PUBLIC

Floraguard

Geospatial research into social media and online marketplaces

Disaster management, Breaking news and UK illegal plant trade

Stuart E. Middleton

University of Southampton, Electronics and Computer Science

www.ecs.soton.ac.uk/people/sem

UCL Seminar 2018

17th Oct 2018

Overview

- Location Extraction & Geoparsing
  - Use Cases, Algorithm
  - Discussion: Velocity
  - Discussion: Veracity
- Geosemantic Analysis
  - Use Cases, Algorithm
  - Discussion: Veracity
- Open Information Extraction
  - Use Cases, Algorithm
  - Discussion: Variety
- Lessons Learnt

Speaker

- Dr Stuart E. Middleton
  - Senior research engineer
  - University of Southampton, Electronics and Computer Science (ECS), IT Innovation Centre
- Research
  - Computational linguistics and information extraction
- Interdisciplinary
  - Disaster early warning & response (GFZ TRIDEC)
  - Journalists (Deutsche Welle Reveal)
  - Archaeologists (British Museum GRAVITATE)
  - Law enforcement agencies (UK Border Force  Floraguard UK National Crime Agency)

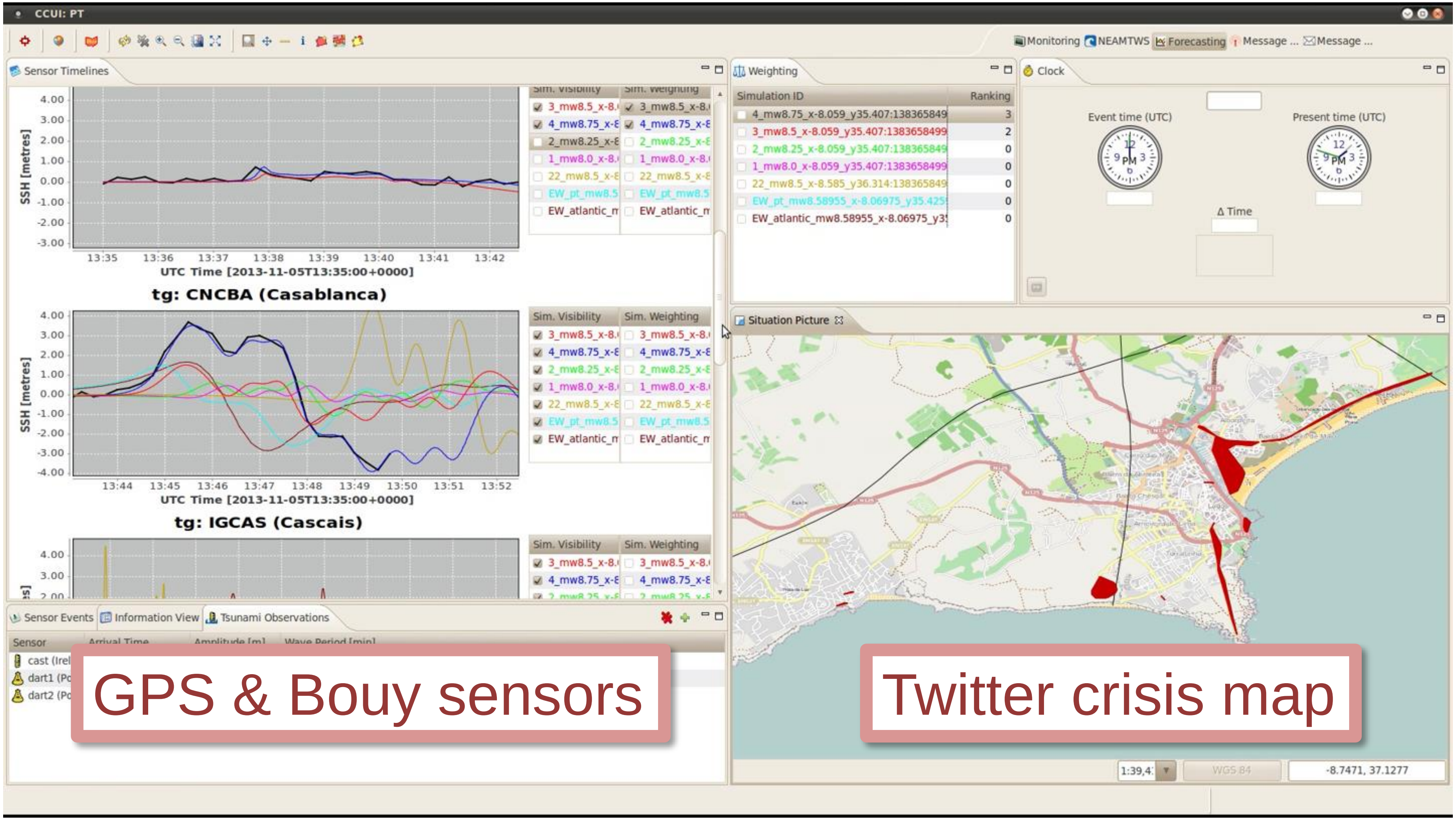
Location Extraction & Geoparsing

- Terminology - my definitions
  - Geocoding
    - Address >> Spatial reference (e.g. coordinate)
  - Geoparsing
    - Free Text >> Location(s) >> Disambiguated location(s)
    - Optionally can also provide spatial reference(s)
  - Geotagging
    - Free Text >> Spatial reference (e.g. coordinate)
  - Location identification
    - Geoparsing without location disambiguation
  - Location estimation
    - Geotagging to a spatial area such as a grid cell
- Location and Toponym used interchangeably

Location Extraction & Geoparsing

- Case studies
  - TRIDEC
    - Geoparsing social media around crisis events
    - Tsunami early warning >> 5 to 60 minutes coastline warnings
    - Earthquake >> Tsunami wave simulation >> Coastline impact SMS warnings via mobile phone system
    - Social media flood maps >> Actual wave impact times >> Adjust Tsunami wave simulation

Location Extraction & Geoparsing

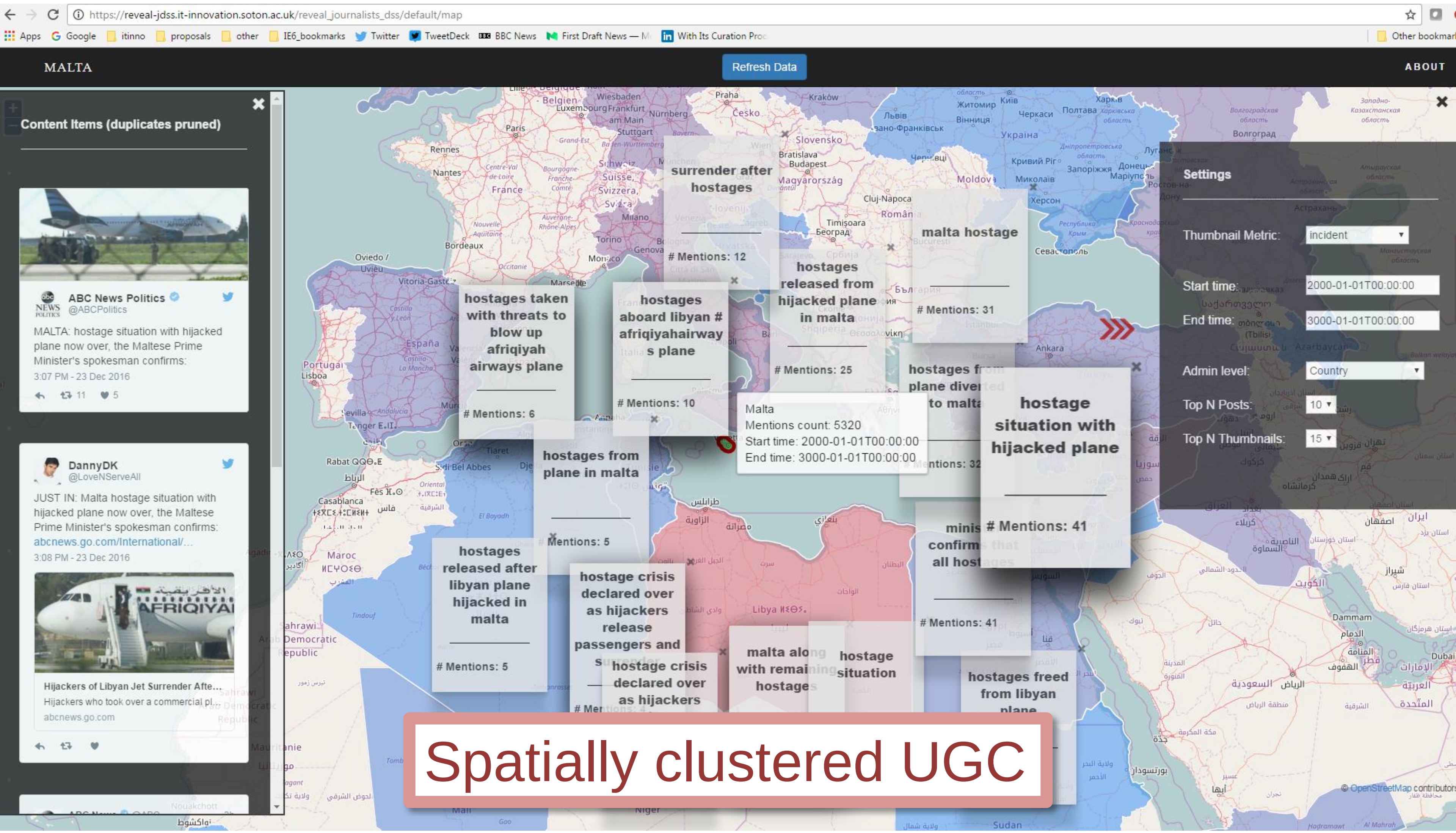


Screenshot from TRIDEC early prototype command and control UI, GFZ German Research Centre for Geosciences
TRIDEC Cloud - https://trideccloud.gfz-potsdam.de

Location Extraction & Geoparsing

- Case studies
  - REVEAL
    - Geoparsing social media for breaking news
    - News event >> 10 to 30 minute breaking news window
    - User Generated Content (UGC) >> Eyewitness images & videos >> Need AI to filter to avoid overloading journalists
    - Interactive map of real-time UGC

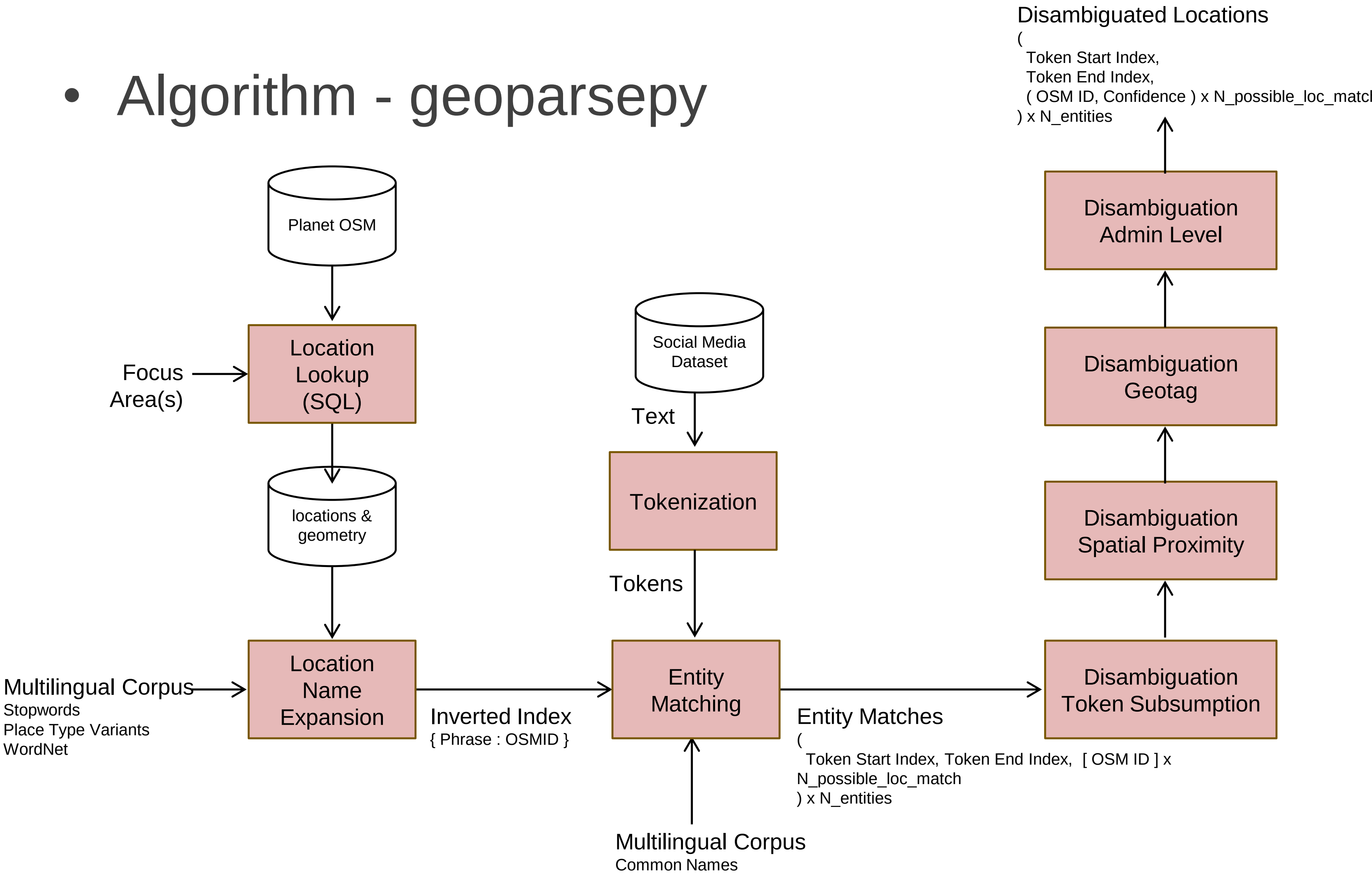
Location Extraction & Geoparsing



Screenshot from Journaliust Decision Support System (JDSS), REVEAL project
JDSS - https://reveal-jdss.it-innovation.soton.ac.uk/reveal_journalists_dss

Location Extraction & Geoparsing

- Algorithm - geoparsepy



https://pypi.org/project/geoparsepy

Location Extraction & Geoparsing

- OpenStreetMap Planet OSM Pre-processing
  - OSM planet > osm2pgsql > PostgreSQL + PostGIS
  - Area of interest for pre-processing
    - Global cities and countries
    - Focus area definition
      - Full name, Set of relation OSMID's, Point & radius, Polygon
      - e.g. Greater Paris
  - SQL query to capture location data
    - SQL WITH >> admin relations >> Lookup index
    - SQL >> polygons (admin) in focus area >> Admin table
    - SQL >> polygons (not admin) in focus area >> Admin lookup >> Polygon table
    - SQL >> lines in focus area >> Admin lookup >> Line table
    - SQL >> points in focus area >> Admin lookup >> Point table
    - Lookup OSM relation, way, node tables to extract OSM metadata

Location Extraction & Geoparsing

- Entity Matching - In-memory Location Cache
  - Load pre-processed focus area tables
  - Token expansion using location name variants
    - e.g. OSM multi-lingual names, short names and acronyms
  - Token expansion using location type variants
    - e.g. street, st.
  - Token filtering against WordNet, stoplists and lists of peoples first names
  - Prefix checking against name list
    - e.g. Victoria Derbyshire != Derbyshire

Location Extraction & Geoparsing

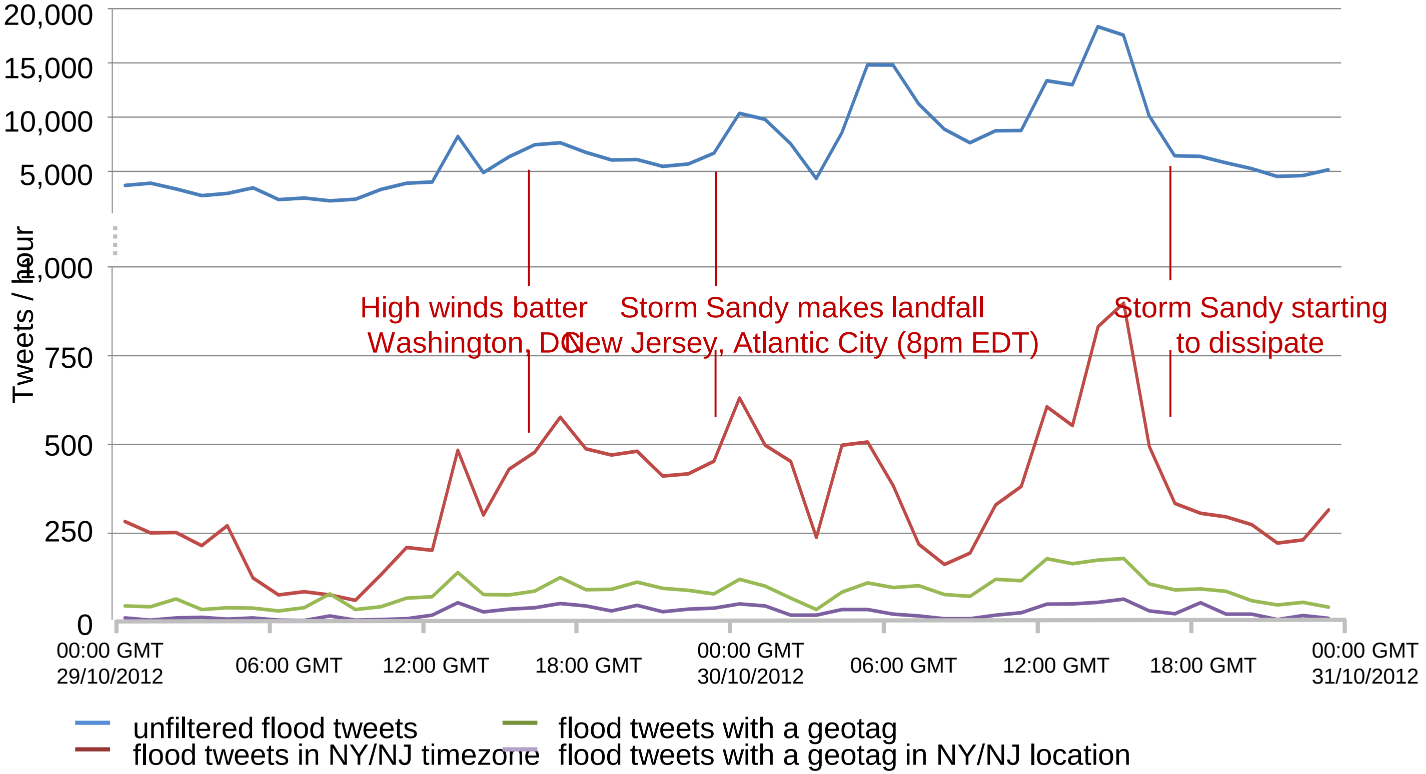
- Location Disambiguation
  - Token subsumption
    - Prefer full location phrases over partial ones
    - ‘New York’ >> [New York, USA] better match than [York, UK]
  - Spatial proximity & Geotag
    - Prefer locations where a parent region OR nearby location OR geotag is mentioned for context
    - ‘New York in USA’ >> [New York, USA] better match than [New York, BO, Sierra Leone]
  - OSM admin level
    - Prefer higher OSM admin levels to lower admin levels
    - ‘New York’ >> [New York, USA, OSM admin level 4] better than [New York, BO, Sierra Leone, OSM admin level n/a as its a suburb]

Location Extraction & Geoparsing

- Discussion: Velocity
  - geoparsepy is naively parallelizable
    - Single machine : Python multiprocessing lib
    - Cluster : APACHE Storm

Location Extraction & Geoparsing

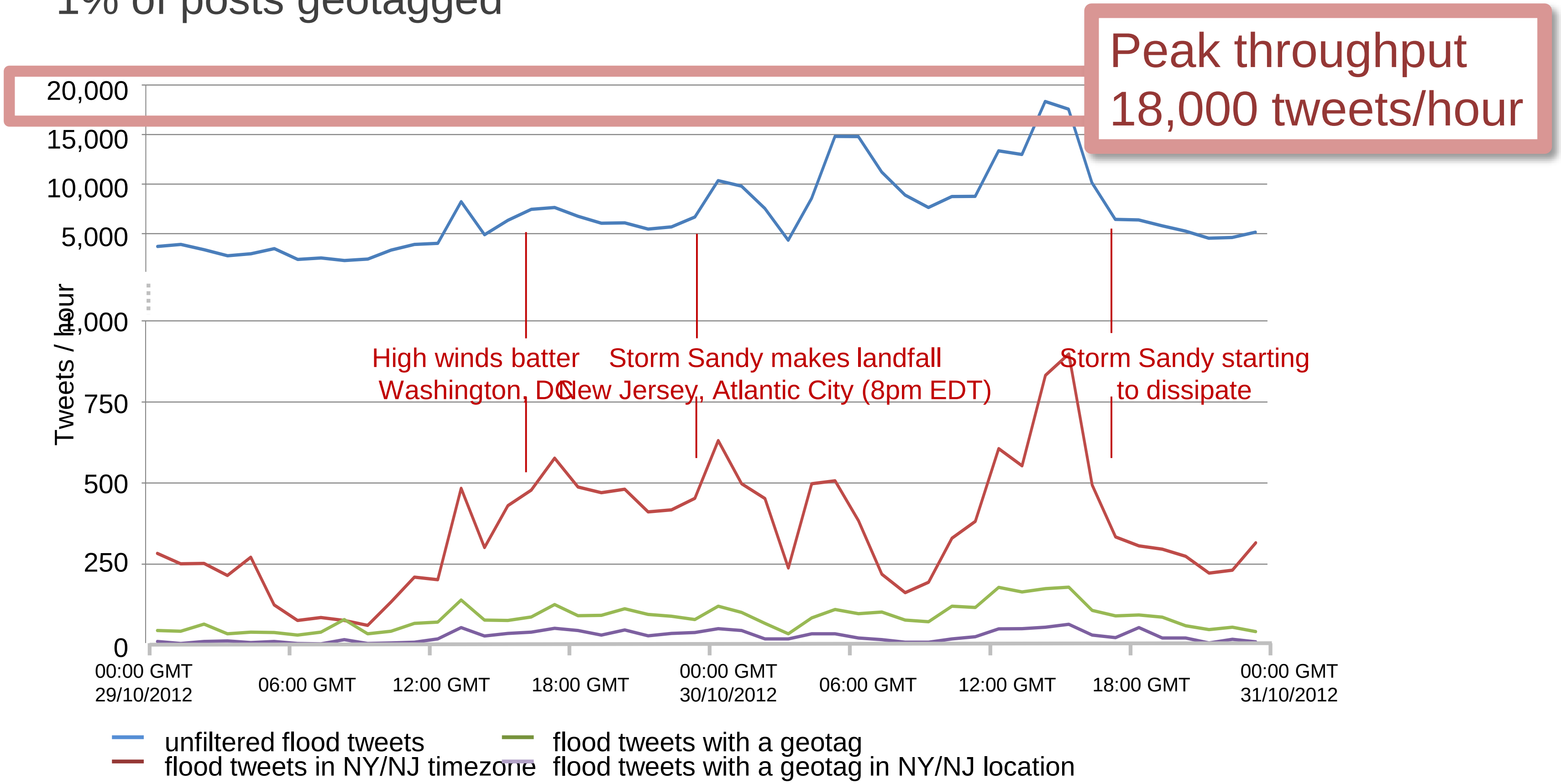
- Discussion: Velocity
  - Hurricane Sandy, Oct 2012, 5 days, Twitter Streaming API (1% sample size)
  - Dataset: 597,000 tweets, 4,300 location mentions, ~170 unique locations, 1% of posts geotagged



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

Location Extraction & Geoparsing

- Discussion: Velocity
  - Hurricane Sandy, Oct 2012, 5 days, Twitter Streaming API (1% sample size)
  - Dataset: 597,000 tweets, 4,300 location mentions, ~170 unique locations, 1% of posts geotagged



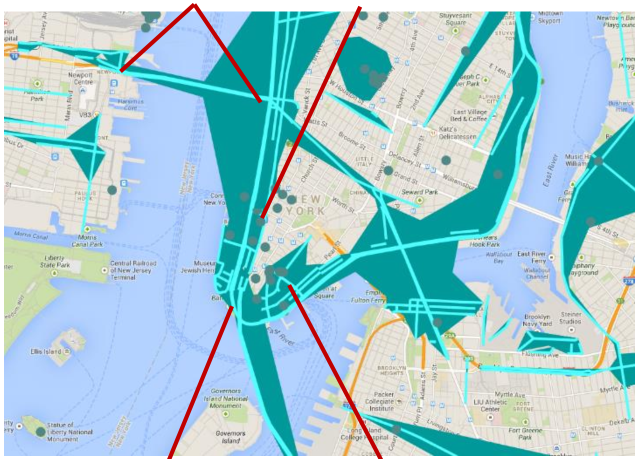
Middleton, S.E. Middleton, L. Modafferi, S. "Real-time Crisis Mapping of Natural Disasters using Social Media", Intelligent Systems, IEEE , vol.29, no.2, pp.9,17, Mar.-Apr. 2014

Location Extraction & Geoparsing

- Discussion: Velocity
  - geoparsepy throughput
    - Equipment setup
      - Single 2GHz CPU core
      - Global country/cities, 422,946 locations, 11 Gbytes RAM
    - Geoparsing throughput on UK election 2015 twitter posts
      - Load cache 0.005s / loc (35min), Geoparse 0.015s / tweet (66/s)
  - Scales up linearly with extra CPU cores
    - Loading location cache is a one-off process setup cost
    - Trade-off - large RAM footprint for higher throughput
  - Options for parallelization
    - Split text between processes, each process has full location set
      - RAM footprint: N x locations, 1 x text
    - Split location set between processes, each process has full text
      - RAM footprint: 1 x locations, N x text

Location Extraction & Geoparsing

- Discussion: Veracity
  - Social media crisis map (right)
  - Ground truth: US Federal Emergency Management Agency (FEMA) storm surge map from aerieal photography (left)



Key Expert post-event assessment storm surge inundation area

Place flooded tweet(s)

Street flooded tweet(s)

Clustered flood reports

Crisis map accuracy for a 8x8 Map segmentation at different map thresholds

Threshold	Places	Streets	Precision	Recall	F1
dev_sma > 0	66	101	0.78	0.76	0.77
dev_sma > 0.0168	19	22	0.81	0.44	0.57
dev_sma > 0.1	4	1	1.00	0.09	0.17

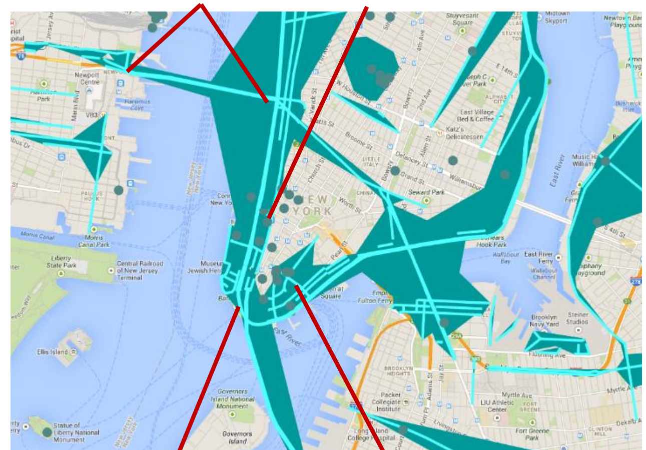
Location Extraction & Geoparsing

- Discussion: Veracity
  - Social media crisis map (right)
  - Ground truth: US Federal Emergency Management Agency (FEMA) storm surge map from aerieal photography (left)

Holland Tunnel Ground Zero

Battery Place Wall Street

Holland Tunnel Ground Zero



Battery Place

Key Expert post-event assessment storm surge inundation area

Place flooded tweet(s)

Street flooded tweet(s)

Map cell precision 0.78+

Crisis map accuracy for a 8x8 Map segmentation at different map thresholds

Threshold	Places	Streets	Precision	Recall	F1
dev_sma > 0	66	101	0.78	.76	0.77
dev_sma > 0.01	6819	22	0.81	.44	0.57
dev_sma > 0.1	4	1	1.00	.09	0.17

Location Extraction & Geoparsing

- Discussion: Veracity
  - Geoparse twitter benchmark dataset
  - Ground truth: Manually labelled locations within dataset
  - https://www.southampton.ac.uk/~sem03/geoparsepy/readme.html

Event	# Tweets	Crawler Keywords	Language	Date	# Regions mentioned	# Streets mentioned	# Buildings mentioned	# Locations mentioned	Spatial mention coverage
New York, USA Hurricane Sandy	1996	flood hurricane storm	Mostly English	Oct 2012	85	18	48	151	US South Coast
Christchurch, NZ Earthquake	2000	earthquake quake #eqnz	Mostly English	Feb 2011	33	24	64	121	New Zealand
Milan, Italy Blackout	391	blackout	Mixture English & Italian	May 2013	17	8	10	35	Milan
Turkey Earthquake	2000	earthquake quake deprem	Mostly Turkish	May 2012	51	0	0	51	Turkey

Middleton, S.E. Kordopatis-Zilos, G. Papadopoulos, S. Kompatsiaris, Y. "Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging", ACM Transactions on Information Systems (TOIS) 36, 4, Article 40 (June 2018)

Location Extraction & Geoparsing

- Discussion: Veracity
  - Geoparse twitter benchmark dataset
  - Ground truth: Manually labelled locations within dataset
  - https://www.southampton.ac.uk/~sem03/geoparsepy/readme.html

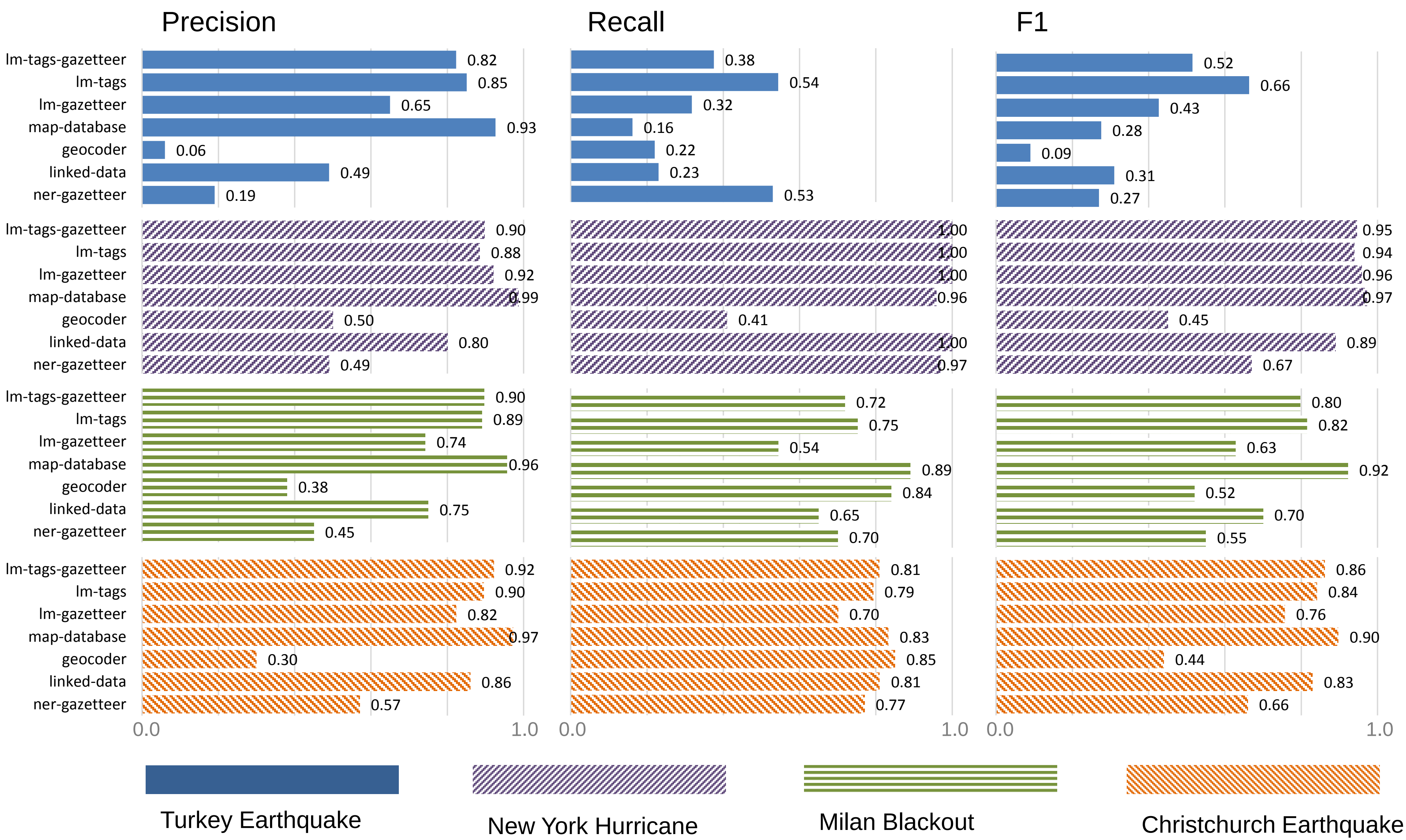
Event	# Tweets	Crawler Keywords	Language	Date	# Regions mentioned	# Streets mentioned	# Buildings mentioned	# Locations mentioned	Spatial mention coverage
New York, USA Hurricane Sandy	1996	flood hurricane storm	Mostly English	Oct 2012	85	18	48	151	US South Coast
Christchurch, NZ Earthquake	2000	earthquake quake #eqnz	Mostly English	Feb 2011	33	24	64	121	New Zealand
Milan, Italy Blackout	391	blackout	Mixture English & Italian	May 2013	17	8	10	35	Milan
Turkey Earthquake	2000	earthquake quake deprem	Mostly Turkish	May 2012	51	0	0	51	Turkey

Geoparse twitter benchmark dataset

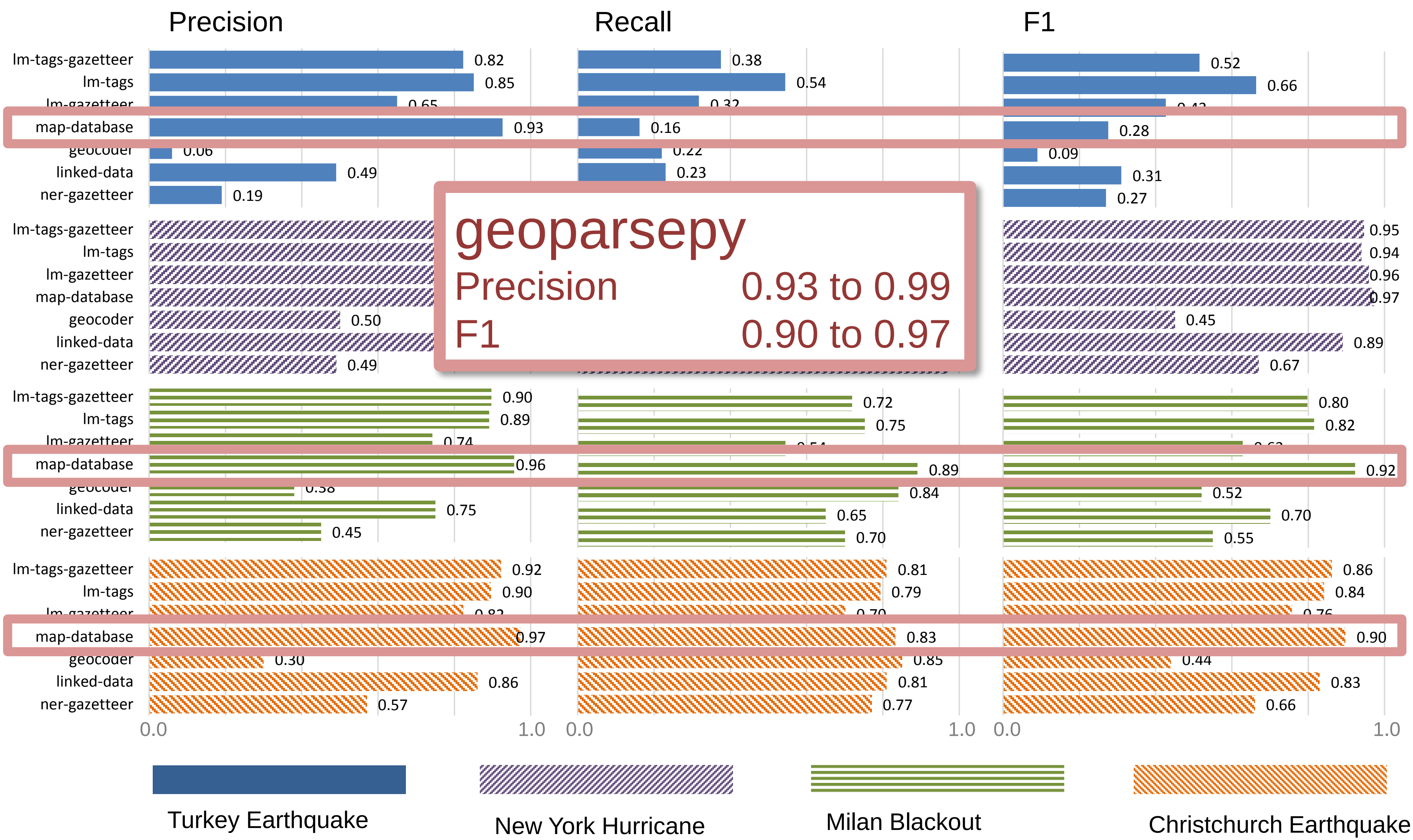
4 events, 6,387 tweets, 358 locations mentioned

Middleton, S.E. Kordopatis-Zilos, G. Papadopoulos, S. Kompatsiaris, Y. "Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging", ACM Transactions on Information Systems (TOIS) 36, 4, Article 40 (June 2018)

Location Extraction & Geoparsing



Location Extraction & Geoparsing



Location Extraction & Geoparsing

- Failure analysis

Pattern *[frequency seen from manual inspection]*	Algorithms which had trouble	Example	Correct Location
Common terms mistaken for location names *[very common]*	geocoder ner-gazetteer	This is the end of my Hurricane Sandy live-tweeting day 1	None. Mistaken location was Hurricane, UT 84737, USA
Peoples names that are also location names *[common]*	geocoder linked-data ner-gazetteer	Webgrrls hosting company is flooded by #Sandy	Sandy, UT, USA
Locations without any context *[common]*	geocoder ner-gazetteer	The city has high winds and flooding by the coastal lines	City of London, London, UK
Not in a well formatted address *[rare]*	geocoder	Street flooding #NYC: 48th Ave	48th St, New York, NY, USA
Spelling mistakes *[rare]*	map-database linked-data ner-gazetteer lm-tags-gazetteer	earthquake in Chrristchurch New Zealand ghastly	Christchurch, New Zealand
Saints and peoples title confused with place type abbreviations *[rare]*	geocoder linked-data lm-tags-gazetteer	I agree with St. Mary on this topic	None. Mistaken location was 1928 St Marys Rd, Moraga, CA 94575, USA
Vernacular names and abbreviations *[very rare on average but depends on event]*	map-database linked-data ner-gazetteer	CHCH hospital has been evacuated	Christchurch Hospital, 2 Riccarton Ave, Christchurch Central, Christchurch 8011, New Zealand
Street names in unpopular locations *[very rare on average but depends on event]*	linked-data ner-gazetteer	Anyone have news of St Margarets Girls College Winchester St Merivale	Margarets Girls College, 12 Winchester St, Canterbury 8014, New Zealand

Lessons Learnt

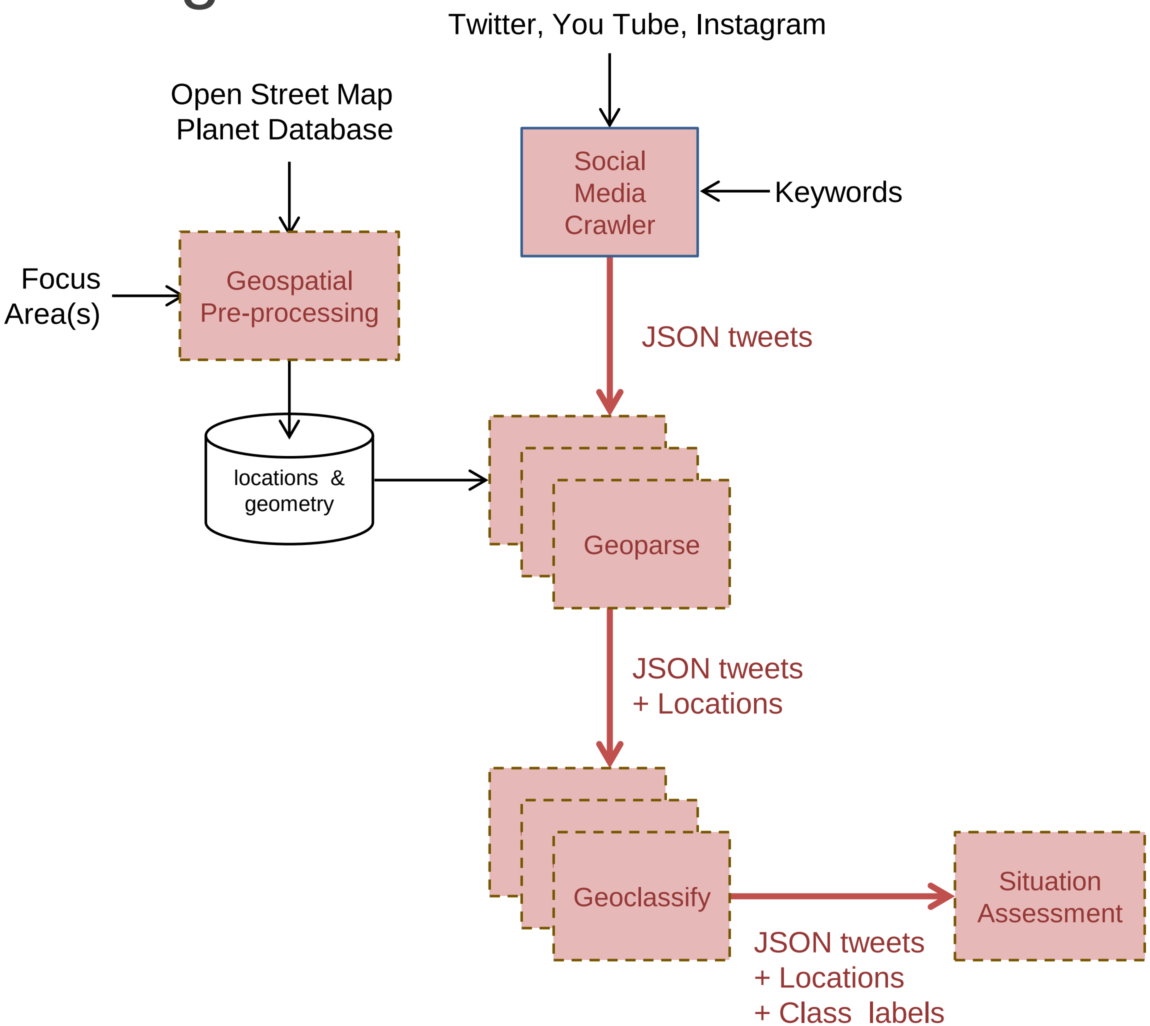
- Geoparsing experience
  - Social media is very noisy
    - Expect spelling mistakes, bad formatting, jargon
    - Eyewitnesses at serious events tend to post clear text descriptions
  - Entity matching algorithms scale well
    - Entity matching algorithms can be naively parallelized
    - Entity recognition algorithms needs POS tagging or dependency parsing, which can be hard to parallelize
  - Language models train on social media tags and pickup vernacular terms well
    - YFCC 100M Flickr image dataset tags e.g. big apple
  - All approaches suffer from variable spatial coverage
    - OpenStreetMap - low population areas often lack data
    - Language models - non-tourist areas often lack data
  - Hybrid models give best overall performance

Geosemantic Analysis

- Terminology - my definitions
  - Geosemantics
    - Use of context in relation to spatial data
    - UGC >> Location mention(s) >> Contextual text >> Classification of how location is being referred to
- Case studies
  - REVEAL
    - Geosemantic classification >> Filter UGC >> Journalist
    - Especially interested in situated and timely UGC
      - Eyewitness UGC for breaking news events

Geosemantic Analysis

- Algorithm - geoclassifier



Geosemantic Analysis

- Context window around location mention
  - 12 terms either side of location >> Text context
  - Text context >> weak stemming (plurals) >> Parts of Speech (POS) tagging >> n-gram features (mix of lexical tokens & POS)
- Example feature extraction

"Oklahoma tornado filmed by Newcastle resident"

▼

▼

Oklahoma/NP tornado/NN filmed/VVN by/IN Newcastle/NP resident/JJ

▼

▼

(Oklahoma tornado filmed), (tornado filmed by), (filmed by Newcastle), ...
(NP tornado filmed), (Oklahoma NN filmed), (Oklahoma tornado VNN), ...
(Oklahoma * filmed), (Oklahoma * by), (Oklahoma * Newcastle), ...
(NP * filmed), (Oklahoma * VNN), (NP * by), (Oklahoma * IN), ...

Geosemantic Analysis

- Feature selection
  - Calculate most discriminating features
    - Remove features below 10% of max TF
    - Top 20,000 features selected after TF-IDF
  - Supervised learning
    - J48 decision tree & IBk classifiers worked best
    - Random forest, LogiBoost and NaiveBayes also tested
  - Labelled training data for 4 geosemantic classes
    - Confirmation >> confirm or deny incident @ location
    - Timeliness >> past, present or future location reference
    - Situatedness >> insitu or remove location reference
    - Validity >> relevant or noise e.g. geoparse error

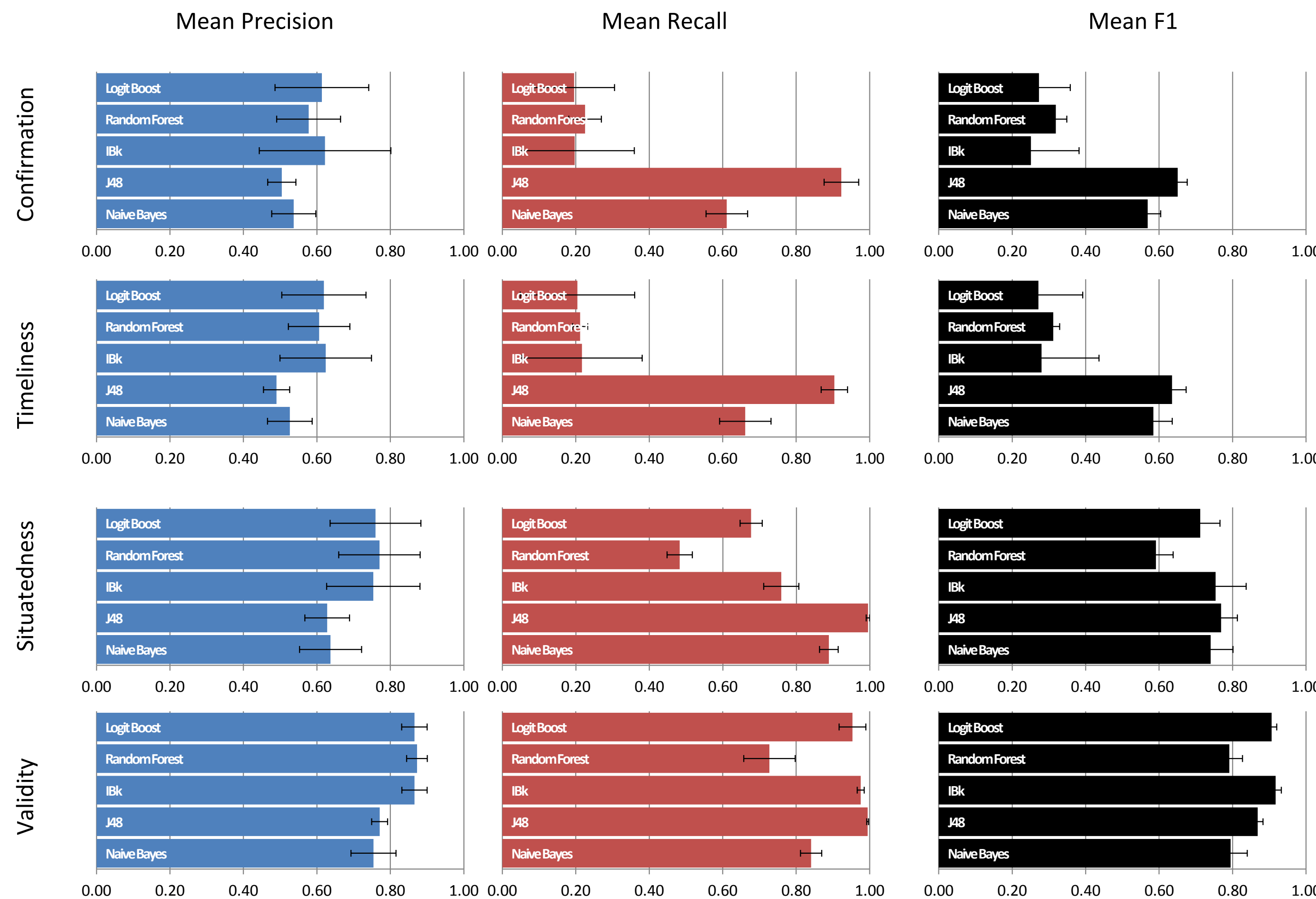
Geosemantic Analysis

- Datasets
  - TREC 2012 microblog dataset (Twitter)
    - Chicago Blizzard 2011
  - UoS crawled events (Twitter)
    - Hurricane Sandy, 2012
    - Oklahoma Tornado 2013
    - Ukraine Conflict 2014
    - Scottish Independence Referendum 2014
- Ground Truth
  - Random sample of each dataset
  - 5,285 total posts, 500 to 1500 each event
  - Manually labelled with 4 geosemantic classes

PUBLIC

Geosemantic Analysis

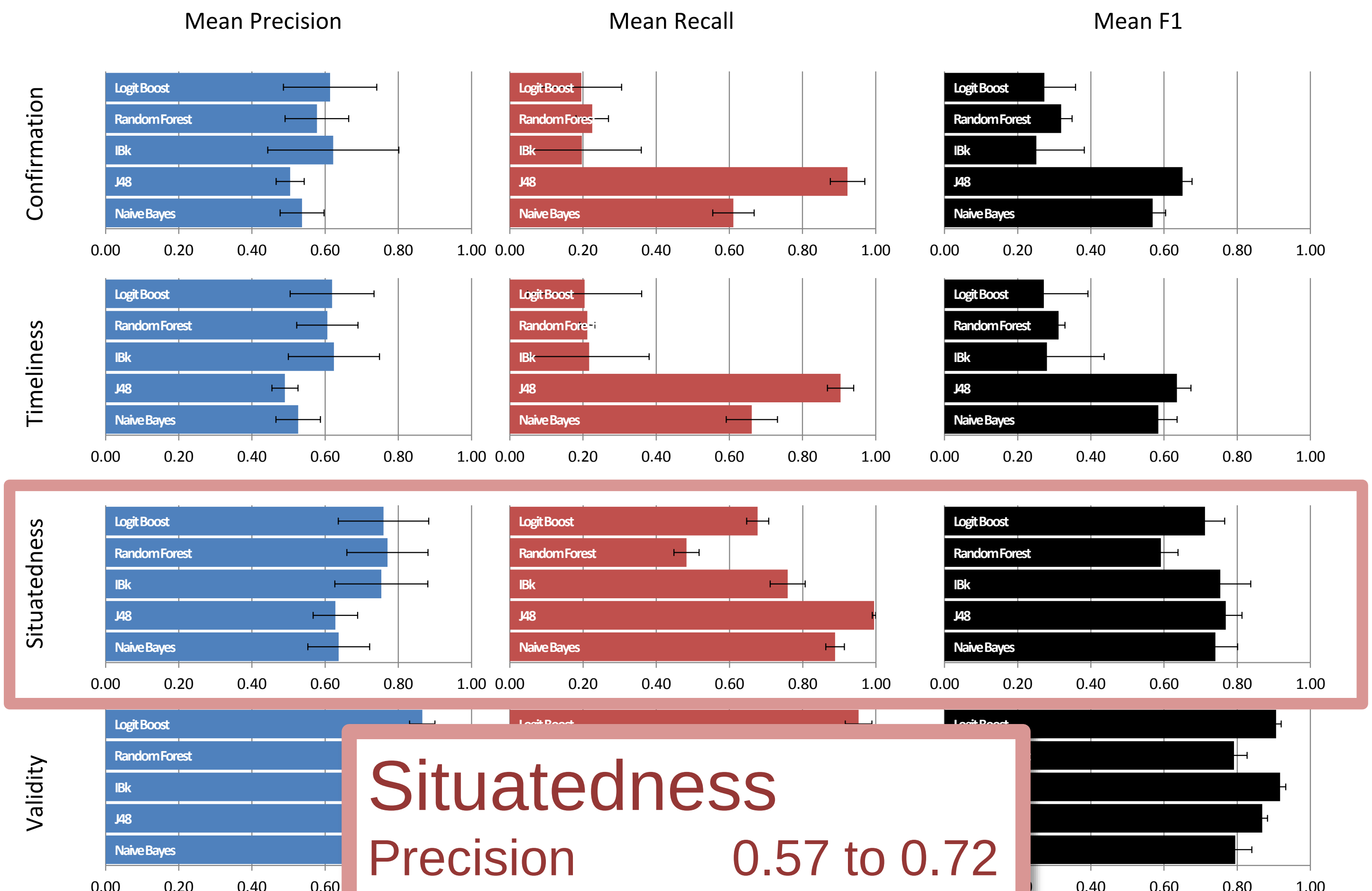
- Discussion: Veracity



Middleton, S.E. Krivcovs, V. "Geoparsing and Geosemantics for Social Media: Spatio-Temporal Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News", ACM Transactions on Information Systems (TOIS), 34, 3, Article 16 (April 2016)

Geosemantic Analysis

- Discussion: Veracity



Situatedness

Precision	0.57 to 0.72
F1	0.73 to 0.83

Middleton, S.E. Krivcovs, V. "Geoparsing a… al Grounding of Content Propagating Rumours to support Trust and Veracity Analysis during Breaking News", ACM Transactions on Information Systems (TOIS), 34, 3, Article 16 (April 2016)

Lessons Learnt

- Geosemantics experience
  - Geosemantic filters are useful for pre-filtering content
    - Precision OK for filtering prior to human assessment
    - Not good enough for fully automated work yet
  - Geosemantics can help location refinement
    - Relative spatial offset e.g. I am 5 miles north of London bridge
    - Negatives e.g. I'm nowhere near London bridge
    - Relevance e.g. I was at London bridge last year
  - Training data is often needed
    - Its expensive to generate and not dynamic

Open Information Extraction

- Terminology - my definitions
  - Open Information Extraction
    - Free Text >> relation tuples e.g. (John, didn't go to, London)
    - Typically unsupervised and able to scale up
- Case studies
  - REVEAL
    - UGC >> OpenIE >> Factual claim extraction >> Journalist
  - GRAVITATE
    - Artifact descriptions from text resources >> OpenIE >> Attribute metadata >> Archaeologist
  - FloraGuard
    - Online marketplaces & Forums >> OpenIE >> Proposition & Entity extraction >> Law enforcement

Open Information Extraction

- Algorithm - moving beyond just location
  - Pre-process
    - Stanford Tagger and Dependency Parser
  - Novel template-based OpenIE algorithm
    - Template-based unsupervised OpenIE algorithm
    - Able to use semi-supervised relevance feedback to incrementally improve over time
    - Propositional extraction
      - e.g. (10 dead, reported in, north of Paris)
    - Attribute extraction
      - e.g. (Left hand, of, statue), (Left hand, missing, three fingers)
  - Naively parallelizable - Python multiprocessing lib
  - Set of Python libraries alongside geoparsepy

Open Information Extraction

- Discussion: Variety
  - Information extraction - context beyond location
    - Locations, Times, Usernames, Products, Financial transaction details, Topics, Actions ...
  - Dynamic language patterns and/or vocabularies are common in many use cases
    - Breaking news >> trending news topics (days)
    - Cybercrime >> jargon in evolving cryptolects (months)
    - Artifact description >> specialist vocabulary for new archaeological digs & exhibitions (years)
  - Unsupervised (or at least semi-supervised) algorithms are needed to handle dynamic variety of language patterns
  - Work in progress - results due early 2019

Summary

- Location Extraction & Geoparsing
  - Entity matching algorithms scale well
  - Database models for areas with good map coverage
  - Language models to capture vernacular terms
  - Hybrid models give best overall performance
- Geosemantics
  - Provides context for pre-filtering and location refinement
  - Training data is often needed
- Open Information Extraction
  - Extracting semantic context beyond location

Thanks you for your attention!

Any questions?

Dr Stuart E. Middleton
University of Southampton, Electronics and Computer Science, IT Innovation Centre

email: sem03@soton.ac.uk
web: www.ecs.soton.ac.uk/people/sem www.it-innovation.soton.ac.uk
twitter:@stuart_e_middle

Acknowledgement
GRAVITATE H2020 grant agreement 665155, FloraGuard ref ESRC ES/R003254/1
TRIDEC FP7 grant agreement 258723, REVEAL FP7 grant agreement 610928