20 years of Text Mining Applications with GATE: from Donald Trump to curing cancer

Dr Diana Maynard
University of Sheffield, UK

GATE and text mining

- Text Mining is the discovery of new, previously unknown information, by automatically extracting information from different textual resources.
- A key element is the linking together of the extracted information
- Text mining lets you investigate what's actually in a collection of documents, and answer questions about the content even when it's not explicitly mentioned
- GATE is the most widely used open source toolkit for text mining in the world, developed in Sheffield over the last 20 years
- Hundreds of thousands of users all over the world, from PhD students to government organisations to multinationals

GATE for text engineering

- GATE – General Architecture for Text Engineering
  - http://gate.ac.uk
  - Started in 1996; established large developer community, incl. industrial committers (Ontotext, Intellius, Text Mining Solutions)
- Tool for developing and deployment of Text Mining technology
- Used worldwide by many organisations to build bespoke solutions, e.g. TNA, Press Association, BBC, NHS
- A free open source framework (LGPL) and graphical development environment
- Includes Information Extraction in many languages
- Component-based, easy mix between OS and proprietary plugins

Why do people use GATE (in the real world)?

- Difficult to **access unstructured information** efficiently
- **Save time and money** on understanding and managing text and data from multiple sources
- Find **hidden links** scattered across massive volumes of diverse information
- Understand what people are thinking and talking about; how they respond to events, how opinions are changing over time….
- Help communication and aid during disasters
- Analyse huge volumes of data, e.g. medical records, patents, consumer intelligence
- Open source means you can see what's going on under the hood, and tweak to your own needs

If it's free it probably isn't very good….

- You pay for what you get, right?
- IBM Watson must be **really** good, since they charge around half a million dollars just to get the basic tool (not to mention thousands of dollars a month for training etc.)
- Well, maybe….but how can you tell if you don't know what's going on under the surface?
- How do you even know if it will do what you want? You have to pay large sums of money just to find out.
- And if it doesn't do what you want, how will you fix it? Can you?
- Open source lets you play with it and see, and you can always ask for training/consultancy to see if it's right for you/adapt it to your needs

Some of our users...

International Agency for Research on Cancer

World Health Organization

BBC SPORT

Nesta

AstraZeneca

NHS
National Institute for Health Research

Ushahidi

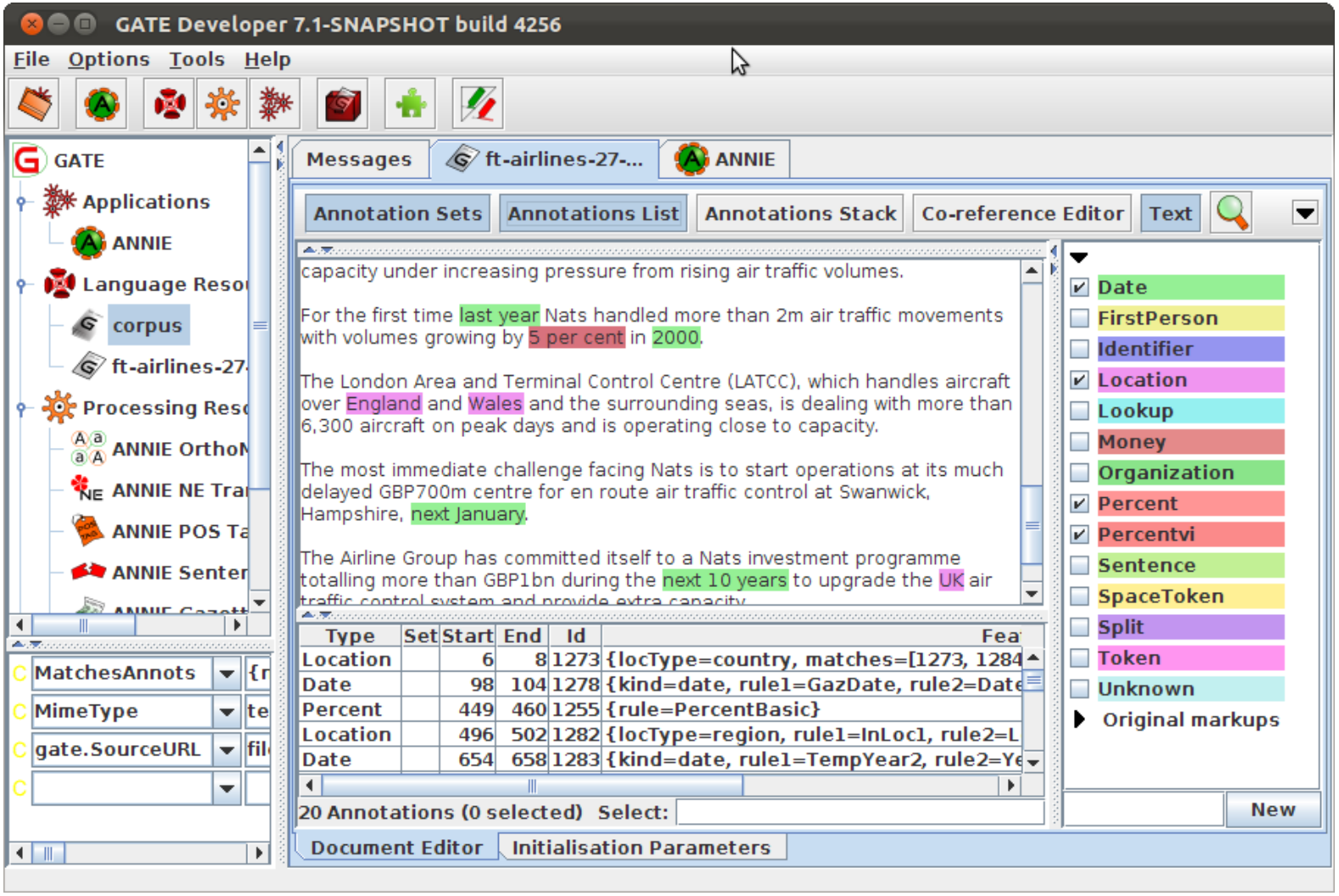
fera
The Food and Environment Research Agency

www.synaptica.com
knowledge organization and discovery

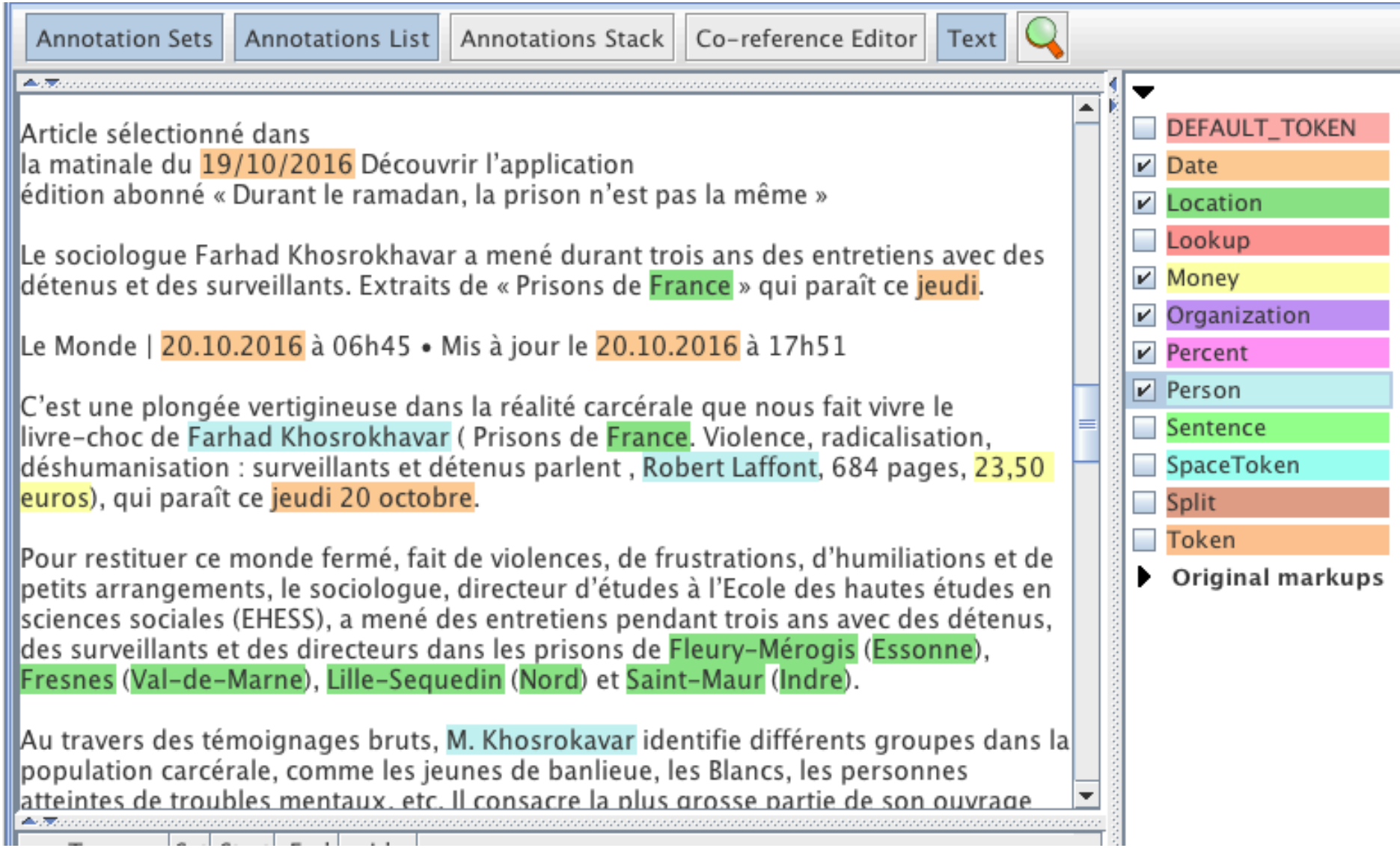
Why do people use GATE (for research)?

- Framework lets you reuse existing tools – no need to rebuild the wheel every time
- Framework allows you to mix and match tools from different sources – we incorporate all our competitors' tools!
- Framework allows for easy comparison and evaluation of tools (your own or other people's)
- Easy to integrate new tools which extend the existing ones
- Separation of programming and linguistic tasks – no need to be an expert in all areas
- Connects with crowdsourcing tools, search and visualisation tools etc.
- Access to a huge network of expertise

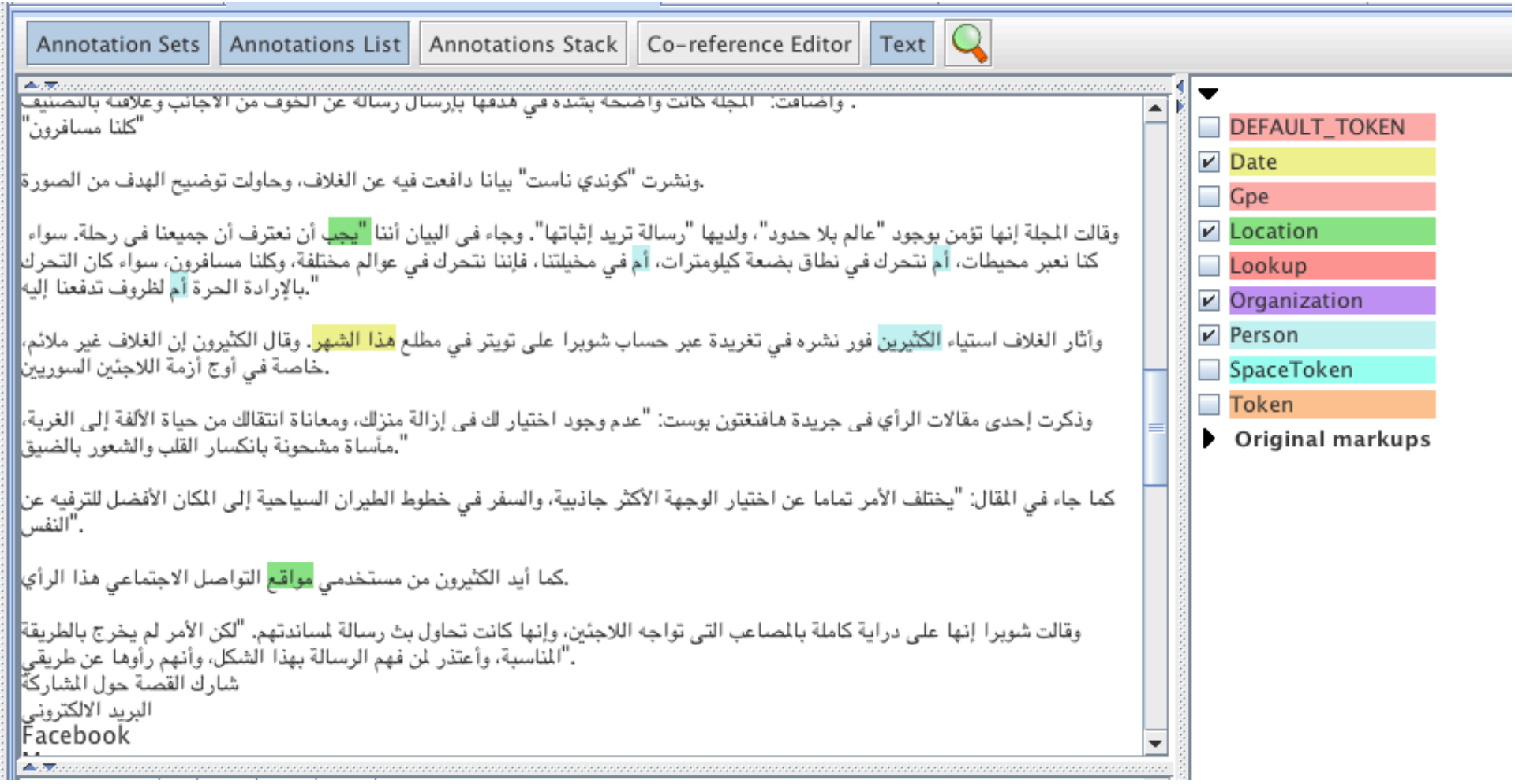
Named entity recognition in GATE



NER in French



NER in Arabic



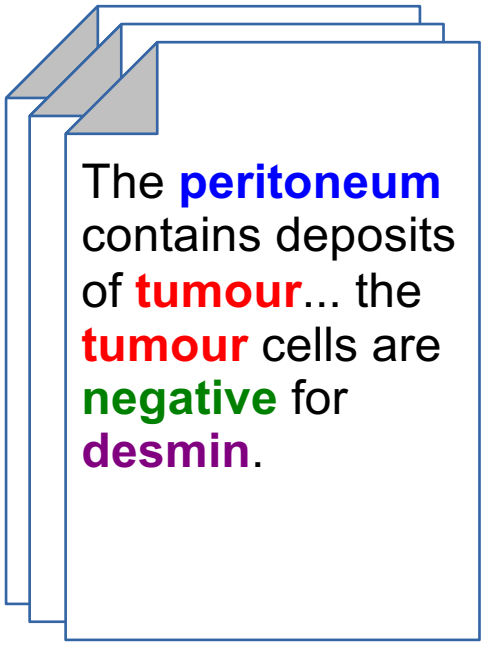
30% of medical information is structured

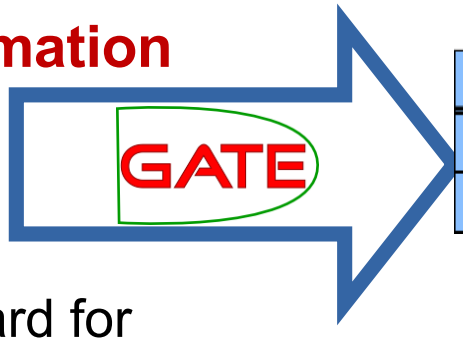
Hb	WBC	RBC	Plt
14.1	5.1	4.5	210
10.2	36.6	5.0	420
13.4	10.1	5.1	180
12.3	8.3	4.6	340

Easy to search and manipulate with computers

GATE structures the free text portion of the medical record, and makes it available for research, management, and clinical care

70% of medical information is in free text





Test		Result
desmin	has finding	negative

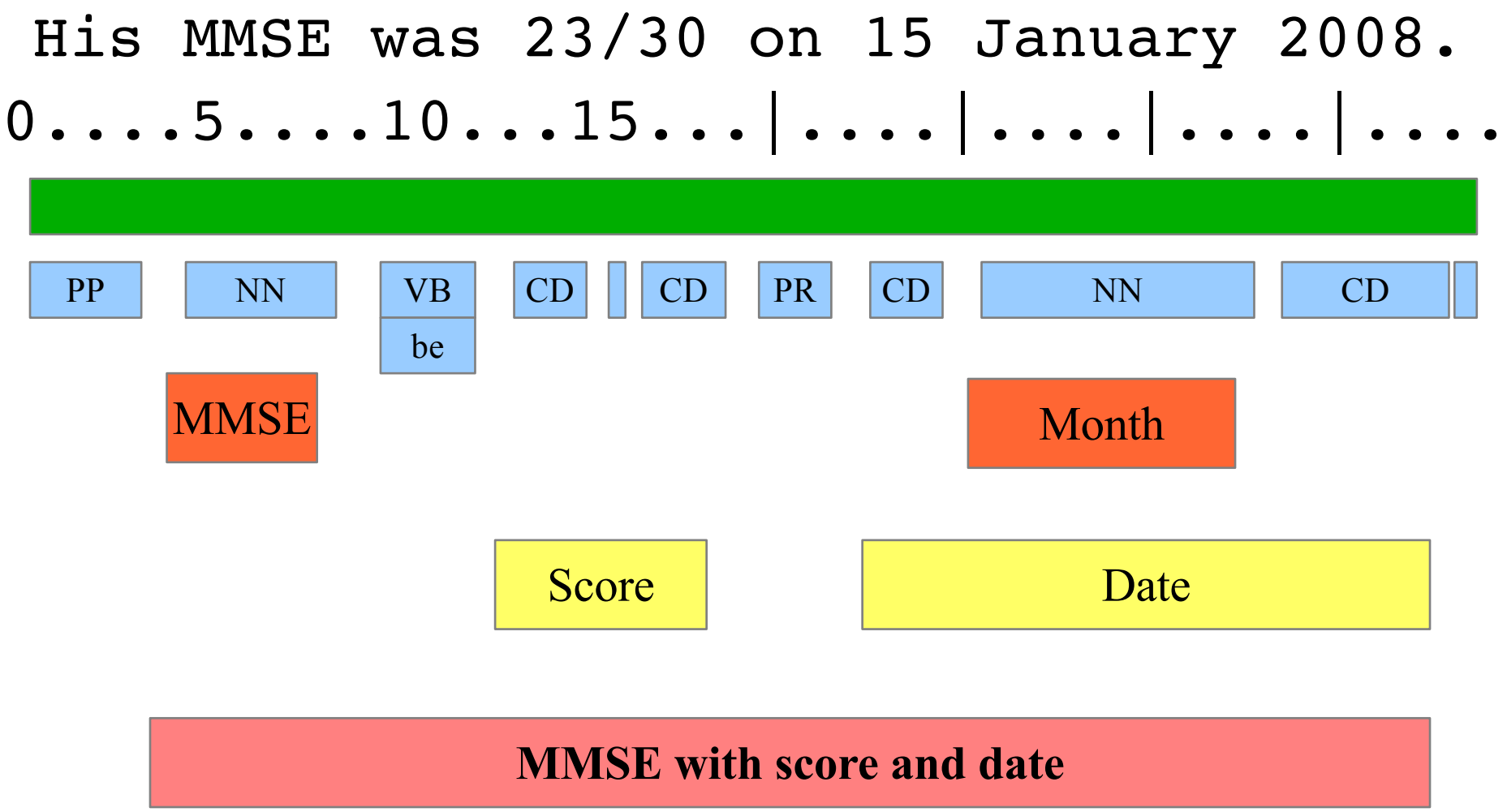
Condition		Locus
tumour	has location	peritoneum
...	...	...

Hard for computers to understand. Ambiguous, nuanced, and complexity

GATE can link the text to external knowledge, terminologies and coding schemes, for intelligent search and analysis

As used by several UK hospitals and world-leading medical record software vendors

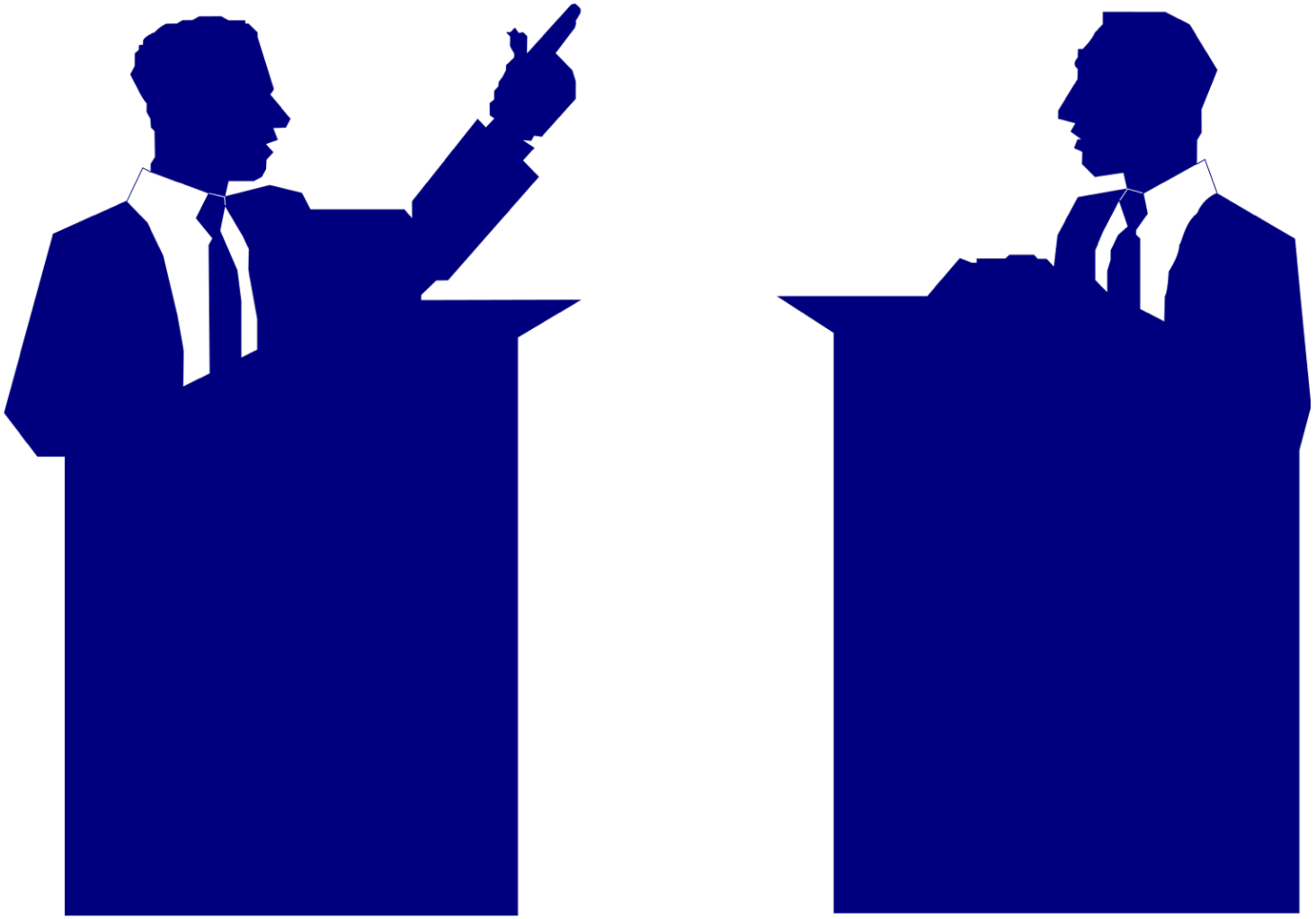
Medical Information Extraction



Text mining can cure cancer!

- Genome Wide Association Studies (GWAS) aim to investigate genetic variants across the whole genome
- With enough cases and controls, this allows them to state that a given SNP (Single Nucleotide Polymorphism) is related to a given disease.
- Can we reduce the costs by analysing published articles to generate prior probabilities for each SNP?
- Using text analysis to mine PubMed, we showed retrospectively that our approach would have saved over a year's worth of work and more than 1.5 million Euros
- We completed a new study in conjunction with the WHO which found a new cause for oral cancer
- Oral cancer is rare enough that traditional methods would have failed to find enough cases to make the study plausible

Analysing polarised societal debates



Polarised Debates

- Public, polarised debates affect policy making, the course of a country, voting intents
- Historically, the public debates took place on TV and involved only politicians or media figures

Motivation

- Today, polarised debates take place publicly on the internet, and anyone can participate
- Analysing these public statements tells us:
  - What participants are thinking
  - What arguments they are making
  - Who is thinking what
  - Gain a “big picture” view of the response to events or policies.
- Literally anyone can join these debates

A Cube
@PrototypeCube

Follow

perhaps @OwenSmith_MP should work on privatizing himself because he's being very publicly owned right now

Pat Condell @patcondell · Aug 30
Good news, Remainiacs (or bad news if you're a petulant stubborn arsehole), #Brexit is boosting the UK economy.

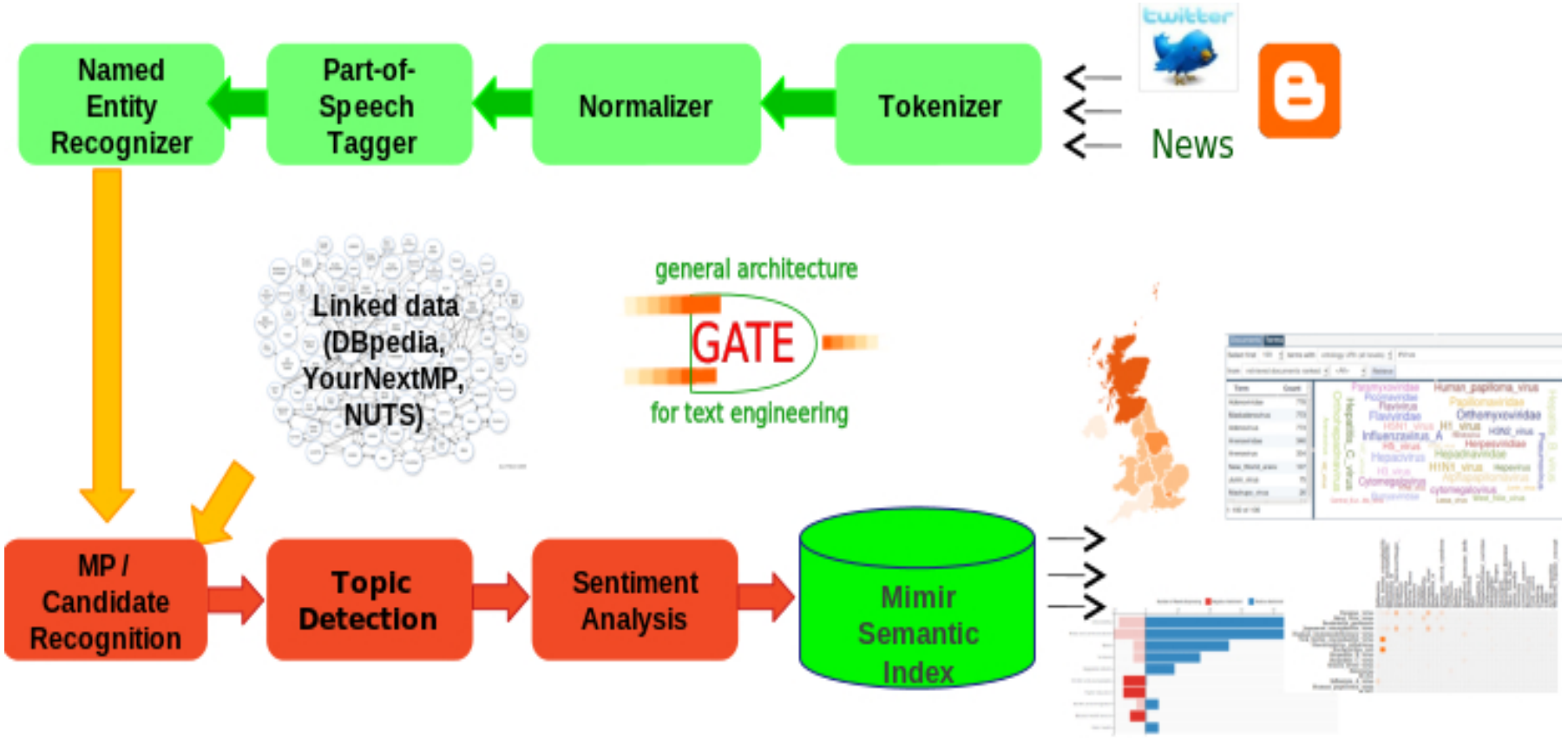
Derek Bateman @DerekBateman2 · 2m
Also. Look at the pollution. I demand a **Brexit** return to horse drawn carriages

> **Sunny Hundal** @sunny_hundal
> Brexiters now demand a return to imperial measurements even though law already allows it. Such big ambitions! politicalscrapbook.net/2016/08/brexit …

The Story

- Analyse social media (Twitter and similar platforms) surrounding polarised societal debates
- Index the results to understand, visualise and explore:
  - What topics are being discussed?
  - What sentiments are being expressed?
  - Who is participating in the debate?
  - How do debates evolve over time?

GATE social media analysis toolkit: analysing the UK elections



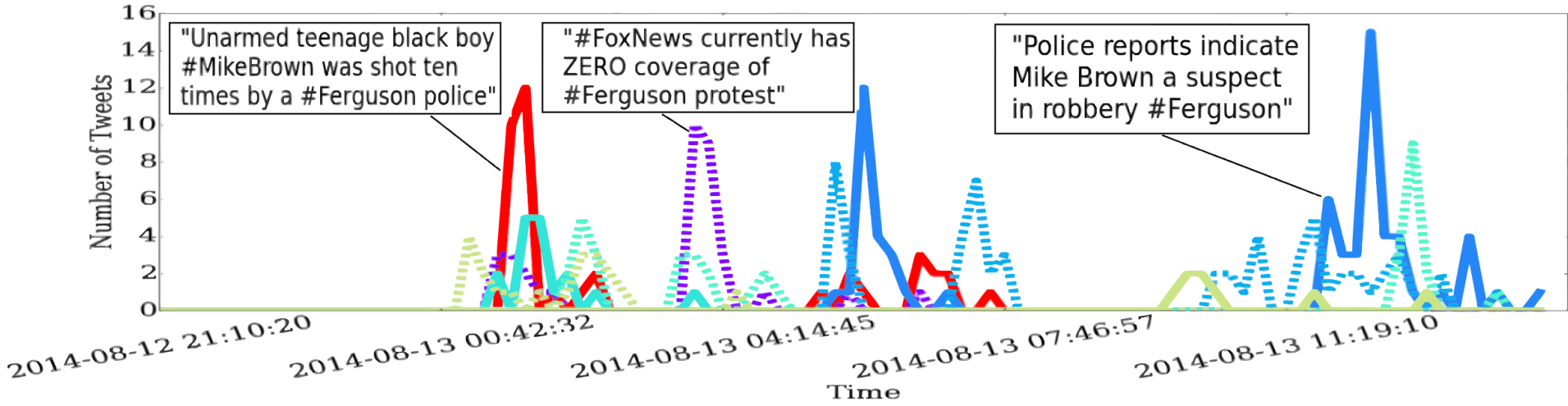
The users

- Policy makers
- News media
- Social science researchers
- General public?

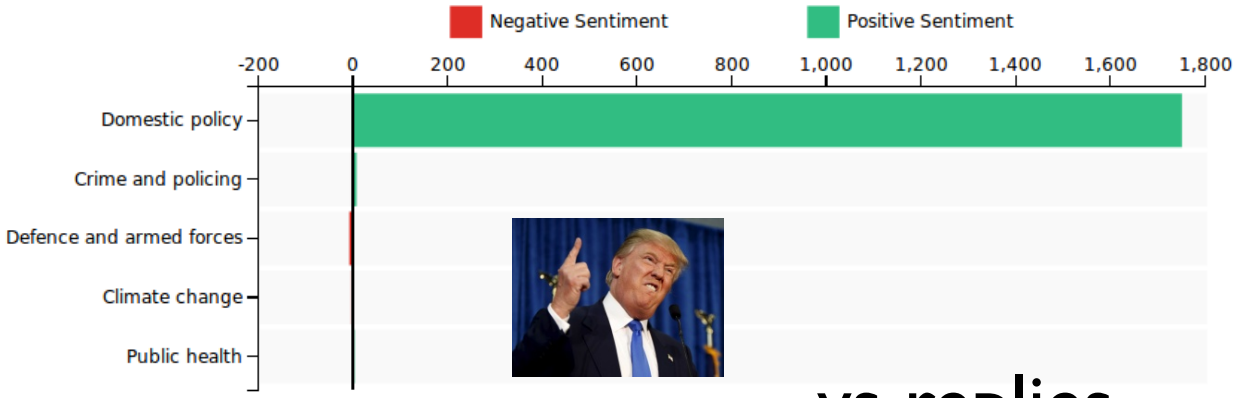
European Parliament

NESTA Making Innovation Flourish

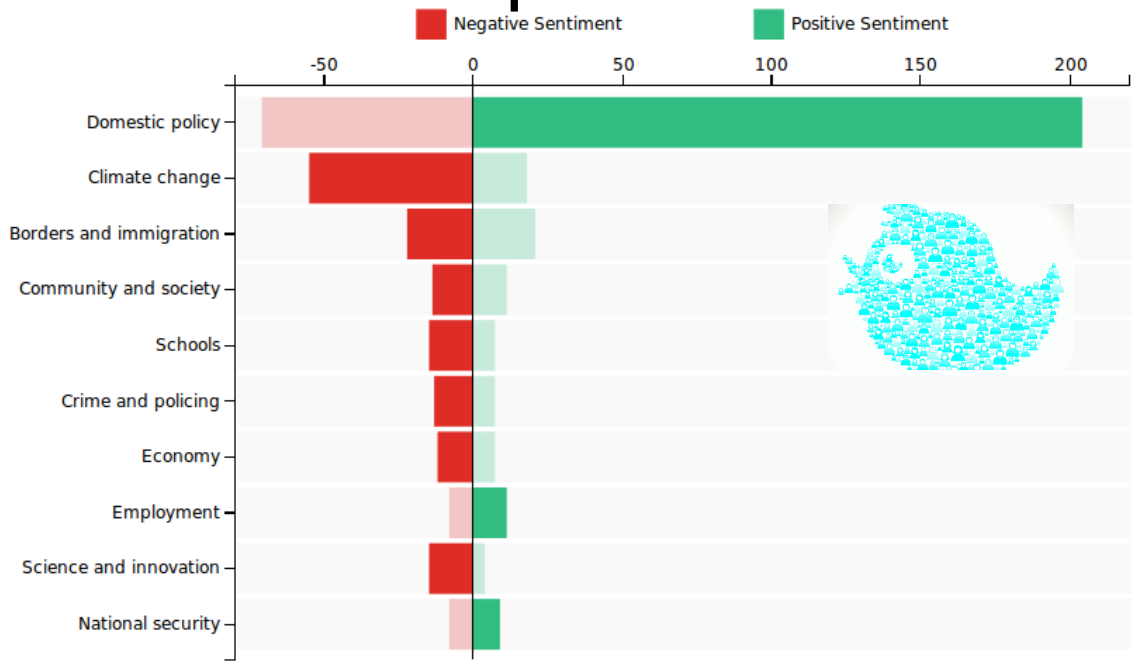
Rumour analysis: The Problem



Real-time Opinion Monitoring



vs replies



Replies to Trump re: climate change

Results 1 to 10 of 57

Tweet from KatieCampson at 2015-10-18T17:21:35.000Z
https://twitter.com/KatieCampson/status/655795876187738112
@realDonaldTrump climate change is a huge problem that rich assholes like you want to ignore so that you can continue unjustly to make money

Tweet from Danny_Sunset at 2015-10-19T00:15:53.000Z
https://twitter.com/Danny_Sunset/status/655900139165384704
@realDonaldTrump Democrats think climate change is bad. We'll run out of oil pretty soon and there will be nothing left to pollute the world

Tweet from goldberg3776 at 2015-10-19T04:57:20.000Z
https://twitter.com/goldberg3776/status/655970966732996609
@realDonaldTrump @joshdill64 this idiot doesnt even believe in climate change..how ignorant can u b

Tweet from goldberg3776 at 2015-10-19T05:03:45.000Z
https://twitter.com/goldberg3776/status/655972582219476992
@realDonaldTrump ur racist against latinos...so u cant take back what u said against them..u deny climate change which makes u stupid

Tweet from robin_kinley at 2015-10-19T13:31:38.000Z
https://twitter.com/robin_kinley/status/656100393026347011
@realDonaldTrump but the cold is caused by global warming don't cha know..

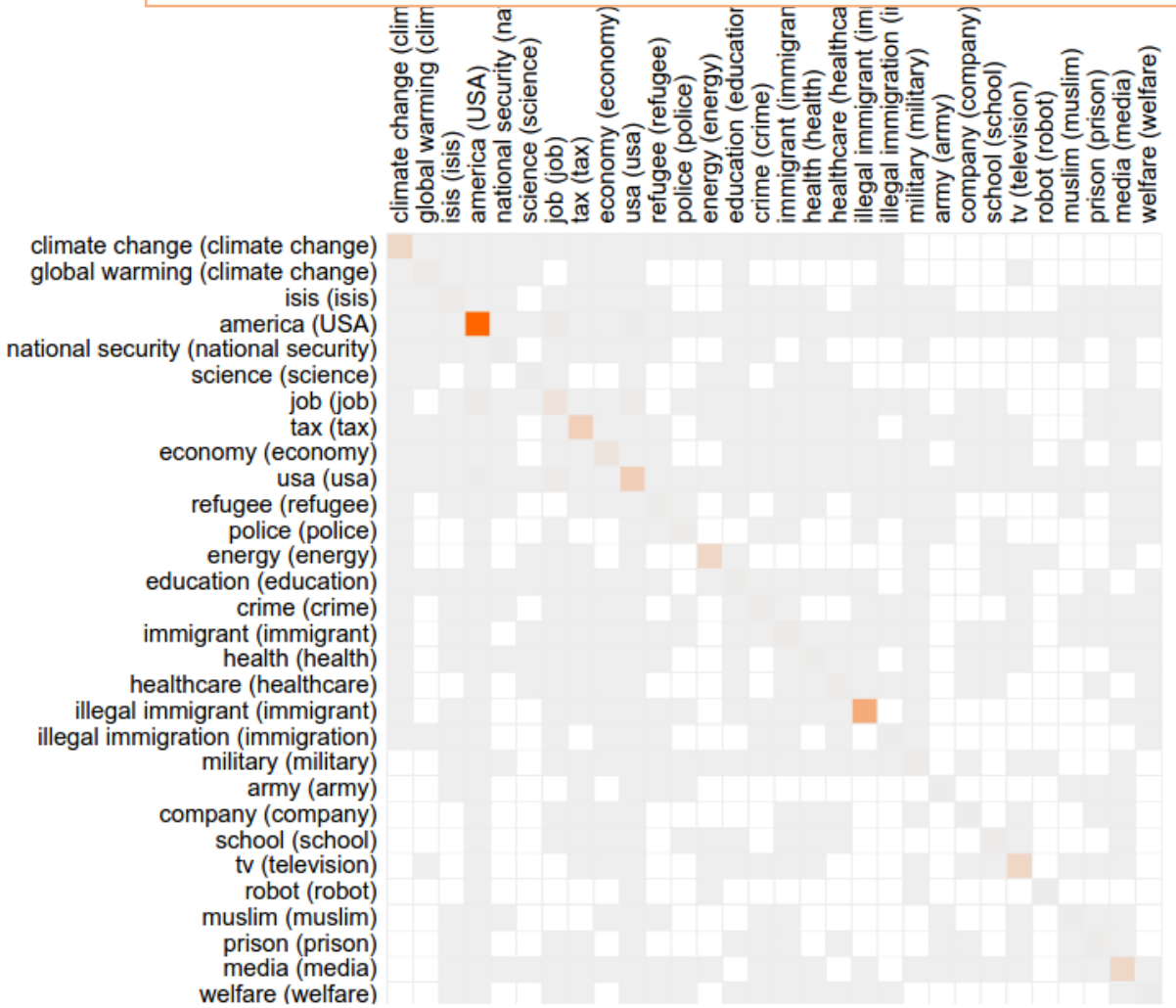
Tweet from MCatlin1984 at 2015-10-19T13:33:43.000Z
https://twitter.com/MCatlin1984/status/656100918237249536
@realDonaldTrump people who mock global warming and use evidence of look it's cold outside show a lack of understanding that is alarming

Tweet from mBTCPizpie at 2015-10-19T13:34:49.000Z
https://twitter.com/mBTCPizpie/status/656101194033590273
@realDonaldTrump Judging the global climate by the weather outside your window is a naive and narrow view of thinking. You are entertaining.

Climate change, ISIS and Trump

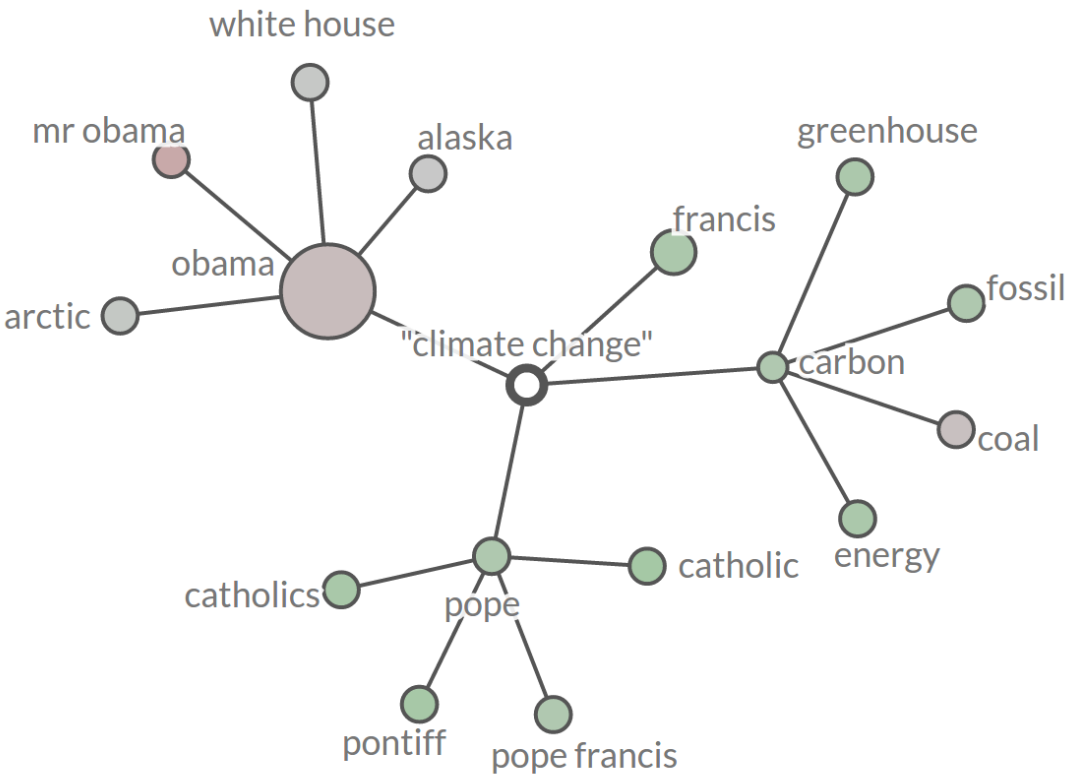
@realDonaldTrump Someone needs to tell Putin Isis and China to beware, the global warming is coming. That will stop them. Not.

@realDonaldTrump WHY IS EVERYONE IN THIS DEBATE BLAMING GLOBAL WARMING!?!? WHAT DOES THAT HAVE TO DO WITH ISIS?!?!!

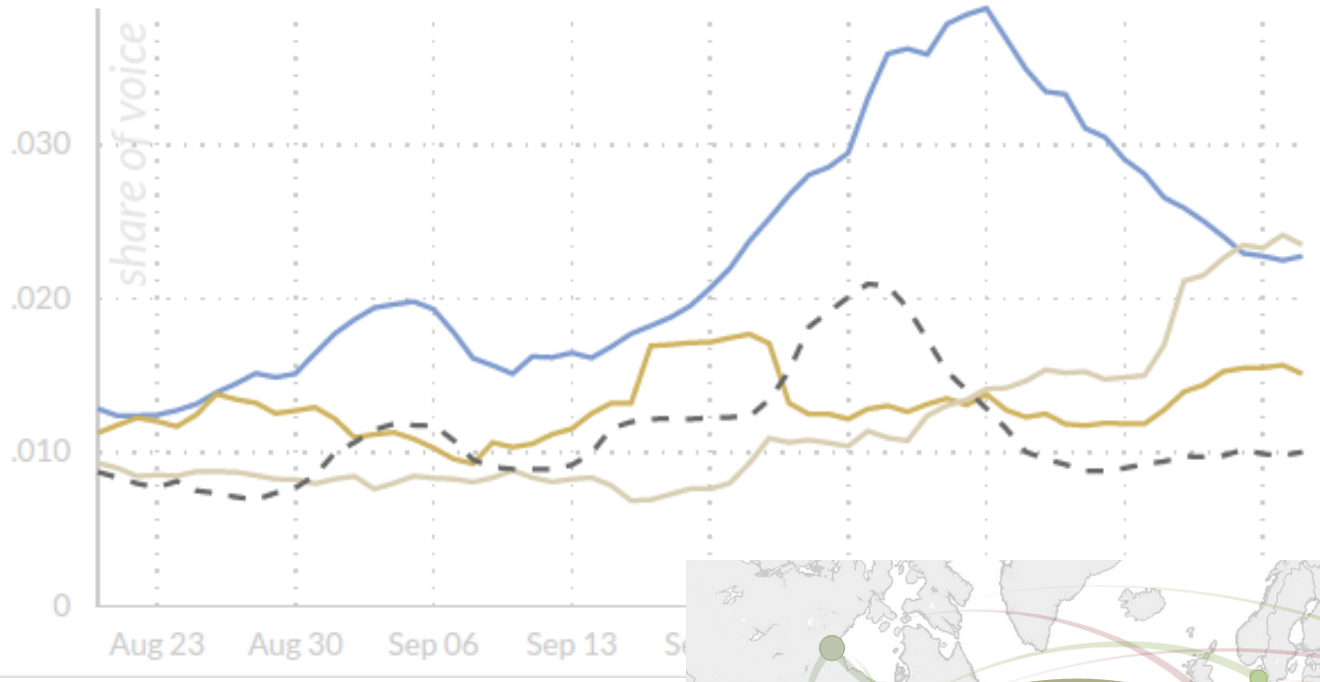


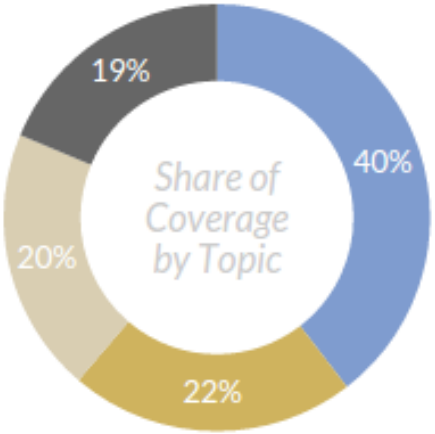
Under the hood

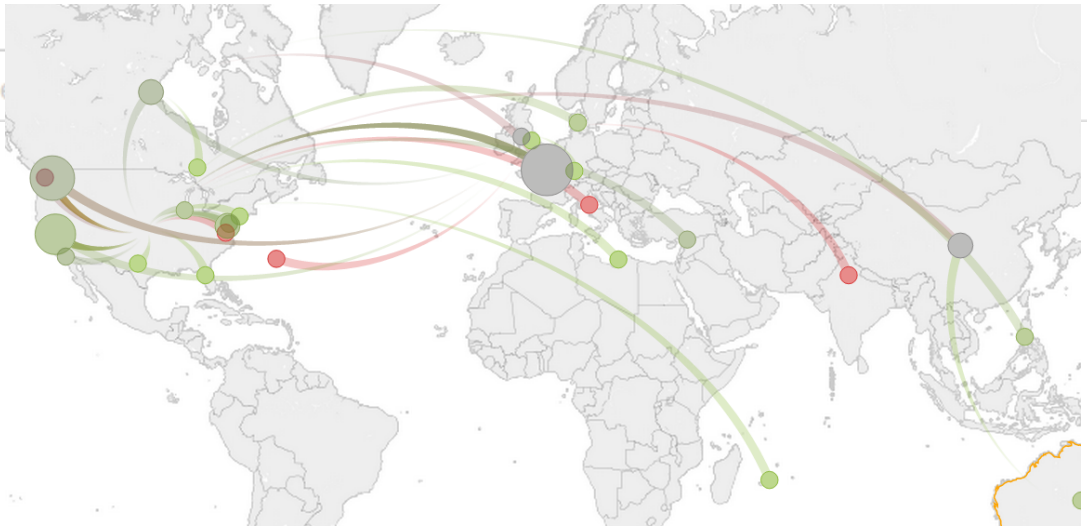
RT @Medieval_React: "Donald Trump wants illegal aliens gone." http://t.co/RBkug2mfyp	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Person	Donald Trump	http://data.ontotext.com/publishing/person/Donald_Trump
RT @dremmelqueen: Pressure Grows On NBC As 115,000+ Sign Petition Demanding SNL Dump Trump http://t.co/tRtevvFao0	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	NBC	pubid:tsk5pw6i8ykg
RT @Splitsider: Watch @TonyAtamanuik and @JAdomian face off in a Trump/Sanders debate at UCB http://t.co/2MmqAlV1uQ http://t.co/qJsQPASB1h	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	UCB	http://data.ontotext.com/publishing/organization/UCB
RT @CNN: .@realDonaldTrump and @RealBenCarson request Secret Service protection https://t.co/31iQ3Dg3Tb https://t.co/c3Dvkls7kZ	"2015-10-19T21:48:12.000+03:00"^^xsd:dateTime	pub:Organization	CNN	pubid:tsk8ragiehog
RT @CNN: .@realDonaldTrump and @RealBenCarson request Secret Service protection https://t.co/31iQ3Dg3Tb https://t.co/c3Dvkls7kZ	"2015-10-19T21:48:1			n/publishing/organization



Dynamics Over Time/Location







Journalism Dashboard Prototype

PHEME

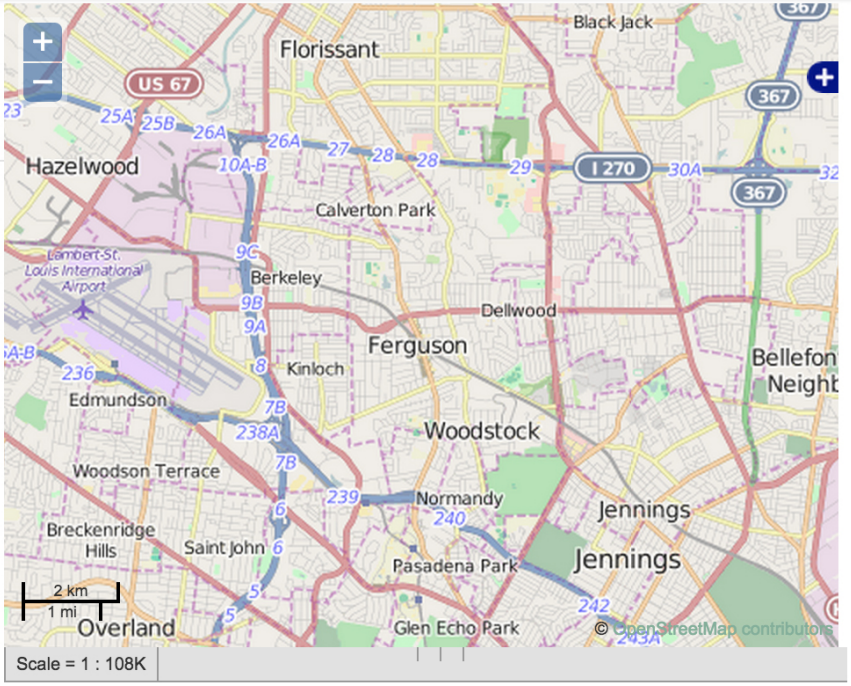
Home Reports Submit a Report Login

Most recent

23/10/2014, 15:10 **80%** True

Police reports indicate Mike Brown a suspect in robbery #Ferguson *@author456*

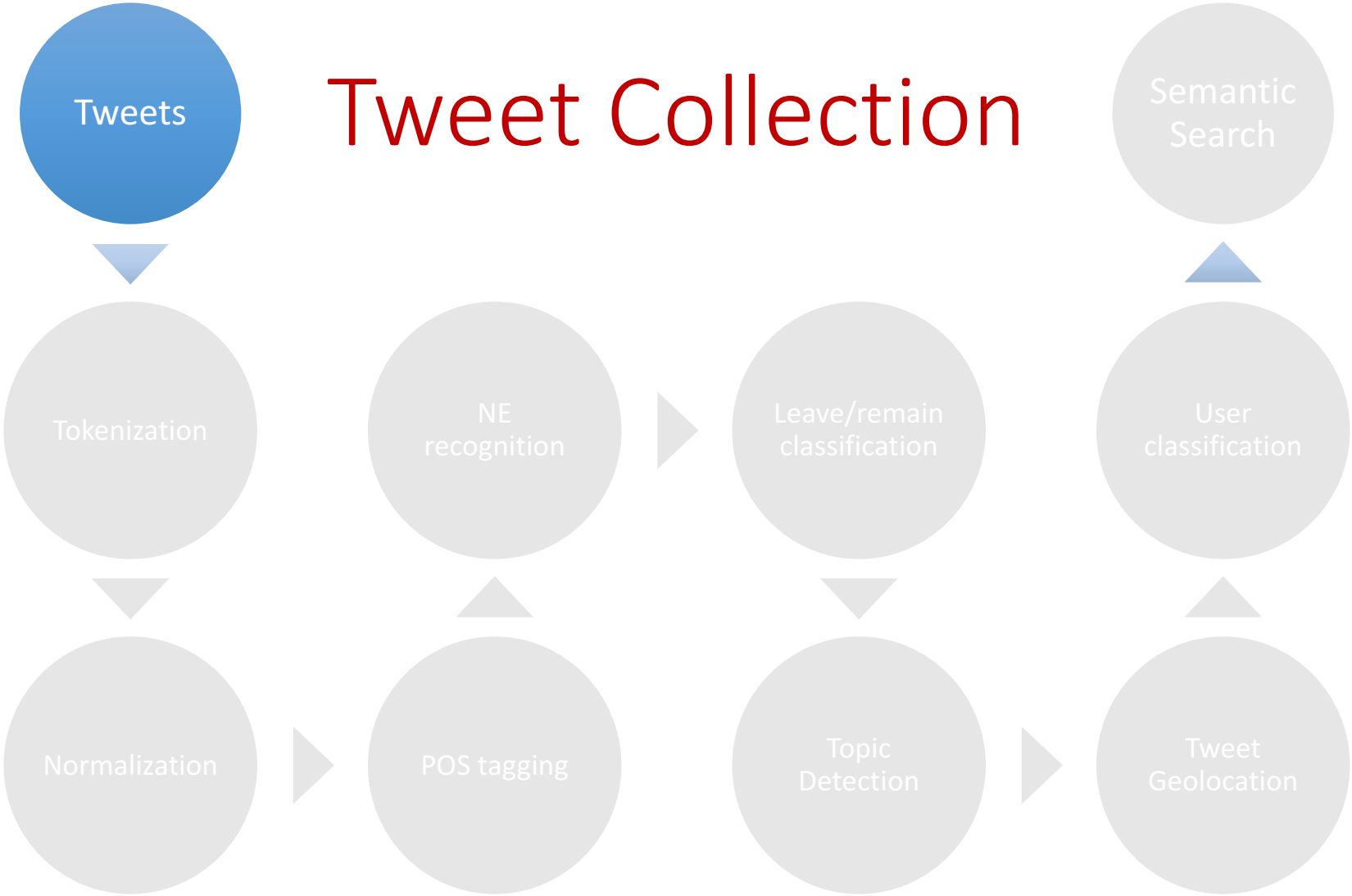
Follow this rumour **100** retweets



brexit: a case study

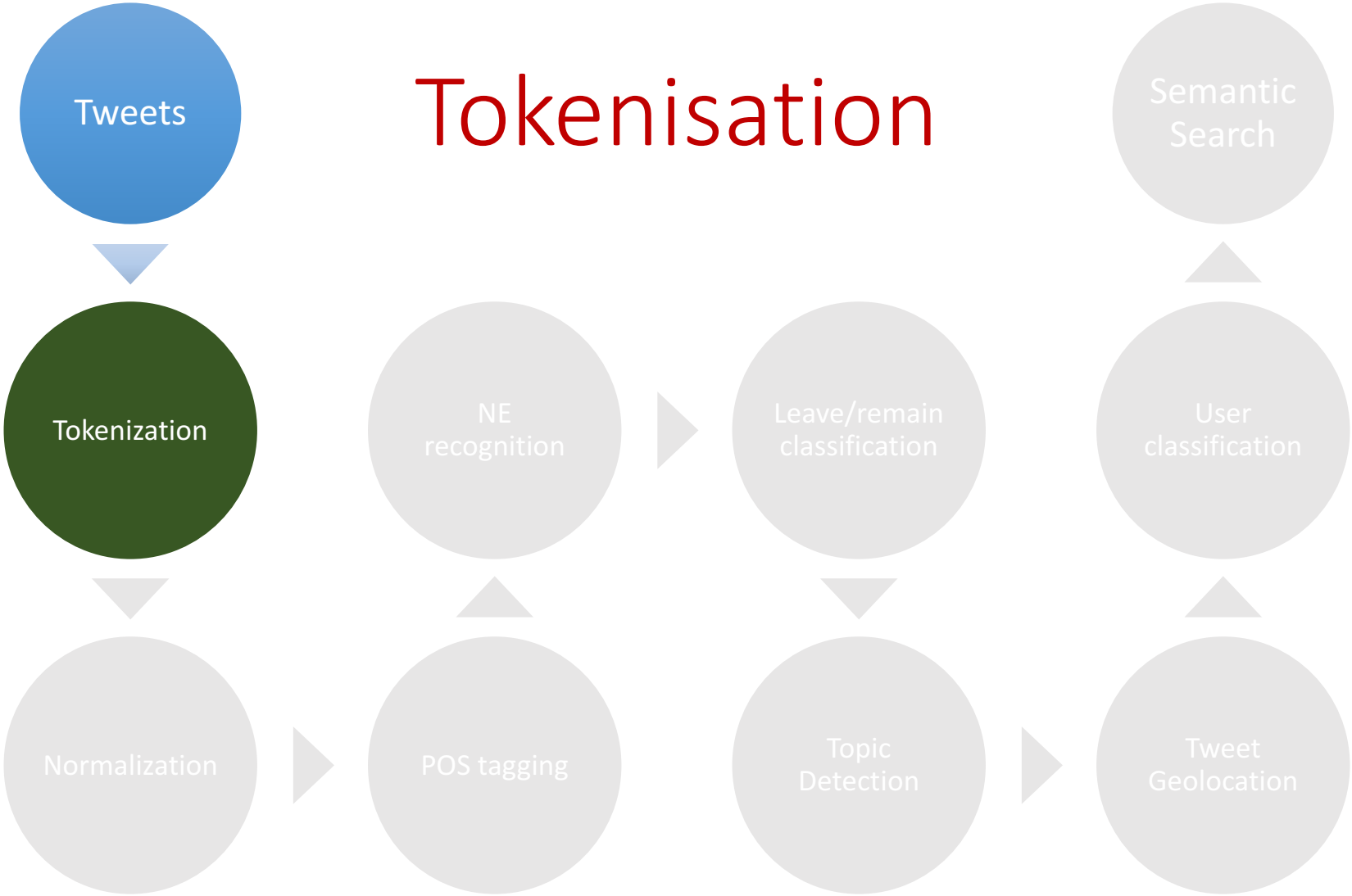
Photo credit: https://www.flickr.com/photos/armydre2008

Tweet Collection

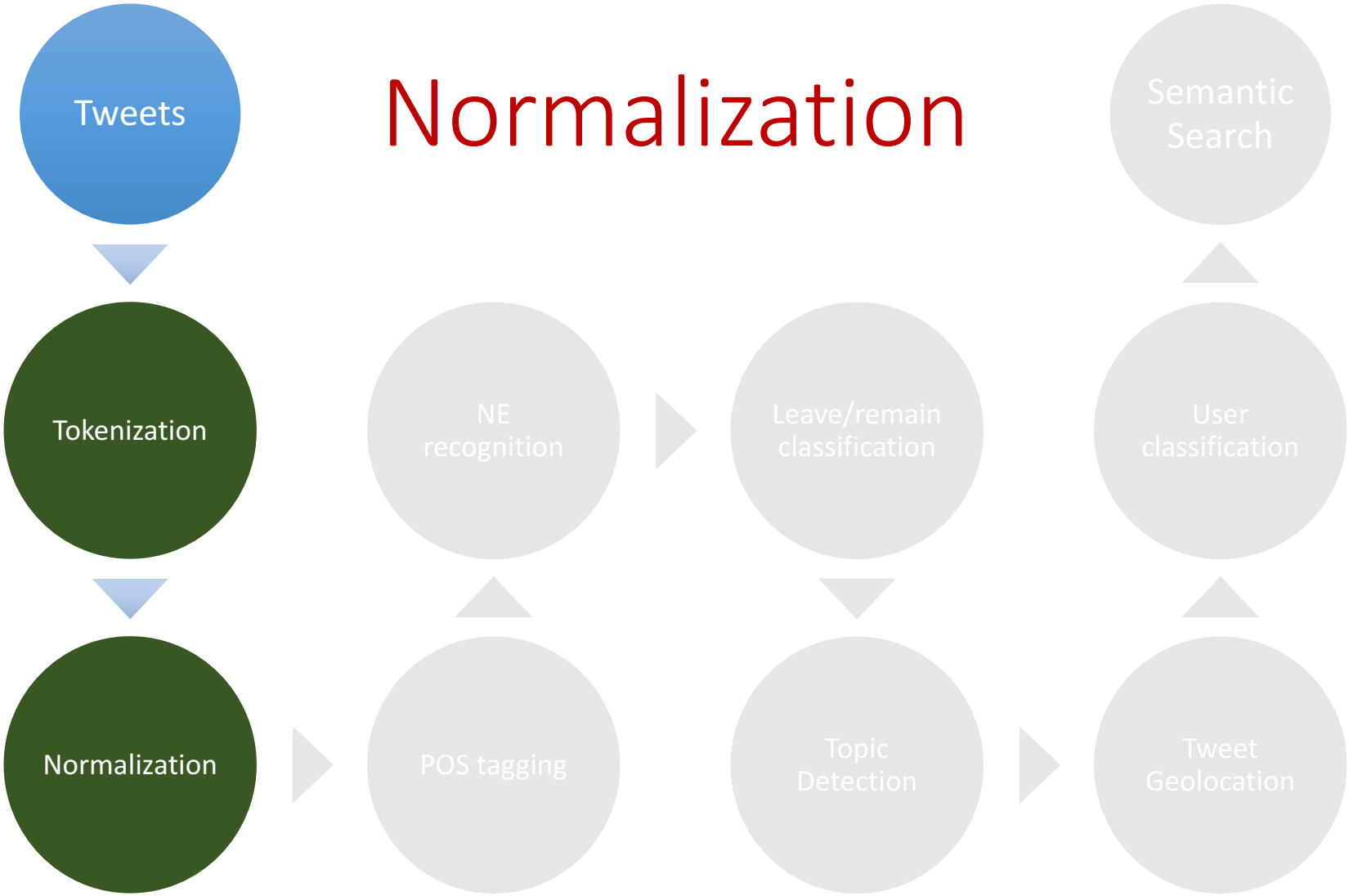


Tweets are tracked in real time using the streaming API

Tokenisation



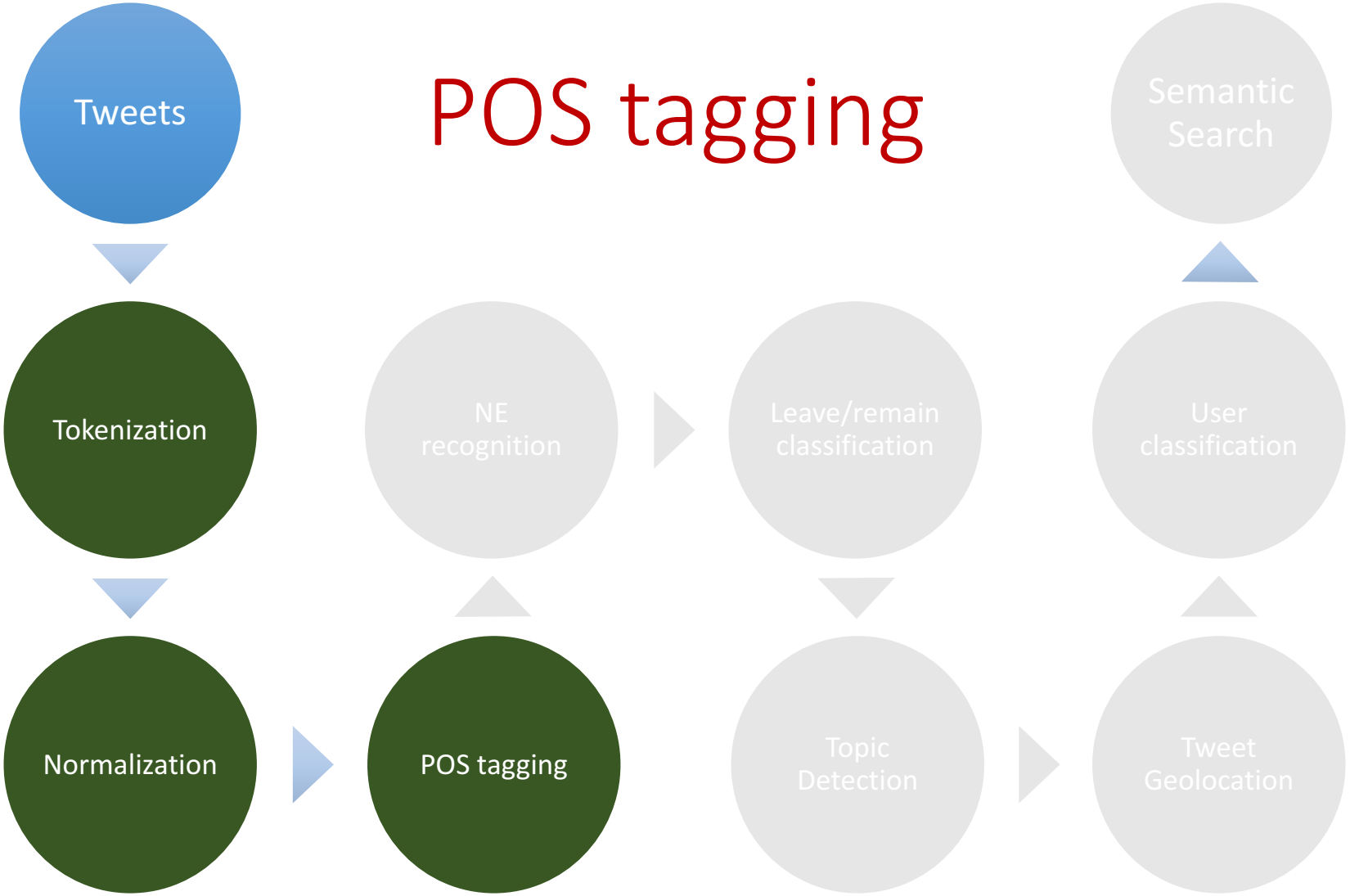
Individual tokens (analogous to terms) are extracted



Normalization

Spelling and abbreviations are normalised to help linguistic processing tools

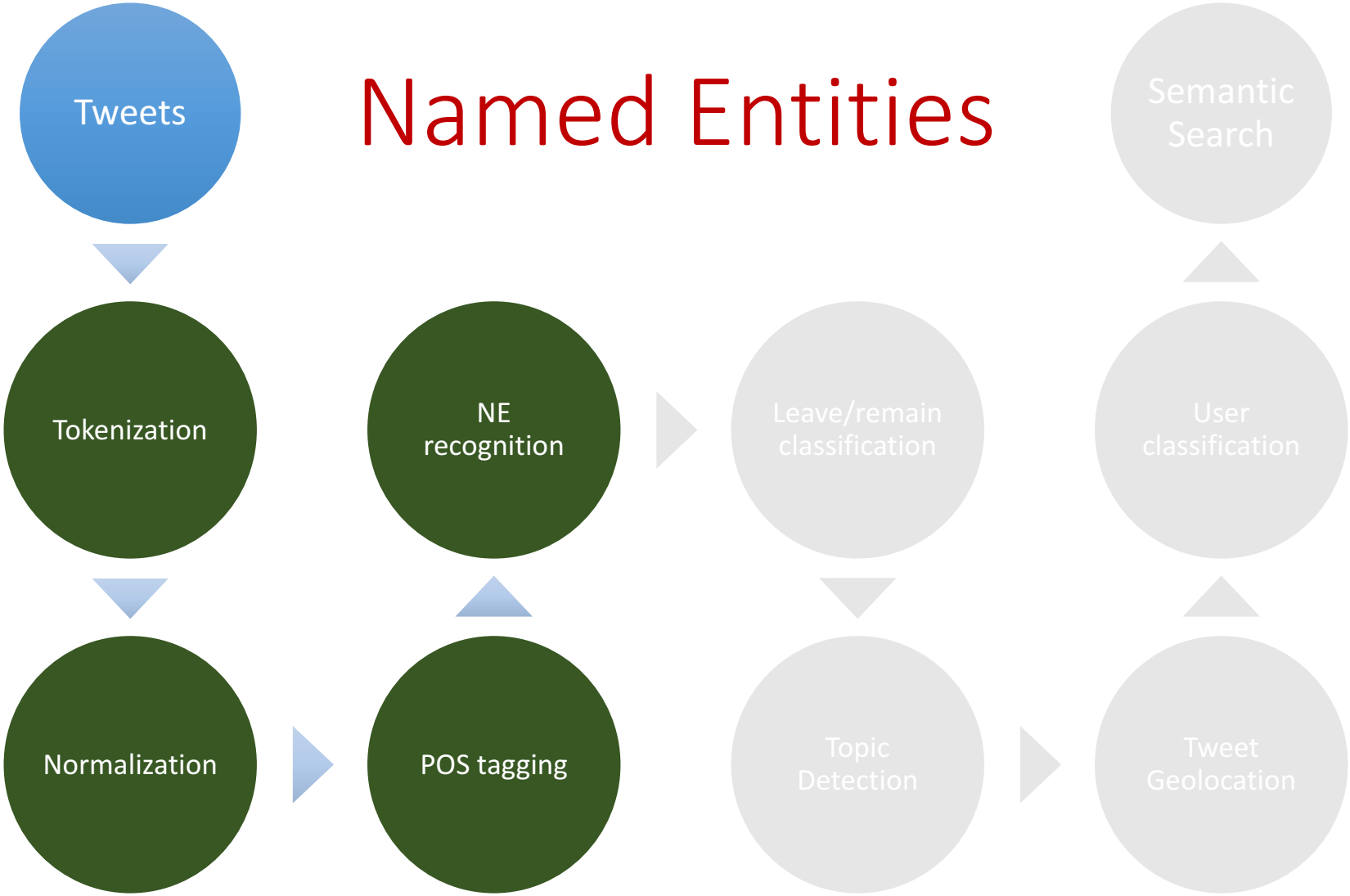
POS tagging



Parts-of-speech (noun, adj etc) are identified.

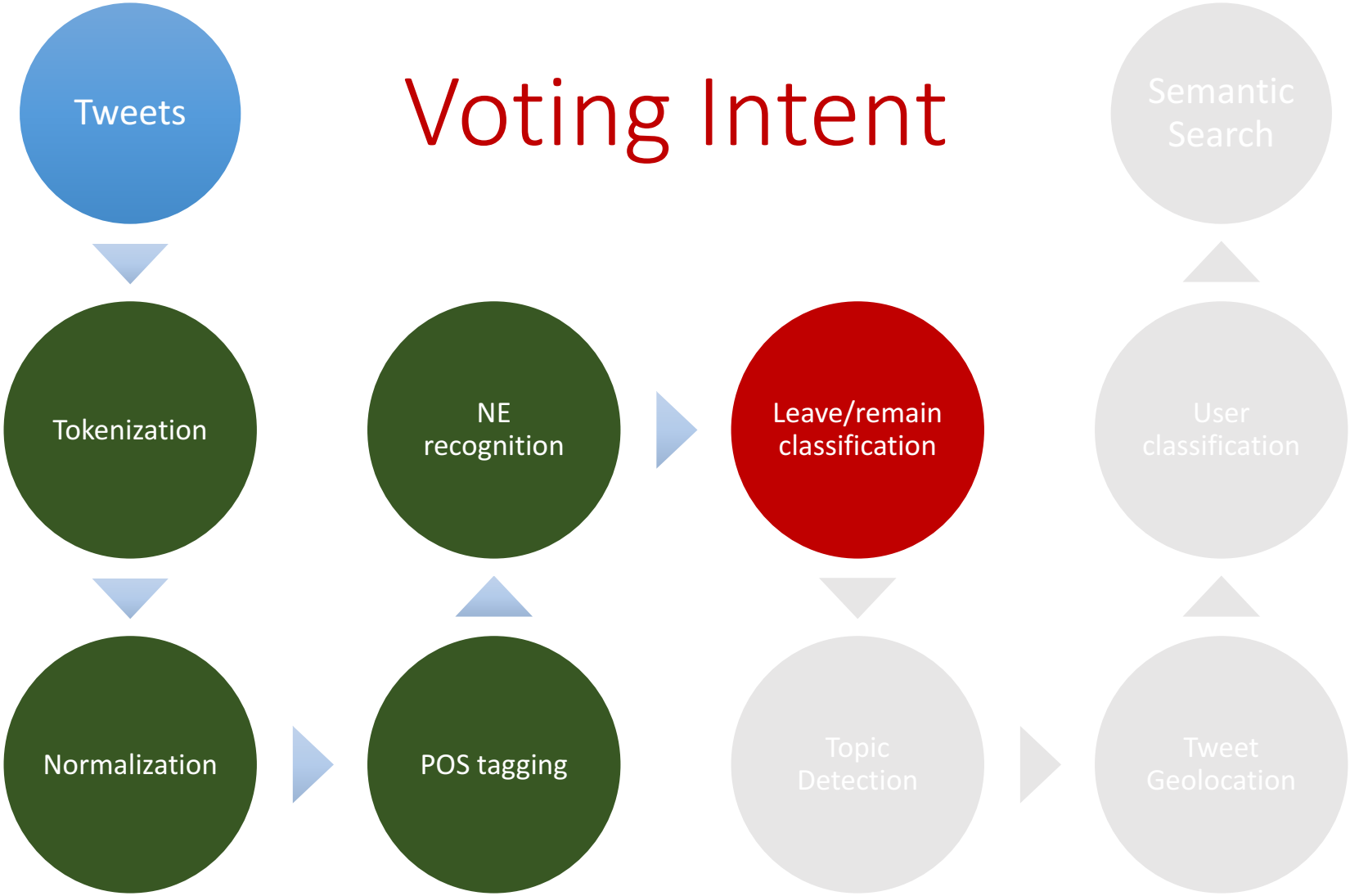
This is necessary to support later processing

Named Entities



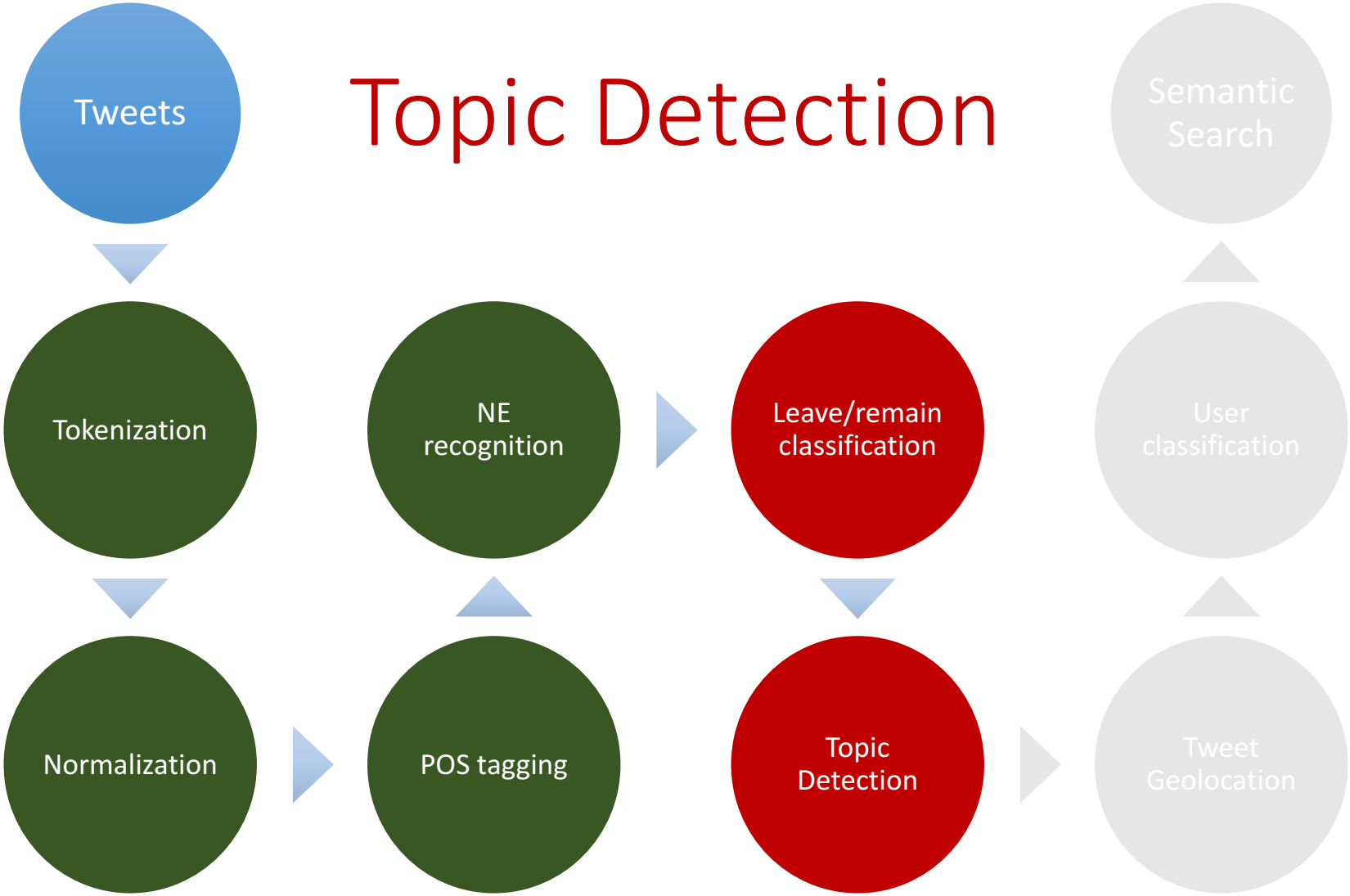
We discover mentions of entities such as people, locations, organisations and products

Voting Intent



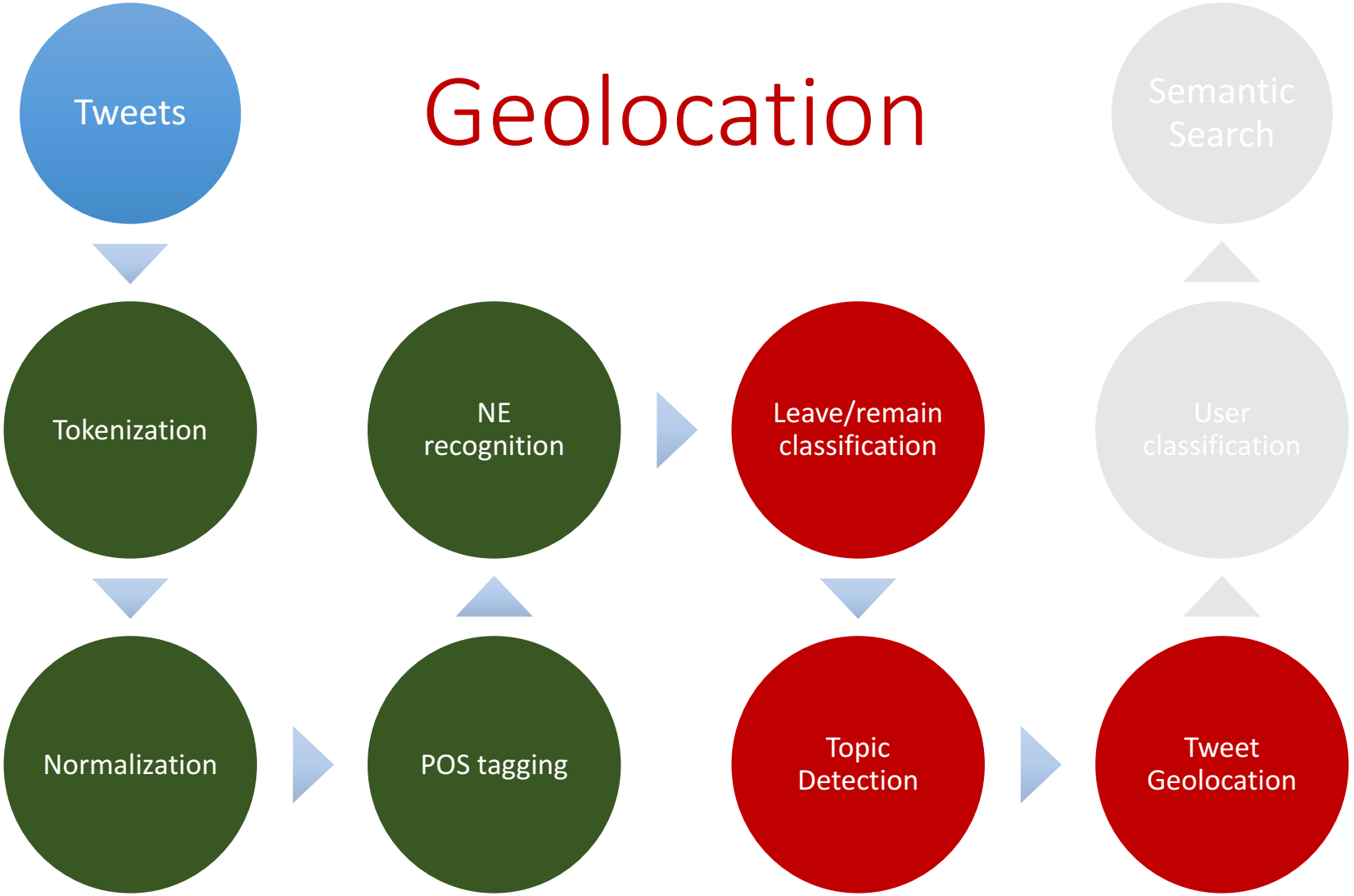
Intent to vote leave or remain classified by #hashtag

Topic Detection



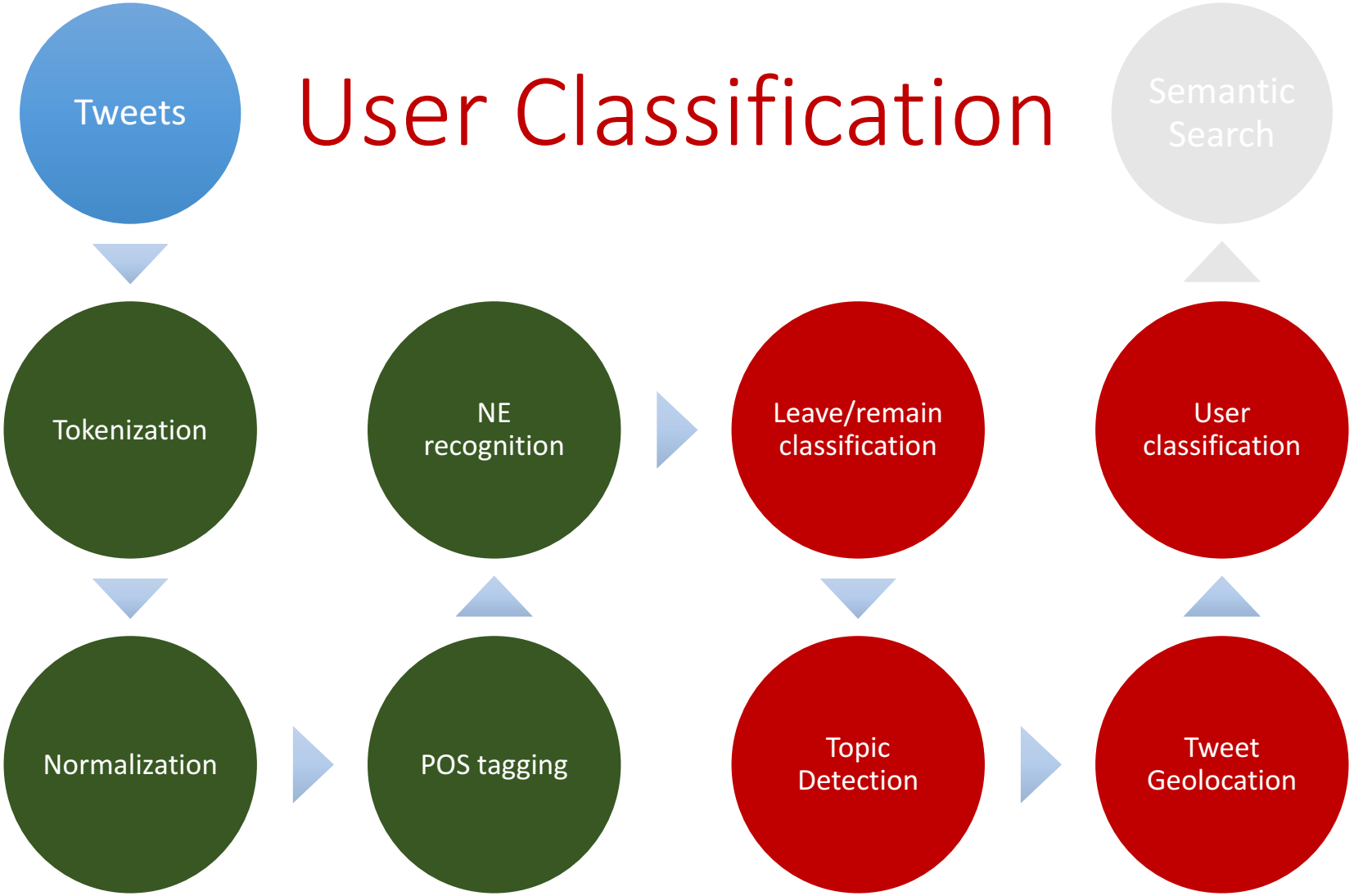
Tweets are matched against a detailed list of topics

Geolocation



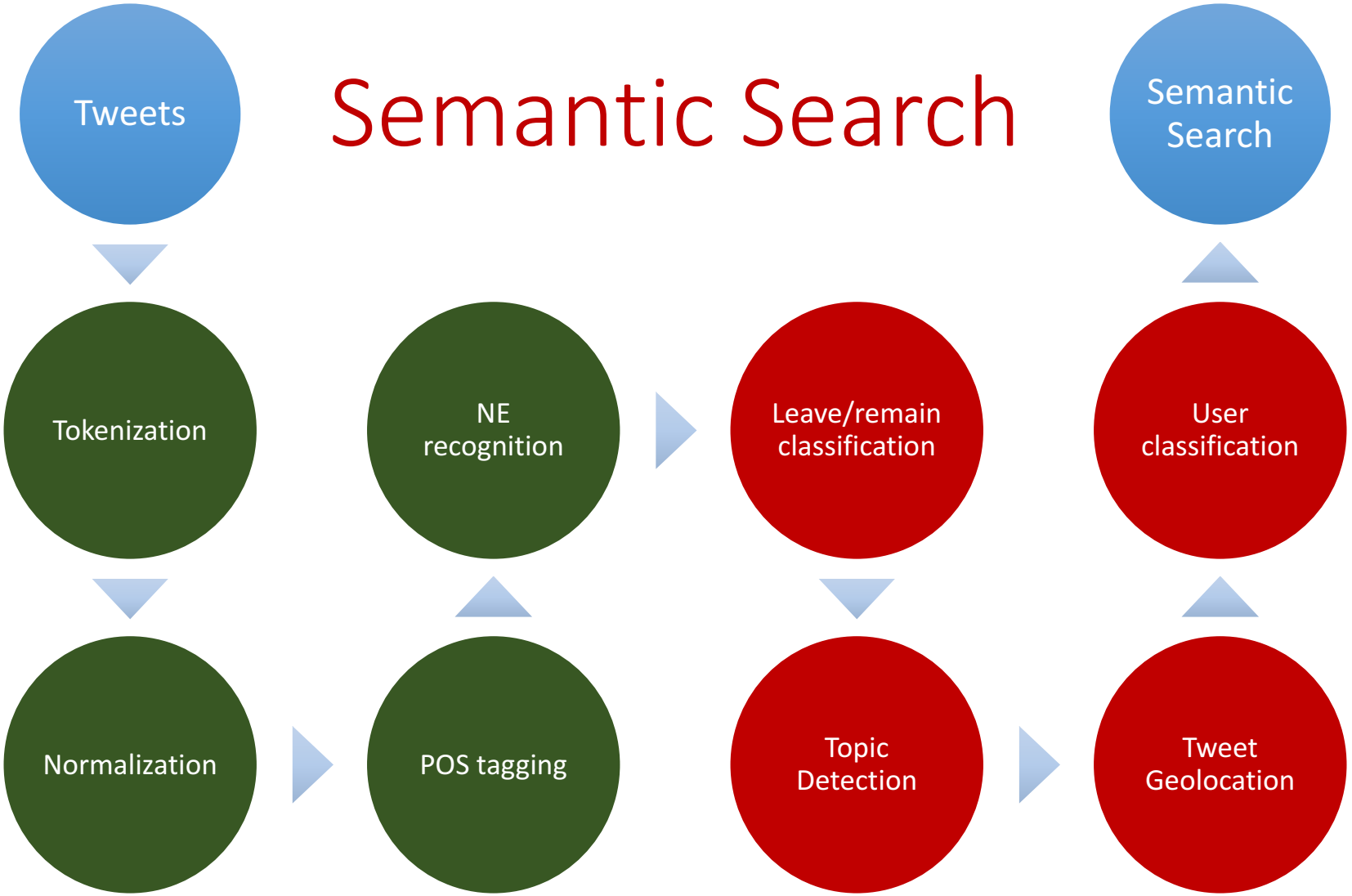
Tweets are linked to NUTS regions based on place tags and user home locations

User Classification



We classify users into eg journalist, charity, member of public

Semantic Search



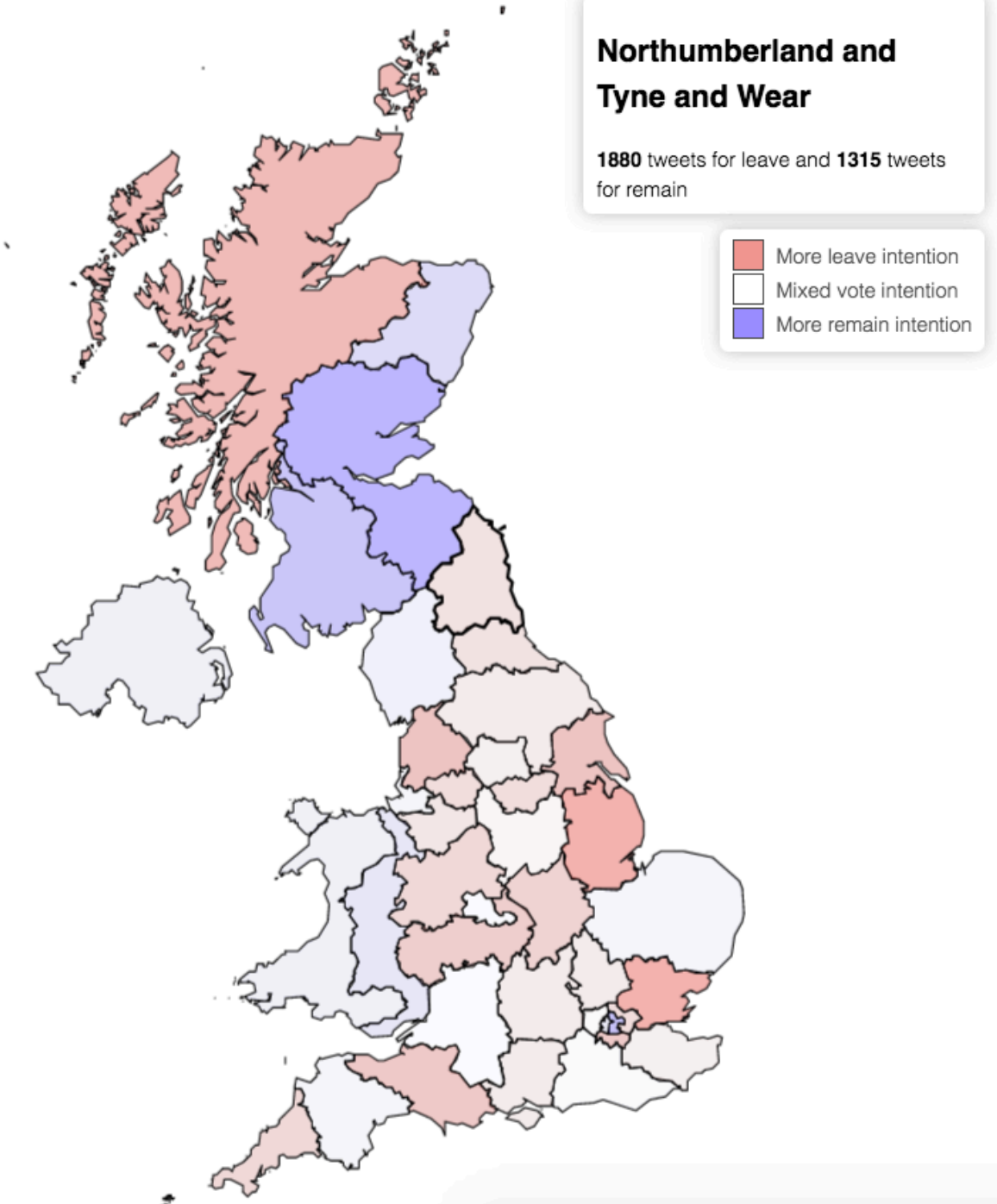
Tweet text and annotations are indexed in semantic search engine Mímir for search and visualisation

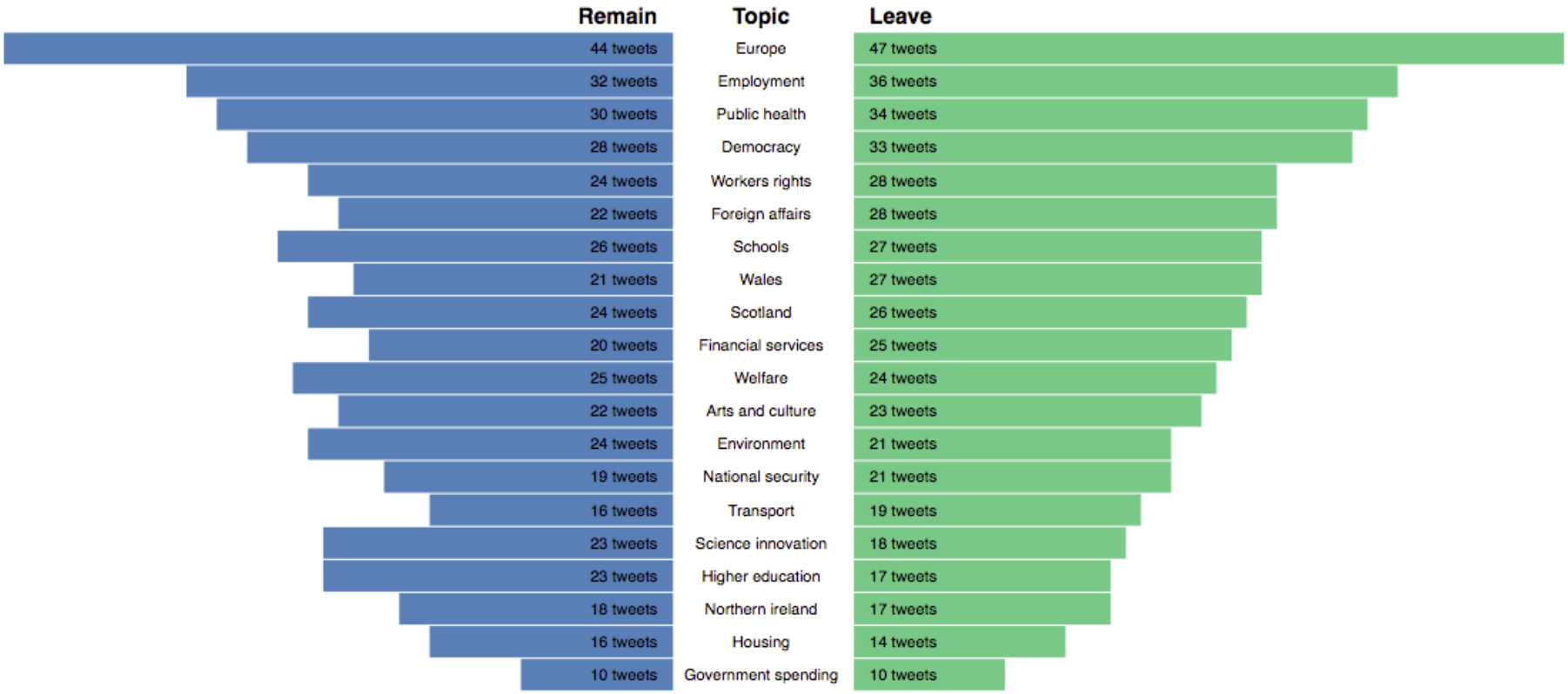
Semantic search with Mímir

- Mímir: Multiparadigm Indexing and Retrieval
- Complex queries – can search over annotations like

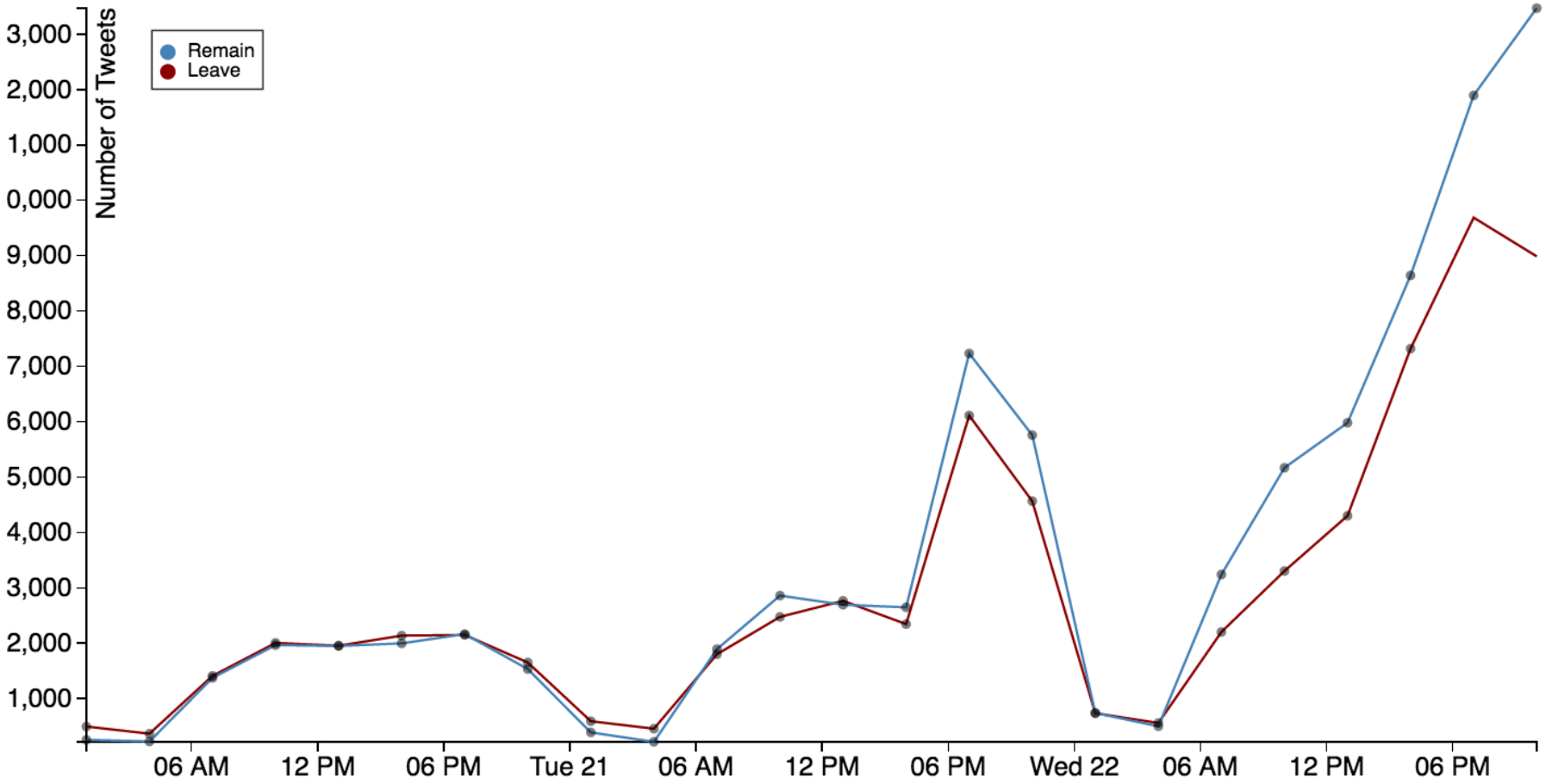
```
{DocumentTimestamp hour_timestamp >= 2016062223 hour_timestamp <
2016062323} OVER ( {DocumentKind tweet_kind = original} AND {NUTS2
NUTS2 = "UKE3"})
```

- Can also mix in full text and semantic queries. Very powerful!
- Which politicians from the North of England over the age of 40 talked most positively about climate change?
- Allows us to drill down and easily see tweets with certain properties


Northumberland and Tyne and Wear
1880 tweets for leave and 1315 tweets for remain
More leave intention
Mixed vote intention
More remain intention


Remain
Topic
Leave
44 tweets
Europe
47 tweets
32 tweets
Employment
36 tweets
30 tweets
Public health
34 tweets
28 tweets
Democracy
33 tweets
24 tweets
Workers rights
28 tweets
22 tweets
Foreign affairs
28 tweets
26 tweets
Schools
27 tweets
21 tweets
Wales
27 tweets
24 tweets
Scotland
26 tweets
20 tweets
Financial services
25 tweets
25 tweets
Welfare
24 tweets
22 tweets
Arts and culture
23 tweets
24 tweets
Environment
21 tweets
19 tweets
National security
21 tweets
16 tweets
Transport
19 tweets
23 tweets
Science innovation
18 tweets
23 tweets
Higher education
17 tweets
18 tweets
Northern ireland
17 tweets
16 tweets
Housing
14 tweets
10 tweets
Government spending
10 tweets

Back - **20/06/2016 00:00 - 23/06/2016 00:00** - Forward



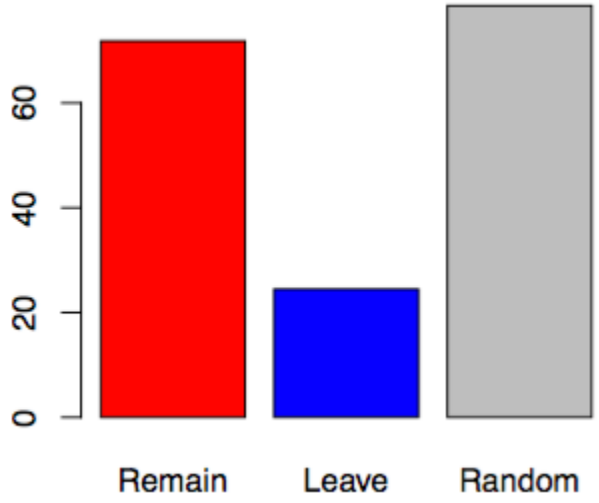
Graph, Data

Ego-network analysis

- Analysis using previous work on selected Brexit data
  - Valerio Arnaboldi
  - Institute of Informatics and Telematics (IIT)
  - Italian National Research Council (CNR)
- Understanding the quantity and quality of relationships of debate participants
- Remain, leave and neutral users were selected by Sheffield using their pipeline and processed by CNR-IIT
- Nice example of bringing several systems together

Remainers are more social

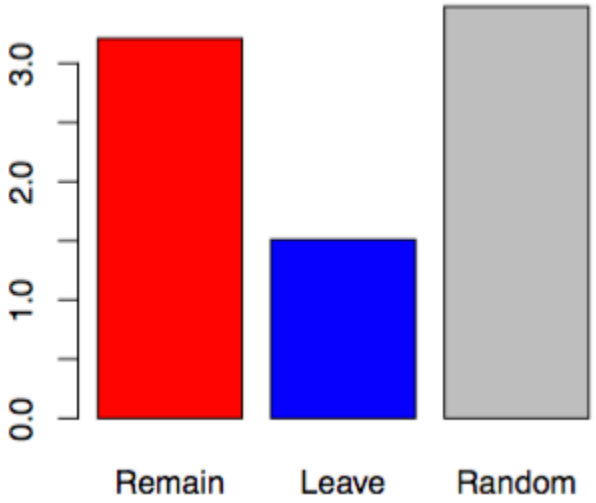
- They maintain larger ego networks
- Larger active network size
  - number of people actively contacted by the egos
  - with a frequency of at least one direct tweet per year
  - considering mentions, replies, and retweets
- Effect remains even after filtering out antisocial accounts



Active network size in users

Leavers had fewer social circles

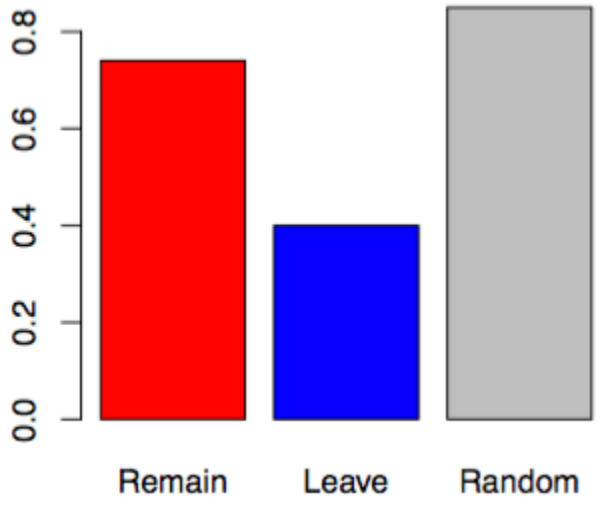
- Social circle:
  - Group of users in an ego-network
  - With similar levels of interaction to one another
  - Most frequent interactions with a smaller social circle
  - More occasional interactions with a larger circle



Average number of social circles

Many “leave” accounts didn’t socialise at all

- A lot of accounts were not used socially
- This is reflected in the number of accounts with social circles
- The leave twitterers were well below the random sample.



% of accounts with at least 1 social circle

The future

- A number of tools have already been brought together around Brexit and other debates
- More tools still being developed
- Tools available through the SoBigData VRE and GATECloud
- Focus on analysis rather than predicting results

Links for more info

- GATE at http://gate.ac.uk
- GateCloud at https://gatecloud.net/
- Brexit Visualisations at http://demos.gate.ac.uk/sobigdata/brexit/
- Brexit study blog post from NESTA at http://www.nesta.org.uk/blog/network-analysis-top-eu-referendum-tweeters
- Brexit study blog posts from Sheffield at http://gate4ugc.blogspot.co.uk/search/label/Brexit
- UK elections monitor at http://gate.ac.uk/projects/pft

Acknowledgements

Current work supported by:

- the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework and H2020 Programmes for R&D
  - DecarboNet (610829) http://www.decarbonet.eu
  - Pheme (611233) http://www.pheme.eu
  - SoBigData (654024) http://www.sobigdata.eu
  - COMRADES (687847) http://www.comrades-project.eu
  - OpenMinted (654021) http://openminted.eu
  - Kconnect (644753) http://www.kconnect.eu/
- Nesta http://nesta.org.uk