PUBLIC

IT Innovation Centre

Crisis mapping and analytics of social media for disaster management and breaking news

Stuart E. Middleton

sem@it-innovation.soton.ac.uk
@stuart_e_middle
www.it-innovation.soton.ac.uk

ECS Seminar, Southampton, UK

21st Jan 2015

Overview

- Geospatial Research @ IT Innovation Centre
- Case Study – Ukraine Crisis 2014
- Scalable Processing Architecture
- Geoparsing
- Geosemantics
- Trust and Credibility Modelling
- Exploitation
- Future Work

IT Innovation Centre

- World-class application-driven R&D
- Applied research and development with and for industry, commerce and the public sector
  - collaborative research (supported by EC and UK programmes)
  - client-funded research, development and consulting

Southampton Science Park, Chilworth
10 mins from main campus

We deliver proofs-of-concept, demonstrators and novel operational systems

We work in a spirit of partnership, aiming to provide effective transfer of knowledge

Chilworth
EASTLEIGH
SOUTHAMPTON
Totton
Bishop's Waltham
Botley
Hedge End
Marchwood
Dibden
Hythe
Netley Abbey
Lyndhurst
Locks Heath
FAREHAM
Hamble-le-Rice
Fawley
Holbury
Blackfield
Stubbington
Lee-on-the-Solent
Lymington
Cowes
East Cowes
Ryde
Wootton Bridge

IT Innovation Centre

- Today, a team of 34
- Over the last five years
  - 42 major projects
  - 25 in the EC Framework Programme
  - over £2.25M of UK funding
  - over €10M of EC funding
  - working directly with tens of Universities
  - over 100 companies as partners and clients

Active Participants in –

OGC, NESSI, NEM, FIA

Big Data Value Association & PPP

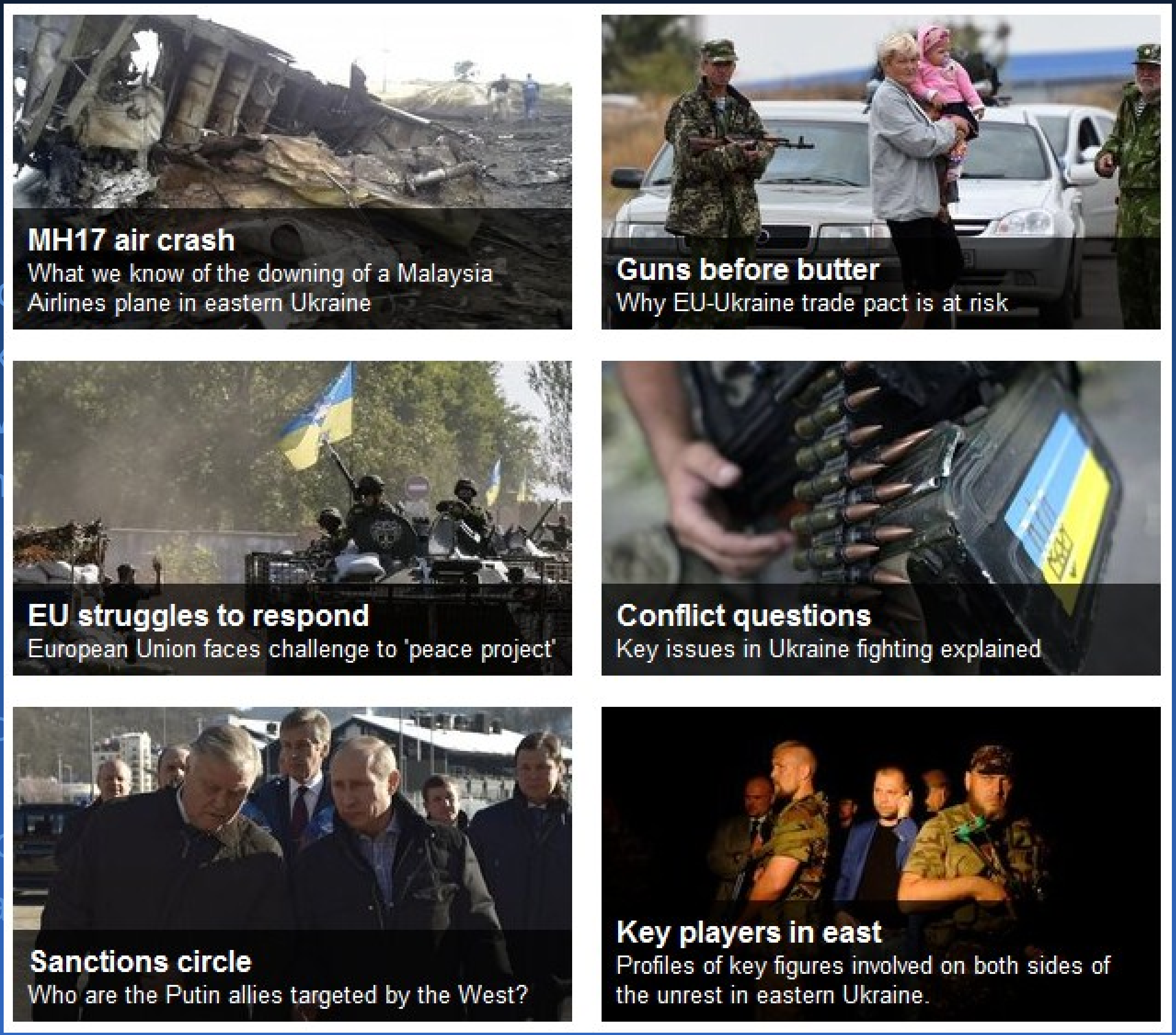
Geospatial Research

- Crisis Management and Environment Monitoring
  - Tsunami Early Warning Systems
  - Twitter, In-situ sensor data, Remote Satellite Data
  - Scalable Real-time Processing ➔ Big Data
  - Geoparsing, Data Fusion, Semantic Interoperability, Decision Support
  - TRIDEC project http://www.tridec-online.eu/
  - *Case Study – Twitter Crisis Mapping e.g. New York Hurricane 2012*
- Social Media and Open Data Analytics
  - Journalists – Breaking News
  - Twitter, YouTube, Instagram, Four Square, Flickr, Facebook ...
  - Scalable Real-time Processing ➔ Big Data
  - Geoparsing, Geosemantics, Data Fusion, Decision Support
  - Analytics, Trust and Credibility Modelling
  - REVEAL project http://revealproject.eu/
  - *Case Study – Breaking News Stories e.g. Ukraine Crisis 2014*

Case Study

- Ukraine Crisis 2014 - Breaking News Story
  - Russia invades Ukraine 2014
    - Russia annexes Crimea
    - Shooting down of Flight MH17
    - Fighting in and around Donetsk airport
    - ...
  - Lots of social media content every day (languages = EN, RU, UK)
  - IT Innovation has been crawling on Twitter, YouTube, Instagram and FourSquare during this period
  - Lots of fake and unverified social media reports
  - Ground truth from verified news agency stores and analysis (BBC News, Deutsche Welle ...)
  - Journalists typically have < 1 hour to verify & collate user generated content (UGC), write a report and broadcast the breaking news story
    - Automation is key to reduce a journalists manual workload
    - Relevance filtering & cross-checking of content
    - Trust modelling of sources & content

- Ukraine Crisi
  - Russian invac
  - Russia annex
  - Shooting dow
  - Fighting in an
  - ...
  - Lots of social
  - IT Innovation FourSquare c
  - Ground truth News, Deutsc
  - Lots of fake a



MH17 air crash
What we know of the downing of a Malaysia Airlines plane in eastern Ukraine

Guns before butter
Why EU-Ukraine trade pact is at risk

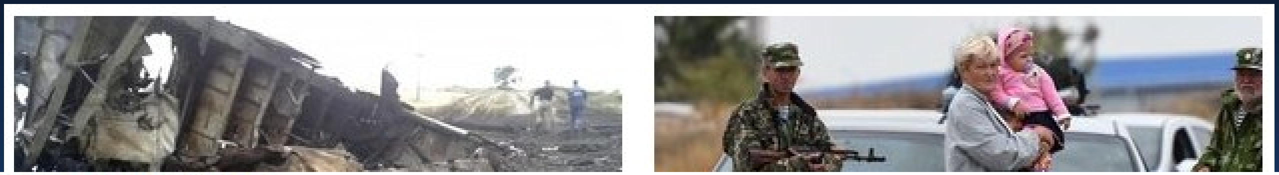
EU struggles to respond
European Union faces challenge to 'peace project'

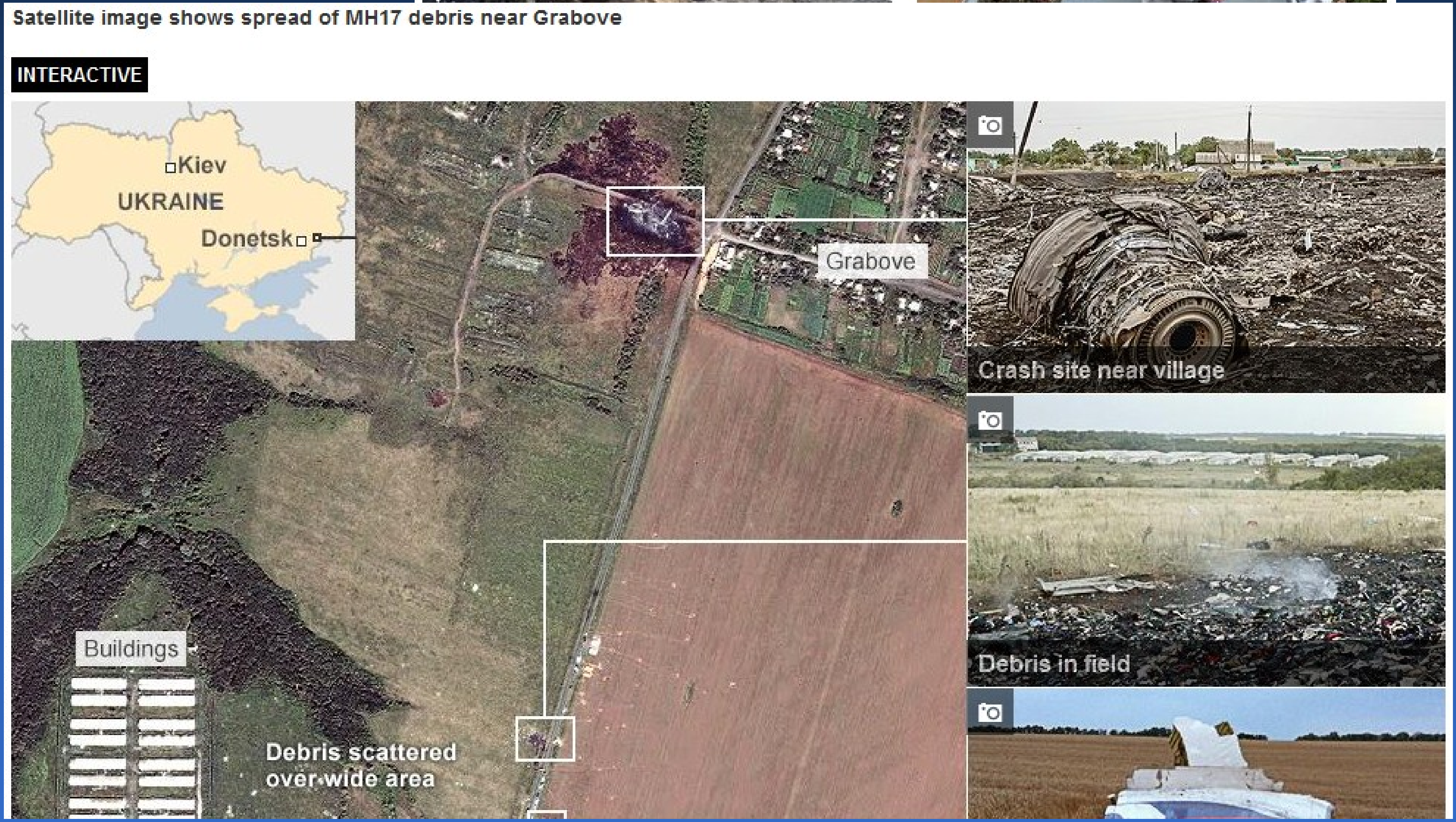
Conflict questions
Key issues in Ukraine fighting explained

Sanctions circle
Who are the Putin allies targeted by the West?

Key players in east
Profiles of key figures involved on both sides of the unrest in eastern Ukraine.

Source - BBC News - © 2014 BBC





Satellite image shows spread of MH17 debris near Grabove

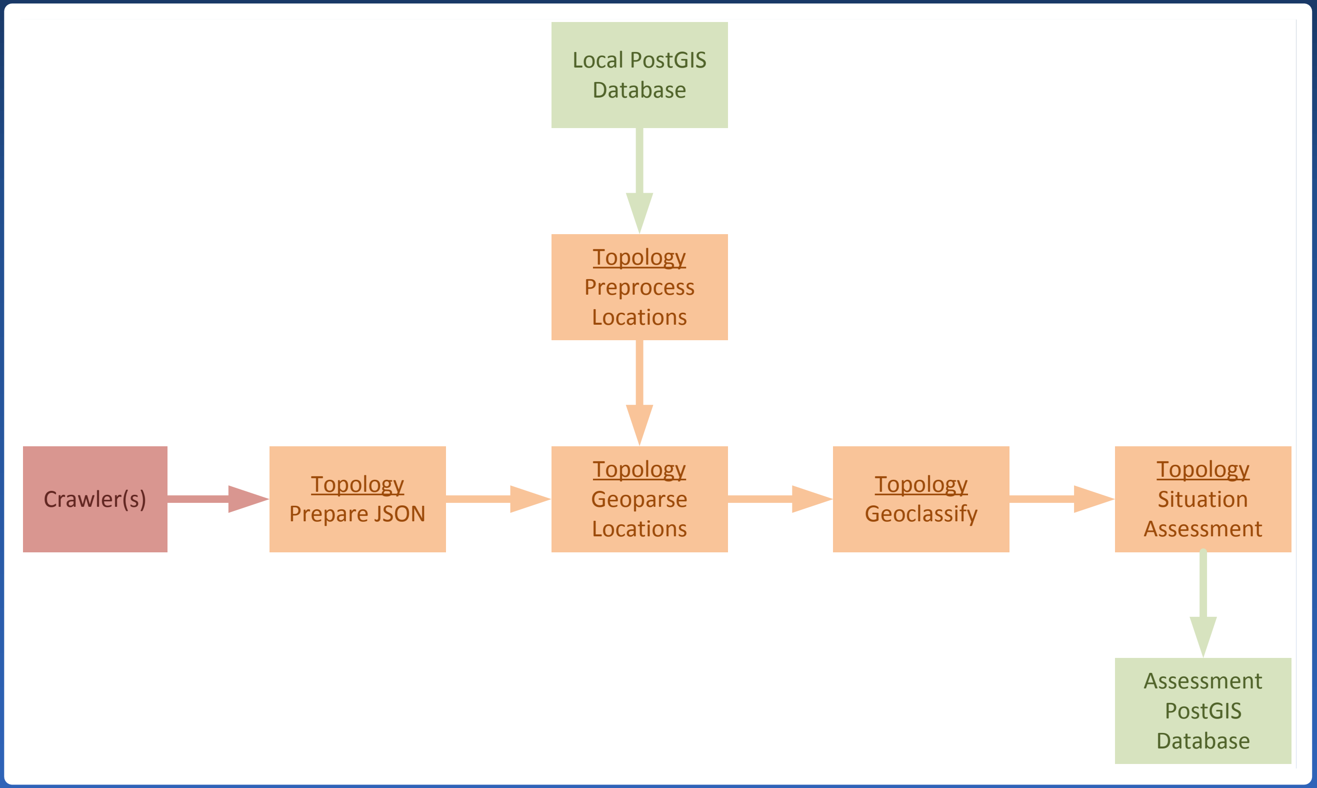
INTERACTIVE

Source - BBC News - © 2014 BBC

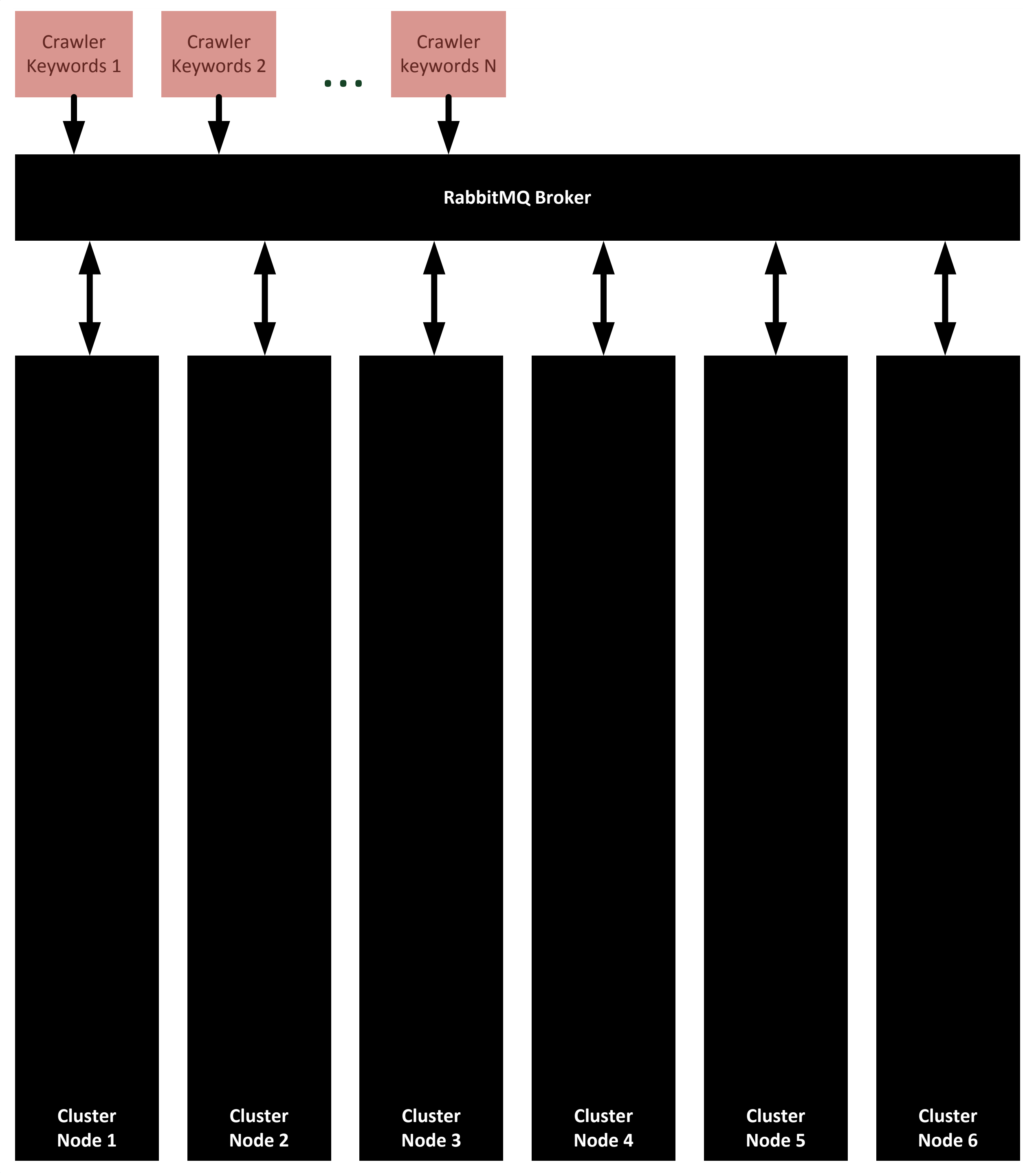
Scalable Processing Architecture

- Storm-based Distributed Processing Framework
  - Physical cluster of 1..N machines (e.g. 10 machine testbed)
  - Storm Nimbus → controller
  - Storm Zookeeper and Supervisors → workers
  - Storm topologies are dynamically deployed
    - Geoparse, Geoclassify, Situation Assessment ...
    - Topologies are a mixture of Java & Python code
  - Local databases for local topology results
    - Including 400 Gbytes of Planet OpenStreetMap PostGIS data
  - Assessment database for aggregated situation assessments
  - Highly scalable approach
    - parallel situation assessments? simply extra topologies to storm
    - more locations? more throughput? just add a few more computing nodes to the cluster
    - we will be testing on a 17 machine cluster

Scalable Processing Architecture


Local PostGIS Database
Topology Preprocess Locations
Crawler(s)
Topology Prepare JSON
Topology Geoparse Locations
Topology Geoclassify
Topology Situation Assessment
Assessment PostGIS Database

S e



Crawler Keywords 1

Crawler Keywords 2

...

Crawler keywords N

RabbitMQ Broker

Cluster Node 1

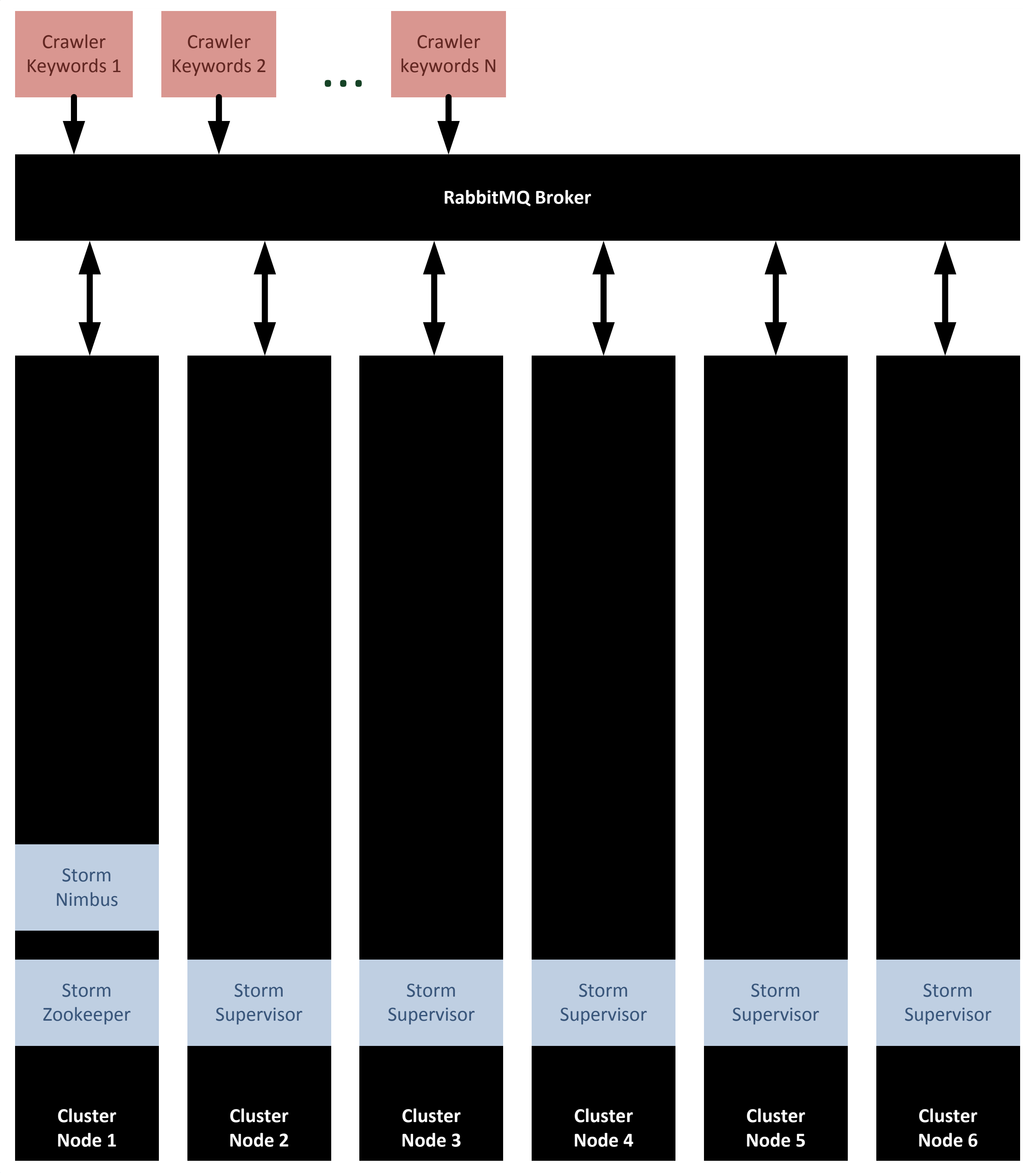
Cluster Node 2

Cluster Node 3

Cluster Node 4

Cluster Node 5

Cluster Node 6



Crawler Keywords 1

Crawler Keywords 2

…

Crawler keywords N

RabbitMQ Broker

Storm Nimbus

Storm Zookeeper

Cluster Node 1

Storm Supervisor

Cluster Node 2

Storm Supervisor

Cluster Node 3

Storm Supervisor

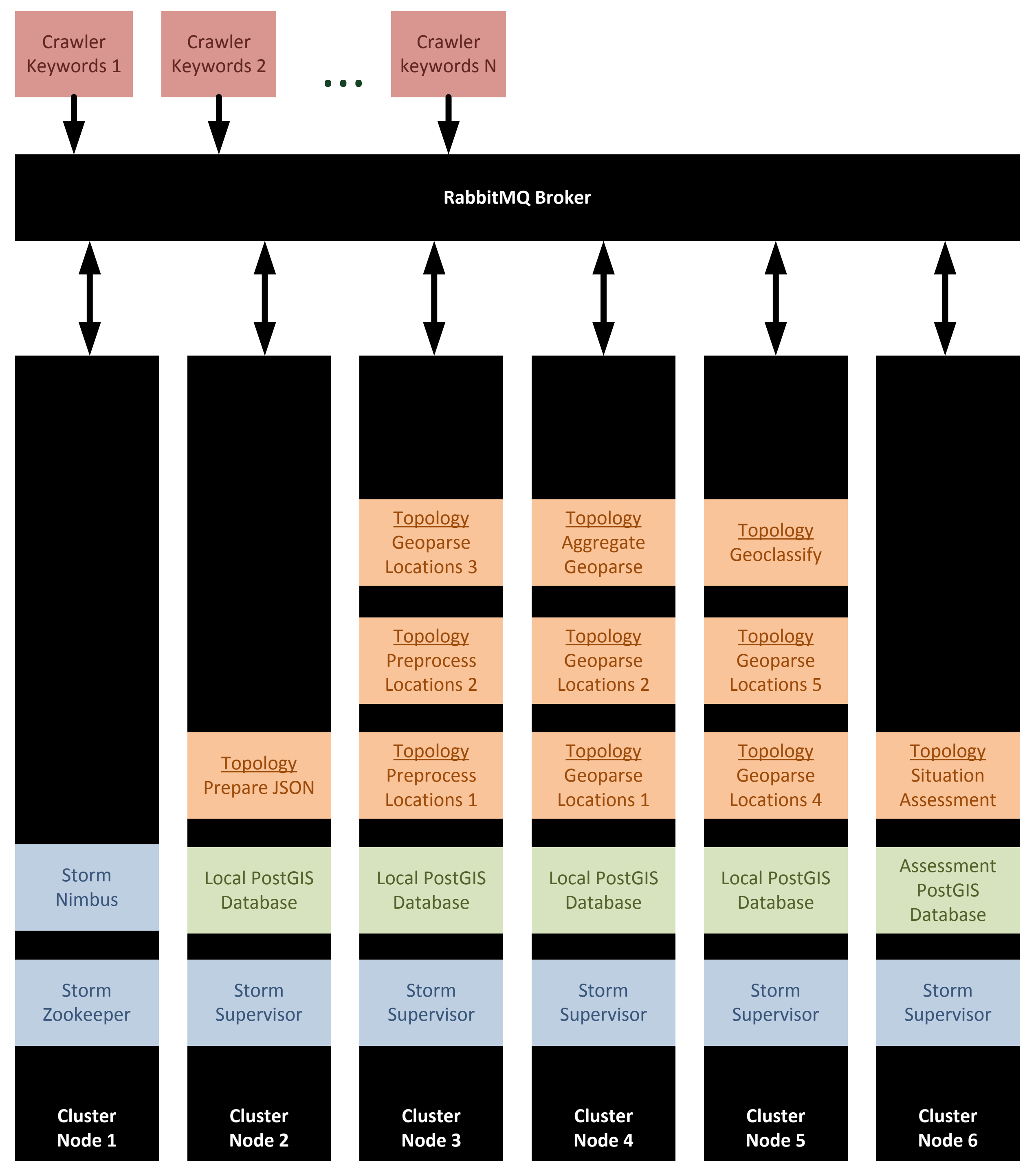
Cluster Node 4

Storm Supervisor

Cluster Node 5

Storm Supervisor

Cluster Node 6



Crawler Keywords 1

Crawler Keywords 2

...

Crawler keywords N

RabbitMQ Broker

Storm Nimbus

Storm Zookeeper

Cluster Node 1

Topology Prepare JSON

Local PostGIS Database

Storm Supervisor

Cluster Node 2

Topology Geoparse Locations 3

Topology Preprocess Locations 2

Topology Preprocess Locations 1

Local PostGIS Database

Storm Supervisor

Cluster Node 3

Topology Aggregate Geoparse

Topology Geoparse Locations 2

Topology Geoparse Locations 1

Local PostGIS Database

Storm Supervisor

Cluster Node 4

Topology Geoclassify

Topology Geoparse Locations 5

Topology Geoparse Locations 4

Local PostGIS Database

Storm Supervisor

Cluster Node 5

Topology Situation Assessment

Assessment PostGIS Database

Storm Supervisor

Cluster Node 6

Geoparsing

- What is ‘geoparsing’?
  - Assignment of geographic identifiers to text
- State of the Art - Geoparsing
  - Parts of Speech (POS) + Named Entity Recognition (NER) + Geocoding
    - NER is trained per language and can be error prone
    - Geocoding is slow and rate limited (e.g. Google Geocoder 1000 requests/day)
  - Named Entity Matching (NEM) + Global Gazetteers (e.g. Geonames)
    - NEM is fast and accurate but can suffer from low recall if token expansion poor
    - Gazetteers work at country & city level, not street & building level
  - Name disambiguation a big problem for both approaches
- Our Approach - Geoparsing
  - Named Entity Matching (NEM) + Planet Open Street Map (OSM)
  - OSM lookup for global regions (~300,000) with translated names
  - OSM lookup for focus areas in the native language of the area (e.g. all streets & buildings in a city)
  - Name disambiguation exploiting OpenGIS super-region SQL queries

Geoparsing

- Algorithm (pre-processing at startup and on-demand)
  - SQL queries of OSM OpenGIS for locations, super regions & tags
  - Heuristics to identify low quality OSM locations names - OSM labels are of variable quality
  - Apply language specific stopwords (e.g. common words & names)
  - Language specific abbreviations ➔ expand tokens sets (e.g. Street, St)
  - Cache and index blocks of locations into memory (e.g. one block of 10,000 locations per geoparse instance)
- Algorithm (runtime)
  - Continually receive JSON social media content (text) from RabbitMQ
  - Clean text ➔ tokenize ➔ compute N-gram phrases
  - Lookup phrases in cached location index ➔ possible matches
  - Aggregate all matches ➔ up-vote locations with super-region mentions
  - Annotate JSON content with location data ➔ publish to RabbitMQ

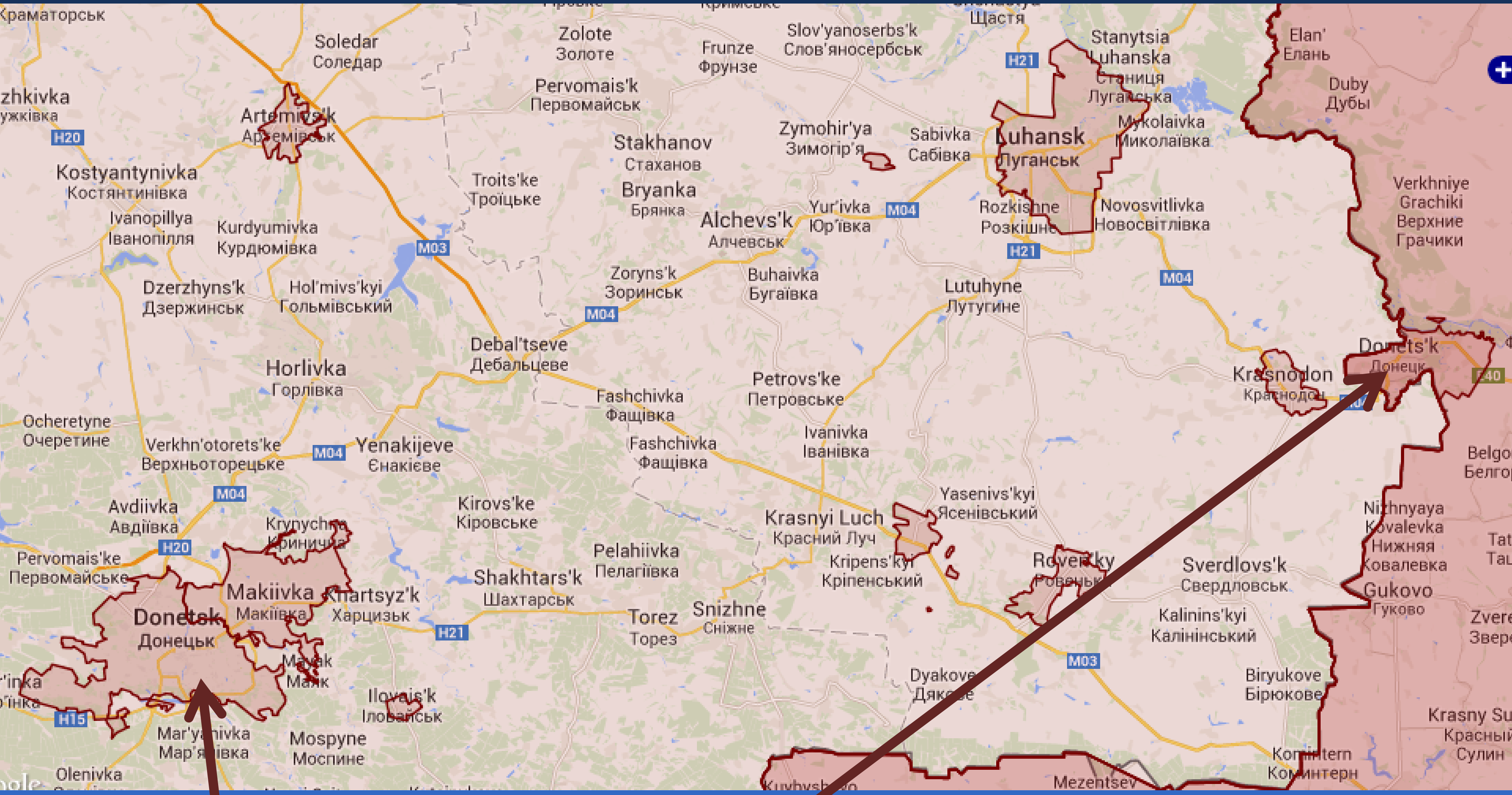
Geoparsing

- Case Study: Geoparsing Donetsk
  - During the Ukraine 2014 crisis the Donetsk area, and airport in particular, has witnessed a lot of fighting. However 'Donetsk' is a region in both Ukraine and Russia
  - This caused news agencies a lot of trouble - with incorrect maps being displayed in news reports until the confusion was resolved
  - To make things worse the Russian Donetsk is geospatially connected to the border of Ukraine (i.e. both 0 distance from Ukraine)
  - Our approach
    - Match all known locations called Donetsk (there are about 10 entries in OSM)
    - Favour high gram phrases over lower gram
      - e.g. **Donetsk Airport** is a preferred match over **Donetsk**
    - Upvote locations where a super region is mentioned in nearby text
      - e.g. See these pictures of rebels fighting in **Donetsk Airport**, **Ukraine** http://...
    - Upvote locations close geospatially to content geotag (if available)
    - Upvote locations geospatially close to other matched locations (e.g. nearby roads)
    - Rank location matches and select best set to report

PUBLIC

Geoparsing

Geospatial clustering
- Twitter
- You Tube, Instagram



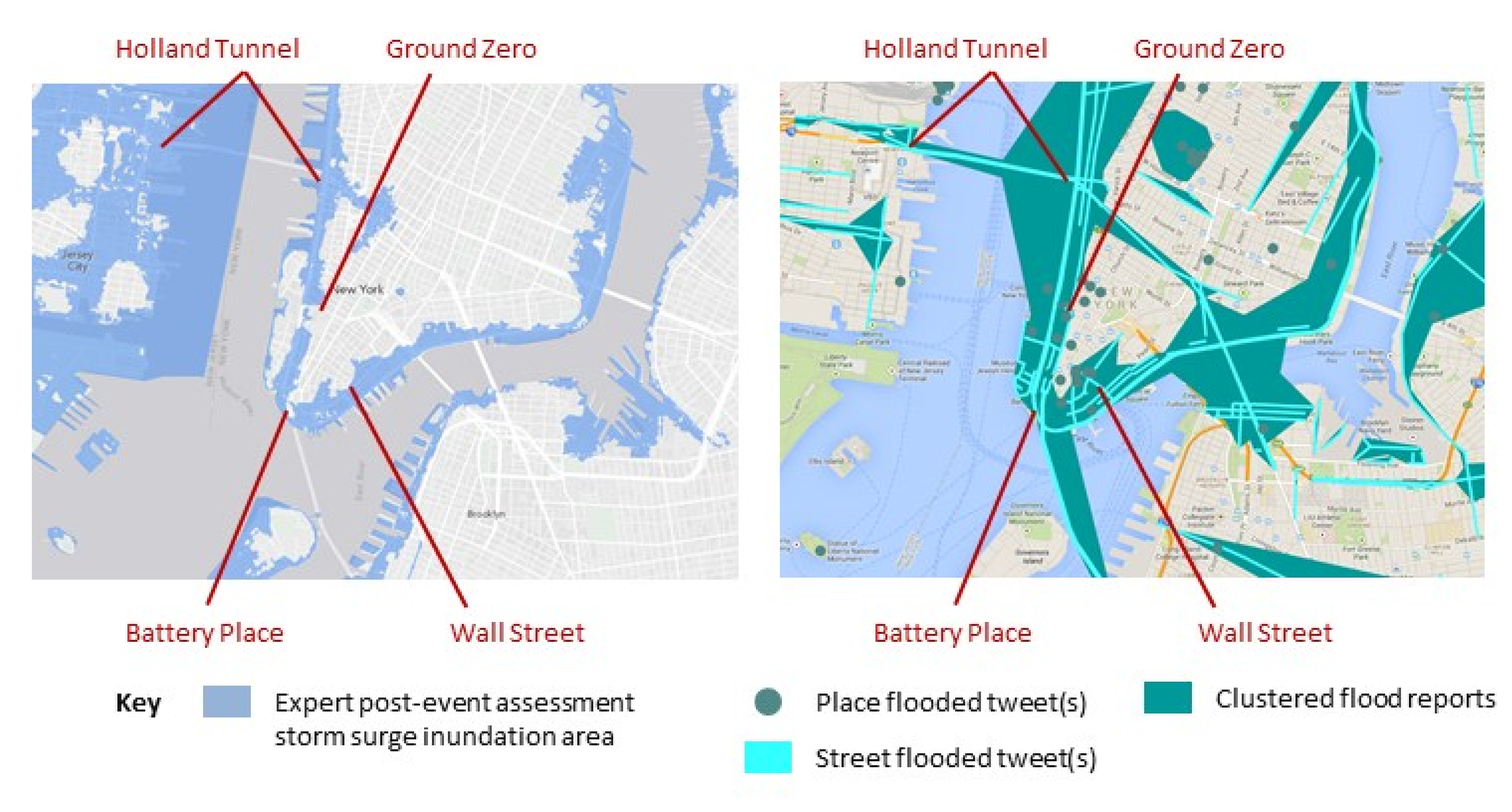
Donetsk (Ukraine)

Donetsk (Russia)

Map Baselayer - GoogleMaps

Geoparsing

- Peer Reviewed Scientific Results
  - Middleton, S.E. Middleton, L. Modafferi, S."Real-time Crisis Mapping of Natural Disasters using Social Media", Intelligent Systems, IEEE , vol.29, no.2, pp.9,17, Mar.-Apr. 2014

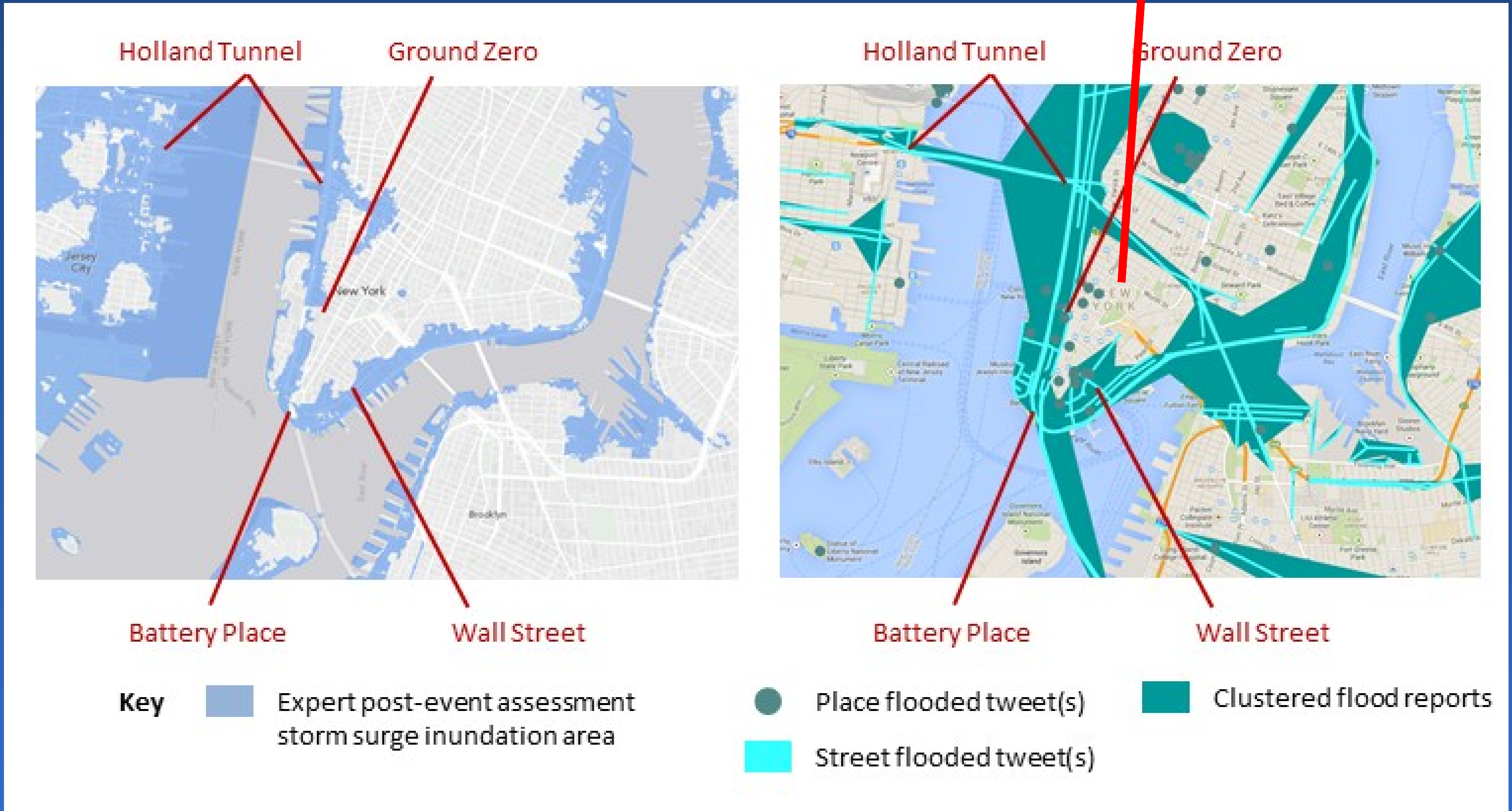


Hurricane Sandy Flooding, New York, Oct 2012

Geoparsing

- Peer Reviewed Scientific Results
  - Middleton, S.E. Middleton, L. Modafferi, S."Rea using Social Media", Intelligent Systems, IEEE , vol.29, no.2, pp.9,17, Mar.-Apr. 2014

Geoparse F1 scores from 0.83 to 0.95
English, Italian, Turkish



Hurricane Sandy Flooding, New York, Oct 2012

Geosemantics

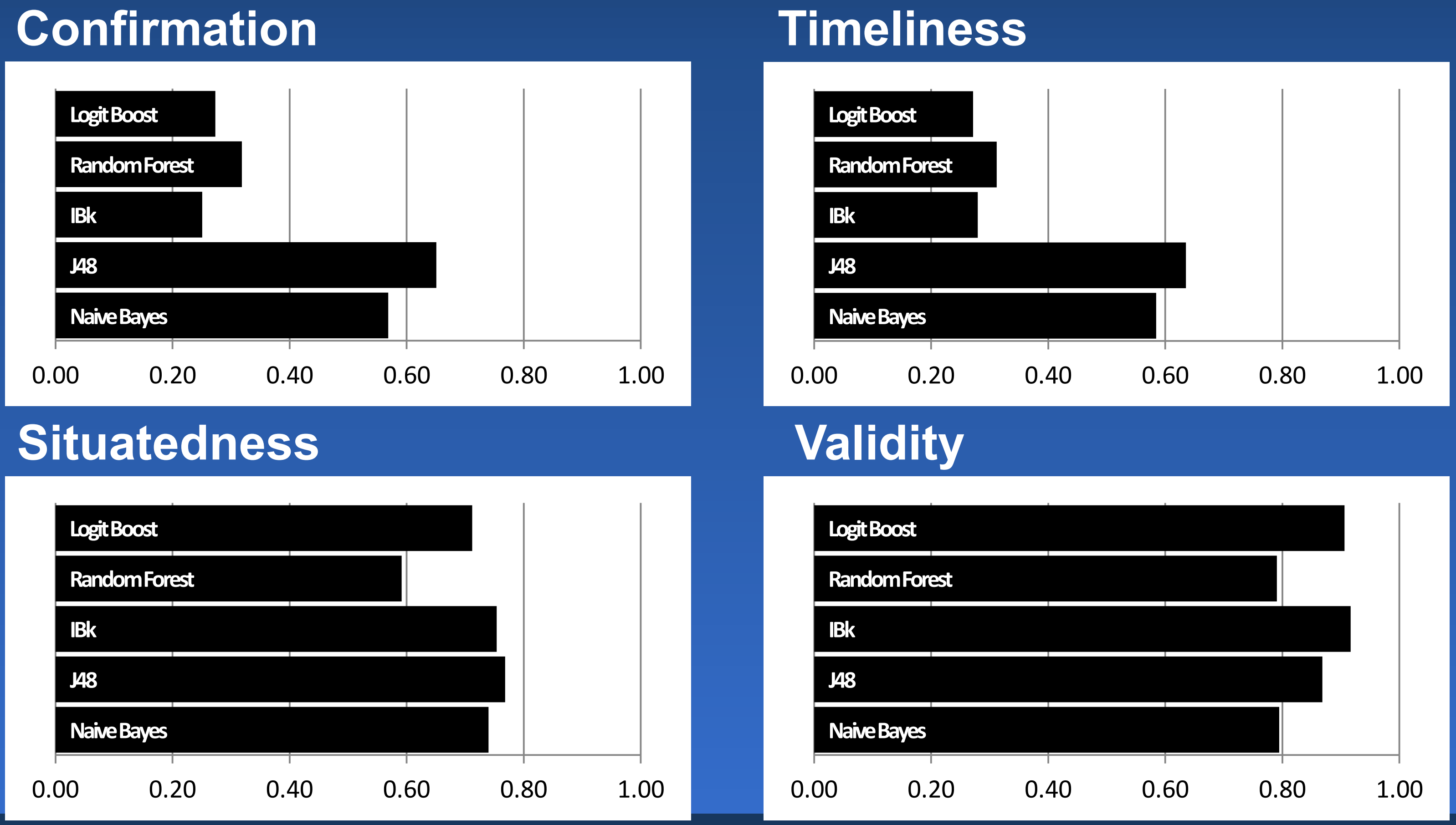
- What is ‘geosemantics’
  - Study of context of spatial data - in our case contextual text relating to mentioned locations
- State of the art – Geosemantic text analysis
  - Text + POS + training examples ➔ classifier ➔ event type
  - Location text ➔ NLP Grammar ➔ direction & distance
    - e.g. trouble spotted 5 miles north of London
  - Location text ➔ sentiment analysis ➔ good / bad opinion
  - Resilience of approaches across event types and languages an issue
- Our Approach – Geosemantic feature classification
  - Text + POS + LOC tag + lang specific training set ➔ calc features ➔ classifier ➔ context of how is location is talked about
    - Stanford and TreeTagger POS taggers used supporting 10+ languages
  - Features based on Text & POS usage close to LOC tokens
  - Classes ➔ past | future | present, insitu | remote, pos/neg report
  - Location matches ➔ class filters ➔ visualization & inference models

Geosemantics

- Case Study: Geosemantics for reports of Donetsk Airport
  - The airport in Donetsk has been the scene of a lot of fighting. There has also been a lot of Twitter chatter and You Tube video uploads.
  - Journalists would like to filter content to see eyewitness reports where author was in-situ within Donetsk, not safe at home commenting (e.g. in America commenting on TV news reports)
  - Our approach
    - Offline: train classifier using IT Innovation's labelled tweet dataset corpus of major events (flood, tornado, conflict & political referendum)
    - Online: stream social media content live
    - Content ➔ geoparse ➔ location set e.g. Donetsk Airport, Donetsk, Ukraine
    - POS tagging ➔ calculate features set ➔ classify feature set ➔ insitu | remote | na
    - ' fighting in airport donetsk – see photo from my mobile ' ➔ INSITU
    - ' News Report: new fighting breaks out in donetsk airport ' ➔ REMOTE

Geosemantics

- Peer Reviewed Scientific Results
  - Middleton, S.E. Krivcovs, V. “Geosemantic Feature Extraction from Social Media for Trust and Credibility Analysis of Breaking News”, draft paper ACM TOIS

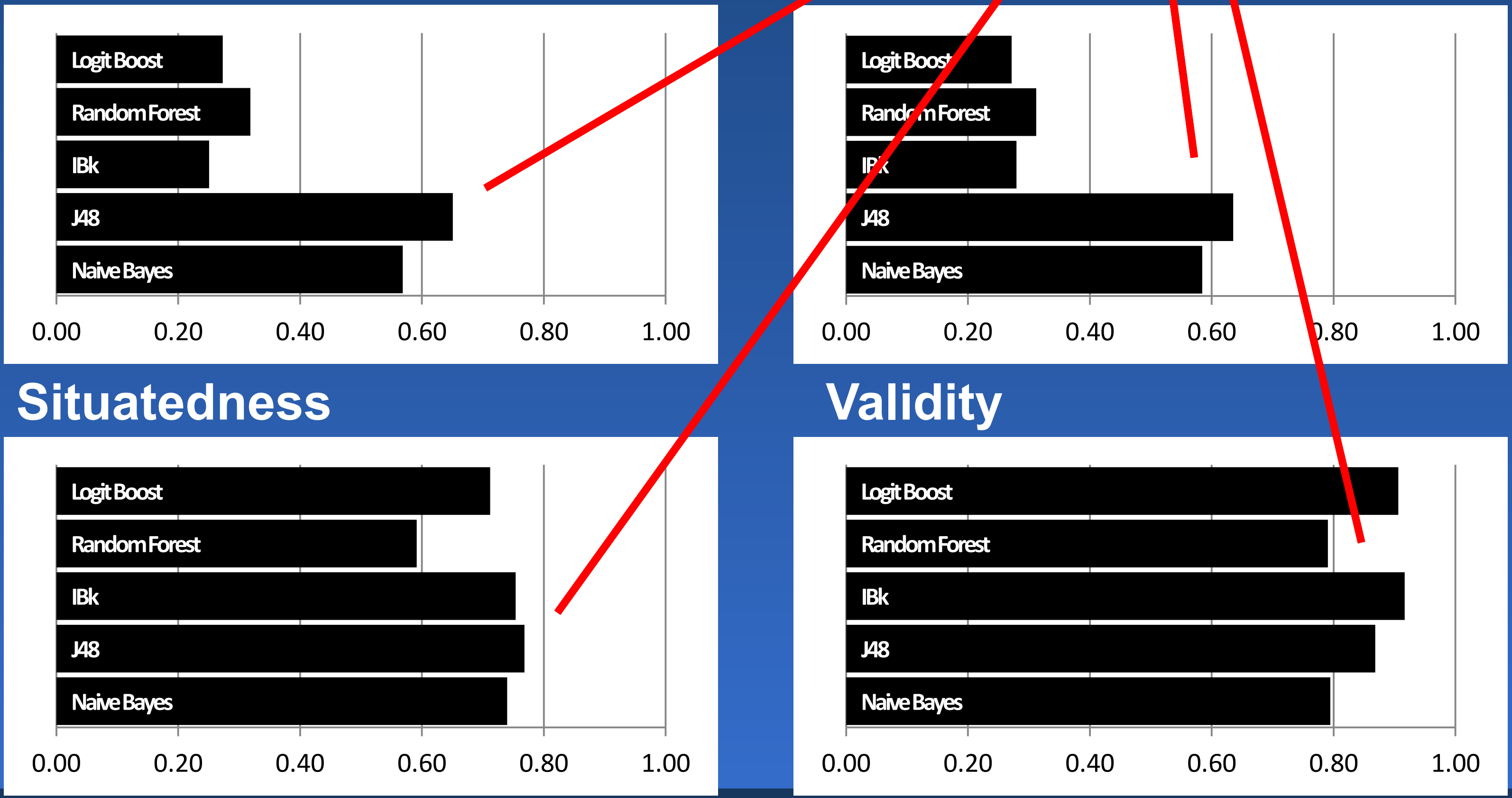


Mean F1 Scores by Class. Leave one out cross-fold validation for 4 news events - NY Flooding 2012, Oklahoma Tornado 2013, Scottish Referendum 2014, Ukrainian Conflict 2014

Geosemantics

- Peer Reviewed Scientific Results
  - Middleton, S.E. Krivcovs, V. “Geosema and Credibility Analysis of Breaking News”, draft paper ACM TOIS

Geoclassify F1 scores from 0.64 to 0.87 English



Confirmation

Timeliness

Situatedness

Validity

Leave one out cross-fold validation for tweet event datasets - NY Flooding 2012, Oklahoma Tornado 2013, Scottish Referendum 2014, Ukrainian Conflict 2014

Trust and Credibility Modelling

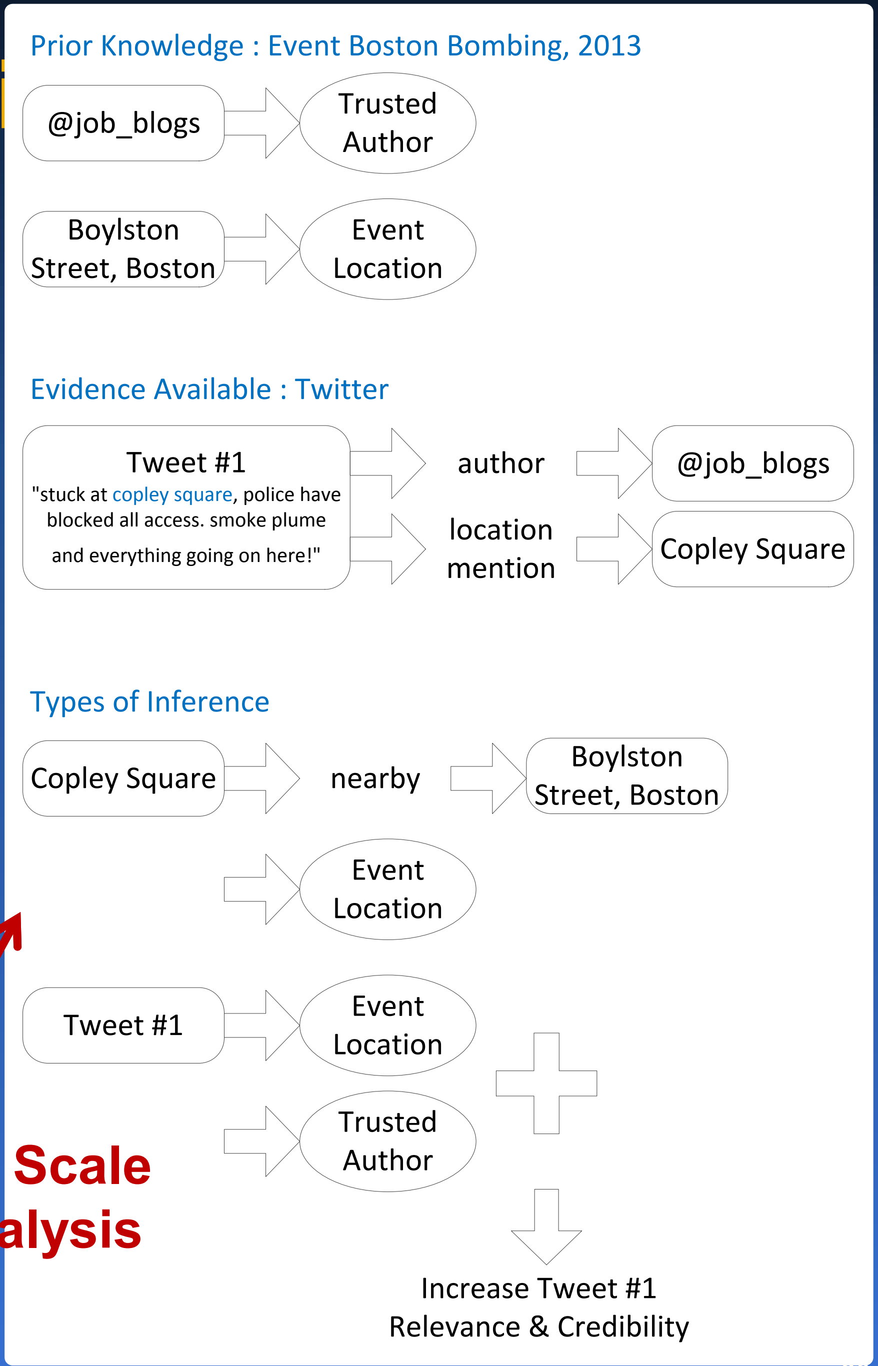
- What is ‘relevance’, ‘credibility’ and ‘trust’
  - Relevance - how well content matches any given search criteria
  - Trust and credibility not well defined – below is our interpretation
  - Credibility - consistency with other content (e.g. similar reports) and contextual information (e.g. local geography)
  - Trust - subjective assessment of likelihood of content being false
  - A credible news report might still be false!
- State of the Art – Trust and Credibility Modelling
  - Unsupervised learning (e.g. Bayesian Network, Damper Shafer) → trust prediction without explanation
  - Supervised reputation models → trust prediction with explanation
  - Heuristics & activity metrics → trust prediction with explanation
  - Features used include text, classified topic, activity metrics (e.g. likes, comments), social network connections and lists of trusted people

Trust and Credibility Modelling

- Our Approach – Knowledge-based Trust Modelling
  - Each journalist has their own personal set of trusted ‘news hounds’ whom they have come to rely upon
  - Knowledge-based approach allows analysts to assert a-priori trusted lists of people and known event context for breaking news stories
  - Geoparse + geoclassification + other ➔ evidence + a-priori context ➔ triple store ➔ OWL inference to classify evidence
    - OWL individuals, OWL restrictions, SPARQL, GeoSPARQL ...
    - Incremental supporting journalist feedback
  - Interactive tools for analysis to use class filters, looking at different combinations of evidence
    - supporting analysts to do their job better
    - NOT 100% automating the task and yielding unexplained results
  - Scalable approach able to represent different viewpoints

Trust and Credibi

- Working with Journalists
  - IT Innovation is collaborating with journalists from Deutsche Welle (German national news agency)
  - Expertise on social media report verification
- Peer Reviewed Scientific Results
  - Middleton, S.E."From Twitter-based Crisis Mapping to Large-scale Real-Time Situation Assessment with Trust and Credibility Analysis ", REVEAL R&D results, July 2014
  - http://revealproject.eu/
  - Work ongoing



Prior Knowledge : Event Boston Bombing, 2013

@job_blogs → Trusted Author

Boylston Street, Boston → Event Location

Evidence Available : Twitter

Tweet #1
"stuck at copley square, police have blocked all access. smoke plume and everything going on here!"

→ author → @job_blogs

→ location mention → Copley Square

Types of Inference

Copley Square → nearby → Boylston Street, Boston

→ Event Location

Tweet #1 → Event Location

→ Trusted Author

+

Increase Tweet #1 Relevance & Credibility

Simple Inference Applied at Scale for Trust and Credibility Analysis

Exploitation

- Prototype system @ IT Innovation
  - Scalable storm cluster deployment
  - Geoparsing and geosemantic support
  - Situation assessment and decision support visualizations
  - Interactive knowledge-based trust models for analysts
  - Prototype system © IT Innovation which we can bring as background to future collaborative projects and commercial prototype work
- Geosemantics library @ IT Innovation
  - Python-based geoparsing and geosemantics library © IT Innovation
  - Functions to geoparse text using a local Open Street Map database and do multi-lingual geosemantic classification of text
  - IT Innovation is considering making library open source at end of the REVEAL project (2016) for non-commercial community applications and evaluation purposes

Future Work

- Roadmap going forward
  - Testing a large scale deployment on 17 machine cluster
  - Journalist ethnographic studies to validate trust and credibility models using real news events VS real journalists as a ground truth
  - Further evaluation of geoparse and geosemantic library
  - Develop and refine interactive visualizations for situation assessment and trust model analysis sessions

- We are always looking for future collaborative and contract R&D opportunities

Thanks for your attention!

Any questions?

Stuart E. Middleton
University of Southampton IT Innovation Centre

email: sem@it-innovation.soton.ac.uk
web: www.it-innovation.soton.ac.uk
twitter:@stuart_e_middle
REVEAL project: www.revealproject.eu