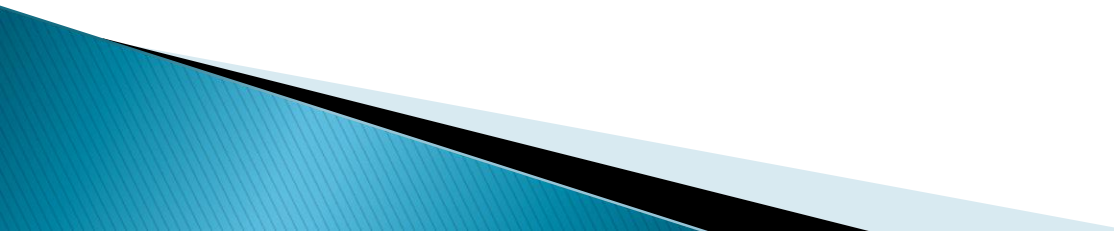


# Social Event Detection using spatio temporal data

Manoj V

# Agenda

- ❑ Overview
  - ❑ Scenarios
  - ❑ Event detection approaches
  - ❑ Literature review
  - ❑ Discussion
- 

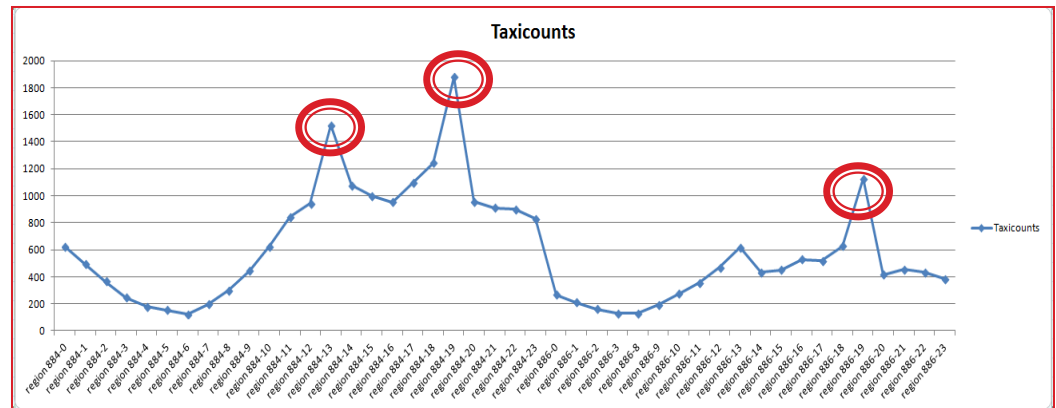
# Overview

## What are Anomalies ?

- ❑ Deviation from the expected behavior.
- ❑ Also known as – Outliers, Exceptions

## Types of Anomaly

- ❑ Point
- ❑ Contextual
- ❑ Collective

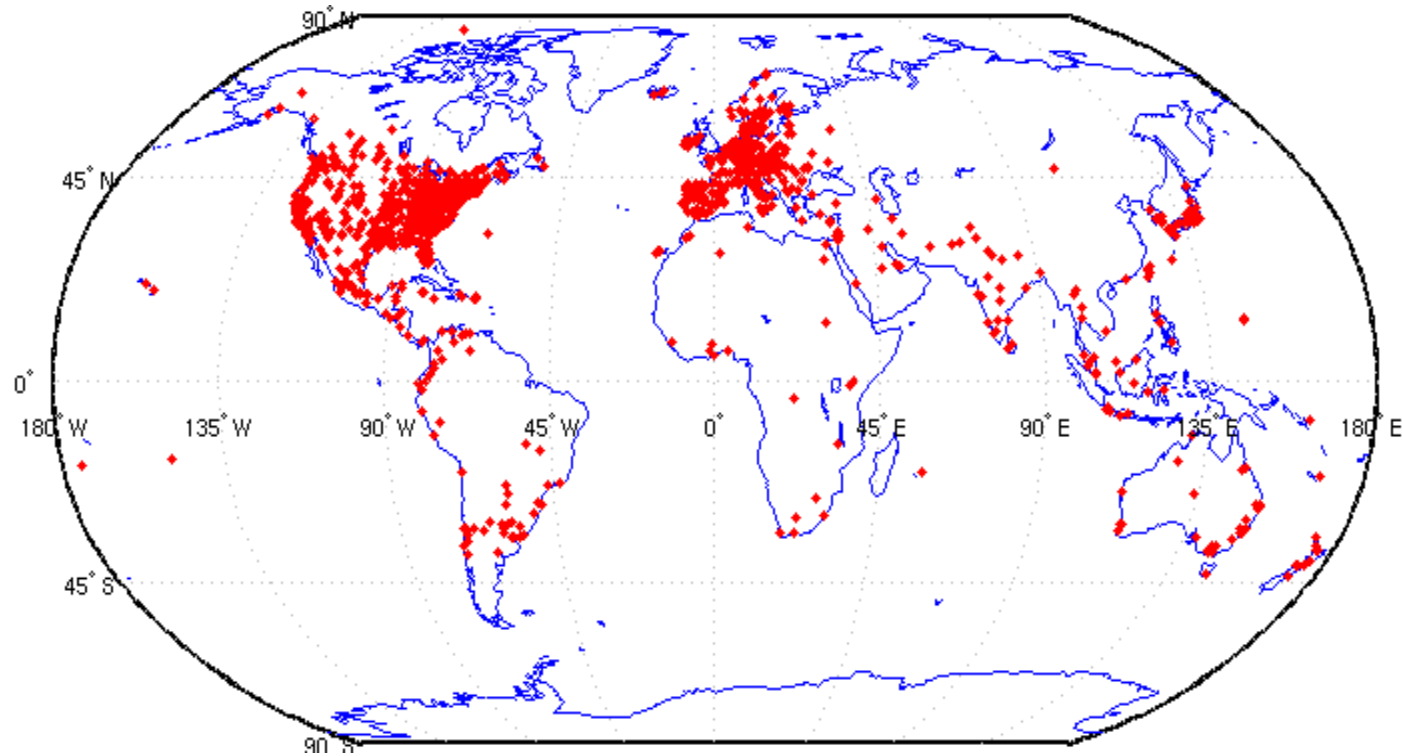


# Overview

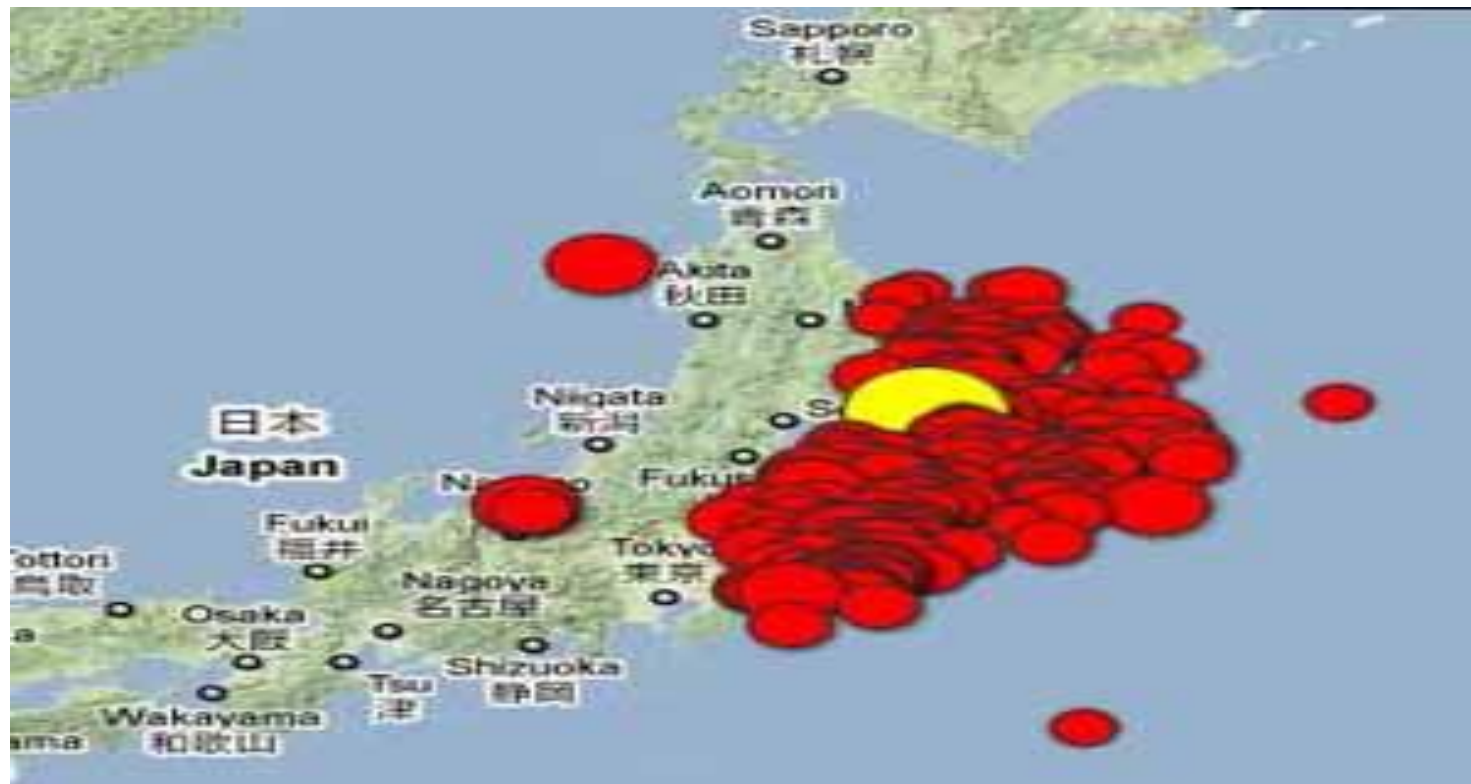
## **Event detection**

- ❑ Identifying anomaly which represents Event.
- ❑ Event Type (Specified vs Unspecified)
- ❑ Detection Task
  1. New Event Detection
  2. Retrospective Event Detection

# Scenarios – Tweets distribution



# Earthquake in Japan 2009

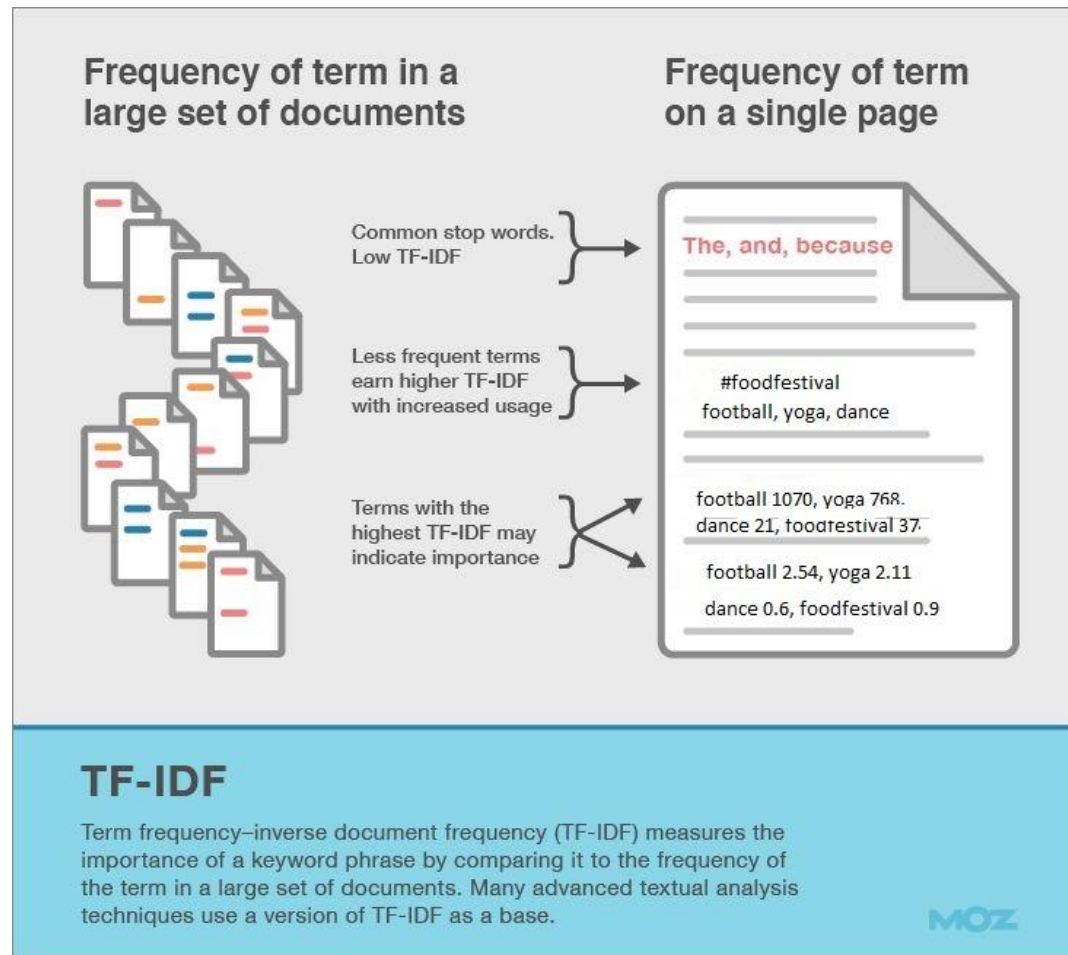


# Event Detection Approaches

## **Document pivot techniques**

- ❑ Grouping similar documents to predict the events.
- ❑ Term Vectors Tf-idf, scoring models
- ❑ Document similarity measured by Euclidean distance, correlation coefficient, and cosine similarity etc.

# Document Pivot technique





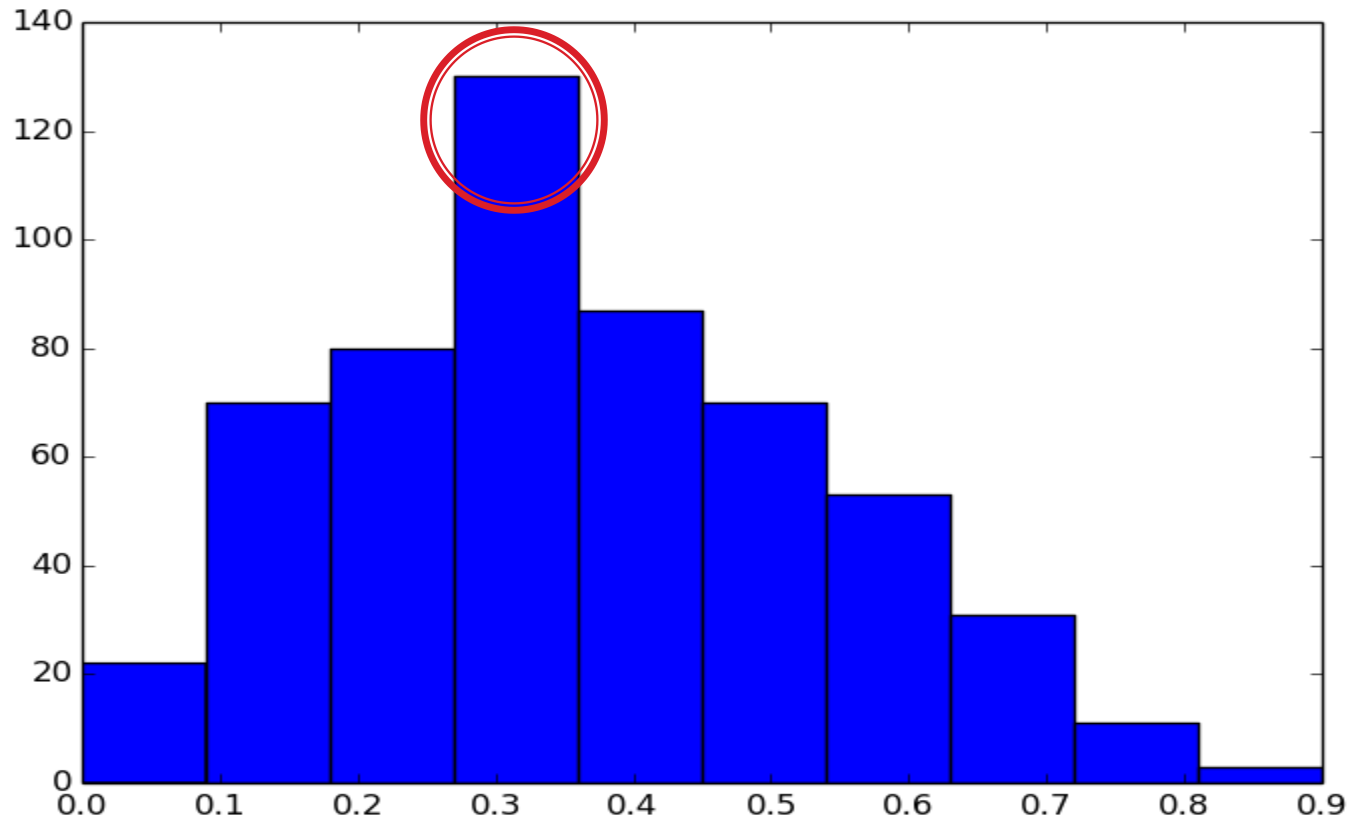
# Event Detection Approaches

## **Feature pivot techniques**

- ❑ Burst of Keyword detect sudden activity.
- ❑ The base assumption is related words would show an increased usage indicating an event occurrence.
- ❑ Keyword frequency and signals.

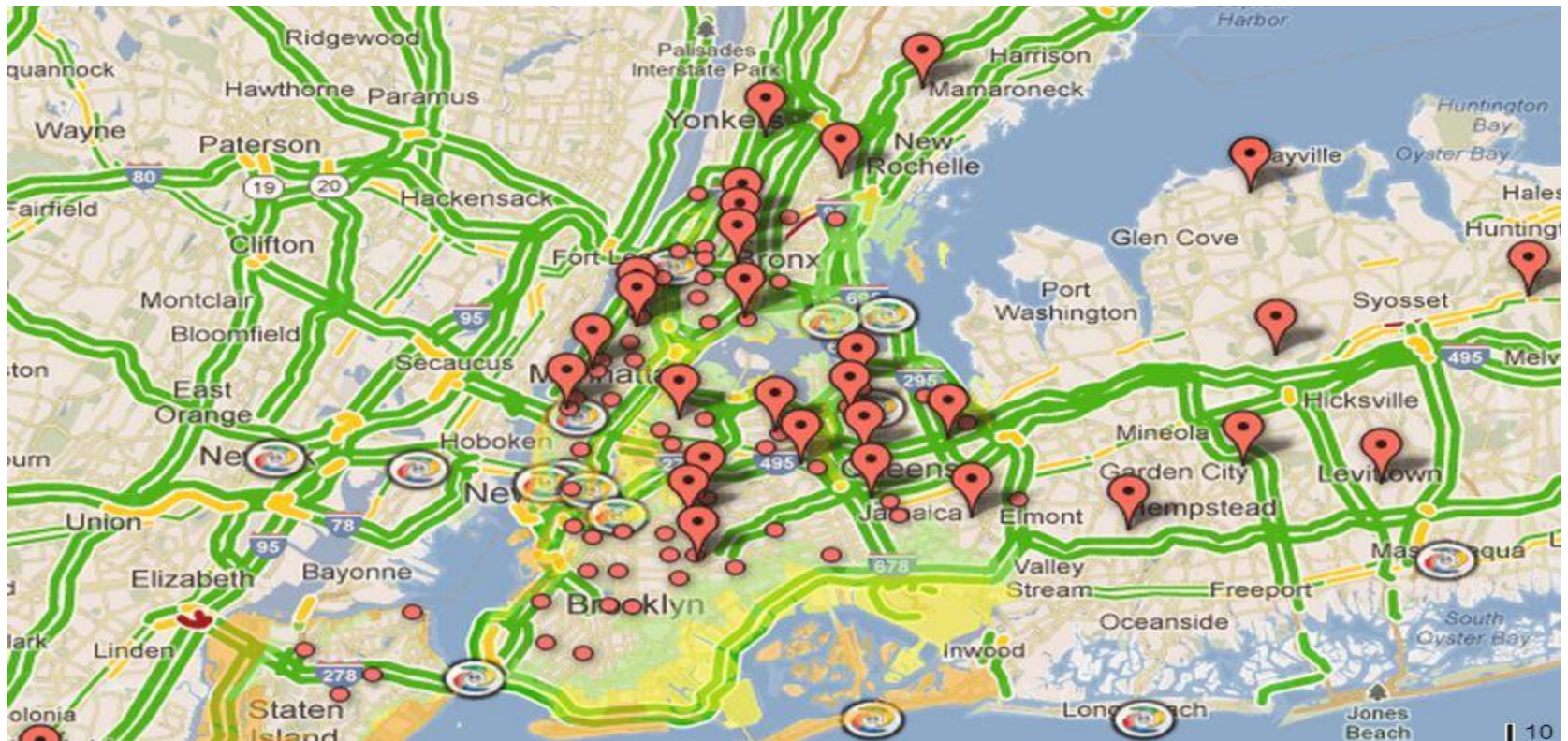
# Feature Pivot technique

Tweets containing keyword “Halloween”



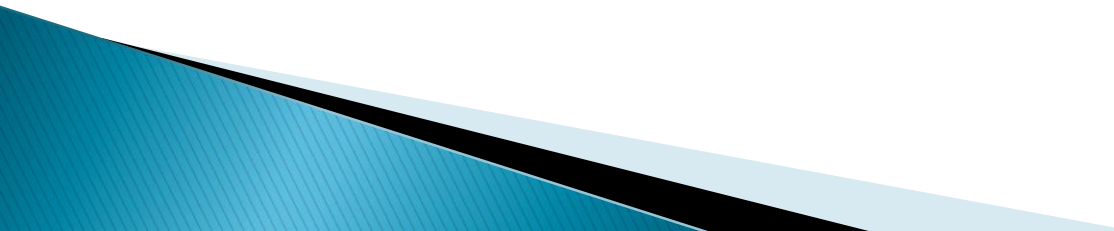
# Event detection

## What other dimensions can be considered ?



# Event Detection Approaches

## **Spatiotemporal techniques**

- ❑ Geotagged information longitude, latitude information and timestamp.
  - ❑ Tweets for example contains the geo tagged.
  - ❑ GPS enabled taxis to monitor the behavior of traffic.
  - ❑ Spatial and the temporal information as sensors.
- 

# Literature Review

- ❑ Spatio temporal approach
- ❑ Region: partition a city into regions  $r = \{r_1, r_2, \dots, r_m\}$  by major roads, such as highways and arterial roads, using a map segmentation method.



A) Raw road network



B) Segmented regions



# Literature Review

## NYC Grid Partition technique



# Literature Review (Contd..)

TaxiCount(I,h,d) - Partitioning based on **(region, hour, day)** for taxi ,  
TweetCount(I,h,d) - Partitioning based on **(region, hour, day)** for tweets

Region  $1 \leq I \leq 1200$

hour (h) in the range of  $0 \leq h \leq 23$  ,

day (d) in the range of  $1 \leq d \leq 31$

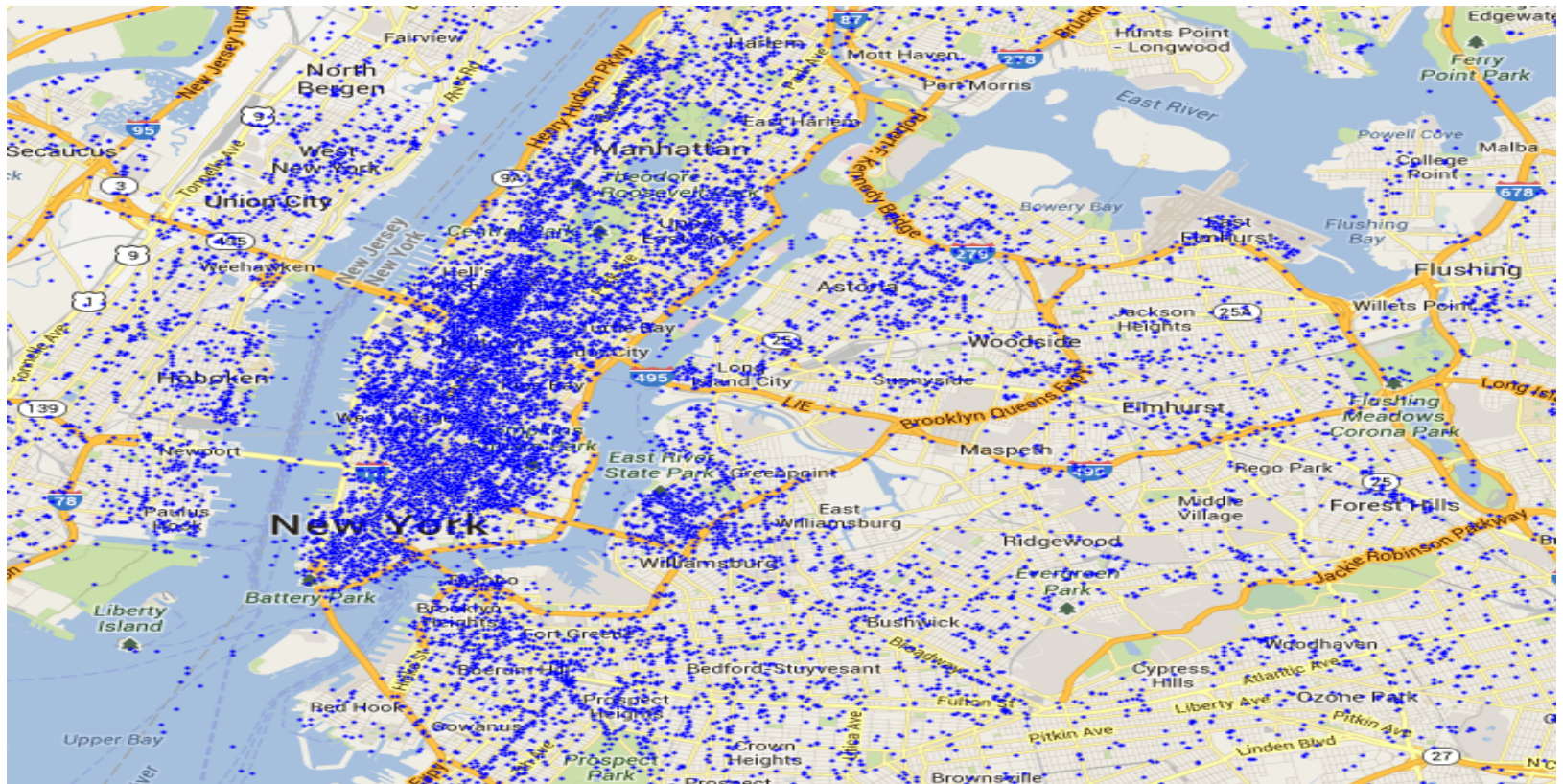
$$\text{AvgTaxiCount}(I,h) = \{ \text{TaxiCount}(I,h) \} / N$$
$$\forall I \in \{1..1200\} \text{ and } \forall h \in \{0..23\}$$

$$\text{StdTaxiCount}(I,h) = [ \text{TaxiCount}(I,h) - \text{AvgTaxiCount}(I,h) ]^2 / N$$
$$\forall I \in \{1,2,3....1200\} \text{ and } \forall h \in \{0,1,2...23\}$$



# Literature Review (Contd..)

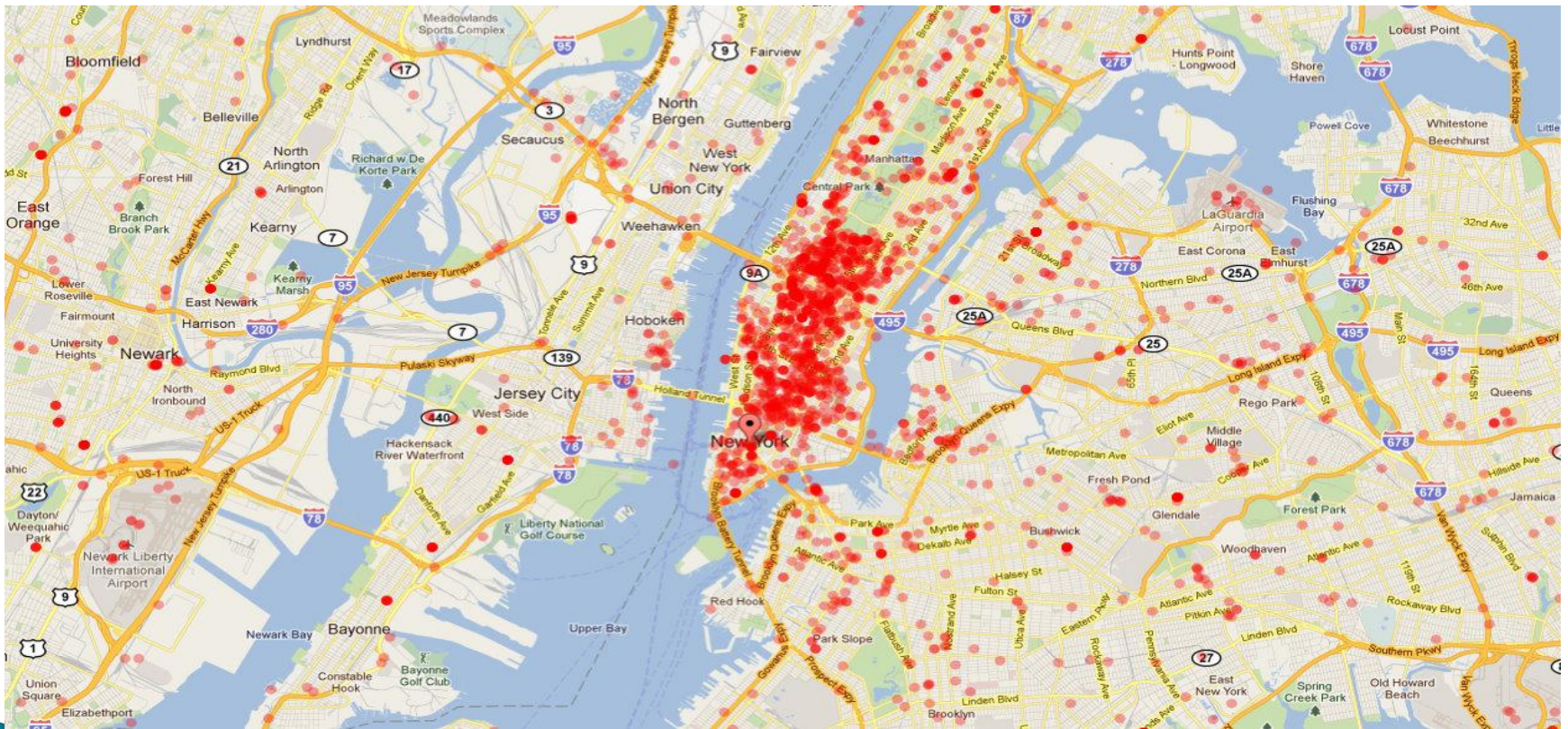
## NYC Taxi Data plot





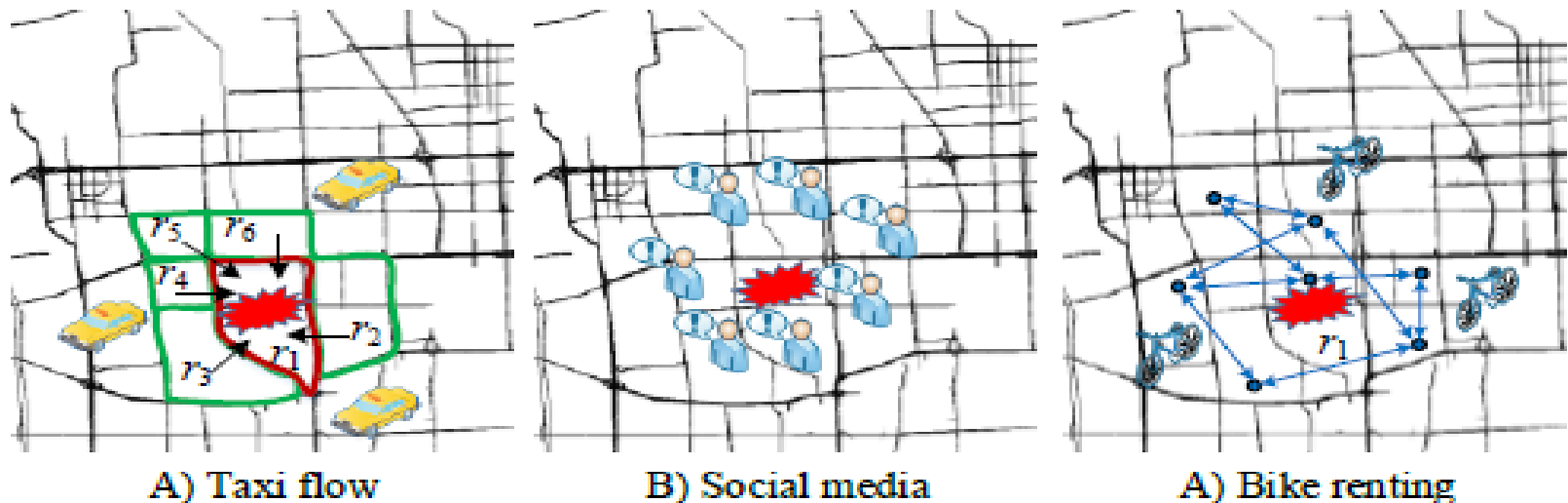
# Literature Review (Contd..)

## NYC Tweets Data plot



# Literature Review (Contd..)

**Collective anomalies** suggest there are underlying problems that may not be identified accurately based on a single data source.



**Figure 1. A collective anomaly witnessed by three sources**

# Literature Review (Contd..)

- ❑ Distance-based (DB) methods
- ❑ **Normal outliers and strict outliers**
- ❑ **1 and 3 time Standard deviation**
- ❑ Single and multiple datasets baselines
- ❑ DB-S-Taxi-S denotes the distance-based (DB) anomaly detection method that identifies an anomaly from a single (S) dataset

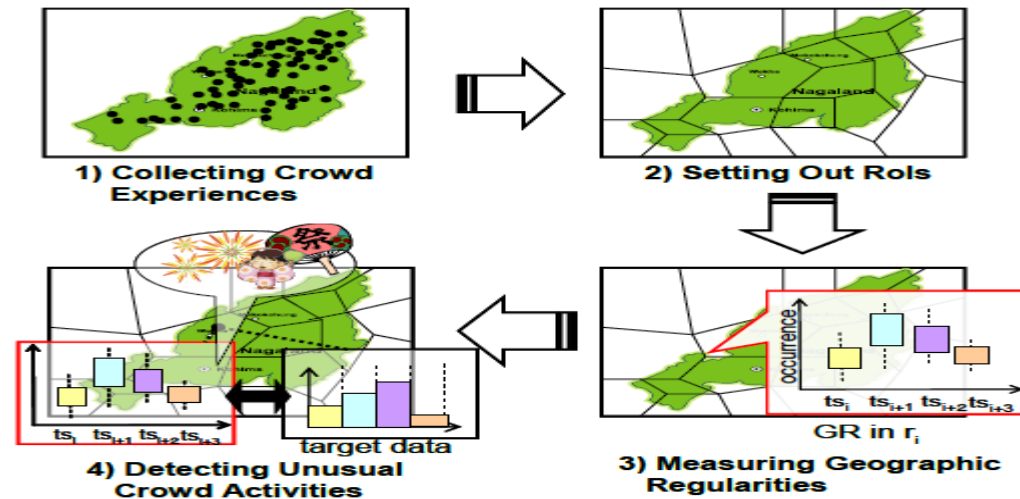
$$\text{AnomalyDetection } F(x) = \begin{cases} X - Y > 0 \\ X - Y \leq 0 \end{cases}$$

where , TaxiCount(I,h,d) = X ,  
StdTaxiCount(I,h) = Y

# Literature Review (Contd..)

## Measuring Geographical Regularities for Geo-social Event Detection

- (1) Collecting crowd experiences via Twitter
- (2) Establishing Region-of-Interests
- (3) Estimating geographical regularity of local crowd behaviors
- (4) Detecting Social events

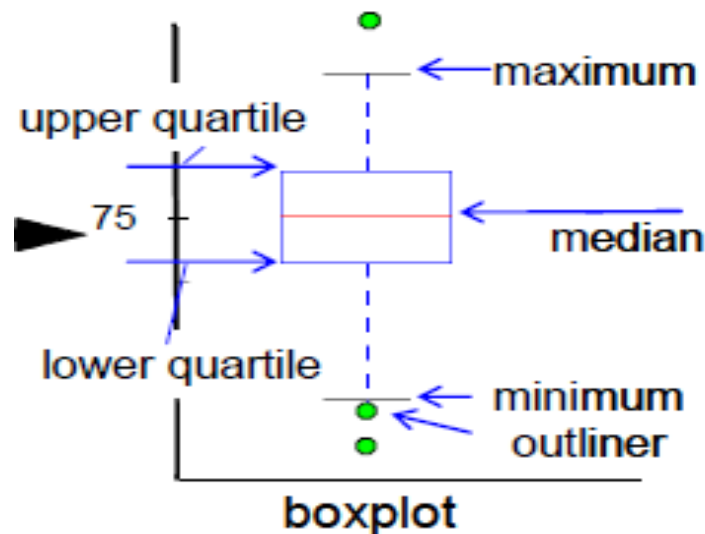


**Figure 1. Processes of unusual geo-social event detection**



# Literature Survey (Contd..)

- ▶ RoIs based on the geospatial and temporal occurrences of tweets.
- ▶ Technique used a boxplot presents five sample statistics—the minimum, the lower quartile, the median, the upper quartile, and the maximum.



# Literature Review (Contd..)

## **Likelihood Ratio Test**

Compare fit of two models

Find outlier degree for single data source

$$\Lambda = -2\log \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}},$$

$$od = \chi^2_{cdf}(\Lambda, fd)$$

Chi-squared distribution cumulative density function

$$L_{null} = \text{Gaussian}(70 | \text{mean}=200, \text{var}=1300)$$

$$L_{alter} = \text{Gaussian}(70 | \text{mean}=70, \text{var}=455)$$

# Literature Review (Completed)

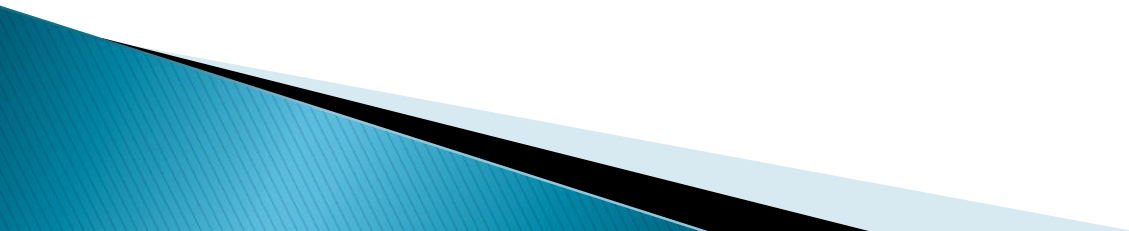
## **Aggregate Multiple data sources**

Skyline detection Algorithm

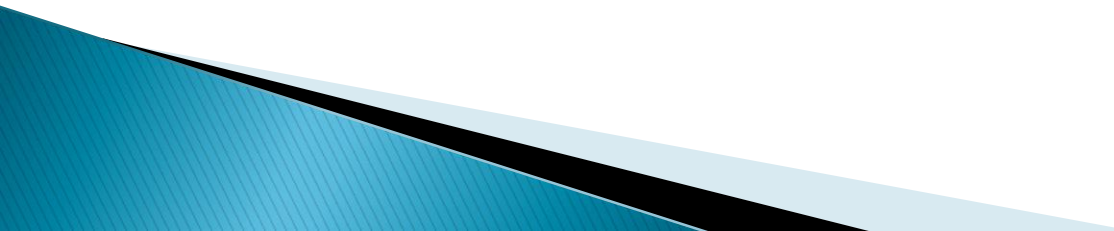
Define threshold for od (setup skyline points )

Weighted scoring model

Detect social Events



# Discussion

- ✓ Spatio temporal technique
  - ✓ Twitter and Taxi dataset
  - ✓ Corresponds to social event in NYC
  - ✓ Standard deviation, Tukey method, LRT
  - ✓ Aggregate multiple sources
  - ✓ Collective anomaly detection
  - ✓ Performance Evaluation
- 



# Questions

