Floraguard

Information Extraction and Analytics from Social Media and Online Marketplaces

UK Illegal Plant Trade, Military Intelligence Analysis, Breaking News

Stuart E. Middleton

University of Southampton, Electronics and Computer Science

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

Cardiff Seminar 2019

21st Feb 2019

Overview

- Speaker
- Research Agenda - NLP and IE
- Geoparsing and Geosemantics
- Attribute Extraction
- Factual Claim Extraction
- Digital Text Forensics
- Challenges in the Future

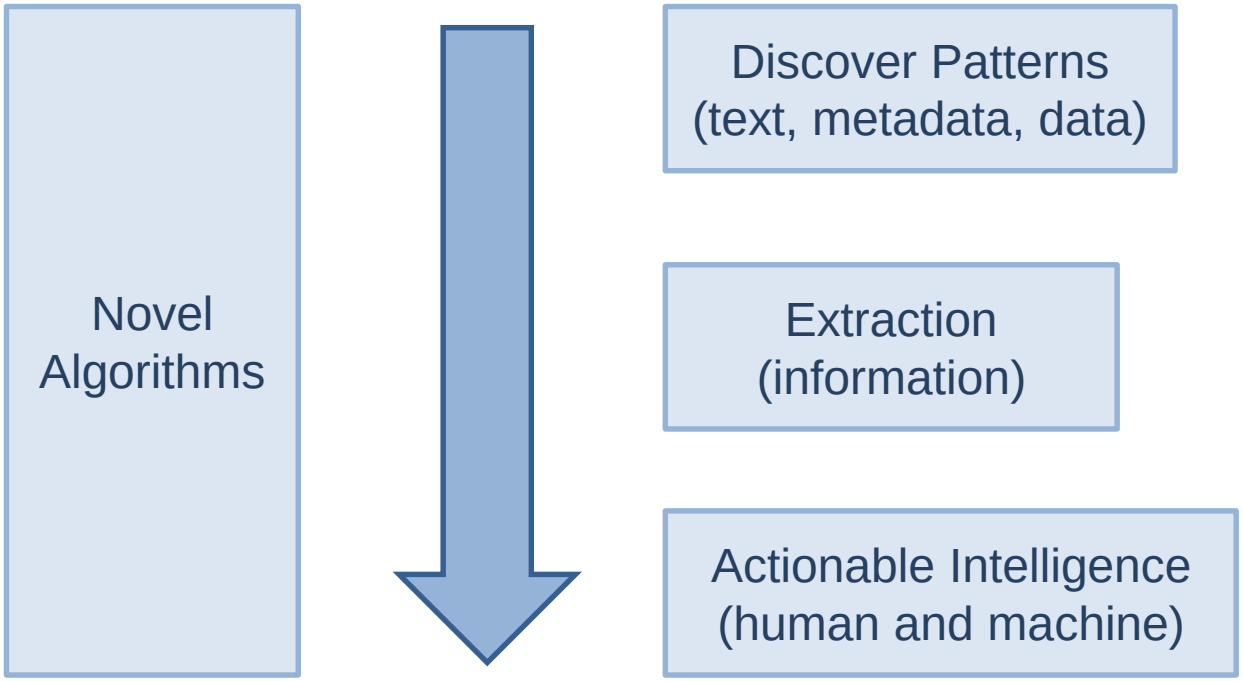
Speaker

- Dr Stuart E. Middleton
  - Senior Research Engineer
  - University of Southampton, Electronics and Computer Science (ECS), IT Innovation Centre
- Research
  - NLP / Computational Linguistics and Information Extraction
- Interdisciplinary
  - Environmental Sciences (GFZ - Tsunami Early Warning)
  - Journalism (Deutsche Welle)
  - Archaeology (British Museum)
  - Criminology
    (UK Border Force, UK National Crime Agency)
  - Law (UK Land Registry)
  - Cyber Security (DSTL, UK MOD, Home Office)

Floraguard

Research Agenda - NLP and IE

- Natural Language Processing (NLP)
  - Machine learning of language structure and use
- Information Extraction (IE)
  - Exploiting patterns in language, metadata and data to extract useful information



Research Agenda - NLP and IE

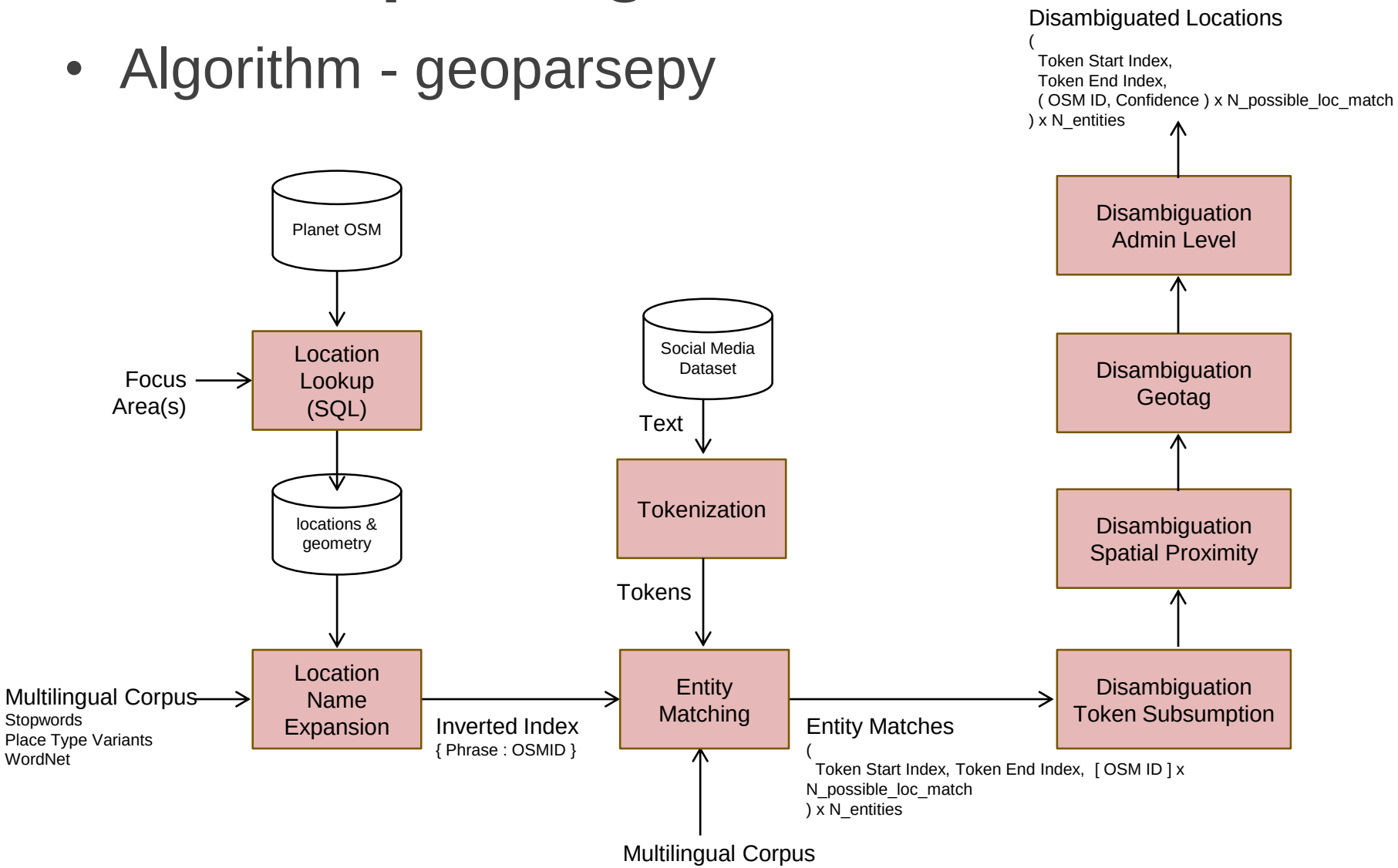
- Juxtaposition to deep learning & big data
  - Algorithms that can work with small, sparse or fragmented datasets
  - Breaking news on social media
  - Emerging topics within online community forums
  - Criminal marketplaces exhibiting deliberate obfuscation
  - Historical datasets where information can be inaccurately recorded, corrupted or lost over time
- Bootstrapping to use expert knowledge & feedback
- Incremental to use data as it appears over time
- Emergent language patterns
- Provenance preserving for explainable AI

Geoparsing and Geosemantics

- Geoparsing
  - Text >> Loc(s) >> Disambiguated Loc(s) + Spatial Ref(s)
- Case studies
  - TRIDEC
    - Geoparsing social media >> Crisis mapping
    - Tsunami early warning >> 5 to 60 minutes coastline warnings
    - Social media flood maps >> Actual wave impact times >> Adjust sensor-based Tsunami wave models
  - REVEAL
    - Geoparsing social media >> Fake news verification
    - 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 User Generated Content (UGC) >> Breaking UGC >> Contextual UGC for verification

Geoparsing and Geosemantics

- Algorithm - geoparsepy

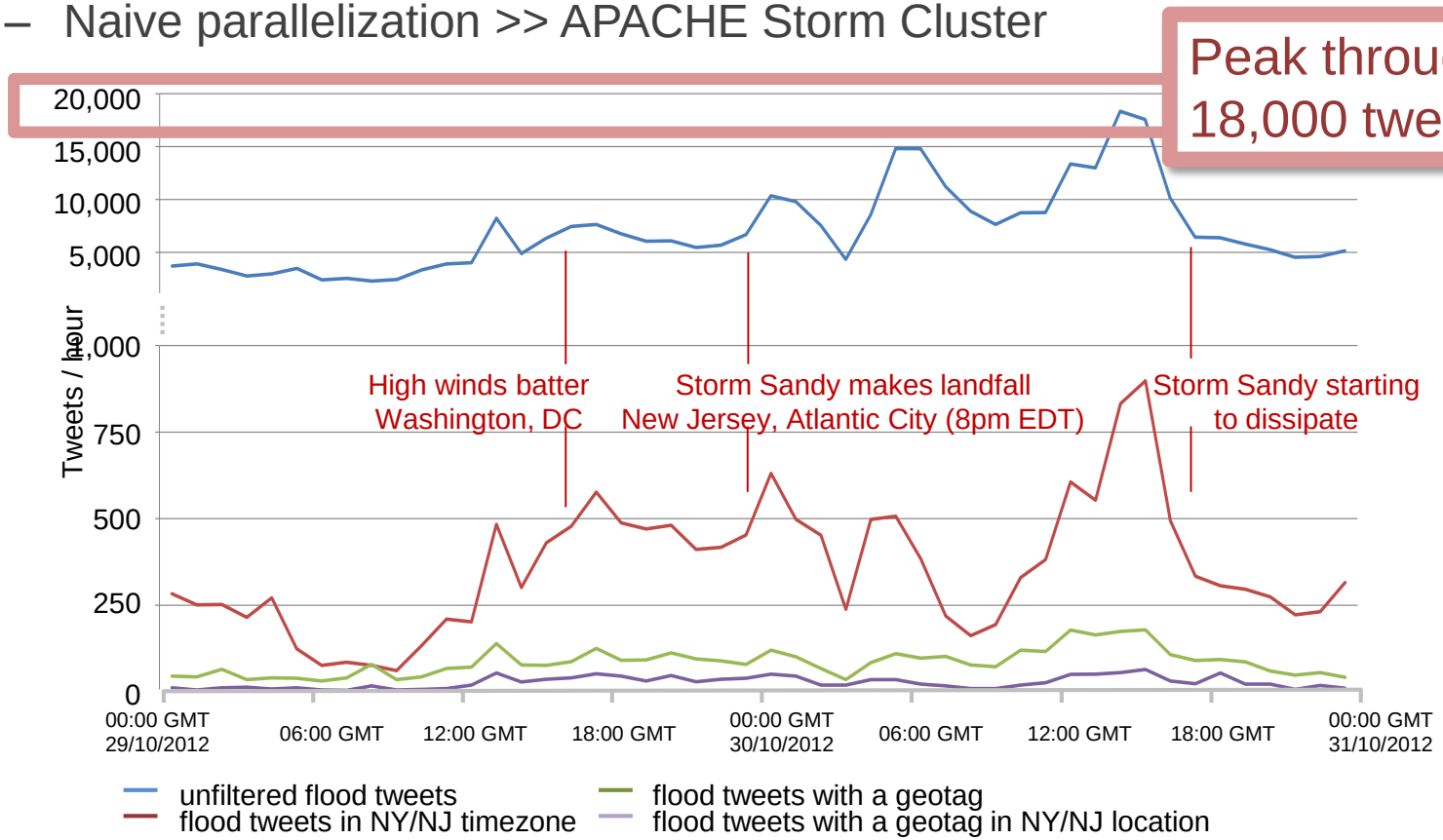


https://pypi.org/project/geoparsepy

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)

Geoparsing and Geosemantics

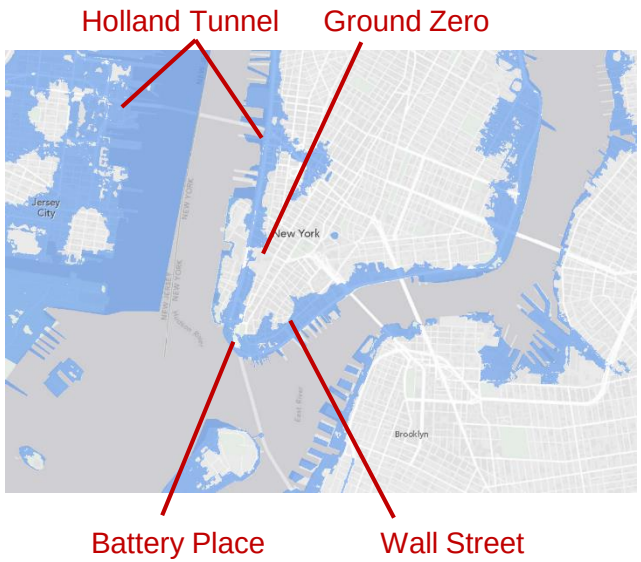
- Scalability
  - Hurricane Sandy, Oct 2012, 5 days, Twitter Streaming API (1% sample size)
  - 597,000 tweets, 4,300 loc mentions, ~170 unique locs, 1% posts geotagged
  - Naive parallelization >> APACHE Storm Cluster

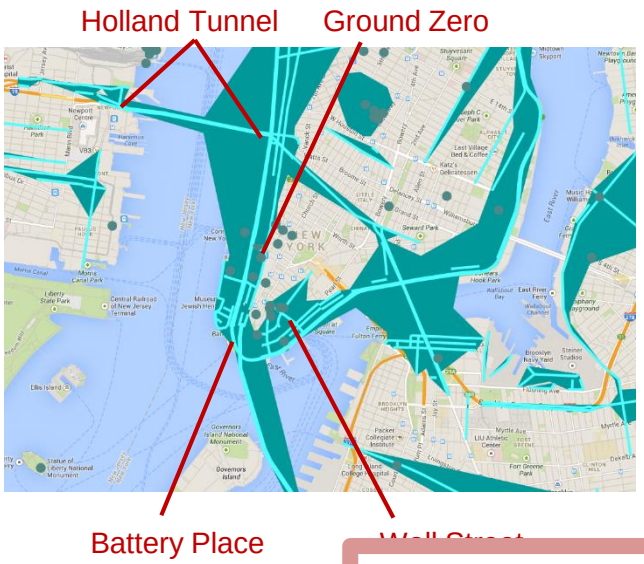


Middleton, S.E. Middleton, L. Modafferi, S. "Real-time Crisis Mapping of Natural Disasters using Social Media", Intelligent Systems, IEEE , vol.29, no.2, pp.9,17, Mar.-Apr. 2014

Geoparsing and Geosemantics

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





Key

Expert post-event assessment storm surge inundation area

Place flooded tweet(s)

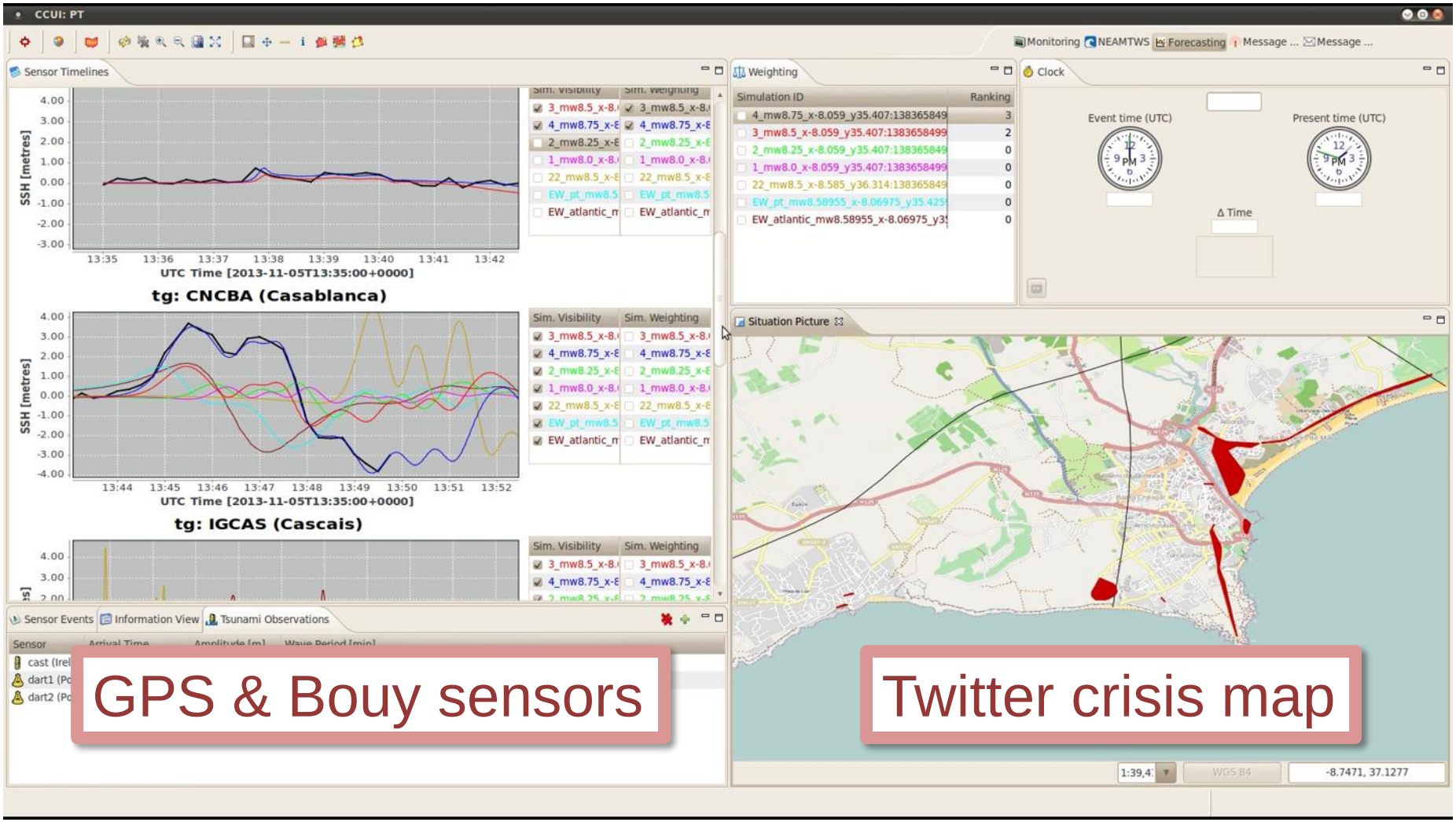
Street flooded tweet(s)

Map cell precision 0.78+

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

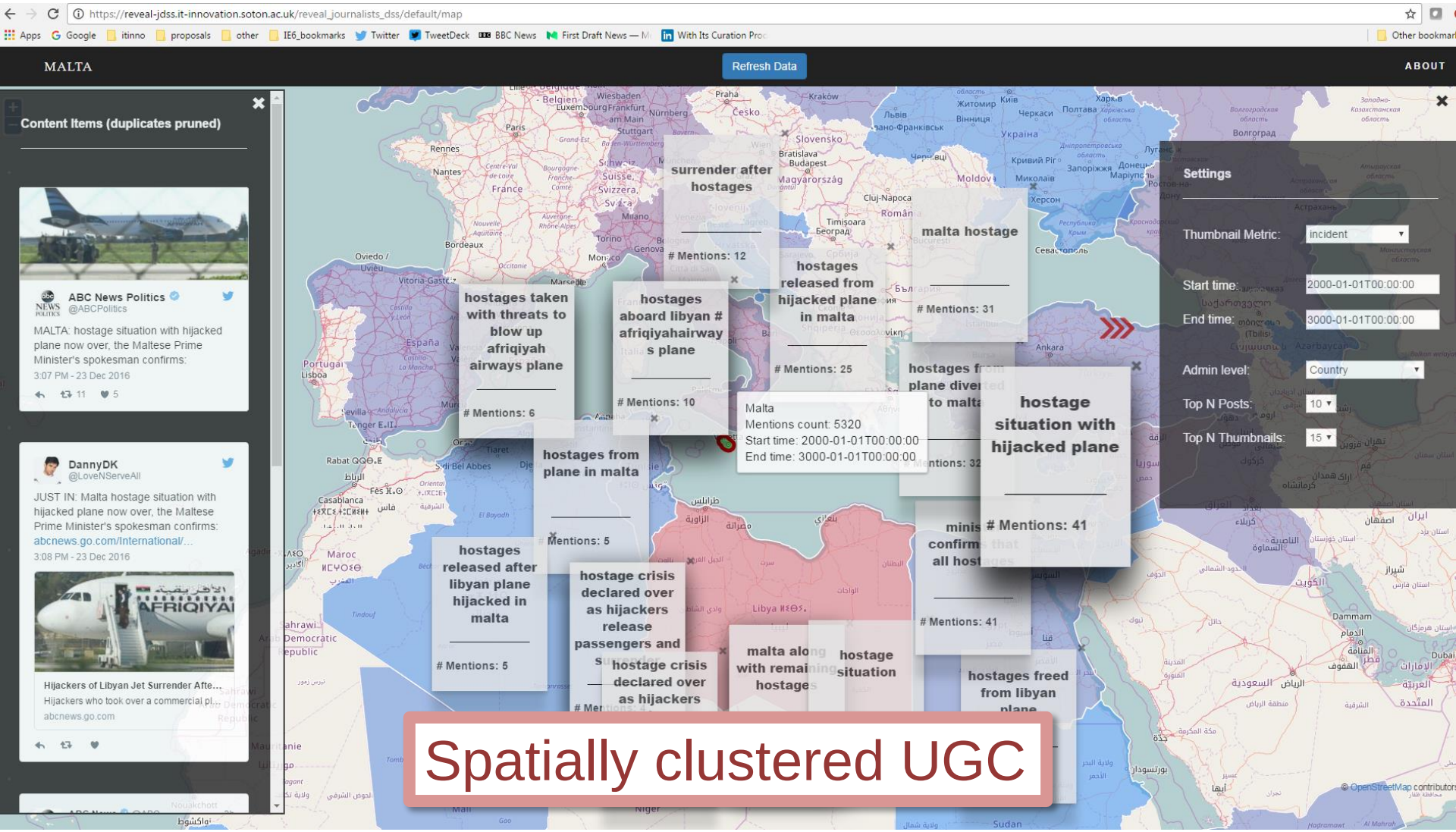
Threshold	Places	Streets	Precision	Recall	F1
dev_sma > 0	66	101	0.78	0.76	0.77
dev_sma > 0.016	19	22	0.81	0.44	0.57
dev_sma > 0.1	4	1	1.00	0.09	0.17

Geoparsing and Geosemantics



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

Geoparsing and Geosemantics



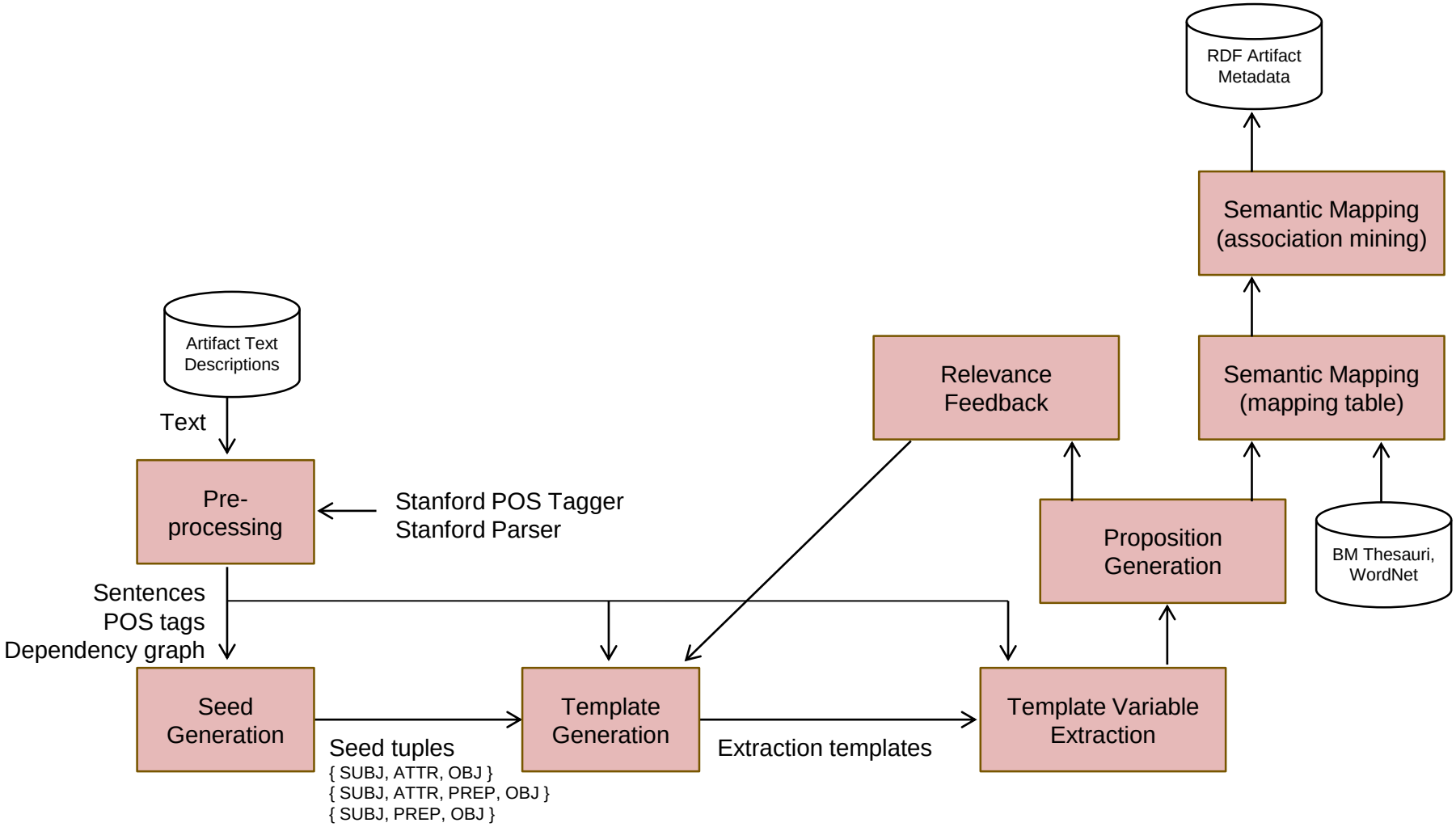
Screenshot from Journalist Decision Support System (JDSS), REVEAL project
JDSS - https://reveal-jdss.it-innovation.soton.ac.uk/reveal_journalists_dss

Attribute Extraction

- Open Information Extraction (OpenIE)
  - Text >> relation tuples
  - Verb-mediated e.g. (John, didn't go to, London)
  - Preposition-mediated e.g. (Statue, of, Zeus)
  - Noun-mediated e.g. (Microsoft, CEO, Bill Gates)
  - Typically unsupervised or bootstrapped, able to scale up
- Case studies
  - GRAVITATE
    - Artifact physical description text >> OpenIE >> Semantic Mapping >> Artifact attributes >> Archaeologist [search, similarity match]
  - Preposition-mediated focus

Attribute Extraction

- Algorithm - OpenIE + Semantic Mapping



Attribute Extraction

- Example extractions
  - { Rim, of, South Ionian pottery cup }
  - { Rim and body sherd of South Ionian pottery cup, with, everted rim }
  - { Isis, rearing, off, the ground }
  - { The goddess, wears, a short-sleeved tunic of uncertain length }

Attribute Extraction

- Artifact RDF graph similarity matching
  - RDF2VEC algorithm
  - Artifact A >> similar to >> Artifact B, C, D ...
  - Ground truth = expert labelled artifact clusters
  - CH Dataset 174 >> artifacts (Salamis collection)
  - DBpedia Dataset 1077 >> books (fiction)

Precision	0.81	0.44	0.26
Yield	100	500	1,000

(N most similar artifacts)

Precision	0.89	0.85	0.85
Yield	100	500	1,000

(N most similar books)

Complex multi-attribute graphs are harder to match

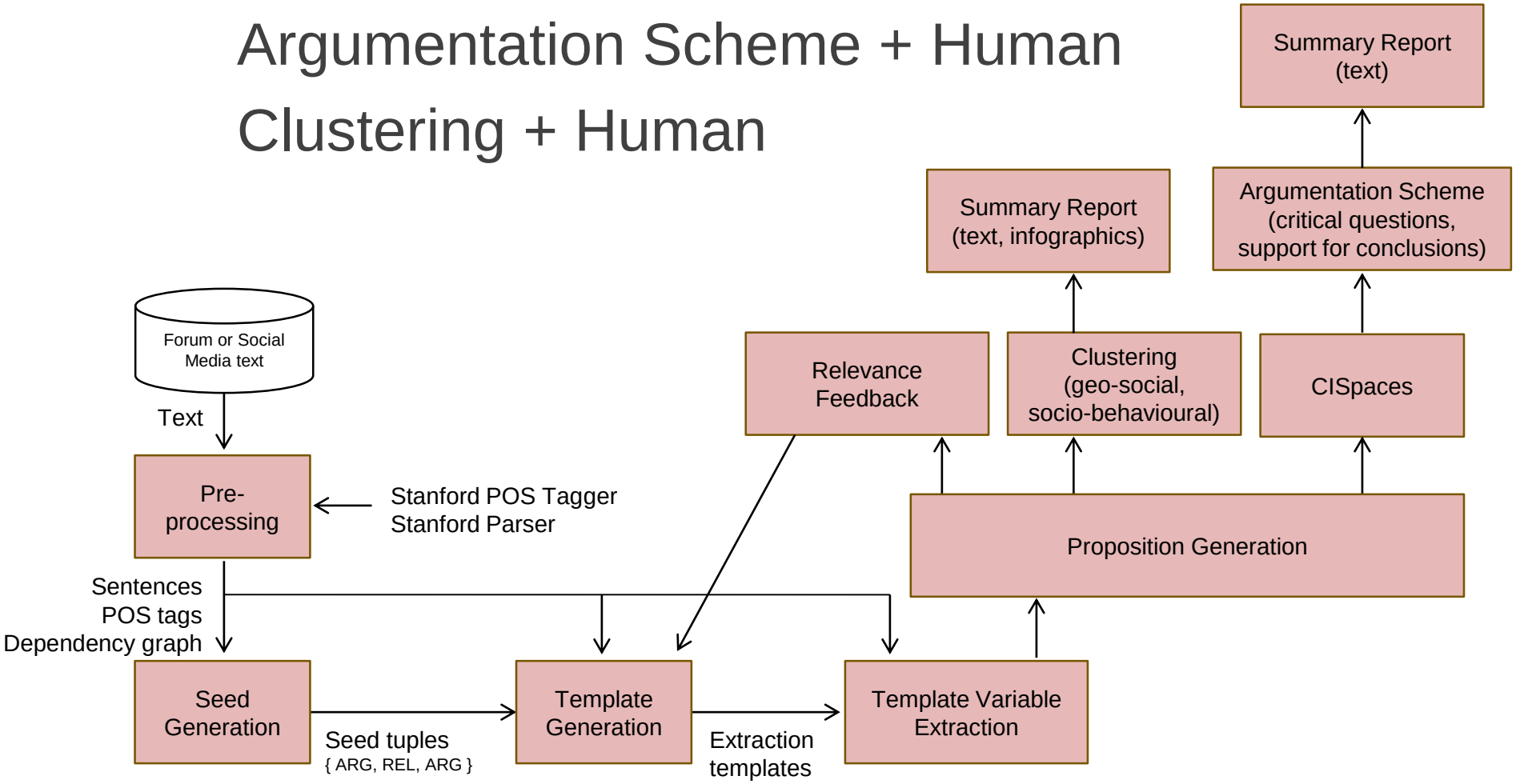
Factual Claim Extraction

- Case studies
  - REVEAL
    - UGC >> OpenIE >> Factual claim extraction >> Journalist
  - DSTL Intel Analysis project - CISpaces.org
    - OSINT (social media) >> OpenIE >> Factual claim extraction >> Sensemaking (argumentation scheme)

      >> Intelligence Analyst
  - FloraGuard
    - Online marketplaces & Forums >> OpenIE >> Factual claim & Attribute extraction >> Law Enforcement Officer
  - Verb-mediated focus

Factual Claim Extraction

- Algorithm - OpenIE + Sensemaking
  Argumentation Scheme + Human
  Clustering + Human



Factual Claim Extraction

- Example extractions
  - { The Embassy, said, 6,700 Americans were in Pakistan }
  - { 6,700 Americans, were in, Pakistan }
  - { Mr. Parkin 's office, was also searched, by the F.B.I. }
  - { they, recovered, to win their final two matches }
  - { Mary McCarty, writes, for the Dayton Daily News }
  - { Sixty people, injured in, Napel Earthquake - source BBC }
  - { Video: gunman, shooting, people in Brussels Airport !!! }

Factual Claim Extraction

- Benchmarking on verb propositions
  - Tested against state of the art OpenIE algorithms
    - ClauseIE, OpenIE5, Stanford OIE, OLLIE, ReVerb
  - Tested on various benchmark datasets
    - NYT news feed, Wikipedia articles, Yahoo web pages
    - OIE benchmark dataset

Precision	0.87	0.82
Yield	100	200

(top N verb-mediated extractions, NYT dataset)

8% improvements over best OpenIE available today

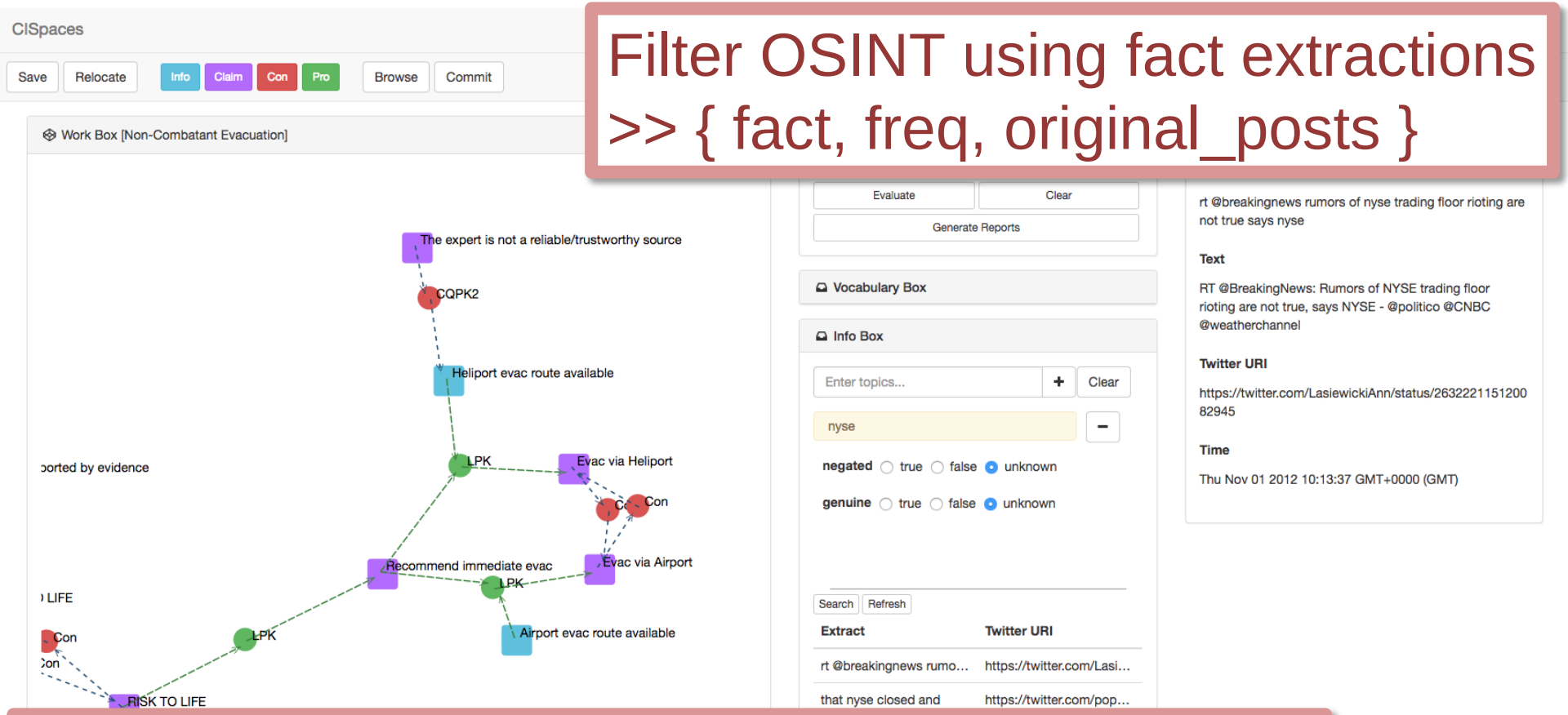
Precision	0.94	0.89
Yield	100	200

(top N verb-mediated extractions with relevance feedback, NYT dataset)

Relevance feedback further improves precision

Factual Claim Extraction

Filter OSINT using fact extractions
>> { fact, freq, original_posts }

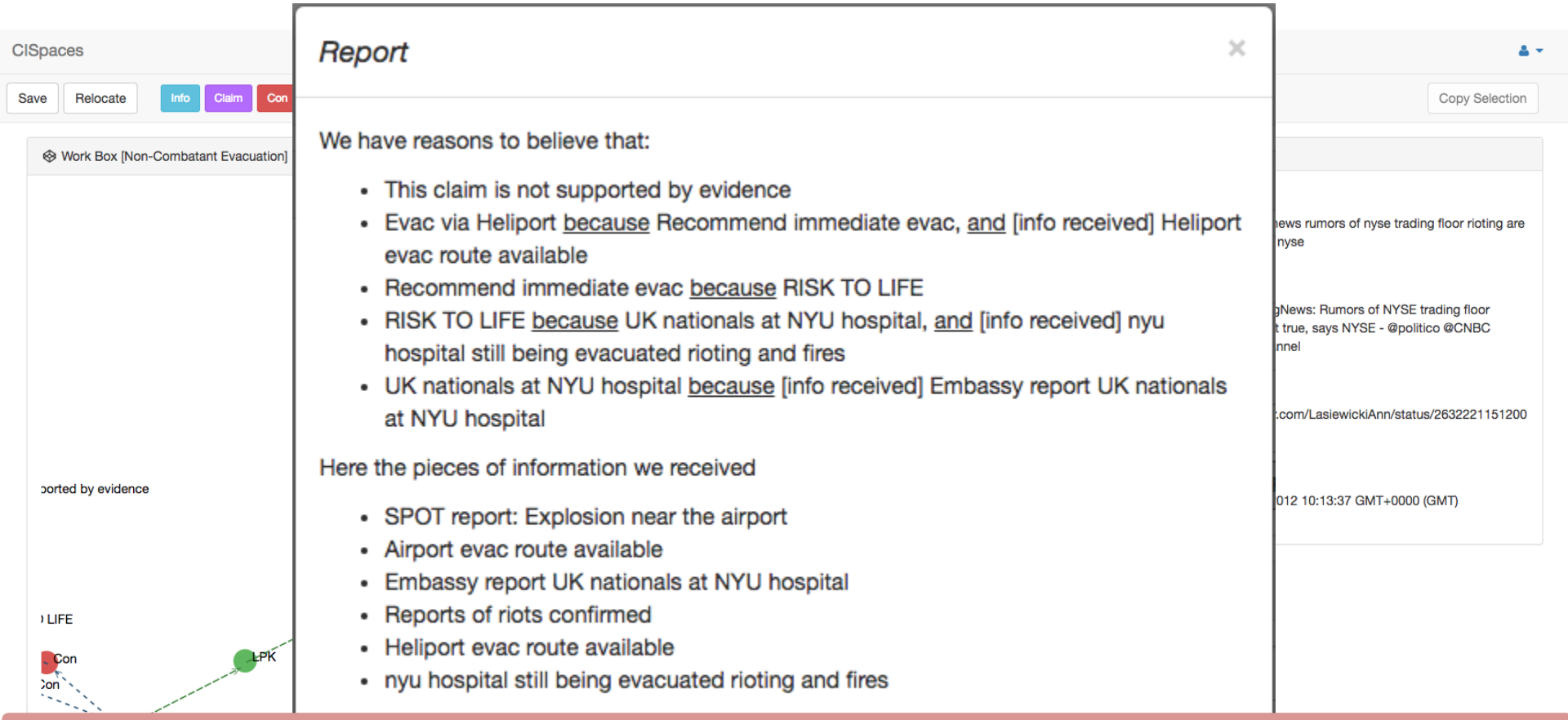


Human >> Add Evidence >> Support / Weaken
Analyst (OSINT) Assertions

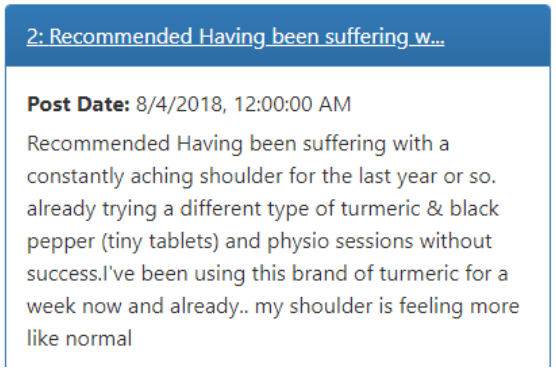
CISpaces.org

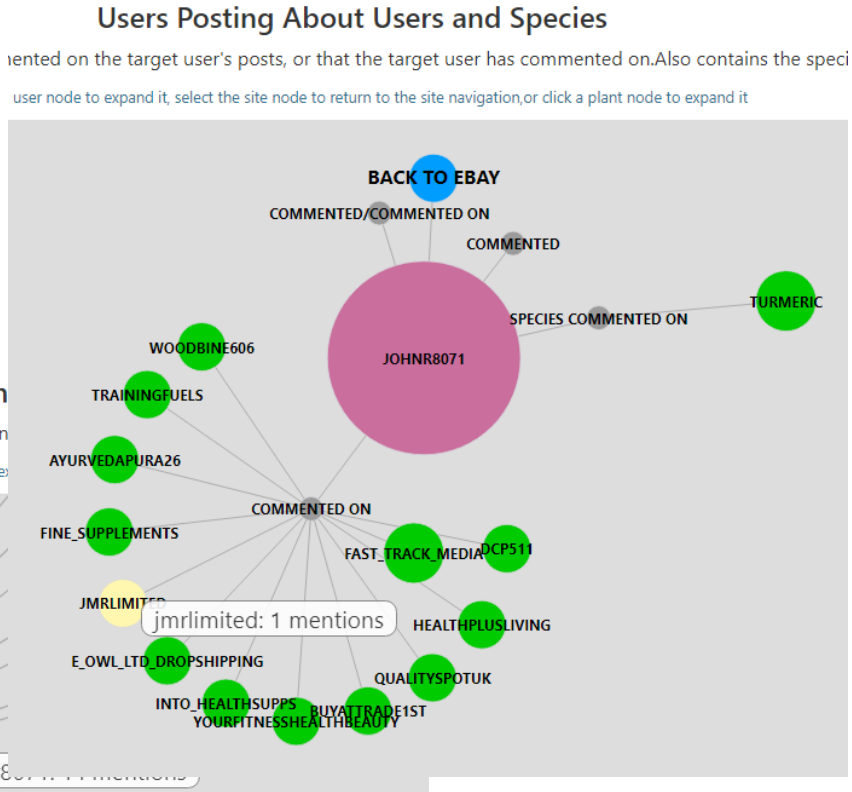
Cerutti, F. Norman, T.J. Toniolo, A. Middleton, S.E. "CISpaces.org: from fact extraction to report generation", 2018 Computational Models of Argument - Proceedings of COMMA 2018. IOS Press, Vol. 305, p. 269-280 12

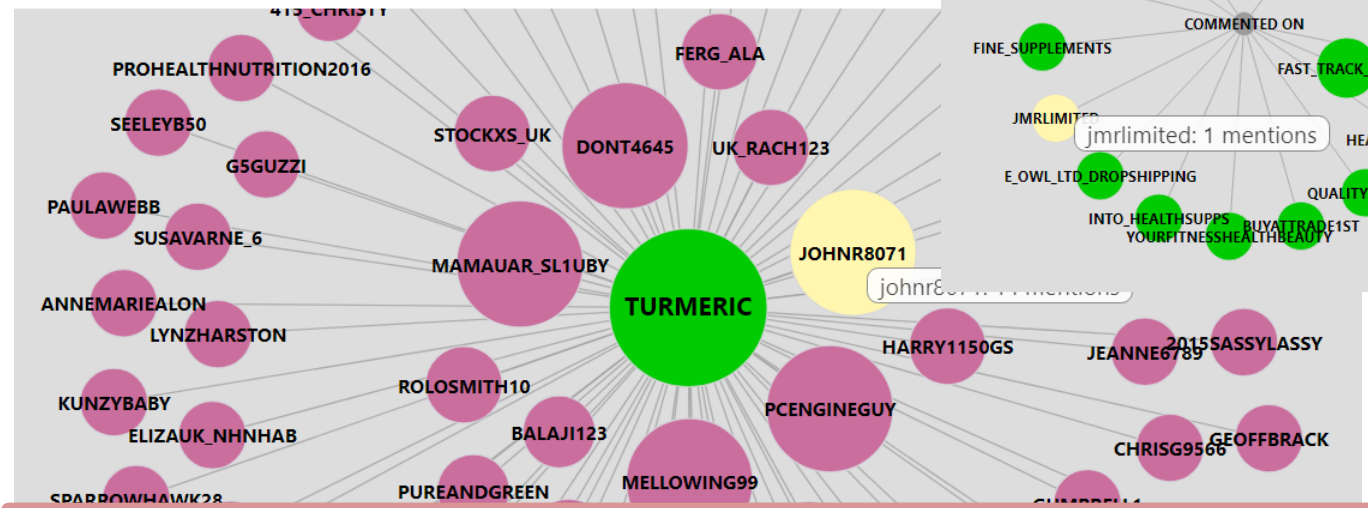
Factual Claim Extraction



Assertions >> Argumentation >> Conclusions >> Report
+ Evidence Scheme + Strength of Evidence







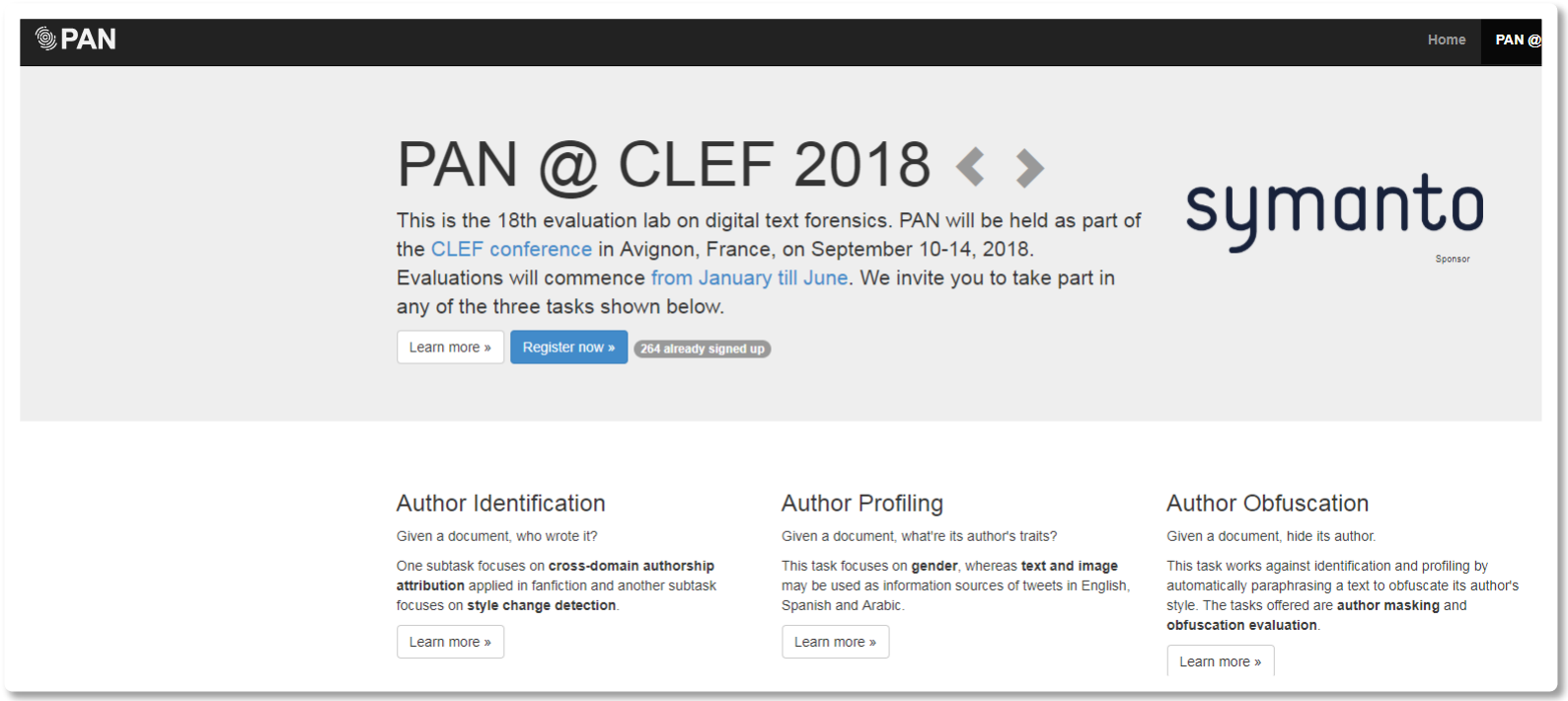
Users, >> Cluster >> Index >> Browse Intelligence,
Plants, Summary Report
Behaviour Types,
Locations

Digital Text Forensics

- Case studies
  - FloraGuard
    - Online marketplaces & Forums >> OpenIE >> Factual claim & Attribute extraction >> Law Enforcement Officer
  - National Crime Agency (NCA)

Digital Text Forensics

- Author attribution, profiling & clustering
  - Academic >> PAN conference series
  - Datasets from News feeds, Wikipedia, Enron emails ...
  - Interest from law enforcement agencies



https://pan.webis.de/clef18/pan18-web/index.html
http://floraguard.org/

Digital Text Forensics

- Author attribution
  - Closed-set attribution task
  - N documents with known authors + M unknown documents >> assign most likely authors to unknown documents
- Author clustering
  - Open-set attribution task
  - N documents with unknown authors >> cluster documents into groups likely to be from same author

Digital Text Forensics

- Author attribution
  - Tested PAN algorithms on posts crawled from TOR-based AlphaBay Credit Card fraud forums
  - PAN datasets >> F1 up to 0.80
  - AlphaBay dataset >> F1 up to 0.51
- Author clustering
  - Tested two successful PAN algorithms on (illegal) plant marketplaces & forums
    - Character n-grams + Word embeddings >> k-means clustering
    - {Word + punctuation} n-grams >> L1-norm distance measure

Cybercrime forums are difficult datasets
>> short posts
>> cryptolect
>> deliberate obfuscation

Challenges in the Future

- Fragmented and cross-platform dialogue
  - Conversations moving cross domain and between apps

    e.g. TOR forum X (advert) >> Private Messaging (payment) >> TOR forum X (recommendation) >> Web forum Y (tech support)
- Fast pace of emerging online threats
  - New criminal products, Breaking news, Behaviour change
  - Often by the time a big dataset (for big data analytics) has been gathered the opportunity for early warning has been lost
  - Big data analytics work best for long running problems or recurring problems where discovered patterns can be re-used
- Anonymity and obfuscation
  - Encrypted private messaging, Reluctance of app providers to cooperate with law enforcement agencies
  - Sometimes published content is all there is to work with

Challenges in the Future

- Working with small datasets
  - Bootstrapping >> seed with domain knowledge
  - Incremental >> improve as more data arrives
  - Gap analysis >> inference from missing data
  - Transfer learning >> applying patterns from similar online communities
- Language modelling
  - Can we identify linguistic markers for emerging cryptolect? e.g. new drugs, new gangs
  - Very fast (breaking news), fast (cybercrime) and slow (historical) moving language changes

Challenges in the Future

- Machine translation and information extraction
  - Machine translation (MT) can overcome scalability issues with multi-lingual problems
  - Can we build in resilience to MT errors within OpenIE algorithms?
- Explainable AI
  - Can we use machine learning and language models where the pattern provenance is maintained back to original features?
  - Helps explain patterns found in understandable domain language
  - Required if relevance feedback is to be provided on discovered intermediate patterns as well as final results
  - Required if results are to be useful as evidence (e.g. court of law)

Challenges in the Future

- Disambiguation and counter-obfuscation
  - Overcoming weak performance of digital text forensic algorithms which are overfitting academic benchmark and challenge datasets
  - Use of wider feature sets than just language use
  - Use of community-specific language markers in digital text forensics algorithms

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
DSTL Human-Machine Teaming for Intelligence Analysis, agreement number ACC102157