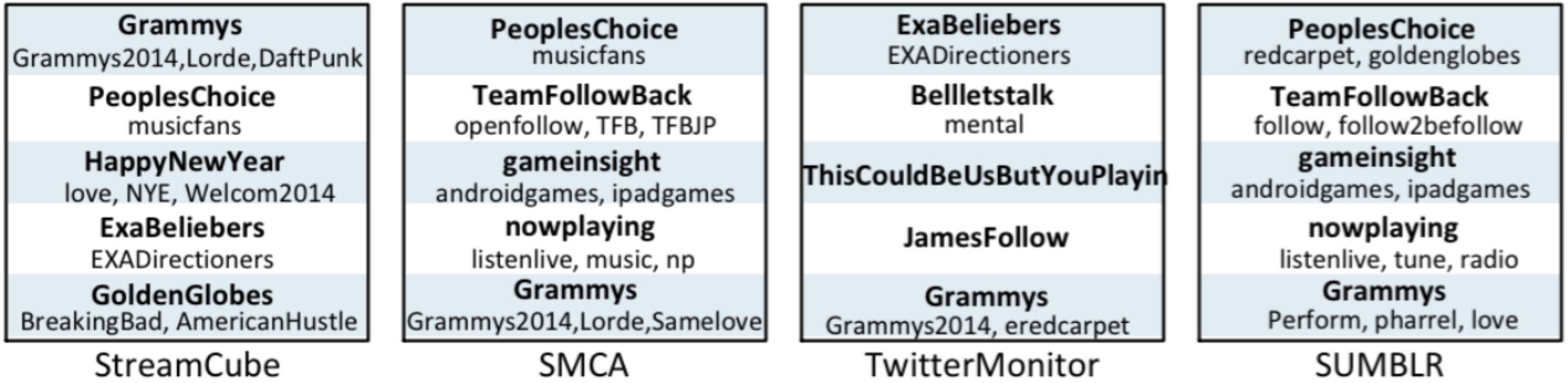


Fig. 5. Detected events by different methods in January 2014[2]

Experimental Study

Qualitative Evaluation

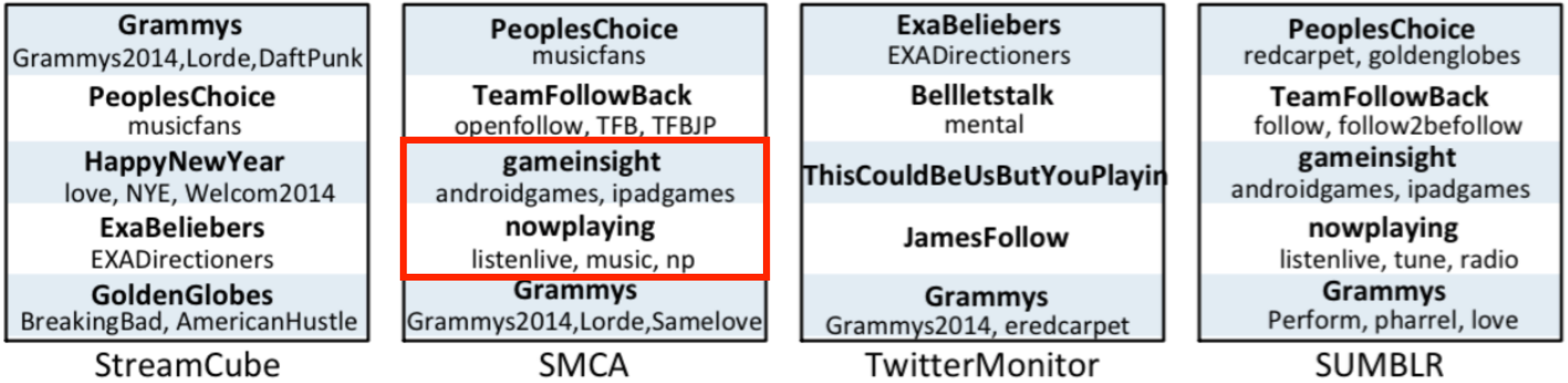


Fig. 5. Detected events by different methods in January 2014[2]

Experimental Study

Qualitative Evaluation

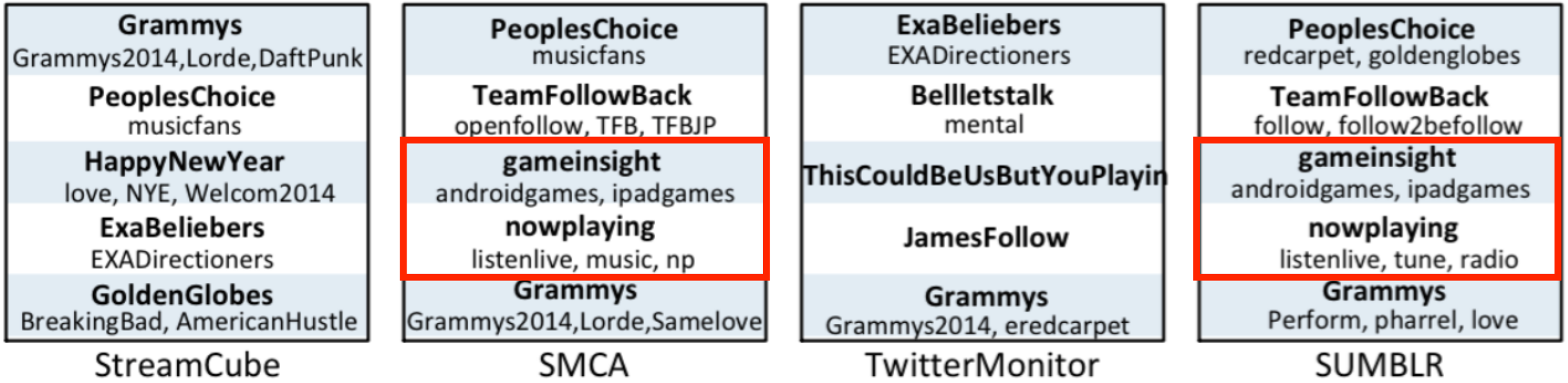


Fig. 5. Detected events by different methods in January 2014[2]

Experimental Study

Qualitative Evaluation

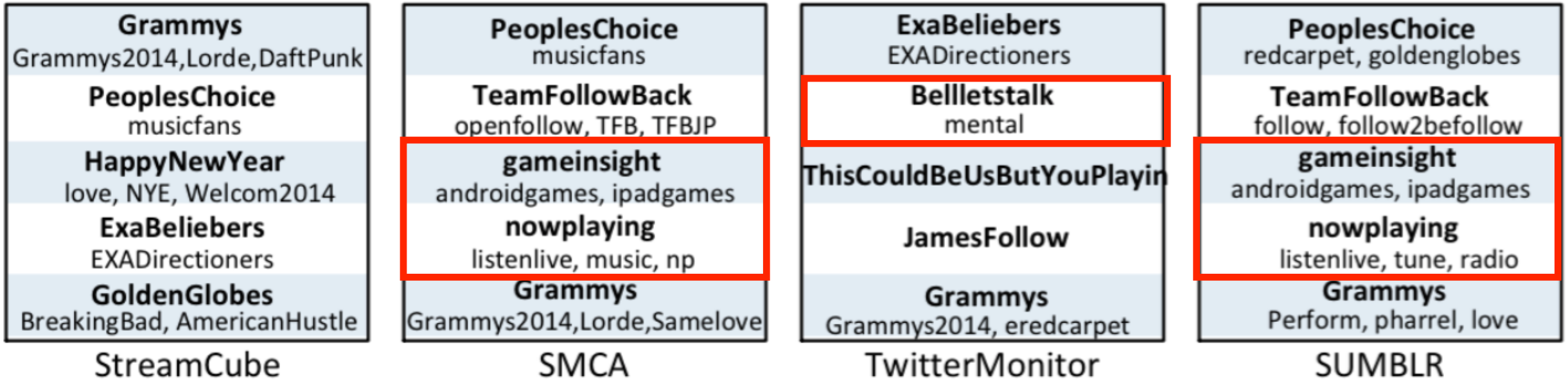


Fig. 5. Detected events by different methods in January 2014[2]

Experimental Study

Qualitative Evaluation

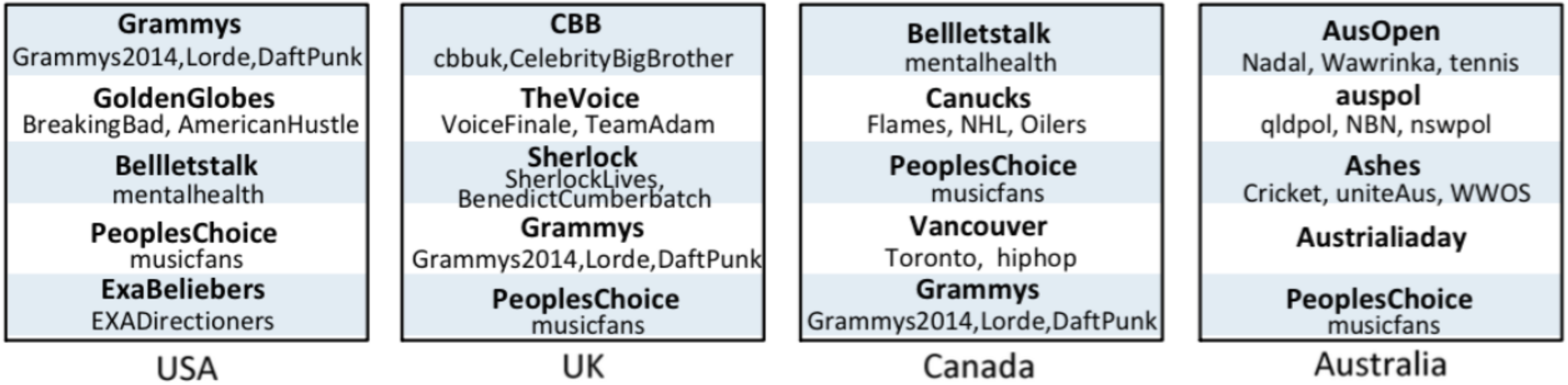


Fig. 6. Events in four different countries in January 2014[2]

Experimental Study

Qualitative Evaluation

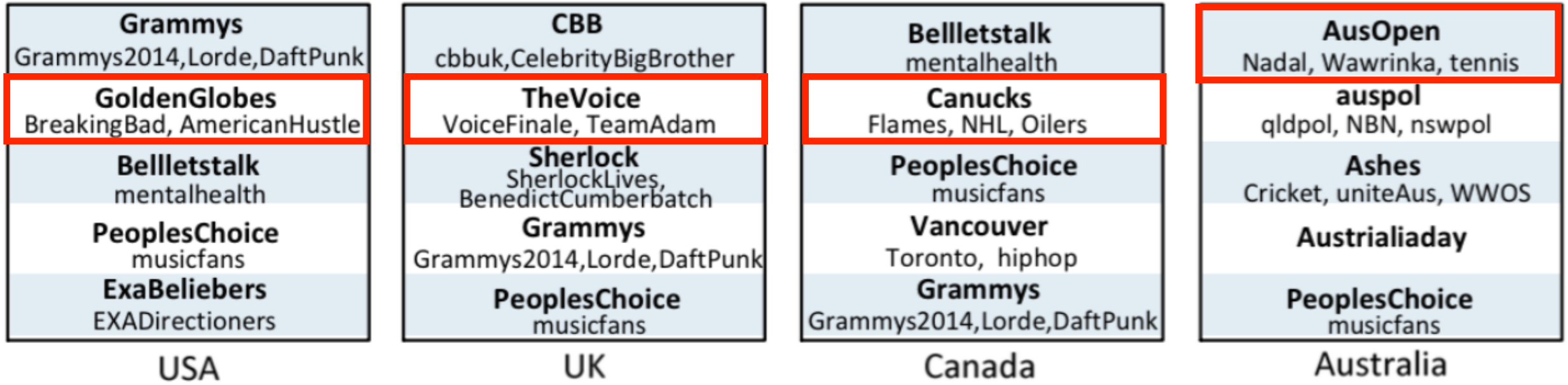


Fig. 6. Events in four different countries in January 2014[2]

Experimental Study

Qualitative Evaluation

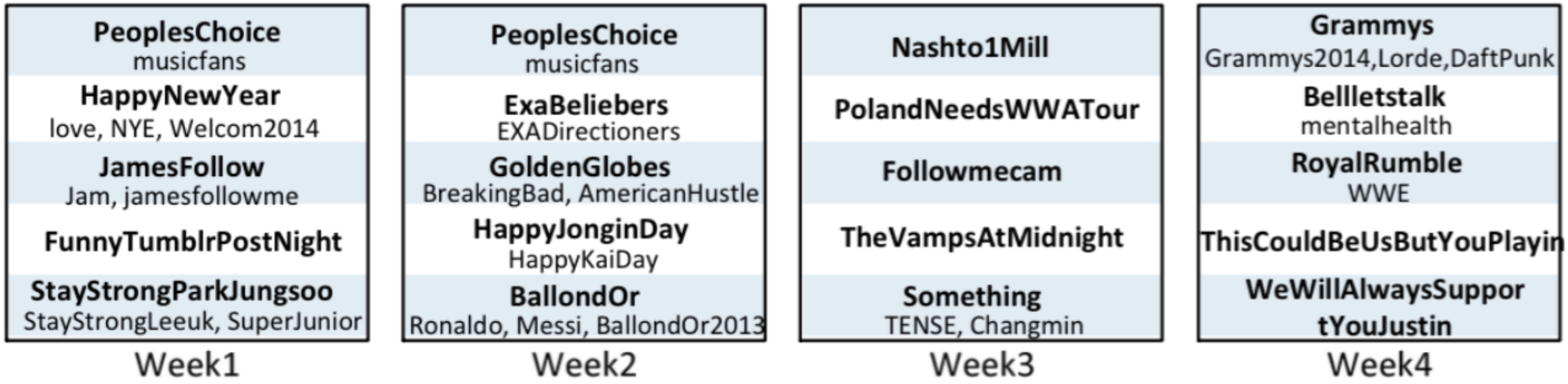


Fig. 7. Events in four weeks of January 2014[2]

Experimental Study

Qualitative Evaluation

Fig. 7. Events in four weeks of January 2014[2]

Experimental Study

Quantitative Evaluation

- Crowdsourcing: 10 People ranked top-10 events in each Candidate Set
- Average Precision (AP)

$$AP = \frac{\sum_{k=1}^{n} precision@k \times isTopEvent(e)}{the\ number\ of\ positive\ events}$$

- Mean Average Precision

$$MAP = \frac{\sum_{i=1}^{N} AP_i}{N}$$

Metric	SUMBLR	SMCA	TMONITOR	STREAMCUBE
MAP	0.511	0.523	0.608	**0.634**

Table 1. Ranking Quality[2]

Experimental Study

Scalability

- **TwitterMonitor:** Slowest as clusters keywords instead of hashtags
- **SCMA & SUMBLR:** Better as batch clustering of tweets performed initially followed by incremental clustering
- **STREAMCUBE:** Shortest running time due to single pass, nearest neighbour search algorithm

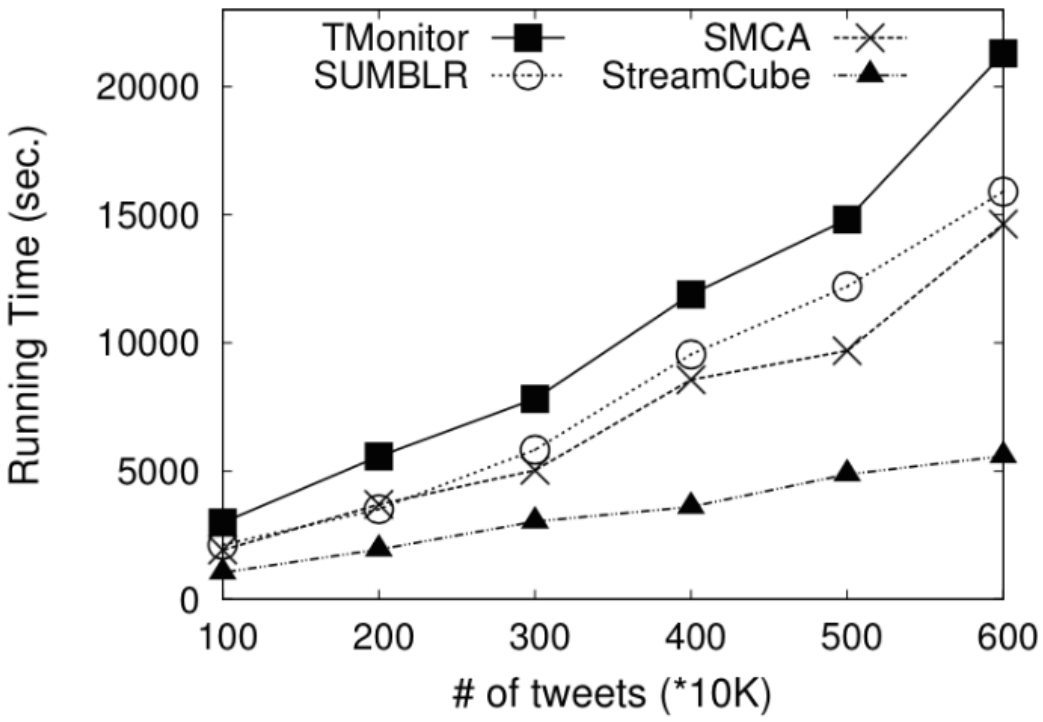


Fig. 8. Scalability[2]

Experimental Study

Memory Usage

- **TwitterMonitor:** Maintains similarity matrix for all pair of keywords, Consumes most memory
- **SCMA & SUMBLR:** SUMBLR performs better than SMCA as its able to remove outdated clusters
- **STREAMCUBE:** Least memory usage due to hashtag clustering performed at early stage and only current 6 hr time frame kept in Memory

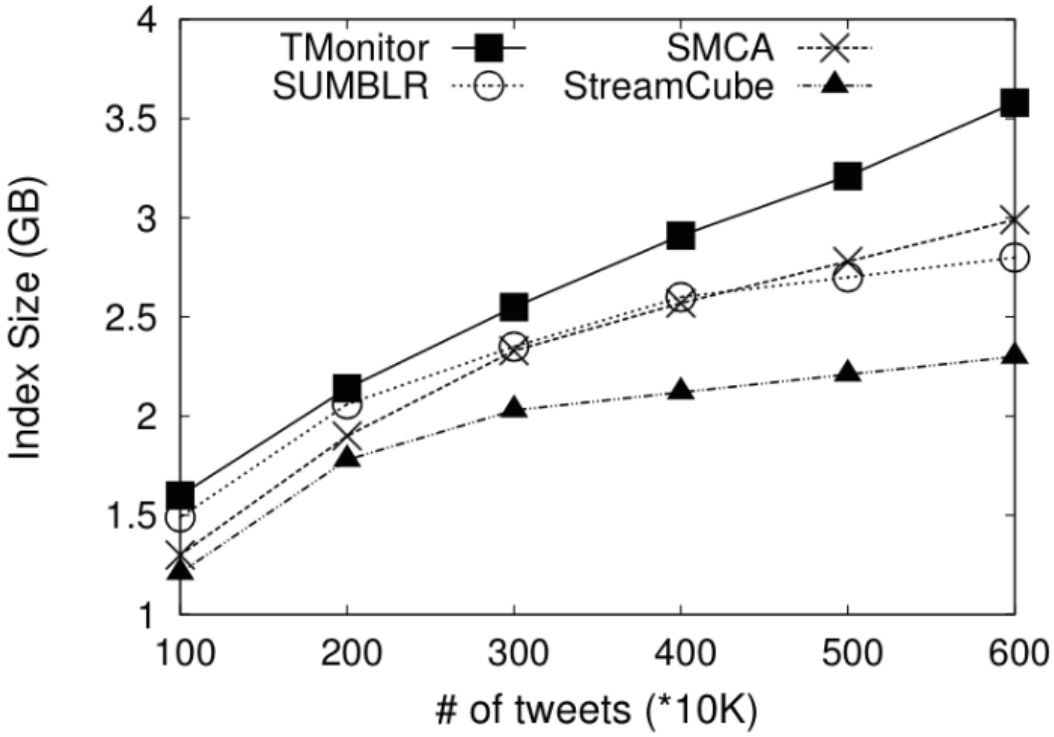


Fig. 9. Memory Usage[2]

Conclusion

- Explore events along different space and time granularity
- Takes into account dynamic nature of hashtags while hashtag clustering
- Considers all tweets, assigns geo-location to tweets not geo-tagged
- **Drawback 1:** Less than 2 per cent of tweets are geo-tagged
- **Solution 1:** Consider all tweets for hashtag clustering but only geo-tagged tweets for event ranking as done by EvenTweet[7] system
- **Drawback 2:** Uses complete global space
- **Solution 2:** Begin with local space (eg. USA) and gradually expand as tweets come in from new countries

Conclusion

- Finest granularity of space restricted to district
- **Possible Extension 1:** Allow any space granularity by specifying co-ordinates of the region in the form of bounding rectangle as done by EvenTweet[7] system

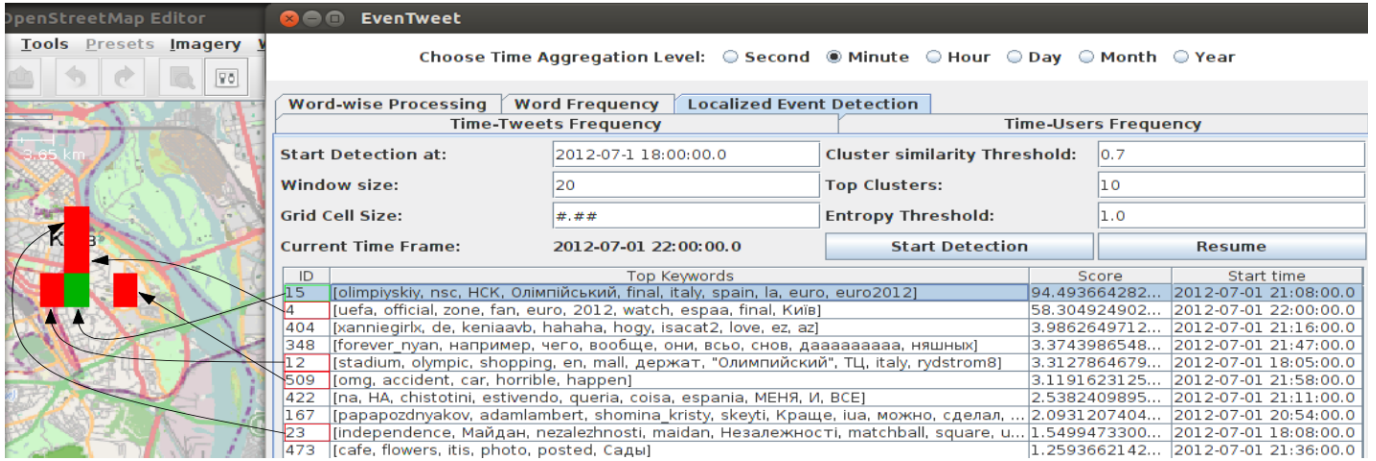


Fig. 10. EvenTweet Interface[7]

- **Possible Extension 2:** Allow users to explore events along Topic dimension
- **Possible Extension 3:** Alert mechanism to regularly push out event information to users

References

[1] Salaheldeen, H.; Nelson, M. L.: “Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?” JCDL, Washington, USA, 2012.

[2] W. Feng et al., "STREAMCUBE: Hierarchical Spatio- temporal hashtag clustering for event exploration over the Twitter stream," 2015 IEEE 31st International Conference on Data Engineering, Seoul, 2015, pp. 1561-1572.

[3] J. Han, Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[4] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,” in SIGMOD Conference, 2010, pp.1155–1158.

[5] O. Tsur, A. Littman, and A. Rappoport, “Efficient clustering of short messages into general domains.” in ICWSM, E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, Eds.,2013.

[6] L. Shou, Z. Wang, K. Chen, and G. Chen, “Sumblr: continuous summarization of evolving tweet streams,” in SIGIR, 2013, pp. 533–542.

[7] Michael Gertz et. al. “EvenTweet: Online Localized Event Detection from Twitter” 39th International Conference on Very Large Data Bases, August 26-30th, Trento, Italy.

Discussion