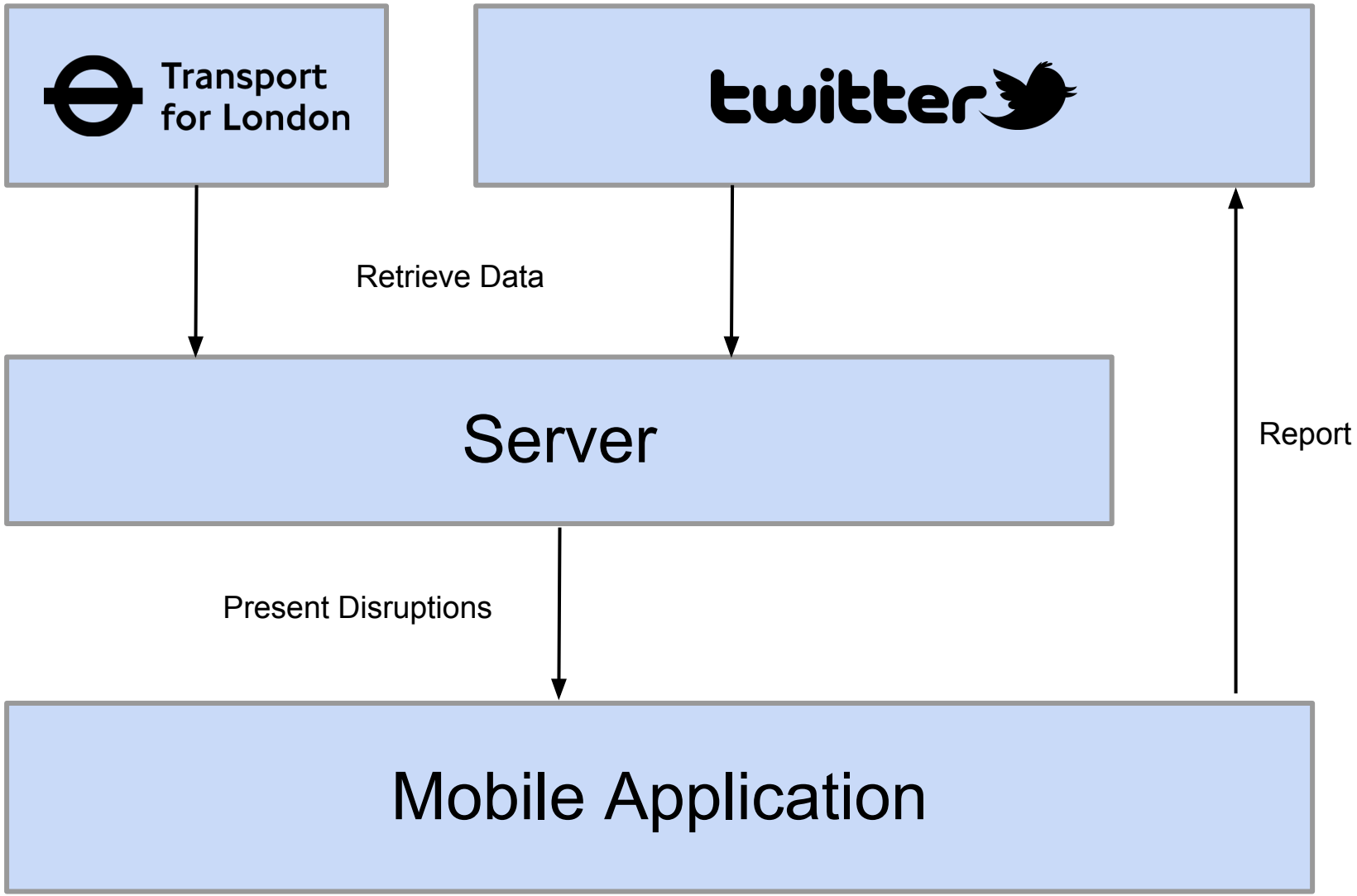
Twitter for traffic

Porfyrios Vasileiou, Marianna Polatoglou, Afxentios Hadjiminas, Panagiotis Tsirigotis, Hanguang Zhou, John Flanagan.

19th March 2012


Transport for London
twitter
Retrieve Data
Server
Report
Present Disruptions
Mobile Application

Challenges

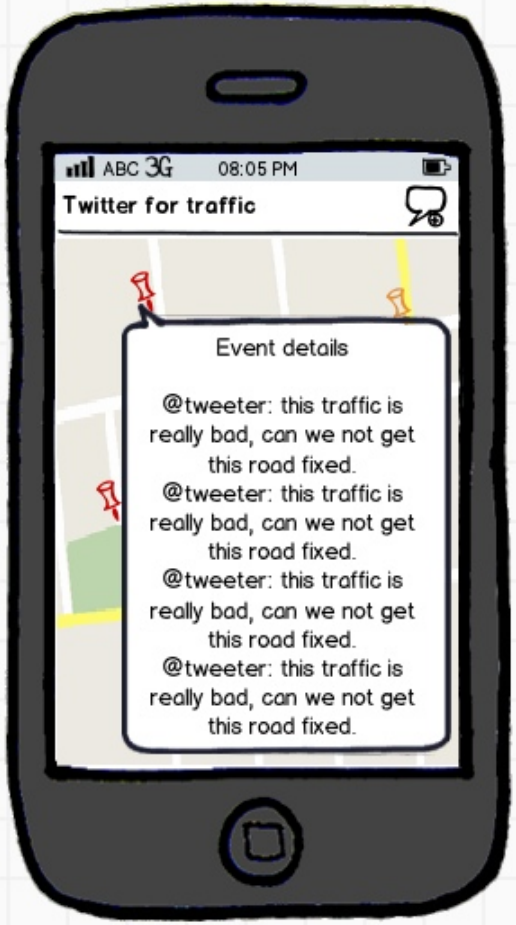
- Retrieve and store data
- Extract traffic information from social data
- Mobile application

Mobile App

Design and build an application with the view to releasing it.

- Google Android based
- Traffic disruptions
- Social conversation
- Traffic cameras
- 'Home route'
- Promote reporting

Main screen



Main screen with popup

Main screen w/ context menu

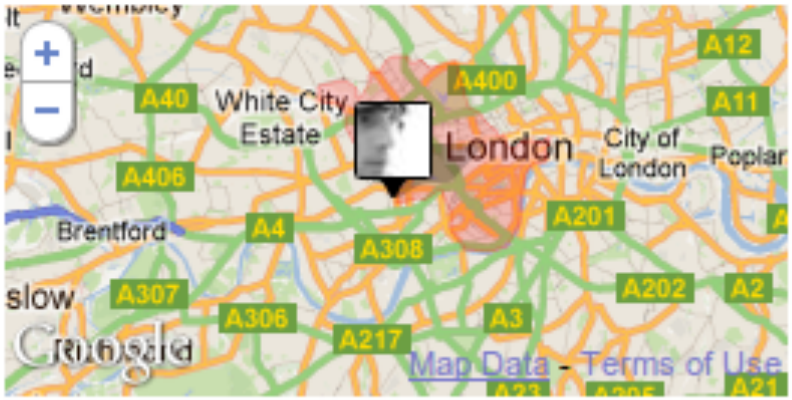

100 12:37
RIGHT TURN
Disruptions map
Regent's Park
London
A4
A3217
Battersea Park
Battersea Park
Google
Winstanley

Disruption Reporting

John Flanagan @johnflano Close

Traffic congestion causing severe problems near Imperial College London (South Kensington Campus), 180 Queen's Gate #RightTurn

from Westminster, London



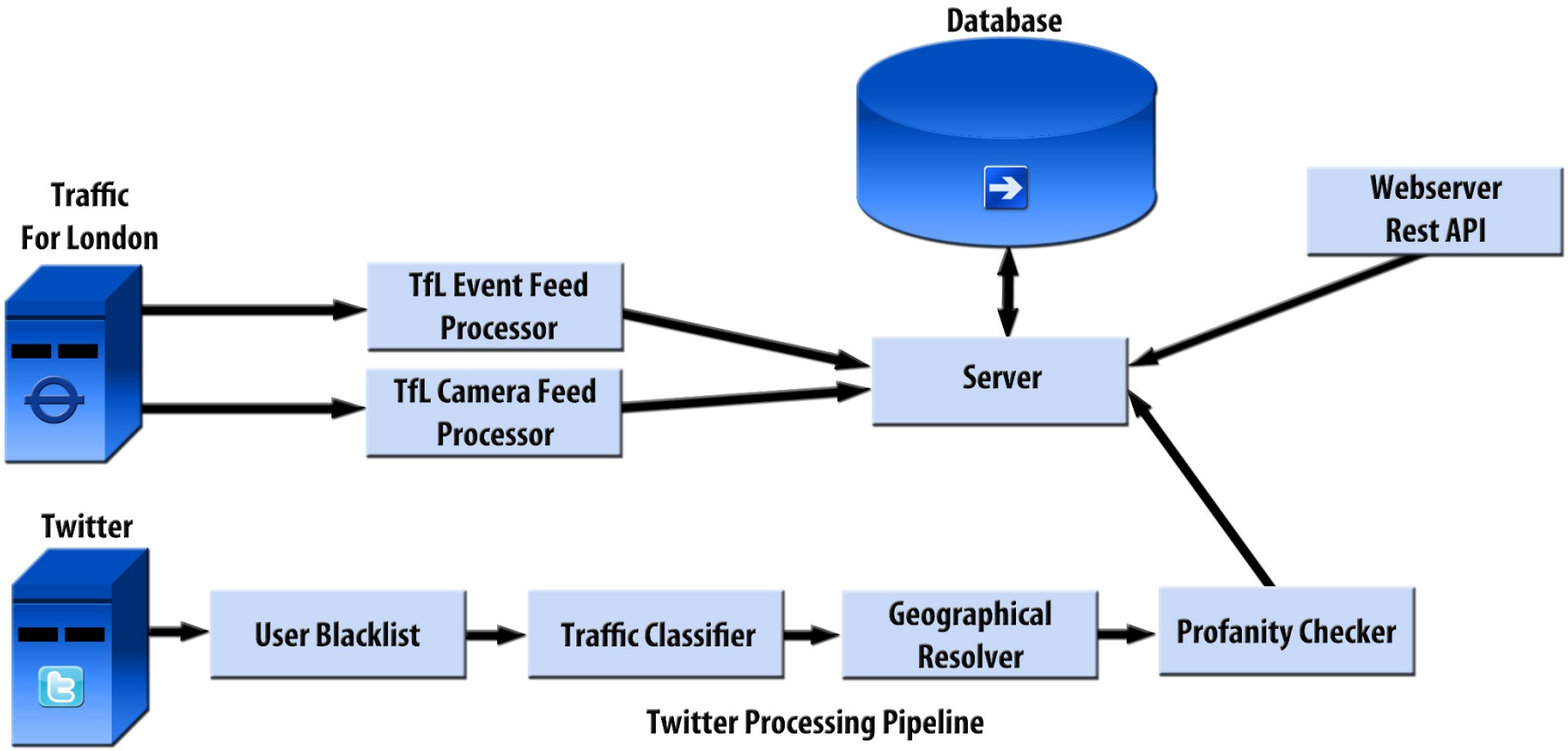
from Westminster
London, England

3:35 PM - 18 Mar 12 via Right Turn · Details

Reply Delete Favorite

Division of work

- Client
  - Mock Server
- Server
  - Modular Programming


Database
Traffic
For London
Webserver
Rest API
TfL Event Feed
Processor
TfL Camera Feed
Processor
Server
Twitter
User Blacklist
Traffic Classifier
Geographical
Resolver
Profanity Checker
Twitter Processing Pipeline

RESTful API

- Organizes everything in resources
- The client can request different resources using GET methods

Types of requests

- Traffic events
- Tweets for an event
- Cameras for an event

PostGIS

- Provides geographic objects for a PostgreSQL database

- Provides functions for these objects
  - all objects that are in an area
  - an object that is closest to another object

Something's going down on #wapping lane. Road blocked by police. Car drove into the butchers?

@jaggeree can't genuinely believe it'll affect internet traffic that much, but hey why not give the ISPs excuse to make more money from us!

First day back to Uni and I am greeted by traffic Just bloody great

Didn't know it cld hurt that bad... stuck home !!

Top tips for increasing traffic to your website http://t.co/HKEQ8mUO Please RT:)

The traffic around here is ridiculous

@PsychFoundation I am adding a links page to my website and would love to add your and add traffic.

Something's going down on #wapping lane. Road blocked by police. Car drove into the butchers?

@jaggeree can't genuinely believe it'll affect internet traffic that much, but hey why not give the ISPs excuse to make more money from us!

First day back to Uni and I am greeted by traffic Just bloody great

Didn't know it cld hurt that bad... stuck home !!

Top tips for increasing traffic to your website http://t.co/HKEQ8mUO Please RT:)

The traffic around here is ridiculous

@PsychFoundation I am adding a links page to my website and would love to add your and add traffic.

Document Classification

- A machine learning technique to categorize documents or pieces of text.
- Supervised Classification
- Training Data = Manually Labelled Tweets

"I am stuck in the WORST traffic ever!!! :(http://www.tfl.gov.uk/tfl/livetravelnews"

"I am stuck in the WORST traffic ever!!! :(
tfl_gov_uk"

"I am stuck in the WORST traffic ever!!! **_sad_** tfl_gov_uk"

"I am stuck in the WORST traffic **ever** _sad_ tfl_gov_uk"

"**i** am stuck in the **worst** traffic ever _sad_ tfl_gov_uk"

{"i", "am", "stuck", "in", "the", "worst", ""traffic", "ever", "_sad_", "tfl_gov_uk" }

{"i", "am", "stuck", "in", "the", "**bad**", ""traffic", "ever", "_sad_", "tfl_gov_uk" }

{"i", "am", "stuck", "in", "the", "bad", "traffic", "ever", "_sad_", "tfl_gov_uk", "**i am**","**am stuck**", "**stuck in**", "**in the**", "**the bad**", "**bad traffic**", "**traffic ever**", "**ever _sad_**", "**_sad_ tfl_gov_uk**"}

Classifier Evaluation

Why there is need for evaluation?

Several evaluation techniques have been used:

- Accuracy
- K-Fold Validation and Confusion Matrix
- Precision and Recall Rates
- F1-Measure

Accuracy:

Before Normalization : ~77%

Classifier Evaluation

Why there is need for evaluation?

Several evaluation techniques have been used:

- Accuracy
- K-Fold Validation and Confusion Matrix
- Precision and Recall Rates
- F1-Measure

Accuracy:

Before Normalization : ~77%

After Normalization : ~87%

Identifying tweet clusters

How to identify there is a traffic disruption through tweets?

- Geographic clustering

But what's important is unusual traffic conditions

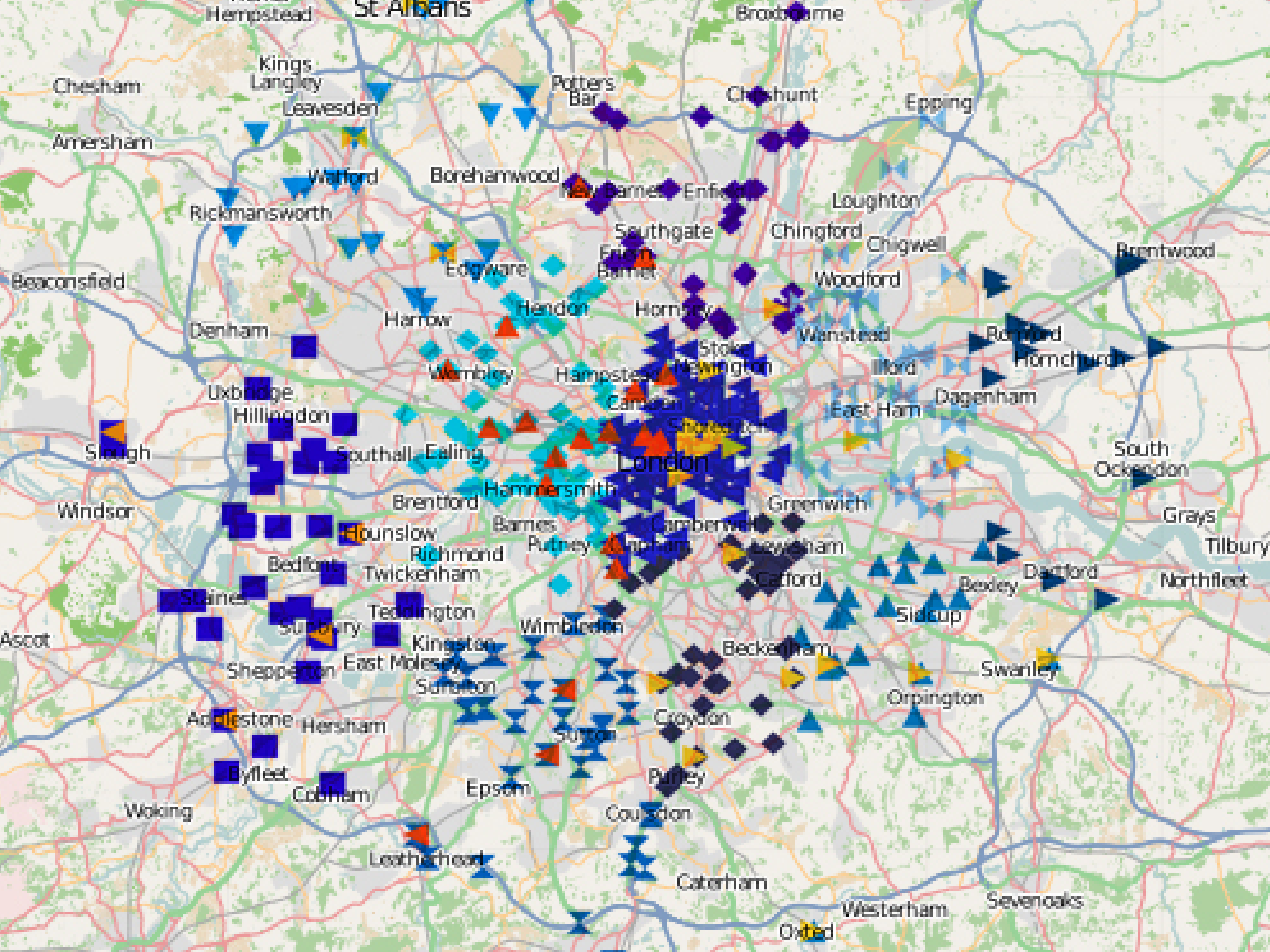
- Static clusters for tweets
- Slice clusters compared to static clusters

- Specific **time frames** of each day have *different behaviour*

Comparing static to slice clusters

- Normal distributions were used to describe the clusters

- Kullback - Leibler divergence
  - Comparing similarity of:
    - Centers of the distributions
    - Variance of the distributions



Hempstead
Kings Langley
Chesham
Leavesden
Amersham
Watford
Rickmansworth
Potters Bar
Borehamwood
Cheshunt
Epping
Loughton
Chingford
Chigwell
Southgate
Edgware
Beaconsfield
Harrow
Hendon
Woodford
Brentwood
Denham
Wanstead
Romford
Hornchurch
Ilford
Wembley
Uxbridge
Hillingdon
Dagenham
East Ham
Slough
Southall
Ealing
London
South Ockendon
Hammersmith
Brentford
Greenwich
Windsor
Barnes
Hounslow
Grays
Putney
Tilbury
Richmond
Twickenham
Bedfont
Catford
Dartford
Northfleet
Bexley
Teddington
Sidcup
Ascot
Kingston
Beckenham
Shepperton
East Molesey
Swanley
Surbiton
Orpington
Hersham
Croydon
Byfleet
Epsom
Cobham
Woking
Leatherhead
Caterham
Sevenoaks
Westerham

Geocoding

Only 5.63% classified traffic tweets with explicit geolocation

How geocoding works?

- Extract street addresses from messages
- Send queries to Google maps to get responses
- Get locations form the responses

Amount:

After geocoding: 10.68% of classified traffic tweets are Geo-tagged traffic tweets

Geocoding

Soundex

Encode homophones to the same representation so that they can be matched despite minor differences in spelling.

i.e.:

'oxford street' → O126S363

'oxfrd stret' → O126S363

Profanity Filter

A list of profanity words which determines if any profanity words are included in tweets

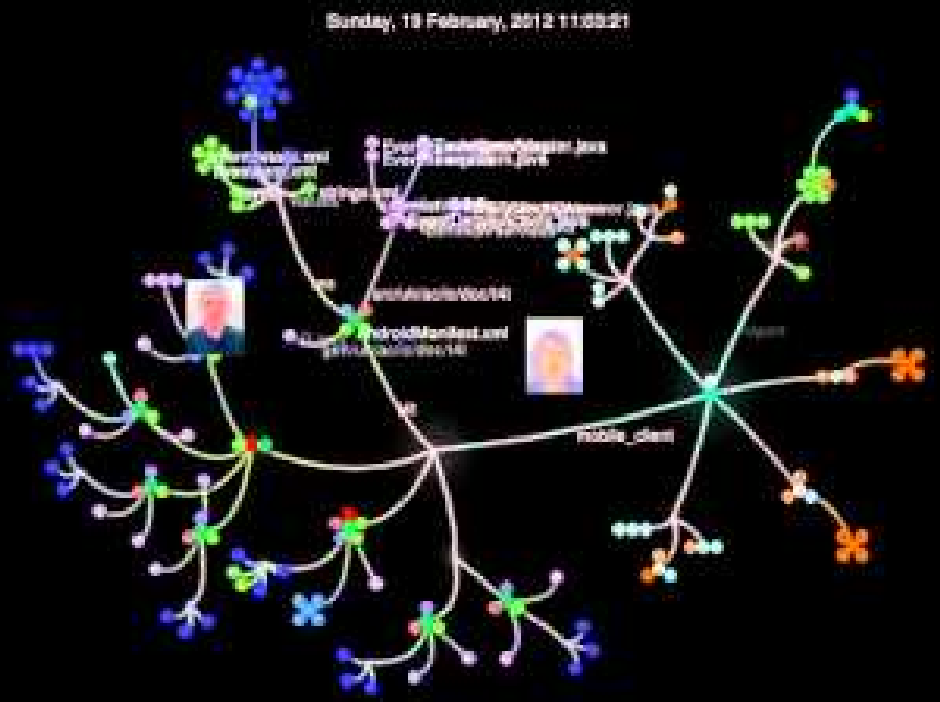
Tweet Metrics

From 90 hours of up-time

Total Tweets	Traffic Tweets	Geo-Tagged
71036	2940 (4.14% of Total Tweets)	314 (10.68% of Traffic Tweets)

Tweets with explicit Geolocation	Geo-Tagged by Geocoding
136 (43.31%)	178 (56.69%)

Questions ?




RIGHT TURN