Social Media & Network Analytics

Event Detection

Overview

- Introduction and Motivation for Event Detection
- Definition of Event
- Trend Detection (Time)
- General Event Detection

Acknowledgements

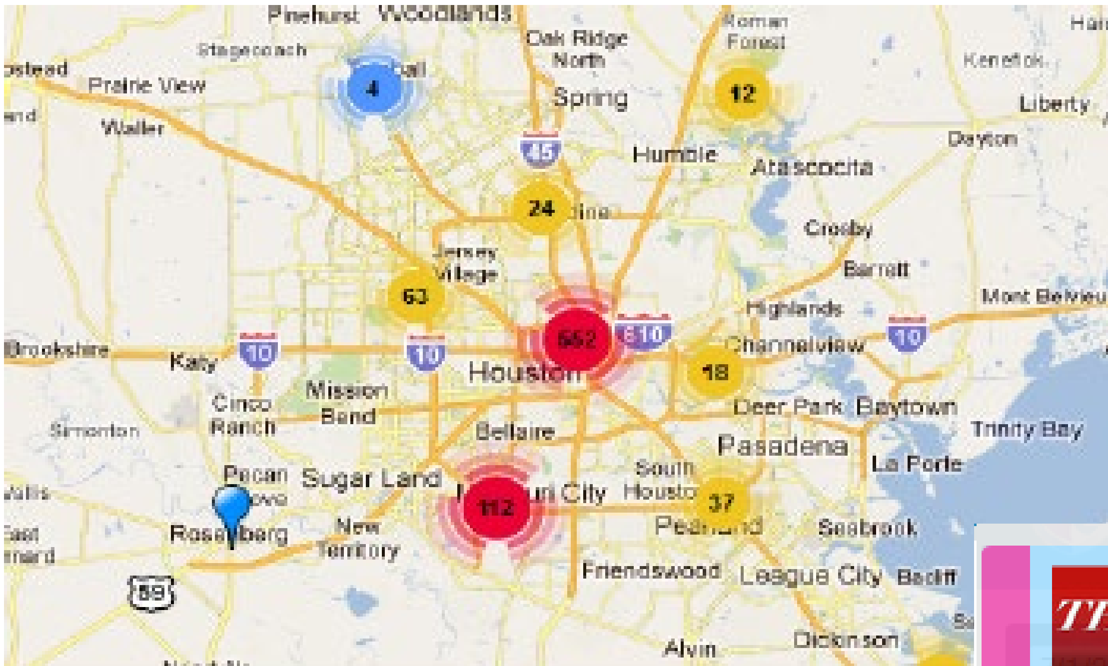
- Part of these slides are based on:
  - Kostas Tsiousiouliklis presentation on “Trend and Event Detection in Social Streams”

Introduction

- We are surrounded by events

Social Media based Event Detection

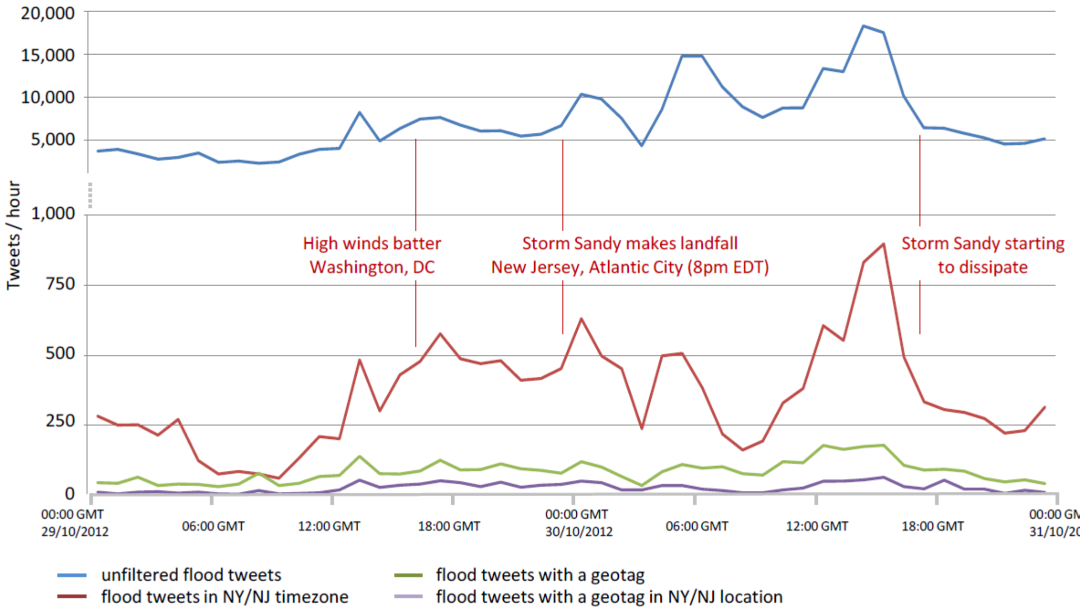
- Can we use social media to help identify them?

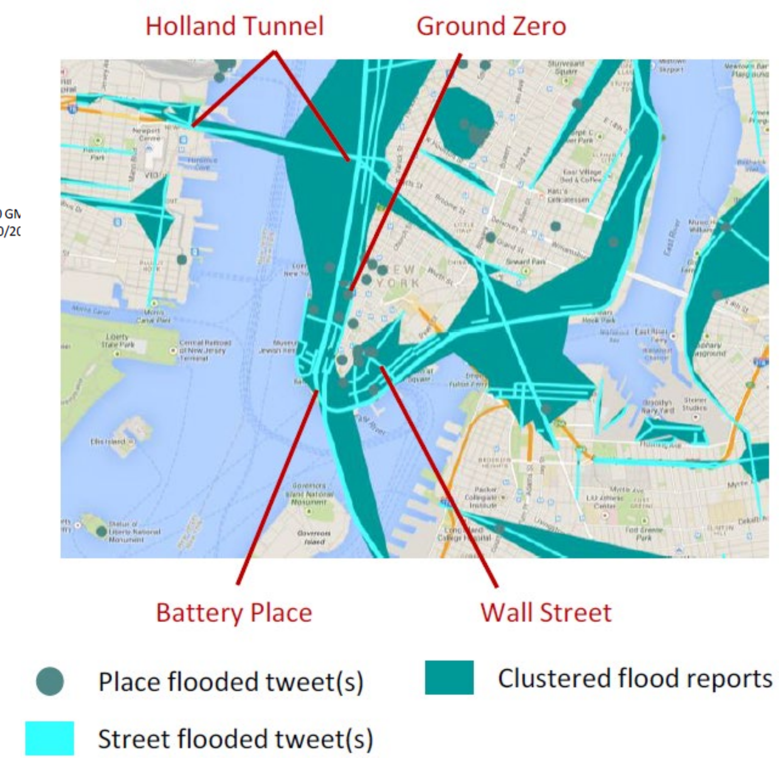


From Li et al, “Tedas: a Twitter based Event Detection and Analysis System”

Social Media based Event Detection

- Can we use social media to help identify them?





From Middleton et al, “Real-time Crisis Mapping of Natural Disasters using Social Media

Applications

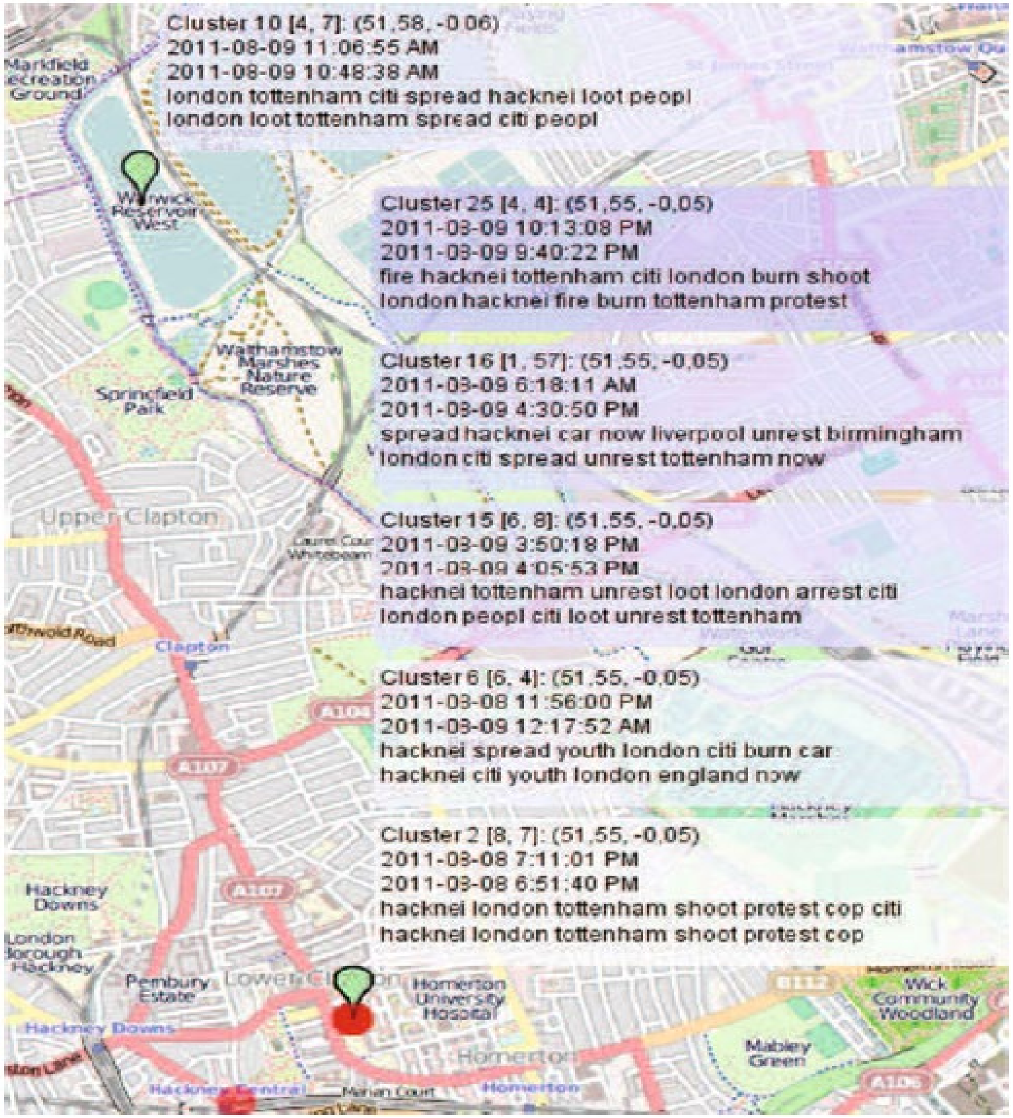
- Crowd-sourced, sometimes more up to date and widespread then traditional sources of events
- Emergency/Crisis management
  - What, where, when, who
  - Traffic accident, natural disasters
- How situation evolves with time
- Reaction
  - VW diesel emissions scandal
- Enhance decision making

Overview

- Introduction and Motivation for Event Detection
- Definition of Event
- Trend Detection (Time)
- General Event Detection

Event Definition

- Collins Dictionary:
  - An **event** is something that happens, especially when it is unusual or important.
- Event:
  - who, when, where, what (4 w's)



From Bouchachia et al, "Social media for crisis management"

Social Media Event Management Tasks

- Event detection
- Event tracking
- Event summarisation

Social Media Event Detection

- Use of social media to detect significant events
- Input:
  - Collection of social media data that arrives over time
  - Tweets, Instagram posts, videos
  - Generally unstructured data
- Output:
  - Event (who, when, where, what)
  - Not always all 4 elements
  - For social media, typically focus is when, where and sometimes what

Social Media Event Detection Taxonomy

- Planned vs unplanned

- Entity vs situation
- Local vs global

- Retrospective vs online (real-time)

2016-2019

Now

Social Media Event Detection Taxonomy

- Topic Specific
  - Track interest about brands, products, entities (e.g., Apple)
  - Planned, known events, interested in effect (e.g., G7 meeting)

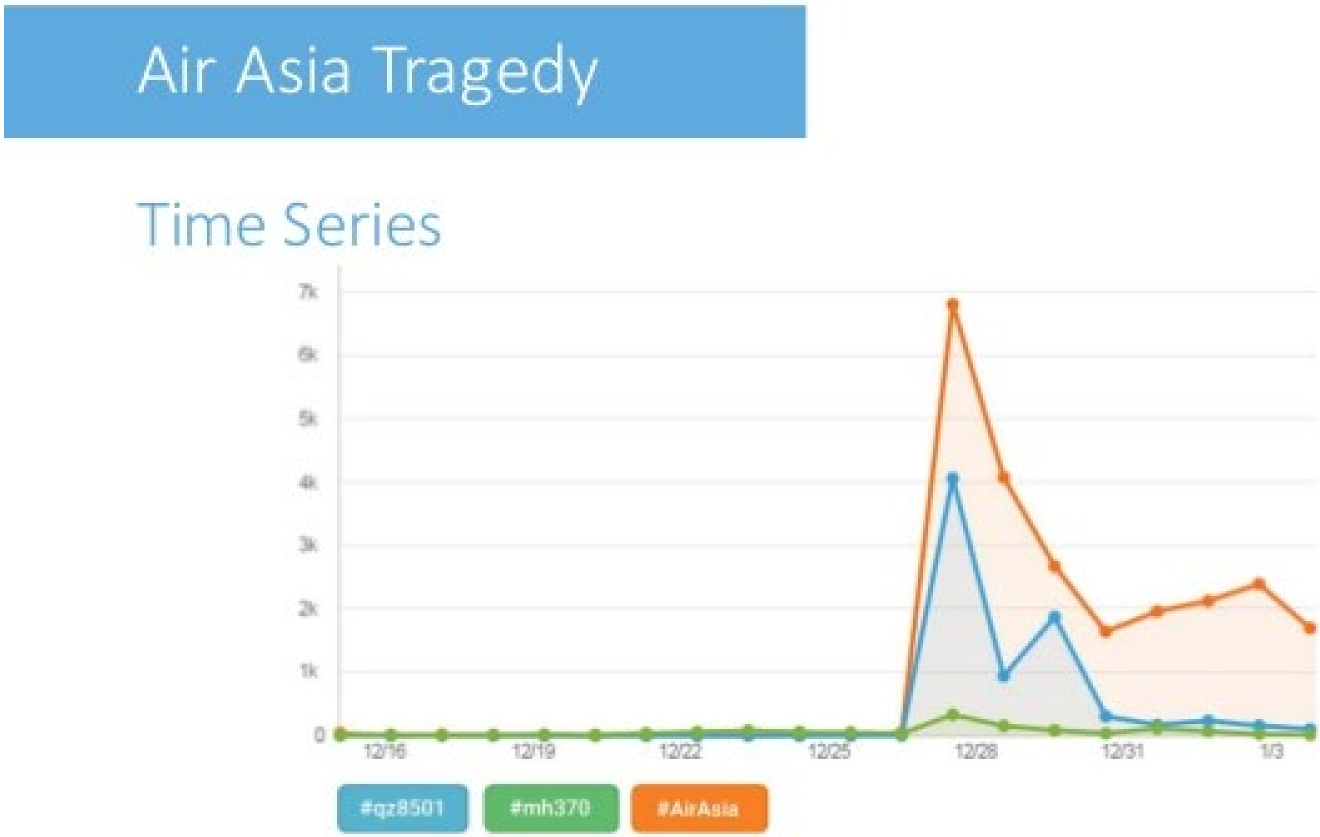
- Unknown Events
  - Unplanned, breaking news (e.g., Amazon fires, HK protests)
  - Discover events - when, where, what.

Overview

- Introduction and Motivation for Event Detection
- Definition of Event
- Trend Detection (Time)
- General Event Detection

Trend Detection (Time)

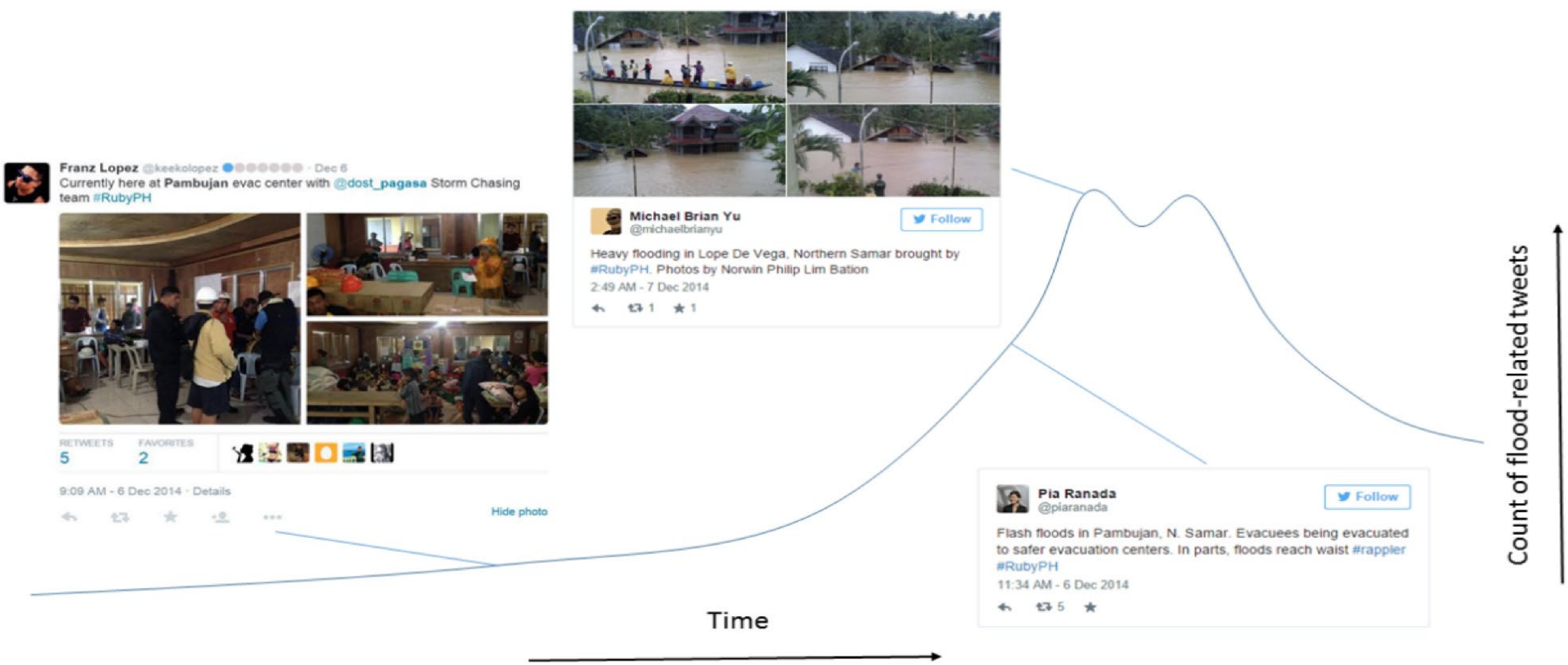
- Interested in detecting when this topic is trending or has breaking news



From Lukas Masuch, “Trend detection and Analysis on Twitter”, slideshare

Trend Detection (Time)

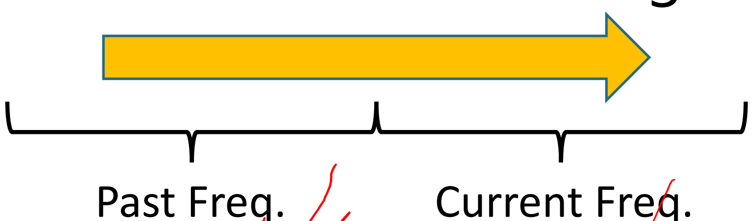
- Generally look for differences of keywords volume over time
- Detect bursts
  - Typically significant events will cause burst like behaviour
  - In raw social media activity/posts (volume)

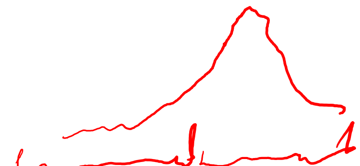


From Jongman et al., “Early Flood Detection for Rapid Humanitarian Response”

Simple Frequency Ratios

- Capture relative growth
- Tokenise social media stream
    - Stream of words
    - Compute frequency
    - Compare some historical average vs current average





freq current

Term	Past Freq. (per time unit)	Current Freq. (per time unit)	Ratio
Bush fire	1	20	20
Hawks	50	500	10
PM	1,000	1,700	1.7
HK Protest	20,000	23,000	1.15

Current
freq

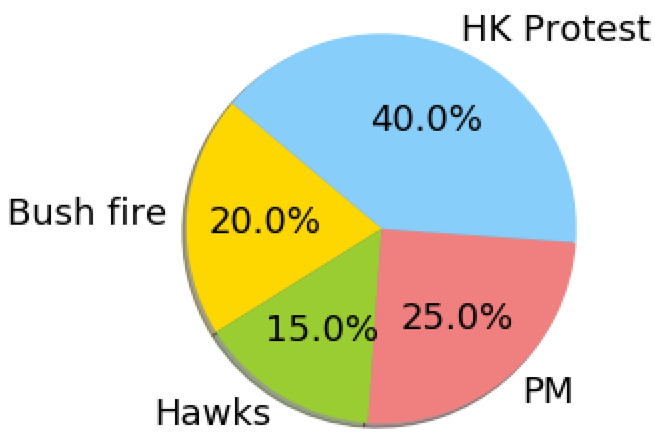
Issues?

- Words with low frequency get artificially inflated ratios
  - If word is new, past frequency is 0
    - E.g., #thisIsTheBestCourseEverAtRMIT
- Ratio
  - Which word is trending more (as an event)?
  - One that goes from 10 to 15, or 10,000 to 15,000?

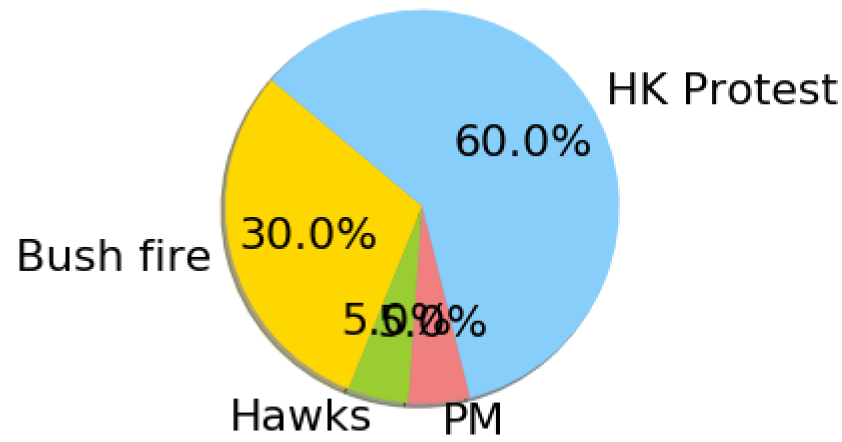
- Need better statistics to capture relative growth

Goodness of fit

- Assume that the terms are drawn independently at random from a static distribution, where each term has a fixed prior likelihood of been selected (multinomial distribution)



Expected distribution



Current/observed distribution

- Is the observed drawn from the same distribution?

Goodness of fit

- Common test of goodness of fit is the chi-squared test

sum across cat

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

expected

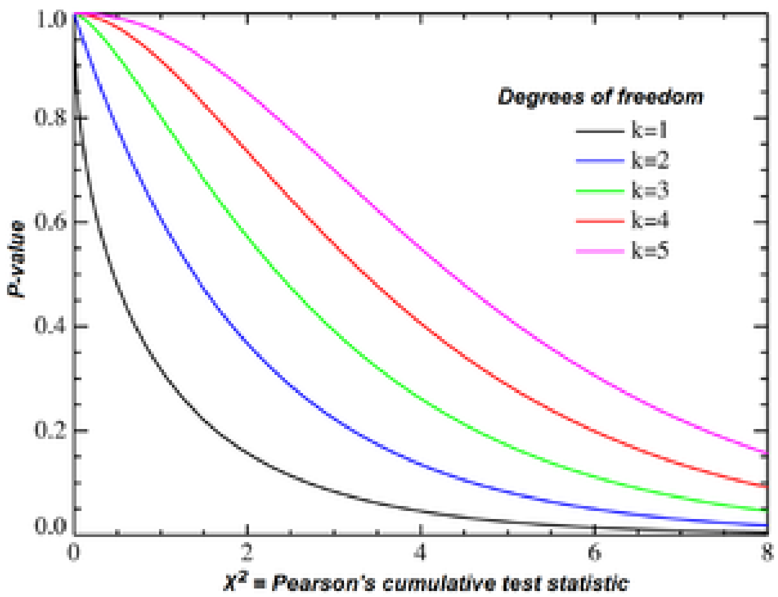
observed

	Previous (Expected)	Current (Observed)
Category 1	70	68
Category 2	30	32
Total	100	100

$$\chi^2 = \underbrace{\frac{(68-70)^2}{70}}_{\text{cat 1}} + \underbrace{\frac{(32-30)^2}{30}}_{\text{cat 2}} = 0.19$$

Goodness of fit

- A chi-square value of 0.19 corresponds to a p-value of 0.663
  - Using hypothesis testing at significance level 0.05, we can say the two distributions difference are not statistically significant (p-value > 0.05)
  - We cannot reject the Null hypothesis (the observed values fit the expected values)



From Wikipedia page on Chi-square test

Chi-Square test

- For burst/trend detection, use the chi-square value to determine burstiness of individual terms
- Example
  - Of N (large) total past terms, 20 where "bush fire". Of N present terms, 30 are "bush fire". 20 is expected frequency, 30 is observed frequency

$$\chi^2 = \underbrace{\frac{(30-20)^2}{20}}_{\text{BF}} + \underbrace{\frac{((N-30)-(N-20))^2}{N-20}}_{\text{all other terms}} = \frac{10^2}{20} + \cancel{\frac{10^2}{N-20}} \approx 5$$

  - Using approximation, if HK protest expected is 40, observed is 60:

$$\chi^2 \approx \frac{(60-40)^2}{40} = 10$$

  - So HK protest is more "bursty"

Chi Square Test

- In a nutshell:
  - If observed > expected, then burstiness score is:

    $$\frac{(O-E)^2}{E}$$

    otherwise 0
  - If E = 0:
    - Add one smoothing

      $$\frac{((O+1)-(E+1))^2}{E+1} = \frac{(O-E)^2}{E+1}$$

    - If low frequencies still dominate, use thresholds or Yate's correction:

      $$\frac{(|O-E|-0.5)^2}{E}$$

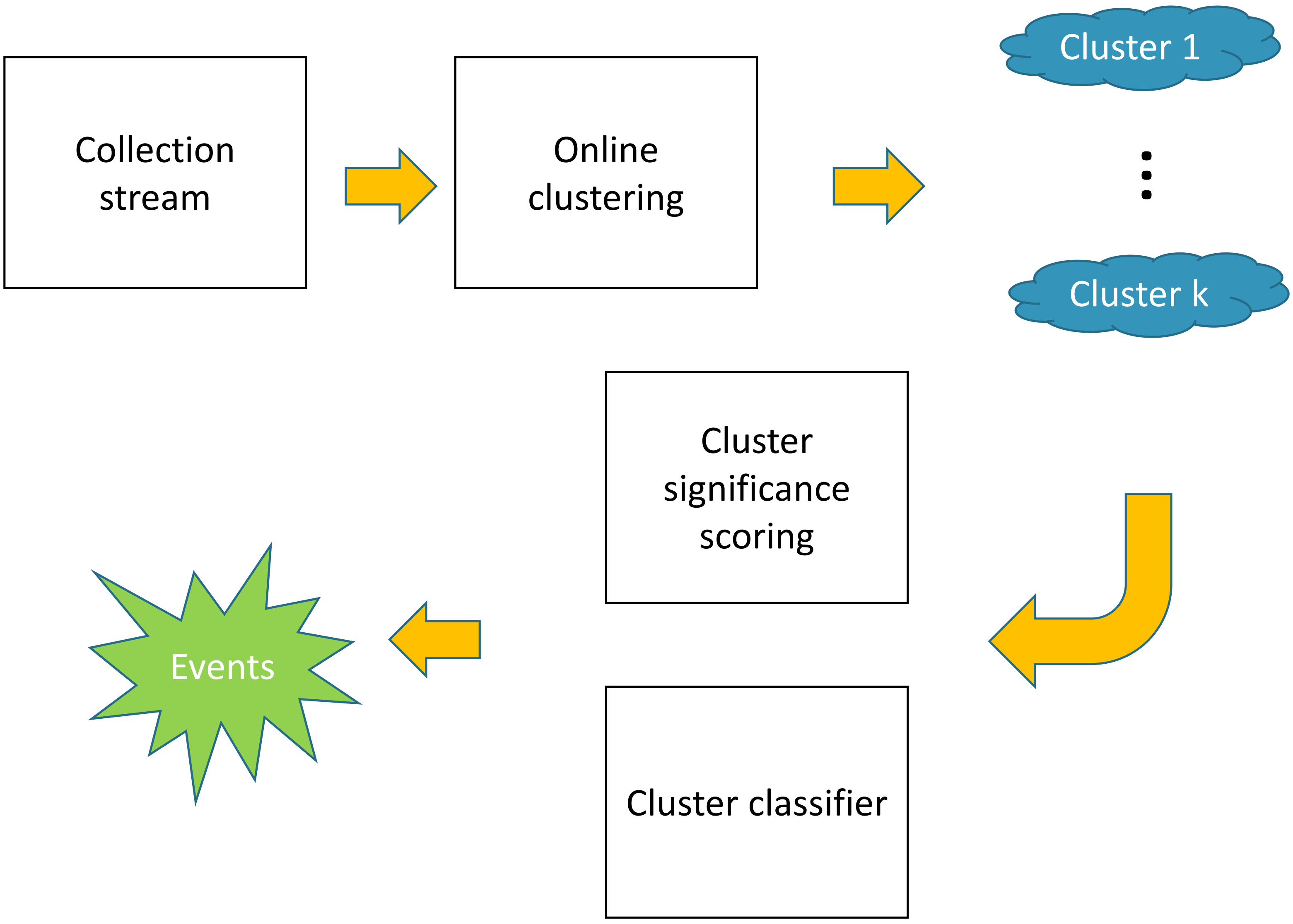
Other possibilities

- Regression to account for seasonality/periodicity + standard deviation + other factors
  - Expected model
- ARIMA type models
- Change point detection statistics
  - Control charts, CUSUM
- State based models
  - Kalman filters, Hidden markov models

Overview

- Introduction and Motivation for Event Detection
- Definition of Event
- Trend Detection (Time)
- General Event Detection

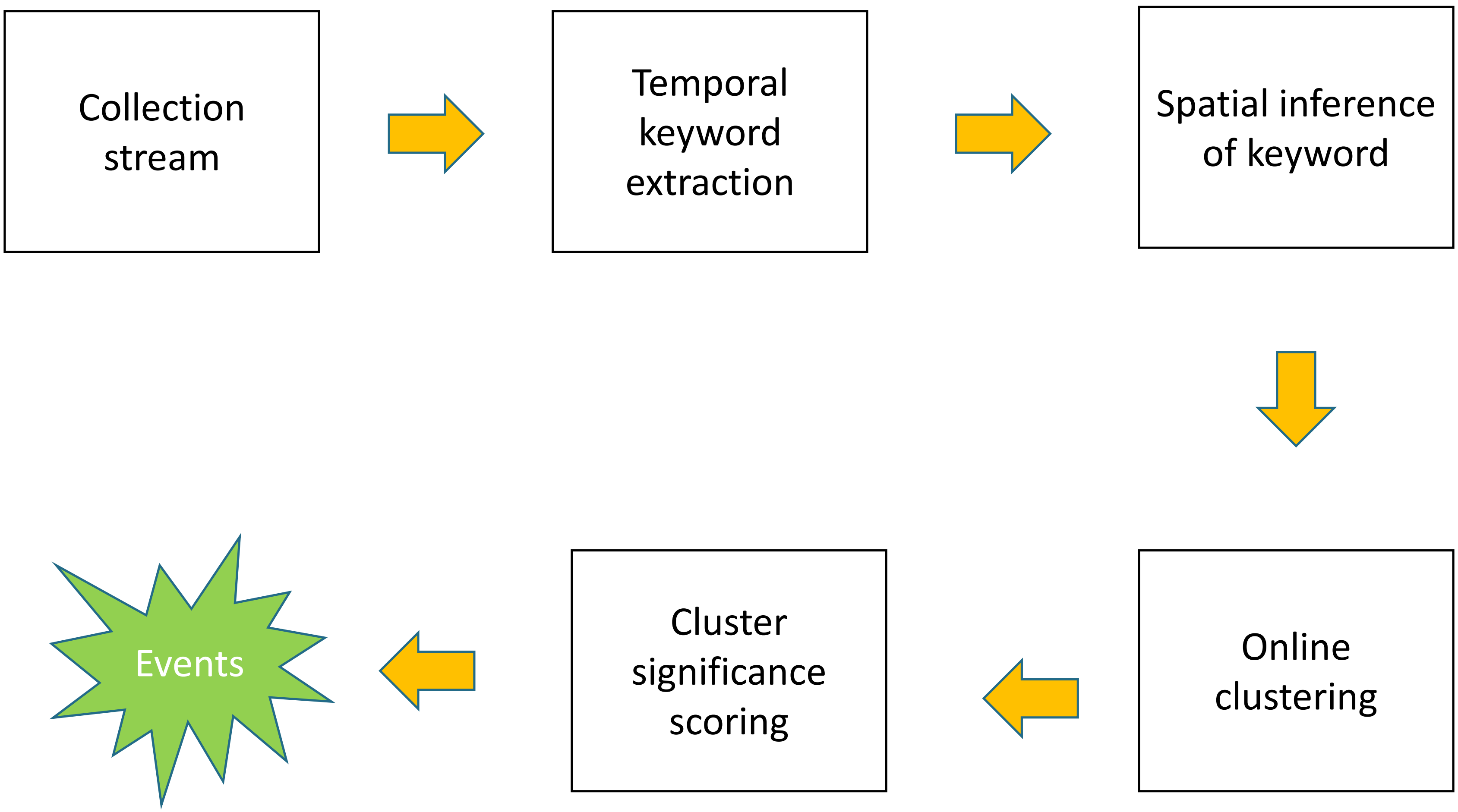
General Event Detection



EvenTweet

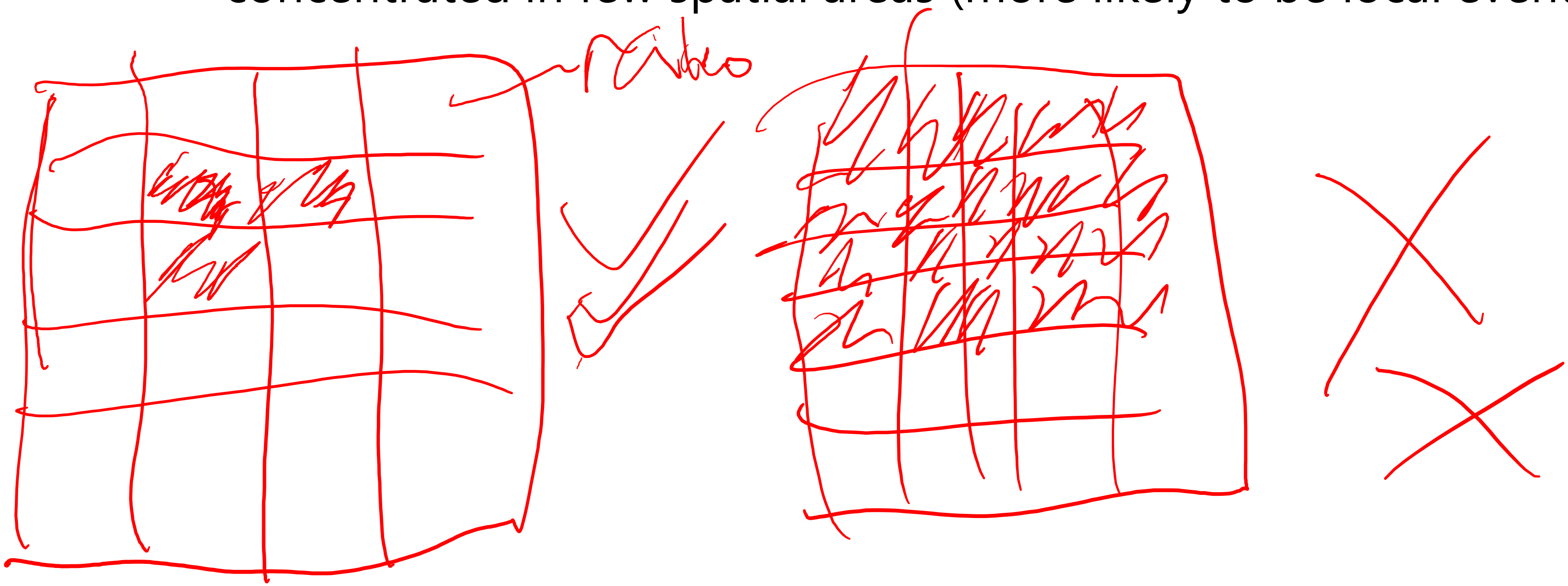
- Events represented by
  - Keywords (what)
  - Start and end times (when)
  - Geographical location (where)
- Tracks (local) events across time
- Provides a significance score for each detected event
- Need geo-tagged tweets (but this can be predicted/learnt)

EvenTweet


Collection stream
Temporal keyword extraction
Spatial inference of keyword
Online clustering
Cluster significance scoring
Events

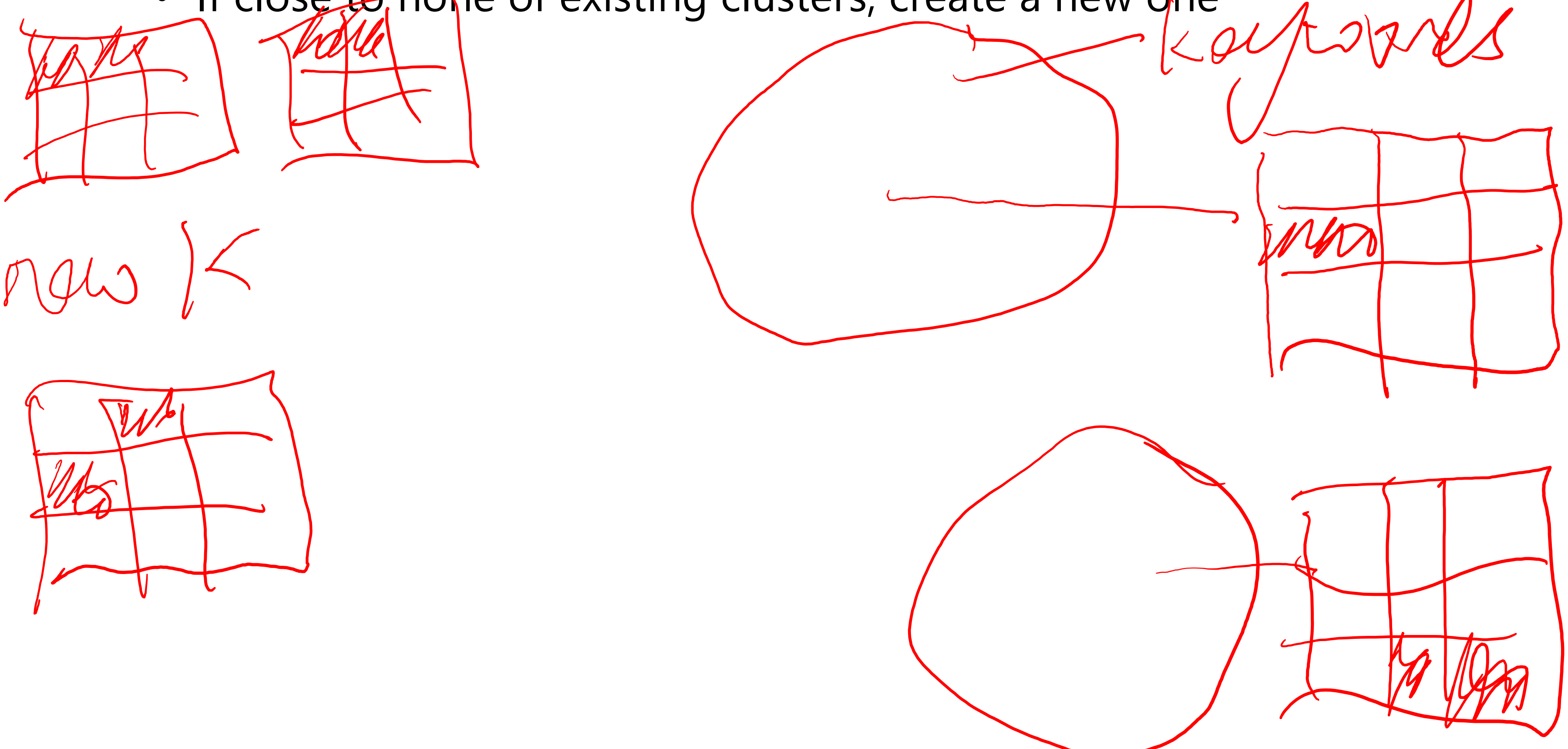
EvenTweet

- Temporal keyword extraction
  - Time window
  - Burstiness detection
- Spatial estimation of usage of keyword
  - Divide map into a grid of cells
  - Usage ratio of keyword:
    - number of users using keyword in cell / number of users tweeting in cell
  - Use entropy of usage ratio to identify keywords that are highly concentrated in few spatial areas (more likely to be local events)



EvenTweet: Online Clustering

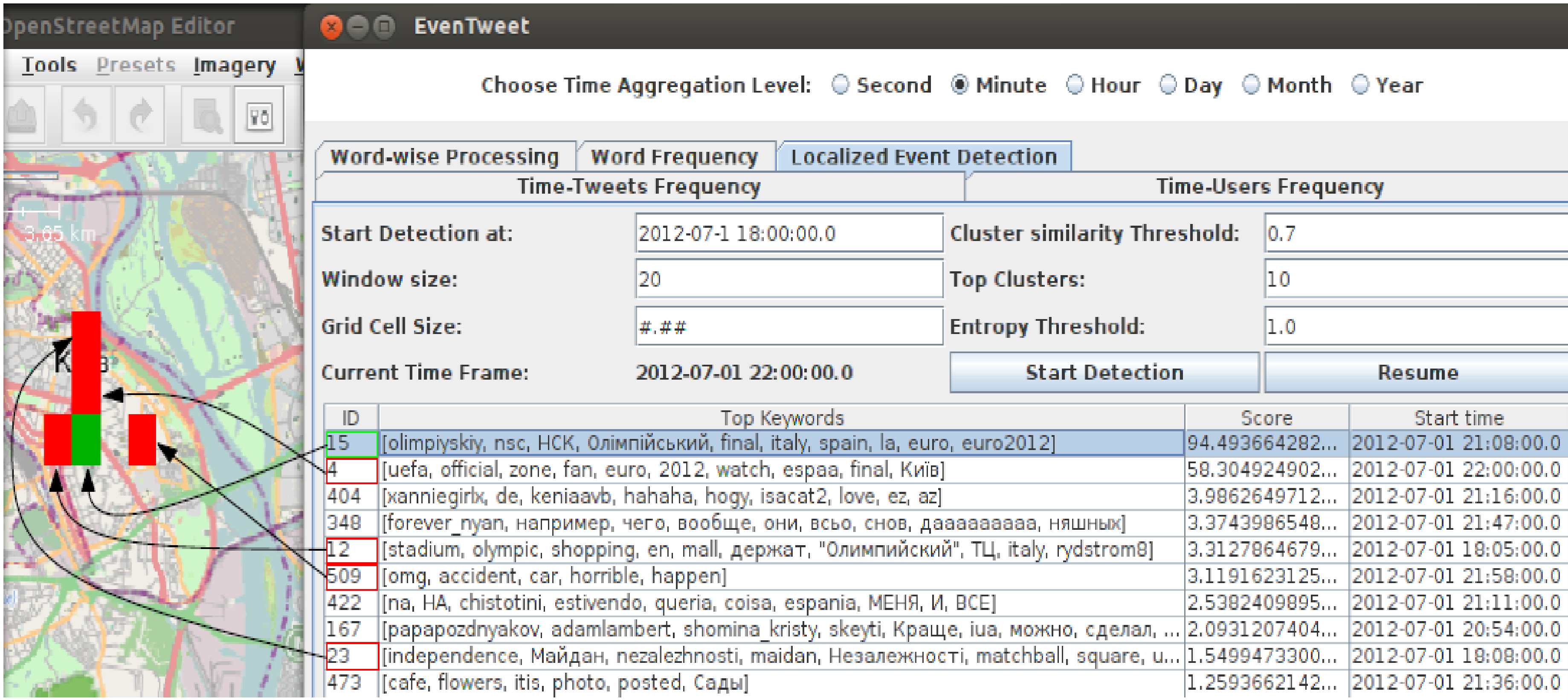
- Online clustering
  - Each existing cluster of keywords has a spatial signature
  - For each new keyword, compute cosine similarity of spatial signature to each centroid
  - If within a similarity threshold, add to cluster (select largest if there are a few possible cluster candidates)
  - If close to none of existing clusters, create a new one



EvenTweet

- Cluster scoring
  - Rank clusters, as there might be many of them and not all relevant
  - Each keyword has a k_score, calculated based on:
    - The last time it had bursty behaviour, its degree of burstiness
    - How long it has been a member of its current cluster
    - Recency of last time it (re)joined as a member of cluster
  - Each cluster then ranked according to the sum of k_score of its keywords
    - Clusters whose keywords has high burstiness, where the busrty keywords have persisted for a while and they also been bursty recently

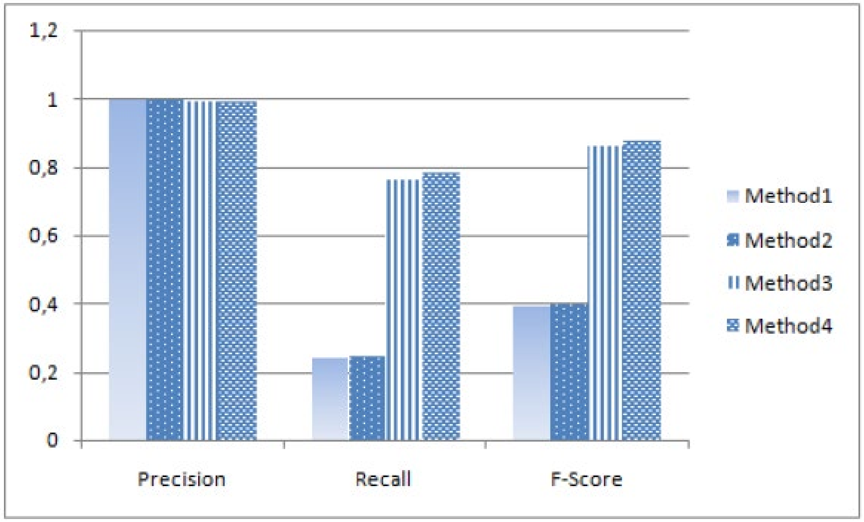
EvenTweet



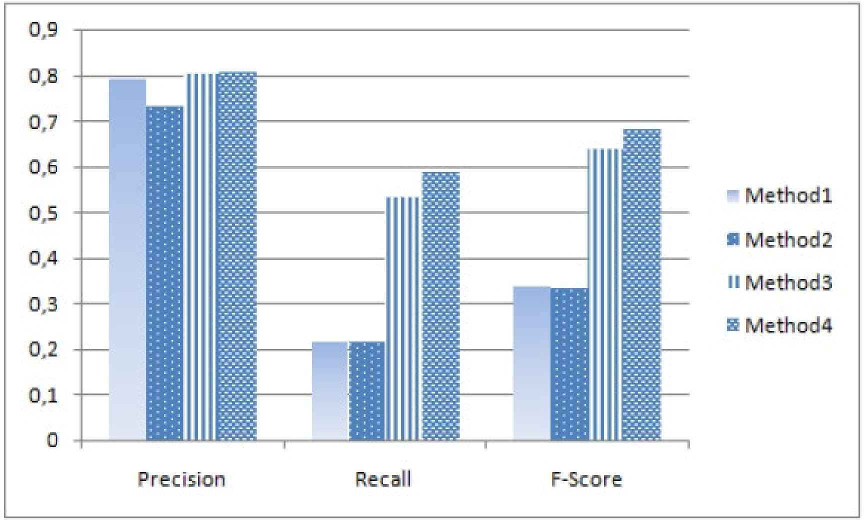
2012 UEFA European Football Championship

HashTag Clustering based Event Detection

- Demonstrate another entity to track
  - Events are assumed to be clusters of related hashtags
- Applied hierarchical clustering to obtain clusters of hashtags
- Only keep clusters whose size is multiple standard deviations above mean cluster size



Gallipoli 2012

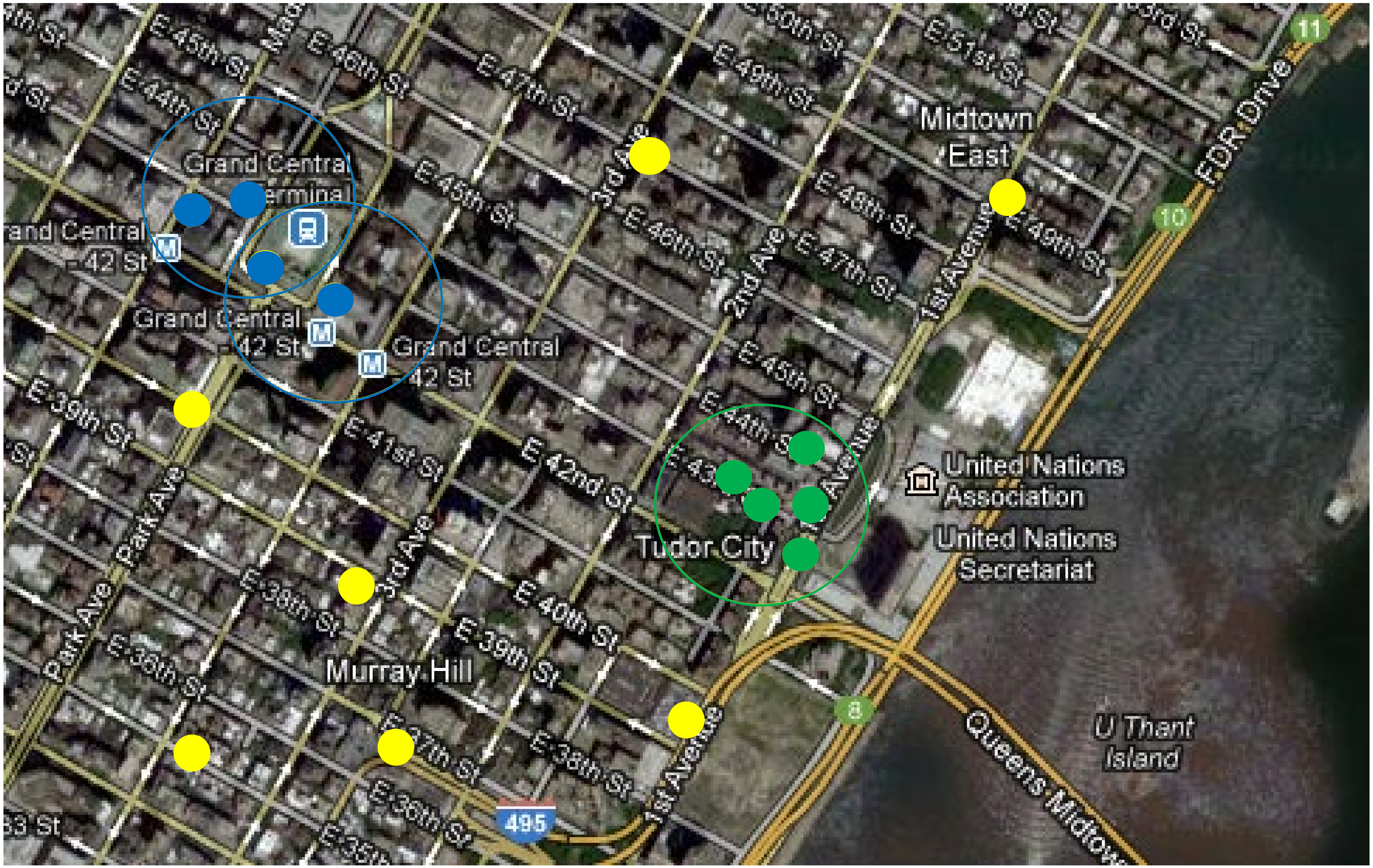


Fenerbahce vs Galatasaray 2012

Geo-spatial Event Detection in Twitter Stream

- An example that uses classifier to determine if identified clusters represent significant events
- 2 step approach:
  - Identify locations with high tweet activity
    - Identify geo-spatial clusters of tweets (e.g. 3 or more tweets in a 200m radius, posted within 30 mins)
  - Evaluate clusters with a classification approach
    - Do these clusters constitute a significant (real-world) event?
    - Use a C4.5 Decision Tree

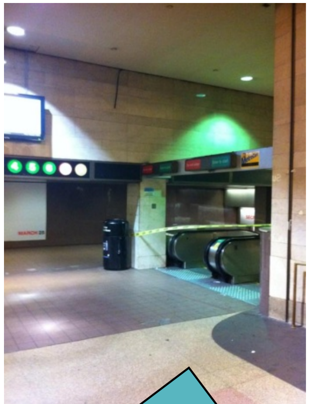
Geo-spatial Clustering of Tweets



= geo-located Social Media post (Tweet)

From 2013 Berlin Buzzwords event

What is the classification task?



Good

- Suspicious package in #GrandCentral #NYC #bomb threat pos not sure?? http://t.co/VwU7SP3X
- Suspicious package found in Grand Central Station... the 456 train..the trains are closed !! [pic]: http://t.co/9YPki4k2
- Something happened in the #456 #trainstation in #GrandCentral #NYC http://t.co/GGKvQura
- Accident on the #456train in #midtown #NYC http://t.co/fj2mJJmf

vs.

Bad

- @refinery29: This image of Madeleine Albright playing the drums will be the best thing you'll see today: http://t.co/rGwQ5RdG
- «@_PrettyPoison Guess ill fill out more job apps today» make punna fill out some 2!
- The Glamour & Glitz at the 2012 Emmy' s that we loved! http://t.co/CiTFszfL
- @IszwanieSyahira: i'm happy and i hope u feel the same too. weeeee ~.~
- How to prepare yourself for Friday's apocalypse http://cnet.co/lPU

We need to automatically determine which of the tweet clusters (tweets issued close to each other in a short time frame) represent real-world events and which are just random chatter.

Classifier Features

- Decision Tree

Tweet cluster:

- Suspicious package in #GrandCentral #NYC #bomb threat possibility not sure?? http://t.co/VwU7SP3X
- Suspicious package found in Grand Central Station... the 456 train..the trains are closed !! [pic]: http://t.co/9YPki4k2
- Something happened in the #456 #trainstation in #GrandCentral #NYC http://t.co/GGKvQura
- Accident on the #456train in #midtown #NYC http://t.co/fj2mJJmf

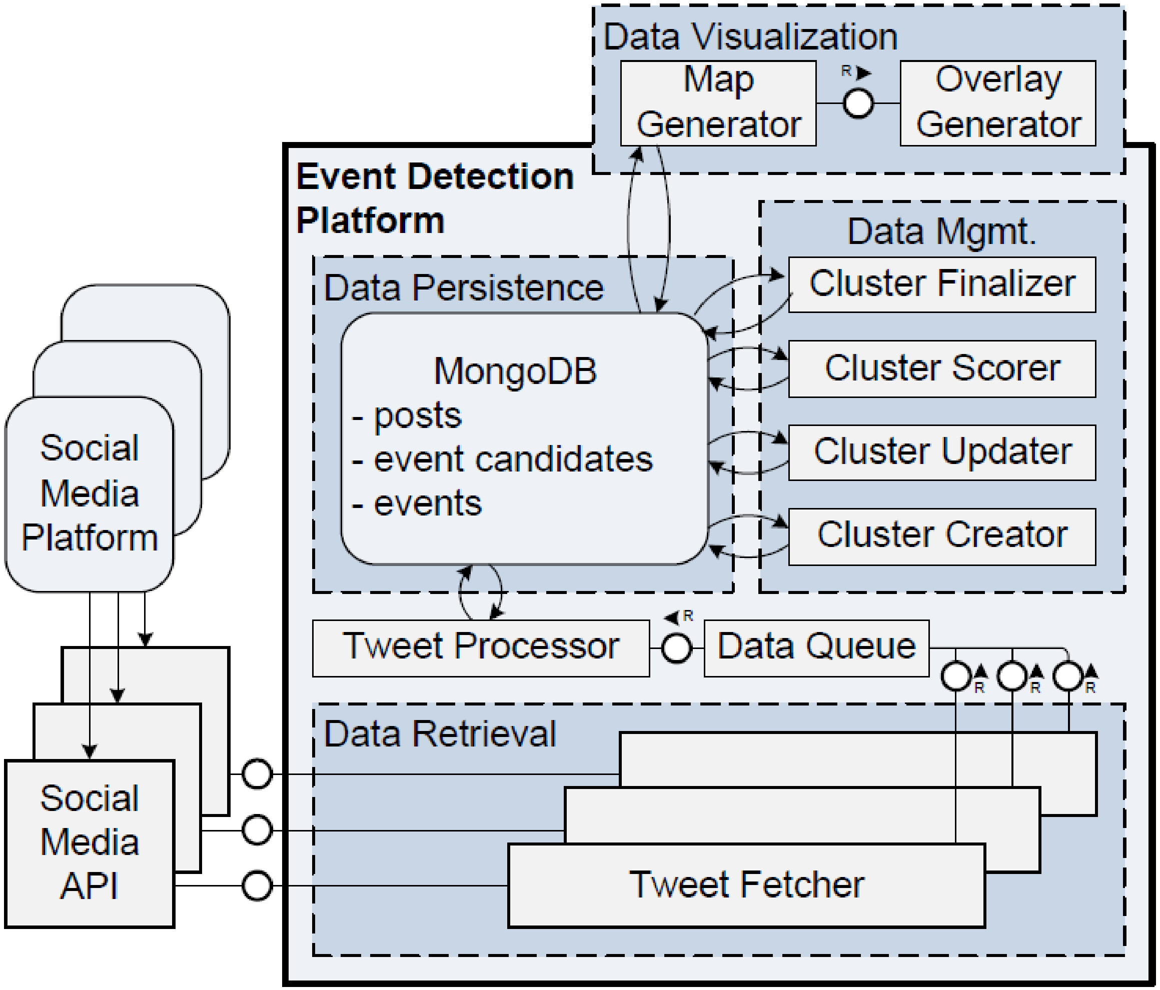
Textual features

Feature Group	#	Brief Description
Common Theme	1	Calculates the n-gram overlap between different tweets in the cluster.
Near Duplicates	1	Indicating how many tweets in the cluster are near-duplicates of other tweets in the cluster.
Positive Sentiment	3	Indicating positive sentiment in the cluster.
Negative Sentiment	3	Indicating negative sentiment in the cluster.
Overall Sentiment	2	Indicating the overall sentiment tendency of the cluster.
Sentiment Strength	3	Indicating the sentiment strength of the cluster.
Subjectivity	2	Indicating whether tweeters make subjective reports rather than just sharing information, e.g., links to newspaper articles.
Present Tense	2	Indicating whether tweeters talk about the here & now rather than making general statements.
# Ratio	1	Number of hashtags relative to the number of posts in the cluster.
@ Ratio	1	Number of @s relative to the number of posts in the cluster.
RT Ratio	1	Fraction of tweets in the cluster that are retweets.
Semantic Category	13	Indicating whether the cluster belongs to certain event categories, e.g., "sport event" or "fire".

Other features

Feature Group	#	Brief Description
Link ratio	1	Indicating the number of posts that contain links.
Foursquare ratio	1	Fraction of tweets originating from Foursquare.
Tweet count	1	Score based on how many tweets there are in the cluster.
Poster count	2	Score based on how many different users posted the tweets in the cluster.
Unique coordinates	2	Score based on how many unique locations the posts are from.
Special location	1	Fraction of tweets that are from a certain known "bad" location, e.g., airports or train stations.

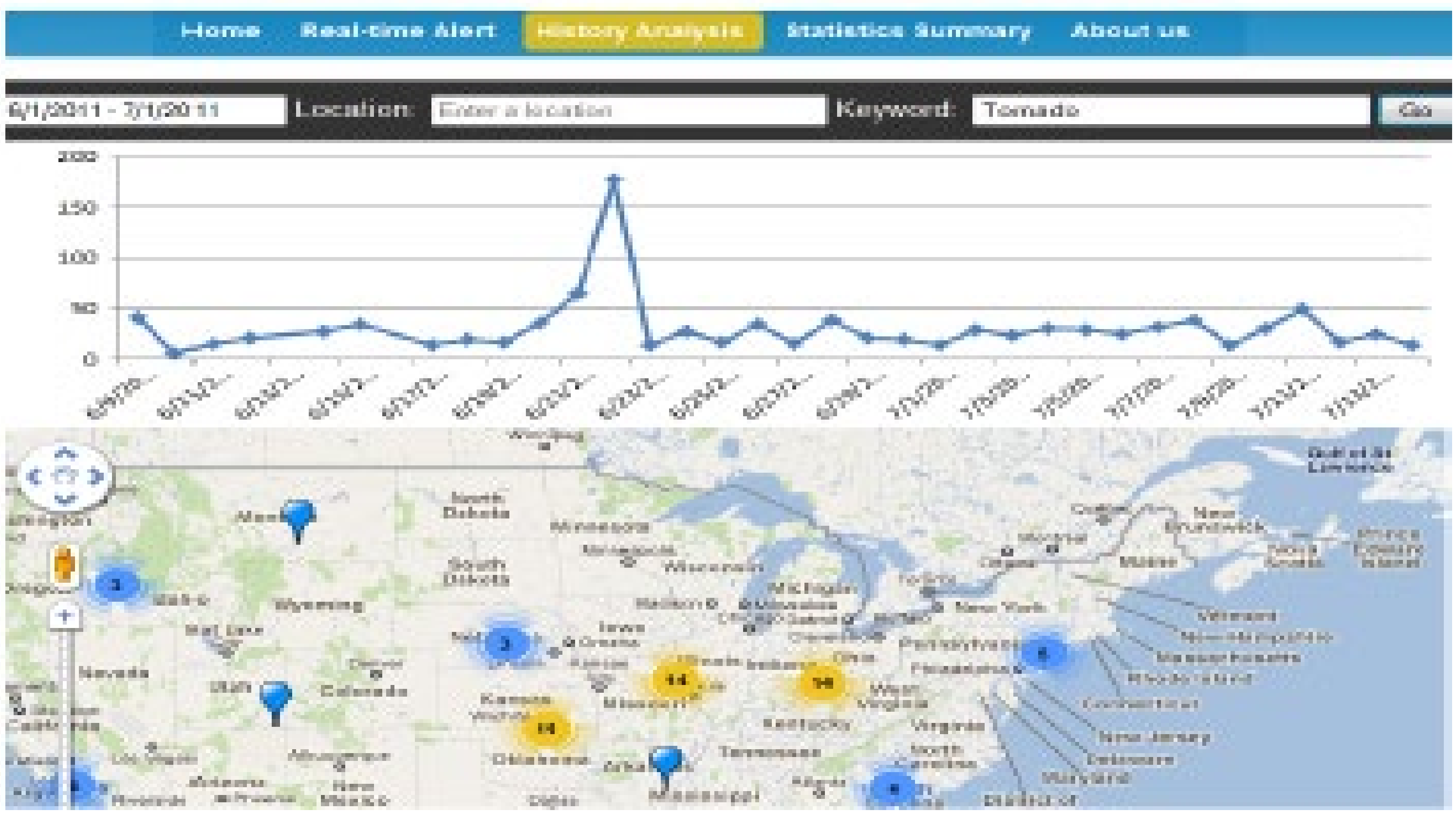
Framework



Data Visualization
Map Generator
R
Overlay Generator
Event Detection Platform
Data Mgmt.
Data Persistence
Cluster Finalizer
MongoDB
- posts
- event candidates
- events
Cluster Scorer
Cluster Updater
Cluster Creator
Social Media Platform
Tweet Processor
R
Data Queue
R
R
R
Data Retrieval
Social Media API
Tweet Fetcher

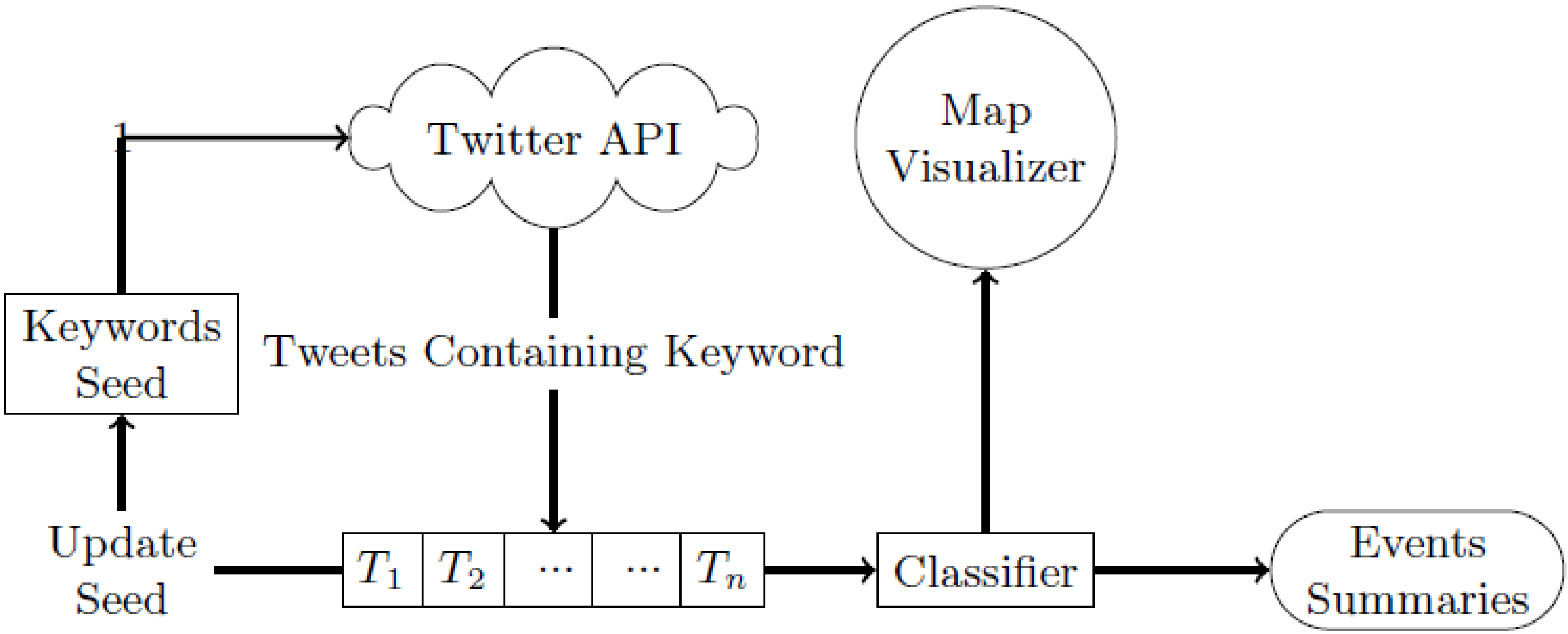
TEDAS: Topic Specific Event Detection

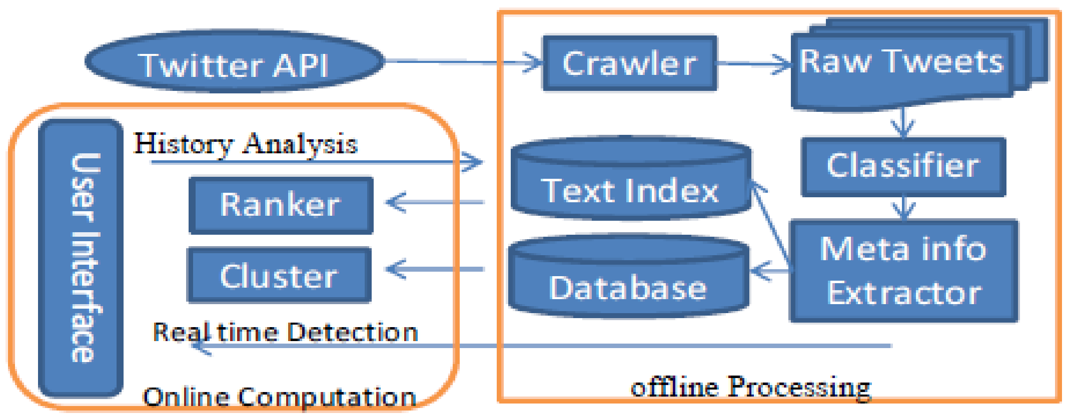
- TEDAS is an example, targeted towards detecting crime and disasters (and traffic accidents)
  - Can query location and time for events
  - Can query keywords for events



From Li et al, “Tedas…”

TEDAS Architecture

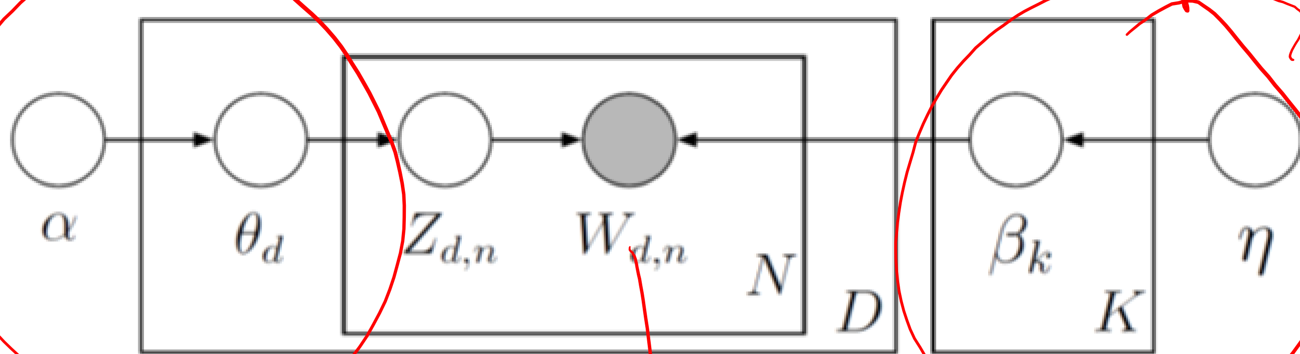




Temporal and Spatial Topic Models

- Recall last lecture we discussed the LDA topic model topic mixture

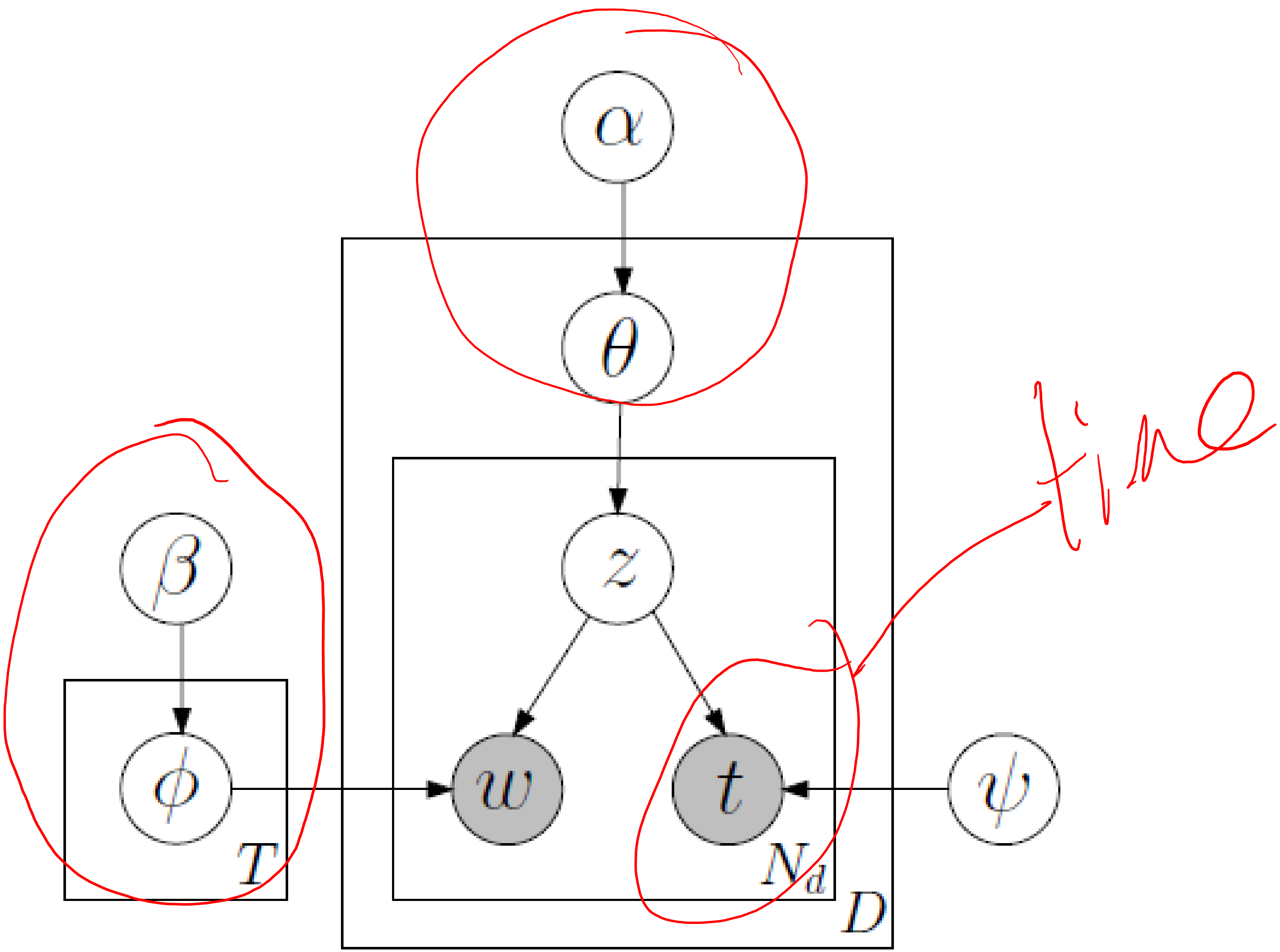
topics



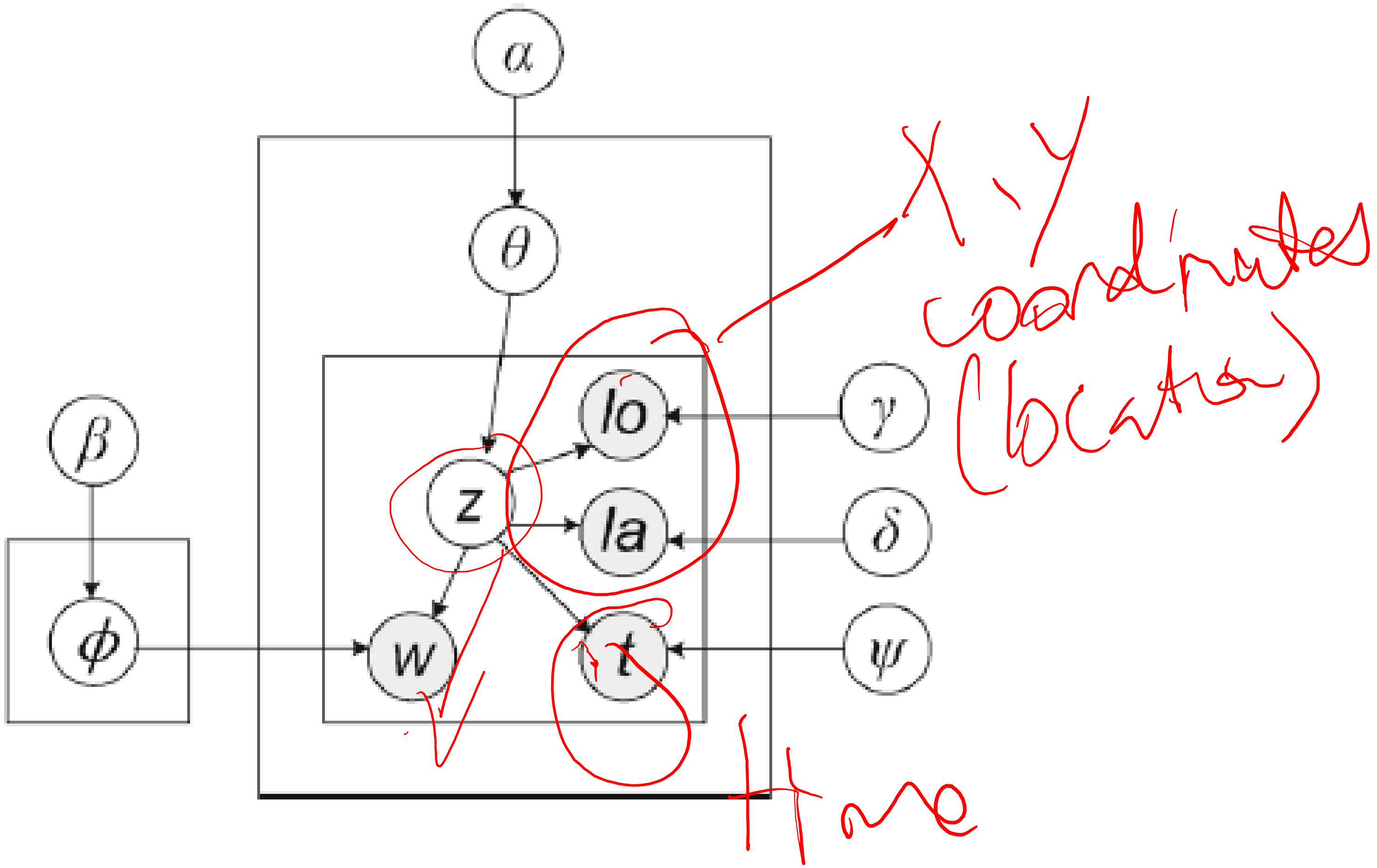
words

- They have been extended for many other purposes, including for event detection
- We restrict ourselves to topic models that models time and space.

Topic over Time Model (TOT)



Location Time Constrained Topic (LTT)



Summary

- Introduced event detection
  - Applications
  - Mostly unsupervised method
- Approaches
  - Burst detection
  - General Event Detection
    - Unplanned, topic-specific, topic models