SOCIAL DATA SCIENCE LAB.

Gaining Insights from Social Media Data: Collection, Analysis and Interpretation

Matt Williams (@MattLWilliams) & Luke Sloan (@DrLukeSloan)
Social Data Science Lab

Cardiff School of Social Sciences

Cardiff University

tweet: @socdatalab
web: socialdatalab.net

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Research Landscape

- ‘coming crisis of empirical sociology’ (Savage and Burrows, 2007)
- Social scientists distinctive expertise using sample survey and interview
- Compromised by proliferation of ‘social transactional data’ generated, owned and analysed by corporations
- Advent of ‘big and broad data’
- Exponential growth of social media uptake and the availability of vast amounts of social data
- 6 Vs: volume, variety, velocity, veracity, virtue and value

SOCIAL DATA SCIENCE LAB.

6 Vs

Volume

–Ninety percent of the world's data has been created in the past two years (BIS 2013)

–Expansion of 'digital publics' to an unprecedented level

–2.5 billion non-unique users, with Facebook, Google+ and Twitter accounting for over half of these (Smith 2014)

–Production of hundreds of petabytes of information daily, with Facebook users alone uploading 500 terabytes (5.2 million megabytes) of data daily (Tam 2012)

–UK Data Archive currently holds between 2.2 and 15 terabytes of data

–Within the UK alone there are 15 million registered Twitter users (Rose 2014) posting on average 30 million tweets per day (Sloan et al. 2013)

–Of these online social interactions, a sizable portion are thought to be relevant to social science research questions

SOCIAL DATA SCIENCE LAB.

6 Vs

Velocity

–Rapid and continual production naturally occurring data means researchers can observe events as they unfold, as opposed to retrospectively gathering data months/years after the event

Variety

–Heterogeneous forms of data

–Text, images, audio and video

–Unlike qualitative and quantitative data that are labelled, coded and structured, big ‘social’ data are messy, noisy, complex and unstructured

–In order to make sense of this rich material Burnap et al. (2014) advocate the establishment of interdisciplinary teams of computer and social scientists using parallel computing infrastructure to store, search and retrieve relevant information

SOCIAL DATA SCIENCE LAB

6 Vs

Veracity

- Quality, authenticity and accuracy of these messy data
- Are naturally occurring data considered more authentic than data generated by surveys and interviews?
- Inability to pose questions and probe responses can result in findings that are superficial and lacking real insight
- Edwards et al. (2013) advocate triangulating social media communications with more conventional sources, such as curated and administrative data
- Instead of big social data acting as a surrogate for established sources, they should instead be used to augment them, adding a hitherto unrealised locomotive extensive dimension to existing research strategies and designs
- Allows social scientists to study social processes as they unfold at the level of populations, while drawing upon gold standard static qualitative and quantitative metrics to inform interpretations

SOCIAL DATA SCIENCE LAB.

6 Vs

Virtue

–Facebook study claimed to have altered the emotions of users via tailored content did not obtain informed consent from participants (Kramer et al. 2014)

–Criticism in the international media despite their adherence to the Facebook Terms of Service

–Is it practically possible to seek informed consent from SM users in big 'social' data research? Yes for smaller data (qual)

–Twitter's T&Cs state tweet name and text must remain intact

–Restricted to reporting findings at an aggregate statistical level to maintain users' anonymity

–Whole tweet object, including the text property of tweets, cannot be redistributed for research purposes

–Possible to provide tweet IDs, user IDs, and screen names to other researchers

–Re-users able to re-vivify the IDs into full tweet objects (including the text) –maintaining Right to be Forgotten (deleted tweets)

SOCIAL DATA SCIENCE LAB.

6 Vs

Value

–Links the preceding five Vs – only when the volume, velocity and variety of these data can be computationally handled, and the veracity and virtue established, can social scientists begin to ethically marshal them and extract meaningful information

–Transformative potential recognised by government '*Seizing the data opportunity: A strategy for UK data capability*' (BIS 2013)

–Big data are identified as one of the UK's 'eight great technologies'

–Few academic or government organisations have analysed such data

–Due to the lack of existing partnerships with social media companies and computational infrastructure to support social researchers in gaining routine access to and analysing these data.

SOCIAL DATA SCIENCE LAB.

Design Strategy

		Research Design/ Data	
		Locomotive	**Punctiform**
Research Strategy	**Intensive**	E.g. Ethnography/ Observational studies	E.g. X-sectional qual interviewing
	Extensive	NSM Analysis (capturing naturally occurring data in real/useful time at the population level)	E.g. Surveys (X-sect., Longitudinal) and experimental studies

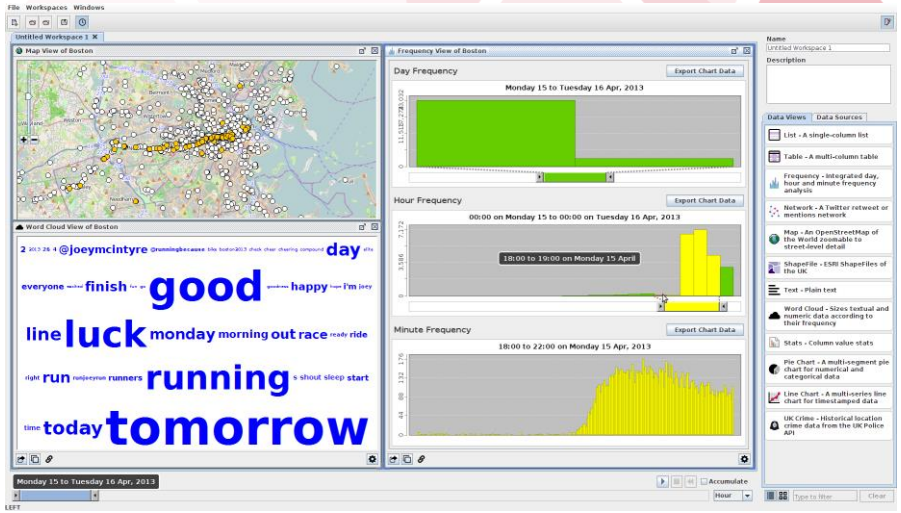
SOCIAL DATA SCIENCE LAB.

Social Data Science Lab

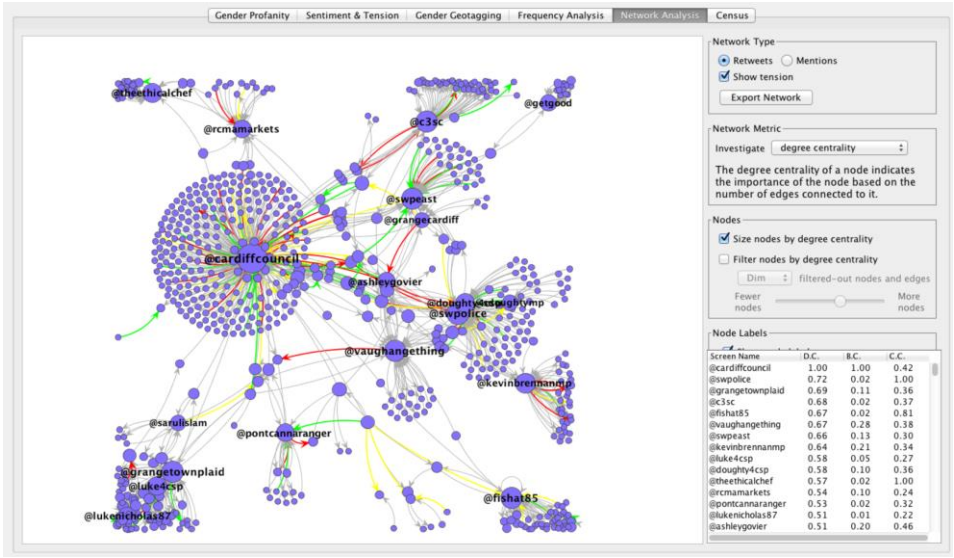
- Aim to establish a **coordinated interdisciplinary response** to “Big Social Data”
- Brings together **computer, social, political, health and mathematical scientists** to study the **methodological, theoretical**, and **empirical** dimensions of Big Data in technical, social and policy contexts
- Developing a **research programme** to help **understand and explain how social processes and interactions manifest on the Web**, with a focus upon the challenges posed by big social data to **government, digital economy and civil society**
- Development of new methodological **tools** and technical/**data solutions** for UK academia and public sector…**a Web Observatory**

SOCIAL DATA SCIENCE LAB.

COSMOS Platform



- Usable – *developed with social scientists for social scientists*
- Reproducible/Citable Research - *export/share workflow*



- Integrated
- Open (“plug and play”)
- Scalable (MongoDB data stores/Hadoop Back End)

SOCIAL DATA SCIENCE LAB.

Burnap, P. et al. (2014) ‘*COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand*’, IJPEDS

Data Observatory Features

- Data Collection and Curation
  - Persistent connection to Twitter 1% Stream (~4 billion)
  - Geocoded tweets from UK (~200 million annually)
  - Bespoke keyword-driven Twitter collections (on crime and security)
  - ONS/Police API
  - Drag and drop RSS
  - Import CSV/JSON
  - …Web enabled so push/pull data from anywhere (i.e. other observatories!)

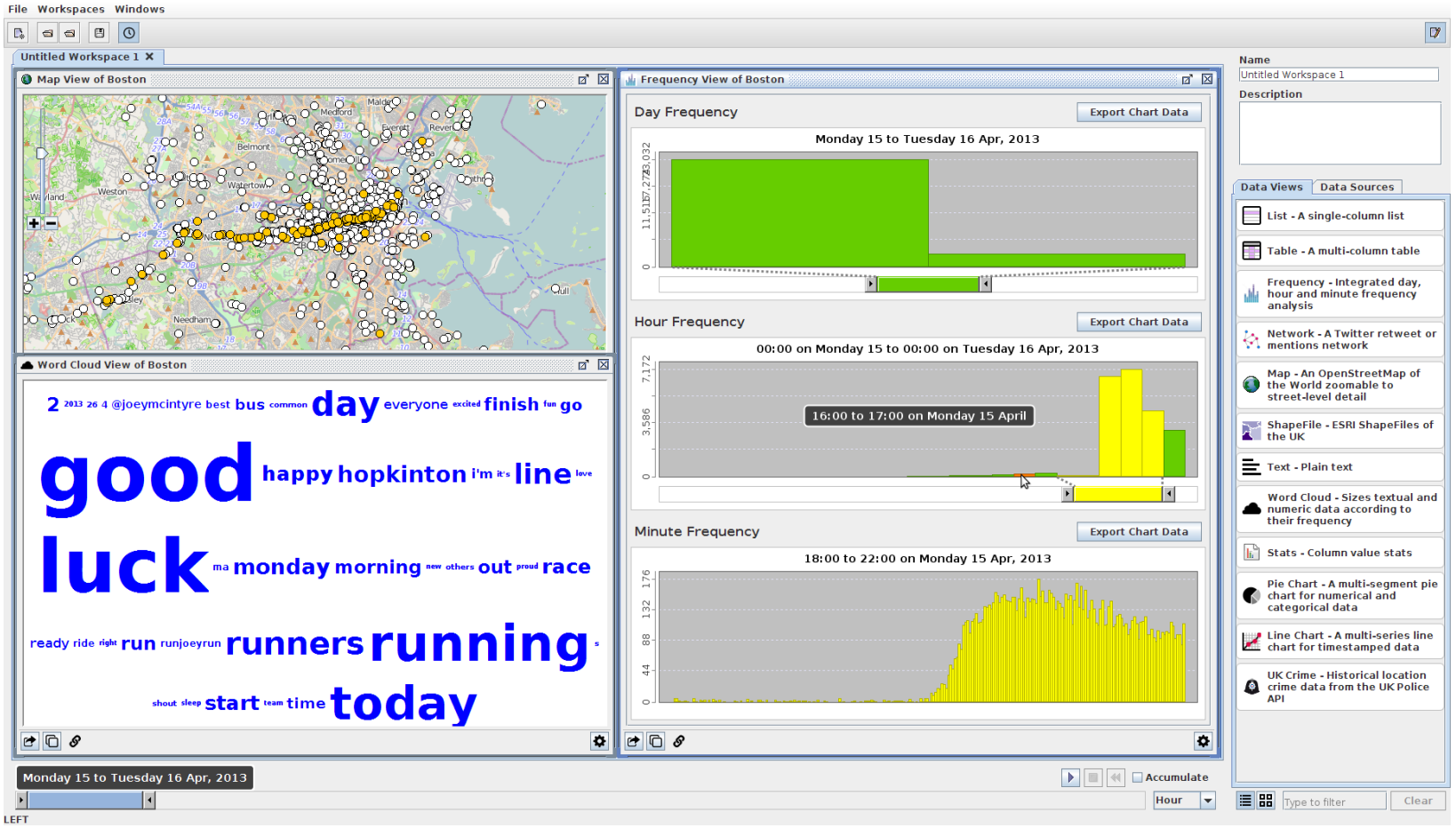
SOCIAL DATA SCIENCE LAB.

Data Observatory Features

- Data Transformation
  - Word Frequency
  - Point data frequency over time
  - Social Network Analysis
  - Geospatial Clustering
  - Sentiment Analysis
  - Demographic Analysis (gender, location, age, occupation/social class)
  - API to plug in new data sources (e.g. ONSAPI, Open Data etc.)

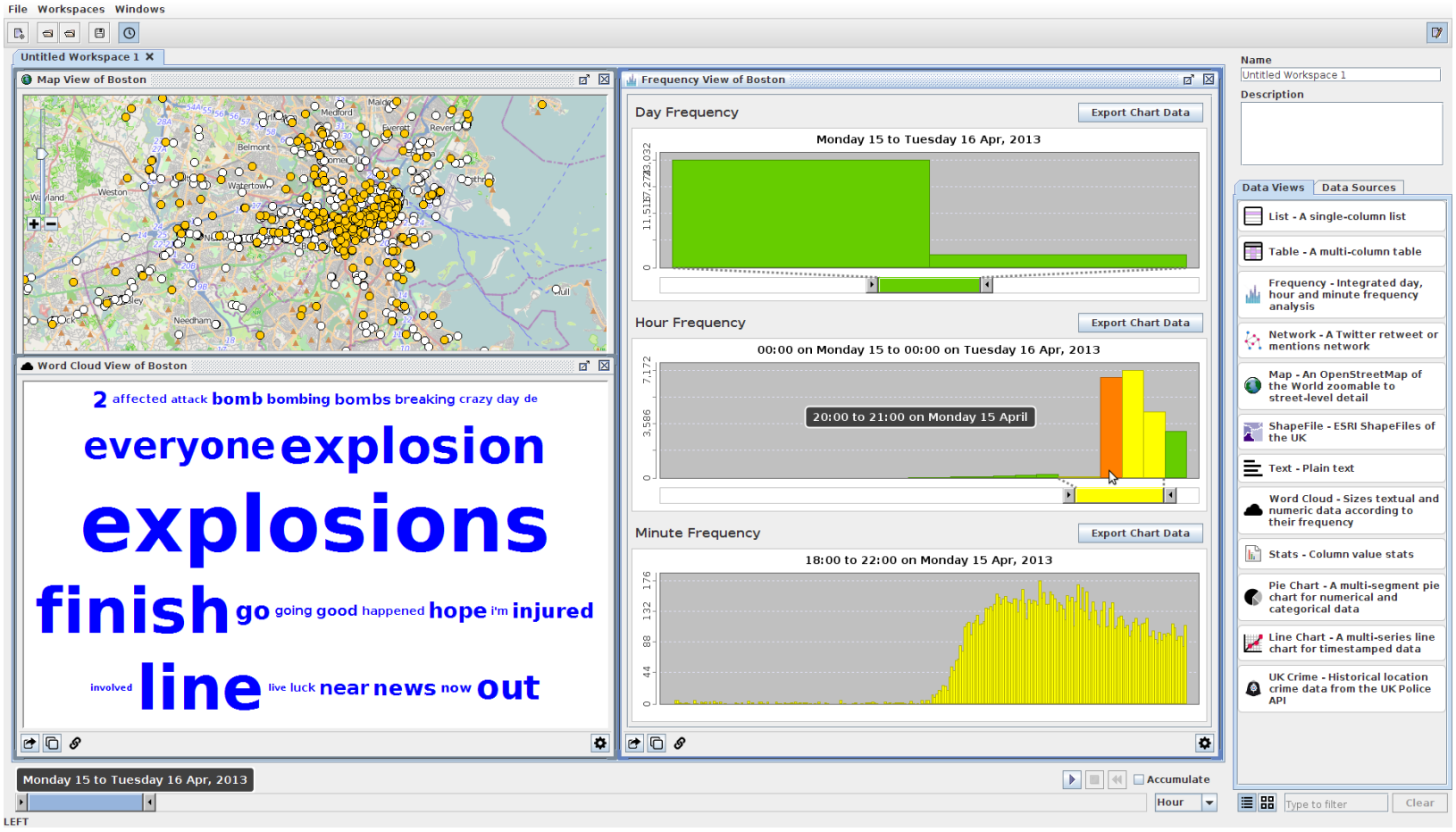
SOCIAL DATA SCIENCE LAB.

Event Monitoring



SOCIAL DATA SCIENCE LAB.

Event Monitoring



SOCIAL DATA SCIENCE LAB.

COSMOS Infrastructure

COSMOS Desktop

- Small local datasets
- Users' API credentials
- Local analysis
- Sept '14 launch (>700 dl's in 17 countries)

HPC WALES · CYMRU

hadoop MapReduce

COSMOS Cloud

- Scalable storage
  - Massive datasets
- Scalable compute
  - On-demand nodes
  - Fast search & retrieve
  - Fast analysis
- Workflow management
- Collaboration support
- 2016/17 launch

SOCIAL DATA SCIENCE LAB.

Lab Projects

2011

2017

Digital Social Research Tools, Tension Indicators and Safer Communities: A demonstration of COSMOS (ESRC DSR)

COSMOS: Supporting Empirical Social Scientific Research with a Virtual Research Environment (JISC)

Small items of research equipment at Cardiff University (EPSRC)

Hate Speech and Social Media: Understanding Users, Networks and Information Flows (ESRC Google)

Social Media and Prediction: Crime Sensing, Data Integration and Statistical Modelling (ESRC NCRM)

Understanding the Role of Social Media in the Aftermath of Youth Suicides (Department of Health)

Scaling the Computational Analysis of “Big Social Data” & Massive Temporal Social Media Datasets (HPC Wales)

Digital Wildfire: (Mis)information flows, propagation and responsible governance, (ESRC Global Uncertainties)

Public perceptions of the UK food system: public understanding and engagement, and the impact of crises and scares (ESRC/FSA)

The 2016 Welsh Election Study (ESRC)

Impact Acceleration Grant: Disruptive Event Detection (ESRC/MPS)

SOCIAL DATA SCIENCE LAB.

Hate Speech on Social Media

Funded from April 2013 - August 2014 (ESRC/Google)

Aims:

1: Analyse social media data for the purposes of monitoring emotive responses following major events of national interest

2: Profile hateful social media networks in relation to user behaviour and type

3: Triangulate the above analysis with other forms of open data, such as Google search metrics and traditional media coverage

4: Study and model forms of counter speech

5: Build a statistical propagation model that forecasts the emergence and evolution of hateful information flows

SOCIAL DATA SCIENCE LAB.

Google

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Motivation

- Research (traditional surveys and interviews) shows that both **crime** and **prejudice** are influenced in the short term by singular events such as:
  - **Widely publicized murders** (Phillips, 1980, on **homicide**),
  - **Riots** (Bobo et al., 1994, **on race relations**)
  - **Court cases & terrorism** (King and Sutton, 2014 on **Hate Crime)**
  - **Terrorism** (Legewie, 2013, on **anti immigrant sentiment**)

- Hate Crimes have been shown to **cluster in time and tend to increase**, sometimes dramatically, in the aftermath of an **antecedent or 'trigger' event** (King and Sutton, 2013)
  - Historic preoccupation with *where* hate crimes happen (risky neighborhoods, demographic factors etc.). Little research that looks at *when* they happen
  - 481 hate crimes occurred with a specific anti-Islamic motive a year following 9/11
  - 58% of them perpetrated 2 weeks following the event (4 percent of the at-risk period)
  - Crimes entailing a prejudicial motive often occur in close temporal proximity to galvanizing events, such as terrorist events
  - Temporal focus allows for the study of escalation, duration, diffusion, and deescalation of crime following events
  - However, there is a limitation in offline data: i) low granularity; ii) under reporting; & iii) retrospective

SOCIAL DATA SCIENCE LAB.

Motivation

•Users of **social media** are more likely to express **emotional content** due to deindividuation (anonymity, lack of self awareness in groups, disinhibition) (e.g. Festinger, 1952)

•There is a case history relating to the expression of hateful sentiment on social media in the UK, providing evidence of a criminal justice response:

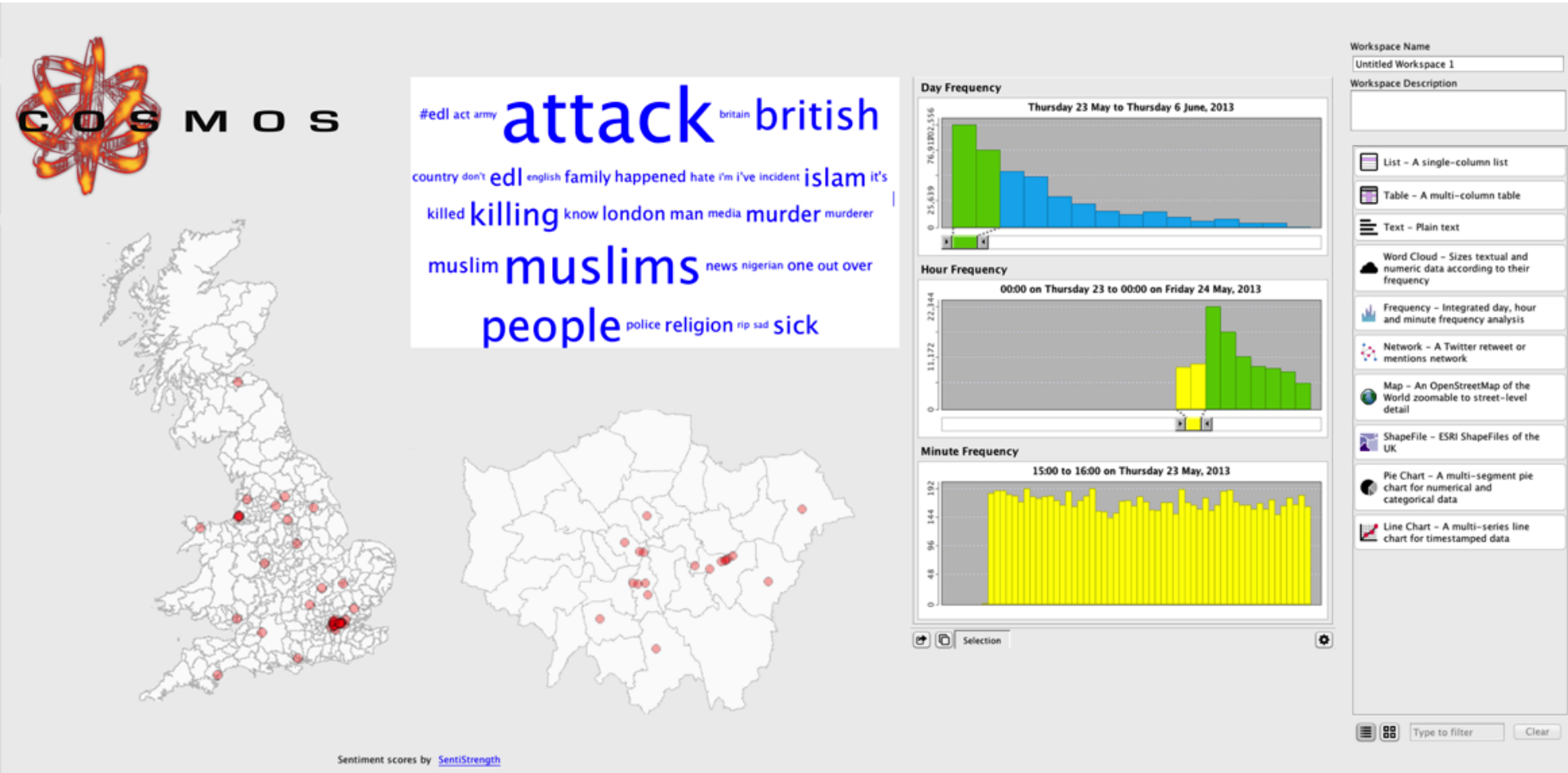
- In 2012, Liam Stacey sentenced to 56 days in prison for posting offensive comments on Twitter after footballer's cardiac arrest;
- In 2012, Daniel Thomas arrested after an abusive message was sent to Olympic diver Tom Daley;
- In 2014, Isabella Sorley and John Nimmo jailed for abusing feminist campaigner Caroline Criado-Perez;
- In 2014, Declan McCuish jailed for a year for tweeting racist comments about two Rangers players.

•As social media data is locomotive, extensive and linked (unlike surveys or interviews) it presents a unique opportunity to study the fine grained (i.e. seconds instead of days, months or years) escalation, duration, *diffusion*, and deescalation of hate speech following events

•We study several case events: Woolwich & Boston terror attacks, Obama & US election, Tom Daley & Jason Collins coming out, Criado-Perez's online harassment and Paralympics

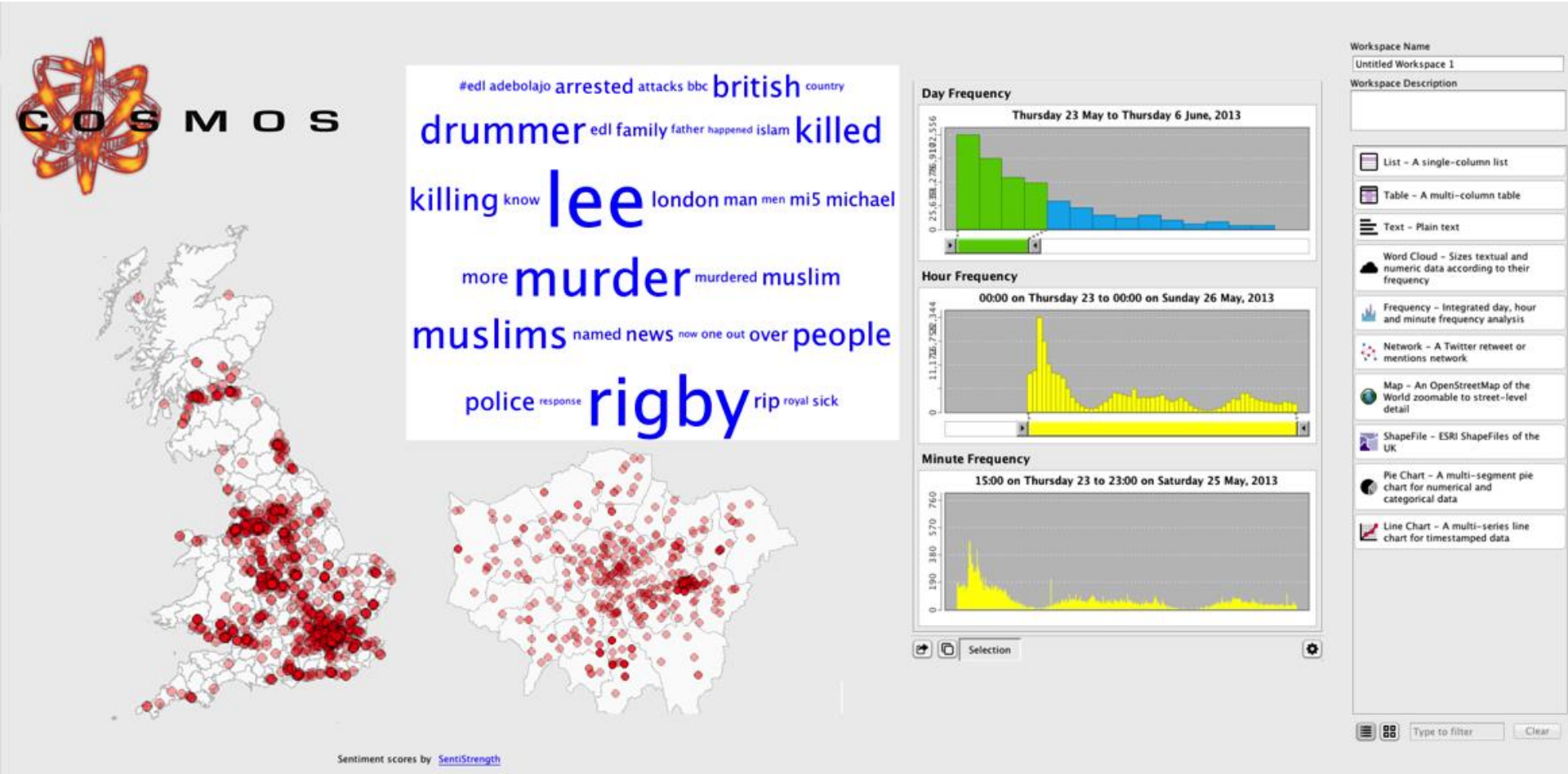
SOCIAL DATA SCIENCE LAB.

Social Media Reaction: Impact



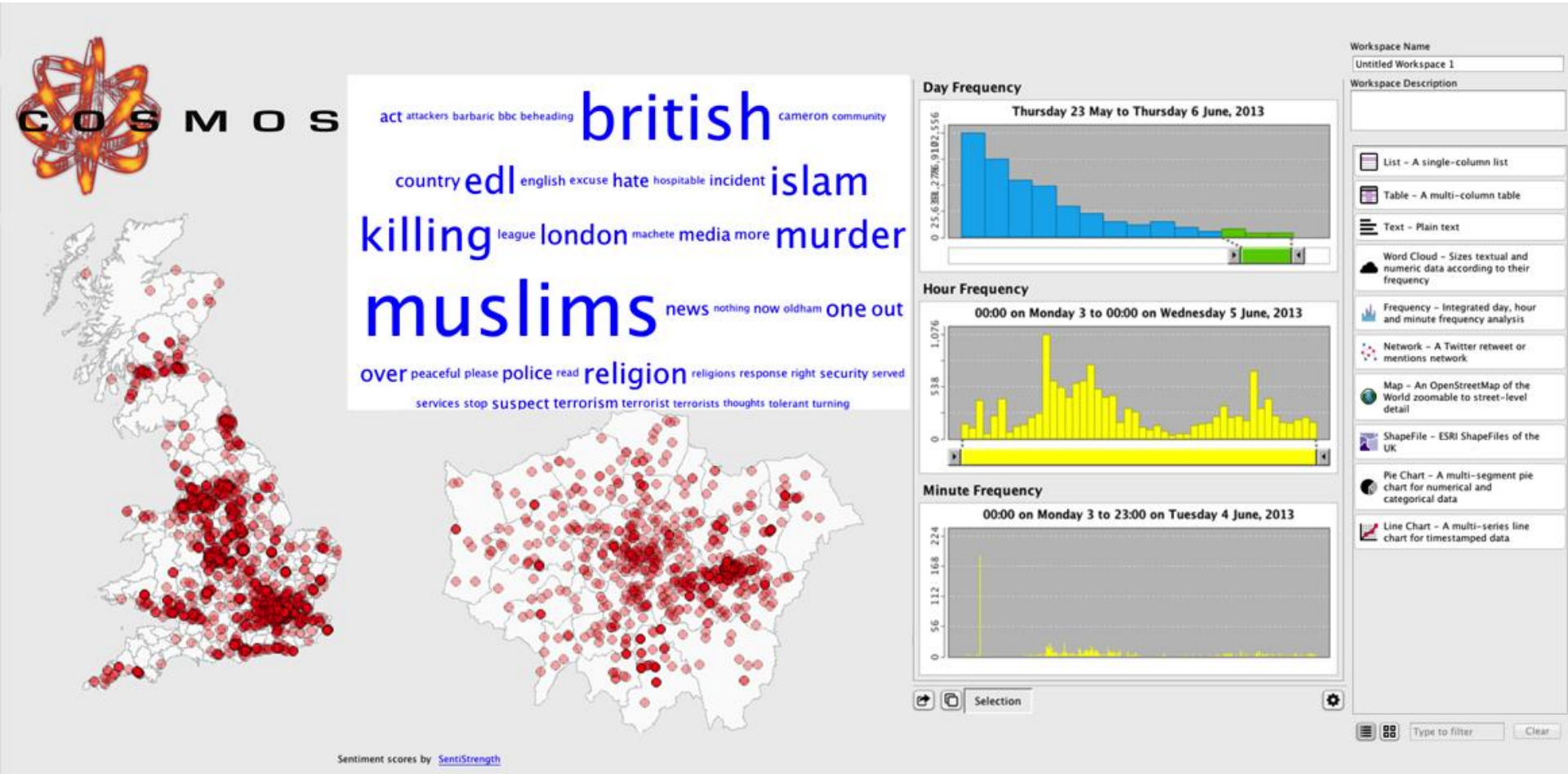
“during which the disaster strikes and the immediate unorganised response to the death, injury and destruction takes place”

Social Media Reaction: Inventory



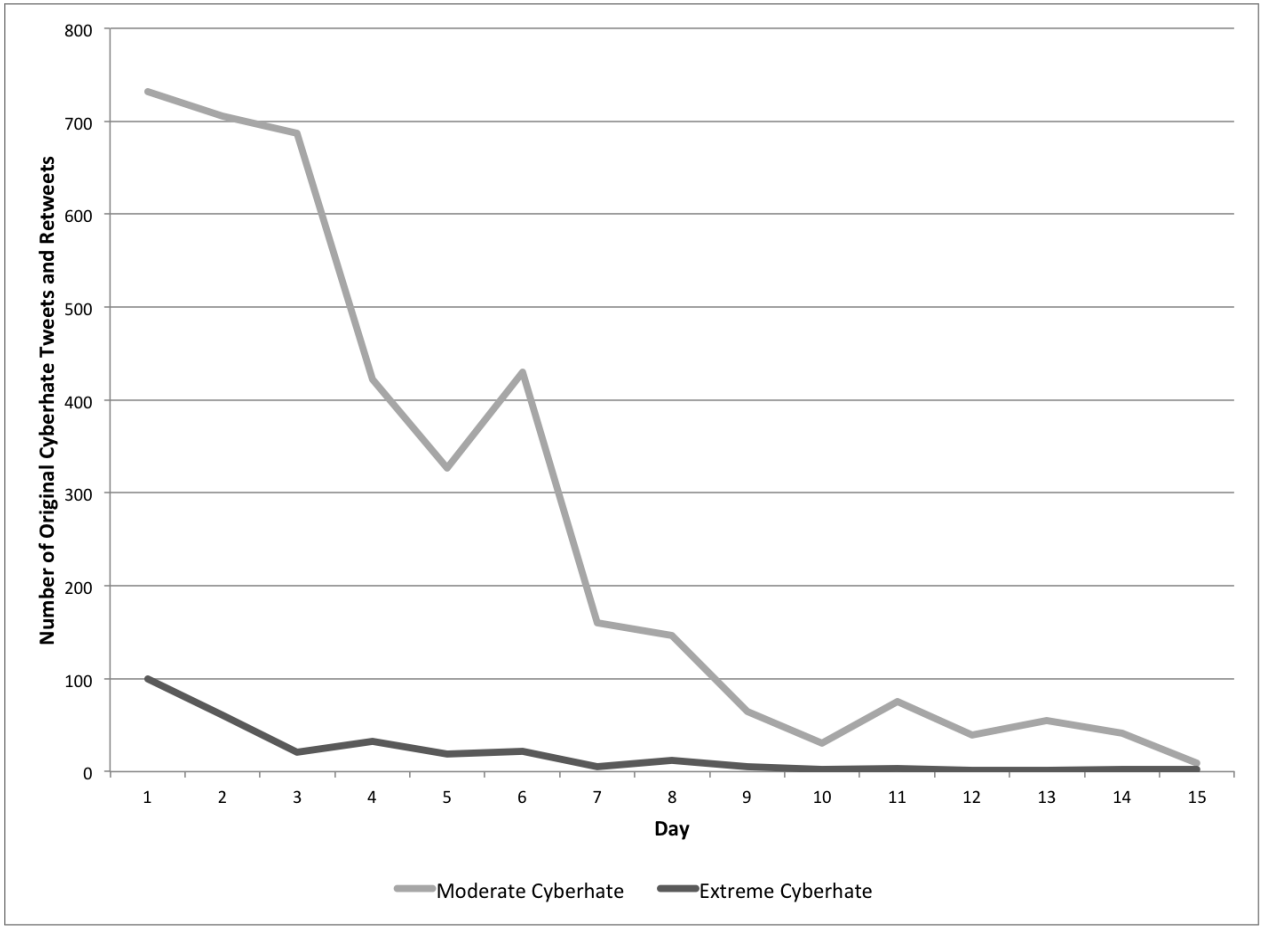
"during which those exposed to the disaster begin to form a preliminary picture of what has happened and of their own condition"

Social Media Reaction: Reaction



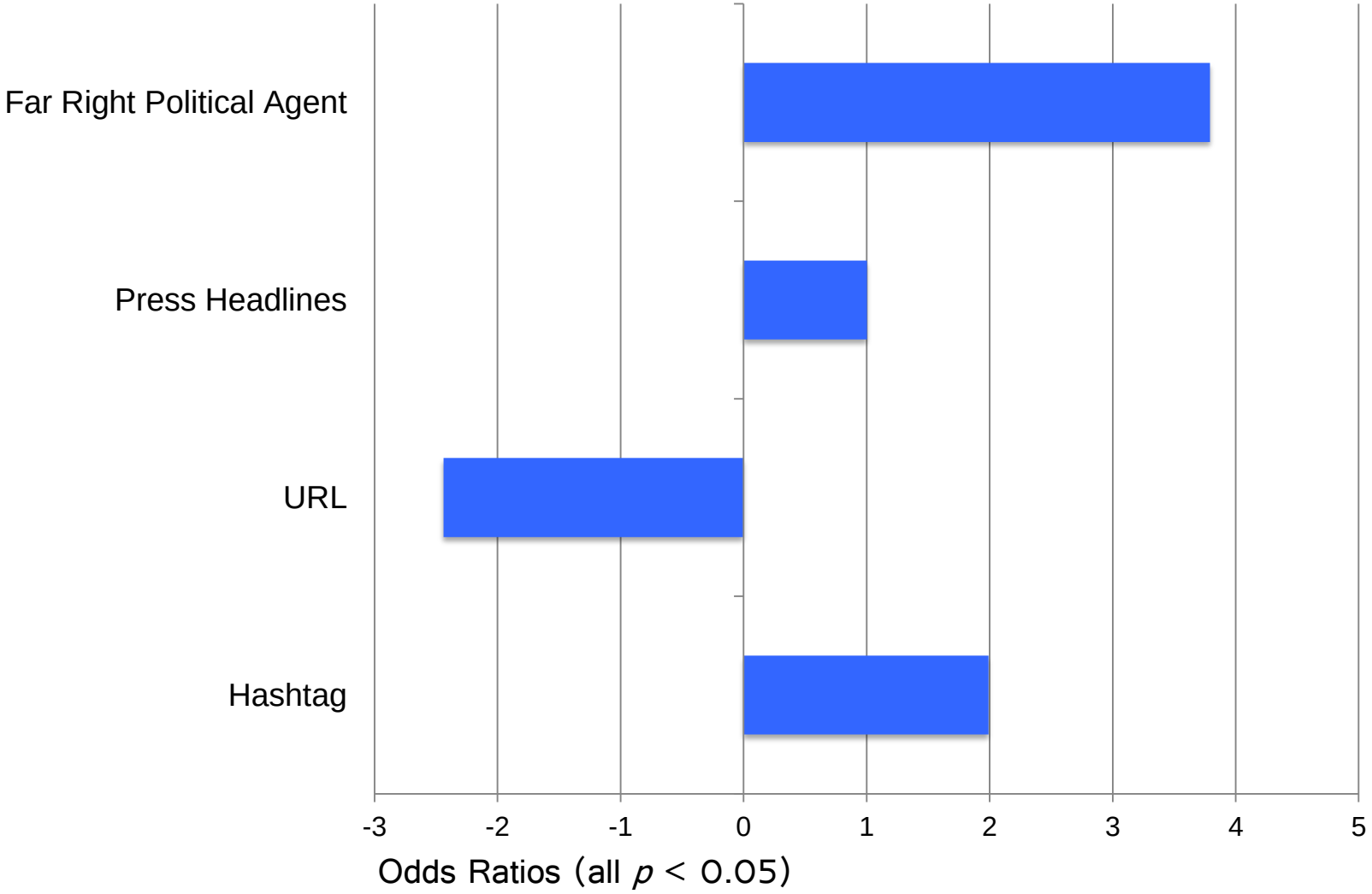
“images in the inventory were crystallized into more organised opinions and attitudes”

Frequency of Cyberhate

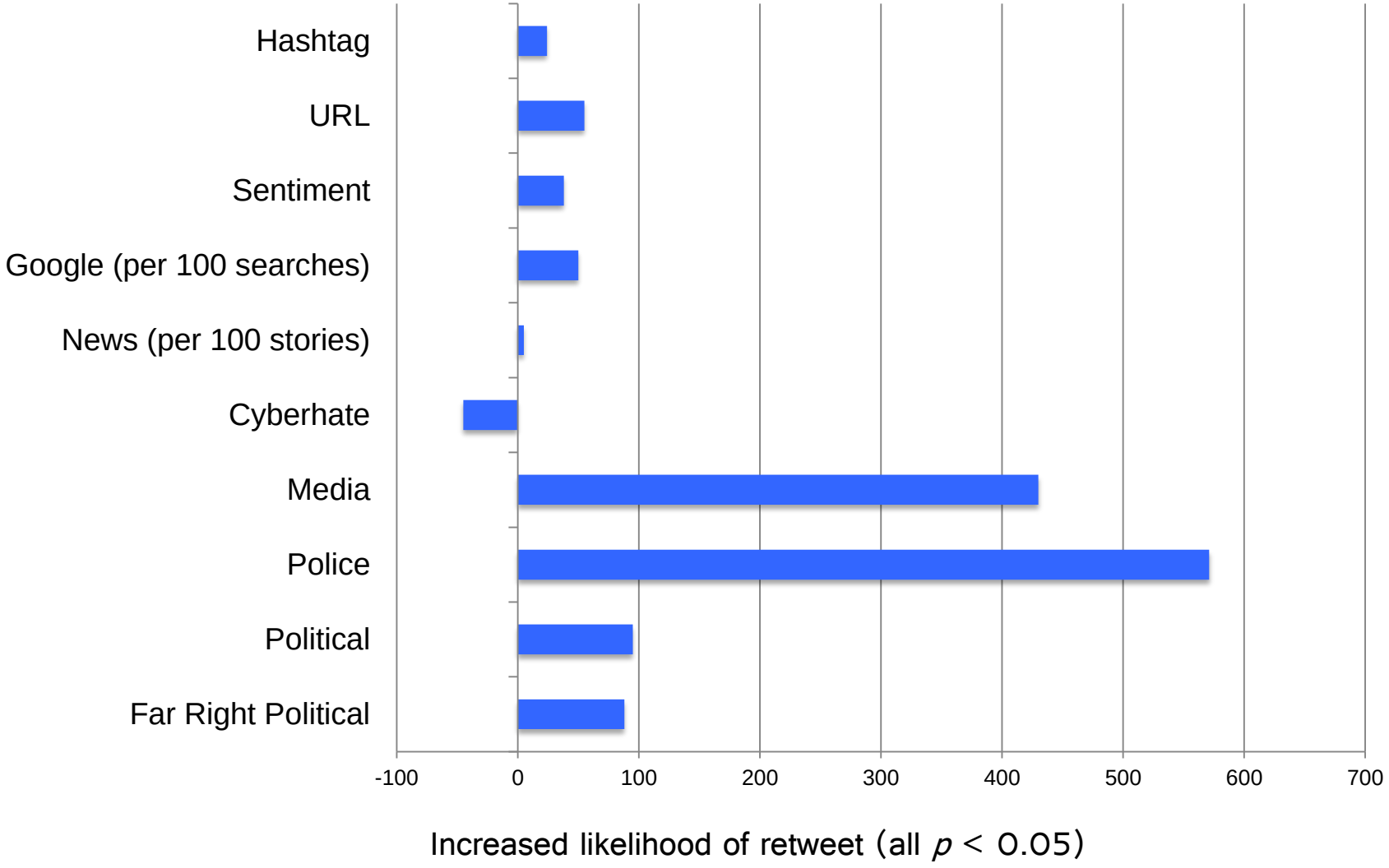


SOCIAL DATA SCIENCE LAB.

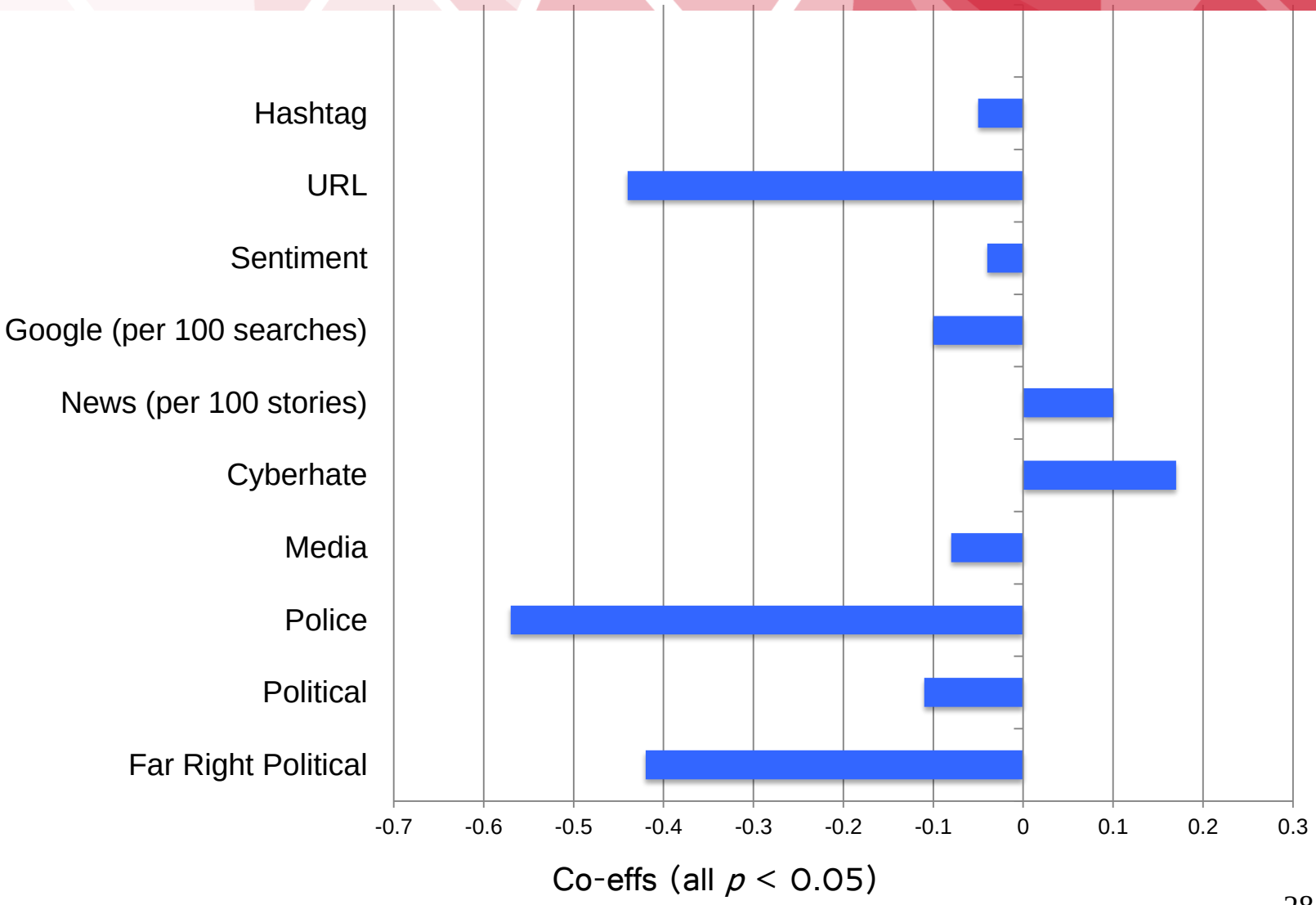
Production of of Cyberhate



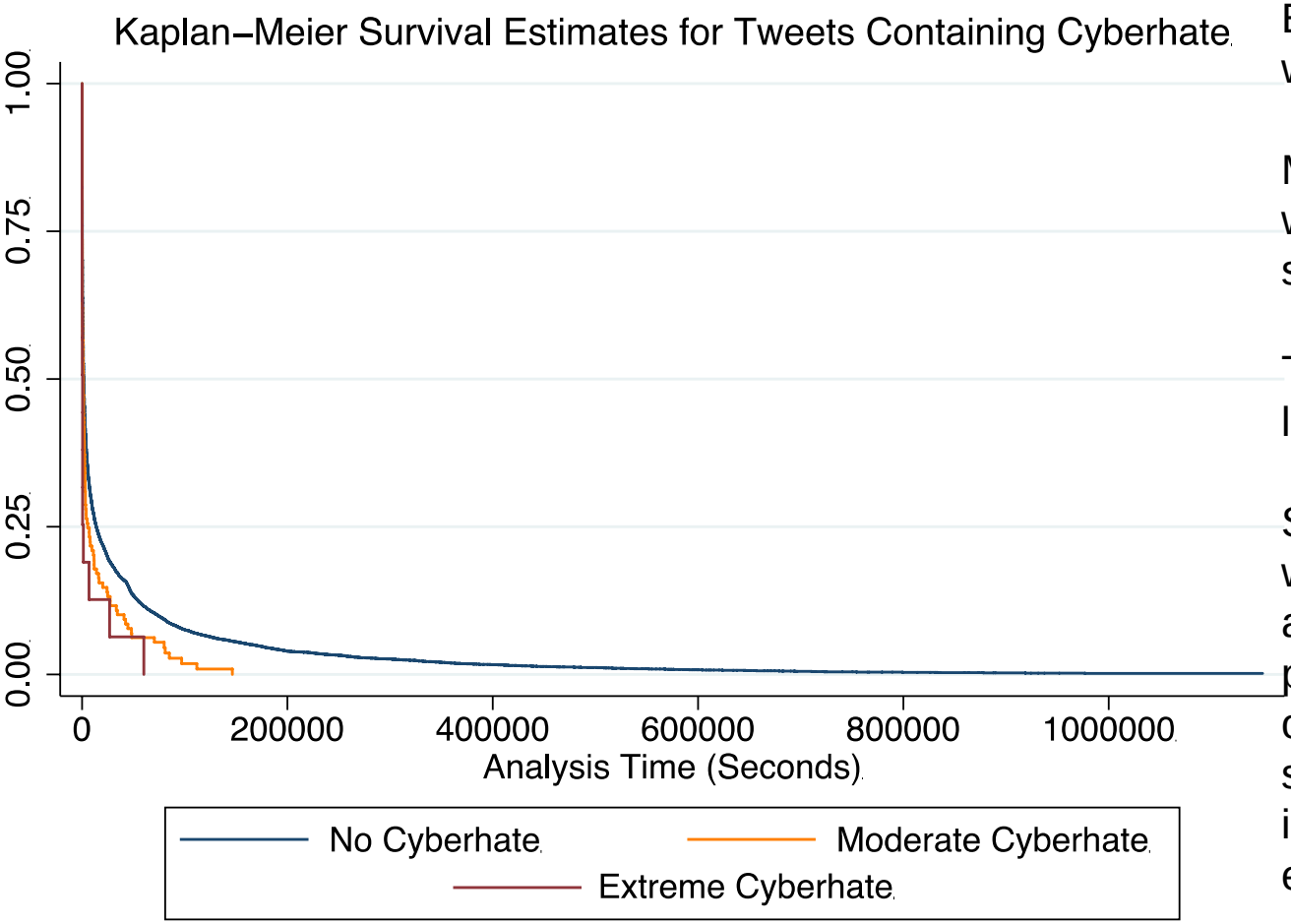
Volume of Cyberhate



Survival of Cyberhate



Survival of Cyberhate



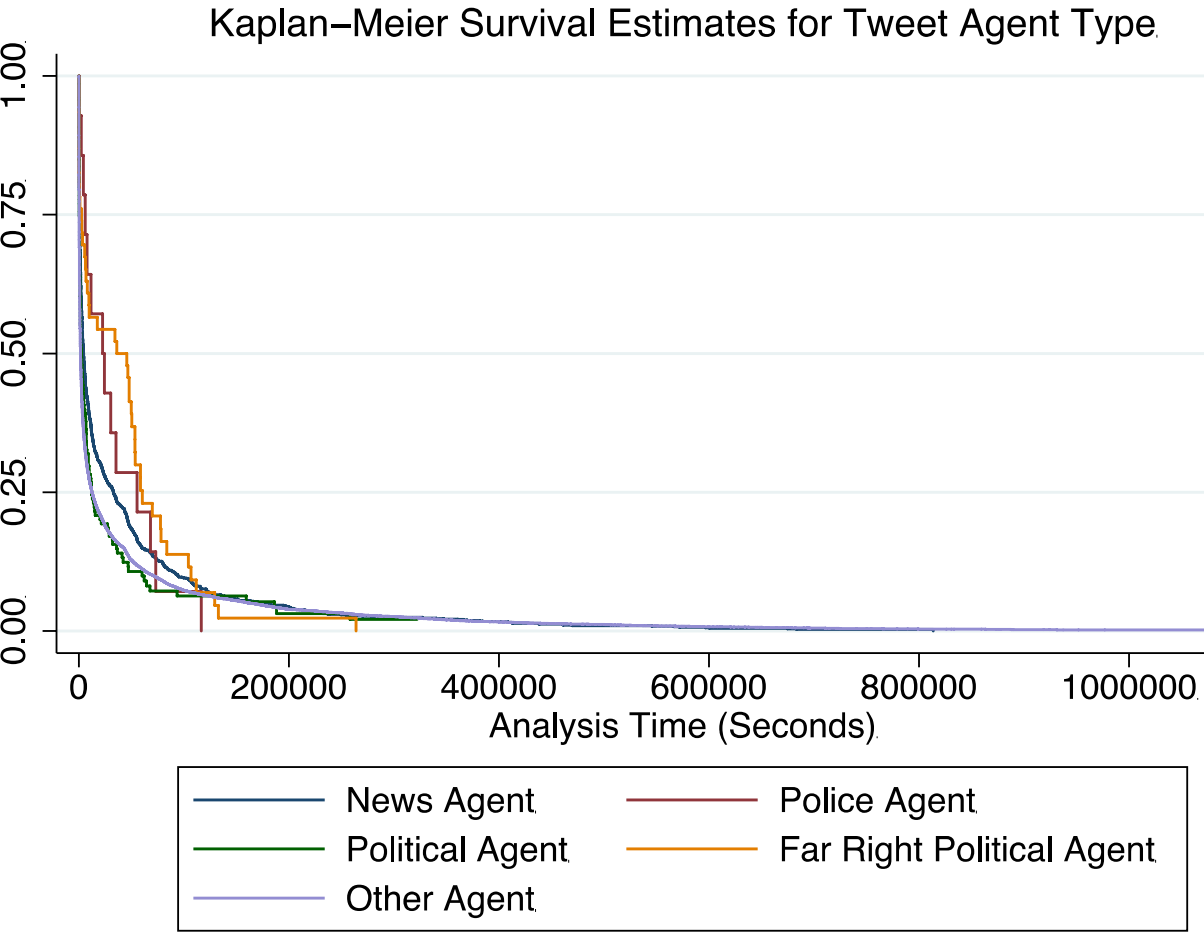
Extreme Cyberhate dies out within 20-24 hrs (impact stage)

Moderate Cyberhate dies out within 36-42 hrs (inventory stage)

Tweets without Cyberhate last longest (into reaction stage)

Sharp de-escalation resonates with the work of Legewie (2013) and King and Sutton (2014) who postulate that the increase in offline anti-immigration sentiment and hate crimes and incidents following terrorist events have a half-life

Survival of Cyberhate



Far Right outlast all other agents up to 36-42 hours (impact/inventory stage)

Police outlast all other agents but the Far Right in the 36-42 hour window (impact/inventory stage)

Both lose ground to Political Agents, News Agents who last longest

Information flows emanating from police were most likely to be large and to be long-lasting (bar the Far Right) in the impact and inventory stages

Information flows emanating from Far Right Political Agents were likely to be small in size, but the most long lasting in the

Demographics

- Social scientists are interested in group differences (gender, age, ethnicity etc)
- Comparative method (groups relative to each other)… but how to identify these groups on social media?
- User generated content can be ‘data light’ (Mislove et al. 2011, Gayo-Avello 2012)
- Facebook is different because it stores baseline demographic information (Schwartz et al. 2013)
- Twitter has signatures, but nothing systematic (Edwards et al. 2013)
- When the data is not available we develop proxies, so why not for Twitter?

SOCIAL DATA SCIENCE LAB.

Demographics

- What insights do demographic proxies offer for behaviour on Twitter?
- Does Twitter behaviour differ by demographic groups?
- Do real-world demographic differences manifest in the virtual world?

SOCIAL DATA SCIENCE LAB.

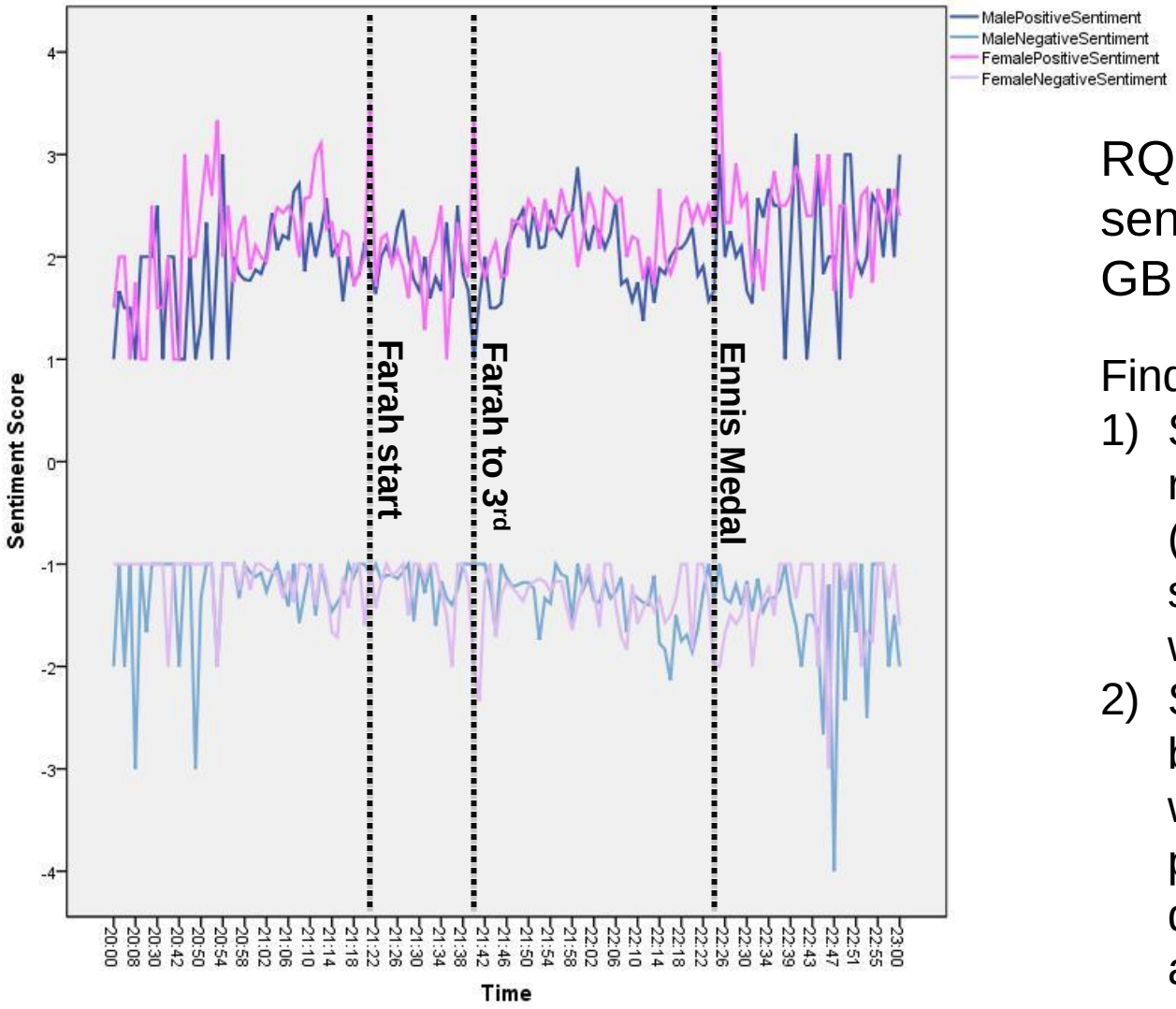
Demographics: Gender

- Use the name field of the Twitter profile
- Clean the data to extract a first name and compare against a large database of first names
- Important to categorise ‘unisex’ and ‘unknown’
- Of those we could identify: 48.8% male and 51.2% female… exactly the same as the 2011 Census

SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online 18(3), article number: 7. (10.5153/sro.3001)

Demographics: Gender



RQ: How does sentiment towards Team GB differ by gender?

Findings:

1) Sentiment peaks reflect real world events (relationship between social media and real world)
2) Sentiment differs between men and women (difference is so pronounced that gender detection method appears to work)

Demographics: Location

- Three primary sources of location:
  - User profile information
  - Content of tweets (inc. ‘mundane geography’)
  - Geo-tagged tweets

- Geo-tagged tweets are the gold standard

- Allows us to locate people at the time they tweeted in existing geographies (output area level!)

- RQ: do people tweet about crime in high crime areas?

SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online 18(3), article number: 7. (10.5153/sro.3001)

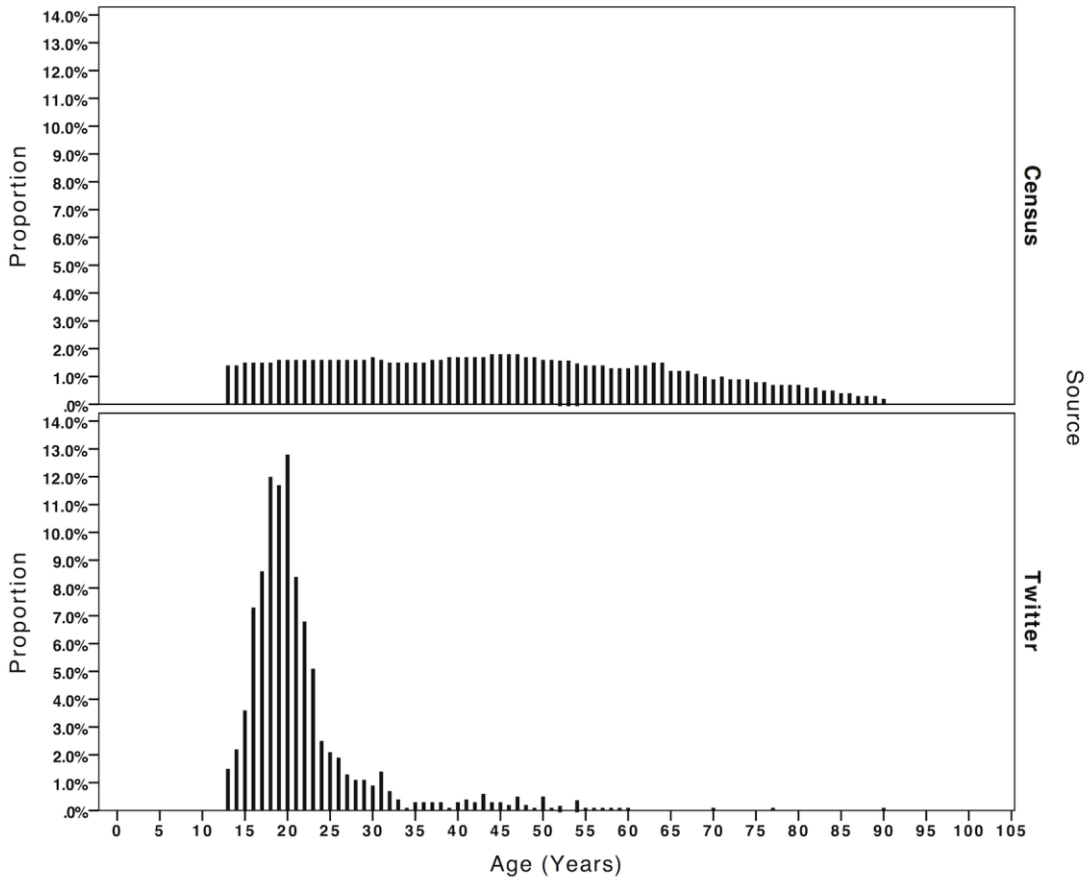
Demographics: Location

SOCIAL DATA SCIENCE LAB.

Source: Sloan et al. 2013

Demographics: Age

- Identifying age from signature data
- Preliminary analysis suggests usable age data for 0.35% of Twitter users
- Note that 0.35% of 645m is 2.25m (approx 40% of which is English language)

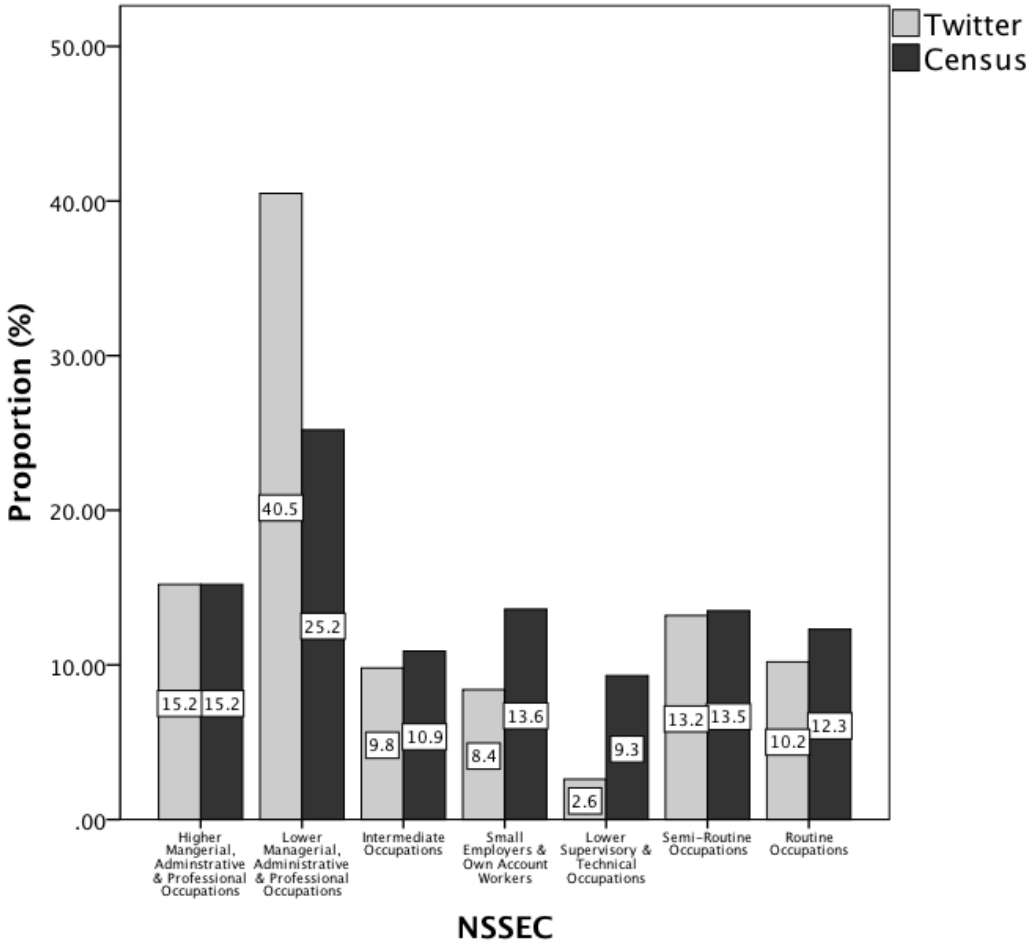


SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

Demographics: Occupation

- Identify occupation from signature data
- Linked to SOC2010 codes
- Enables allocation into NS-SEC groups



SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

Case Study: Ebola

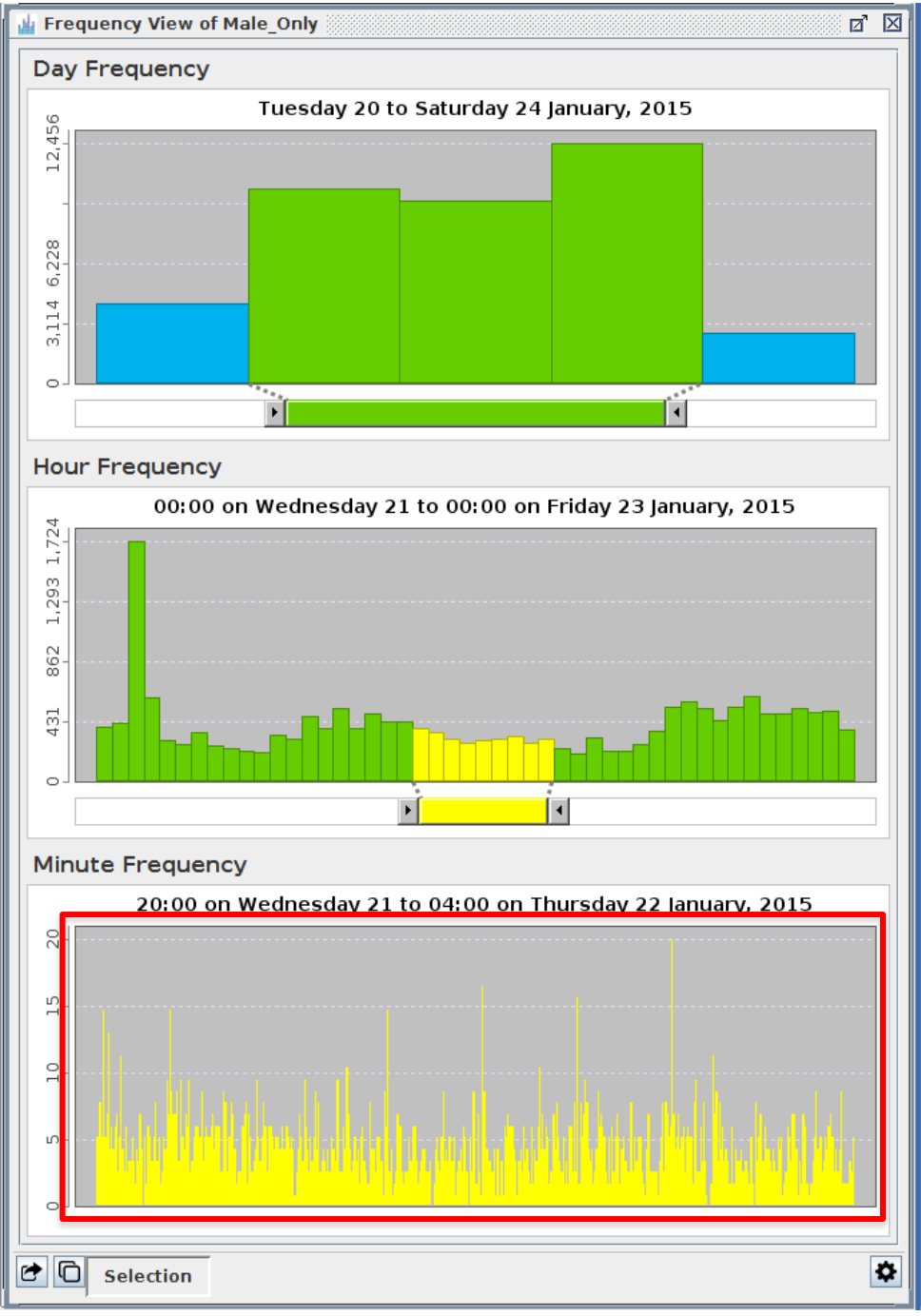
- Twitter data collected via COSMOS Desktop
- Tuesday 20th to Sat 24th Jan 2015
- Condition: contains “ebola”
- 182,517 tweets of which:
  - 39,037 made by male users
  - 31,244 made by female users
  - 1,715 with geo-tagging enabled (0.94%)

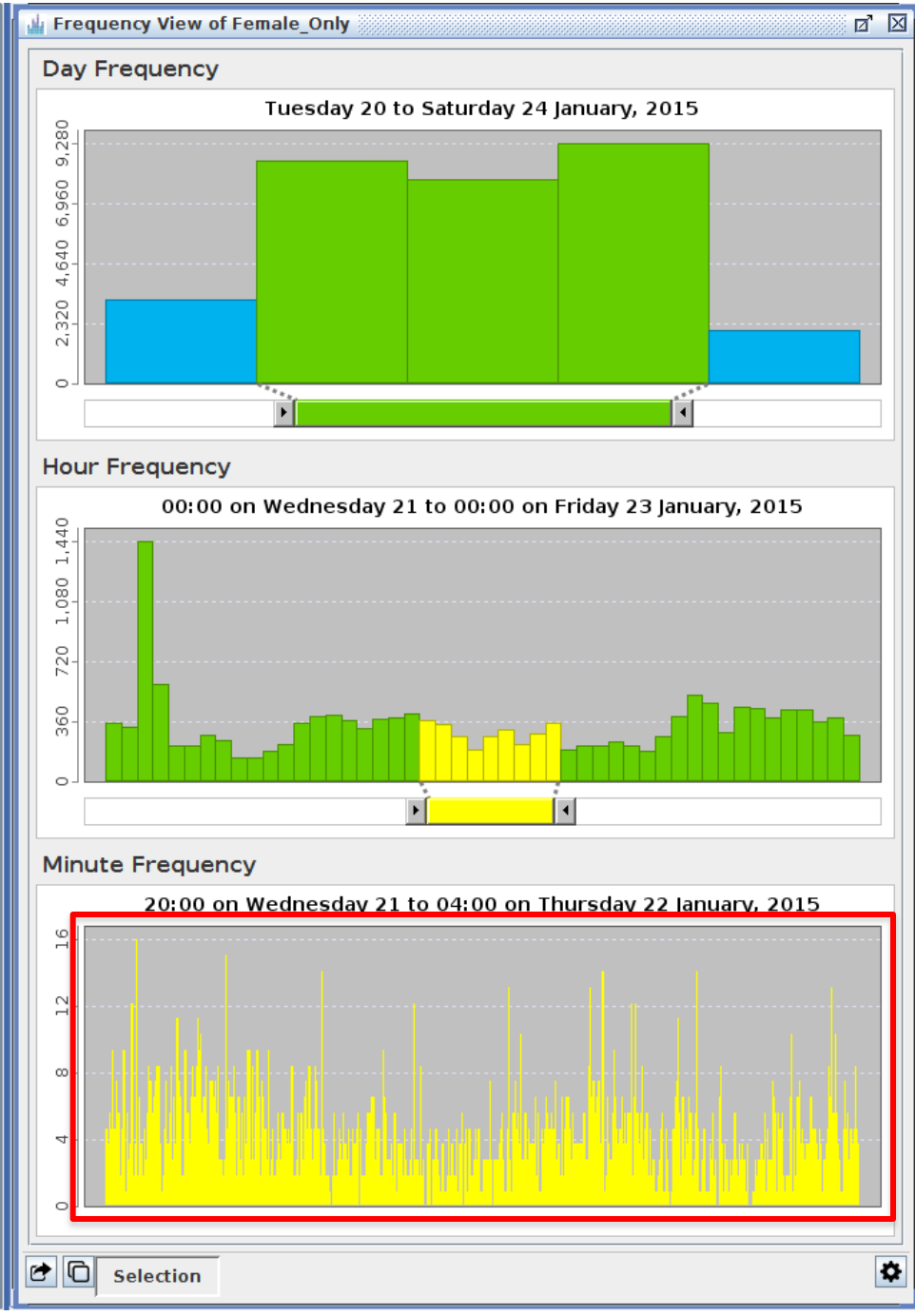
SOCIAL DATA SCIENCE LAB.

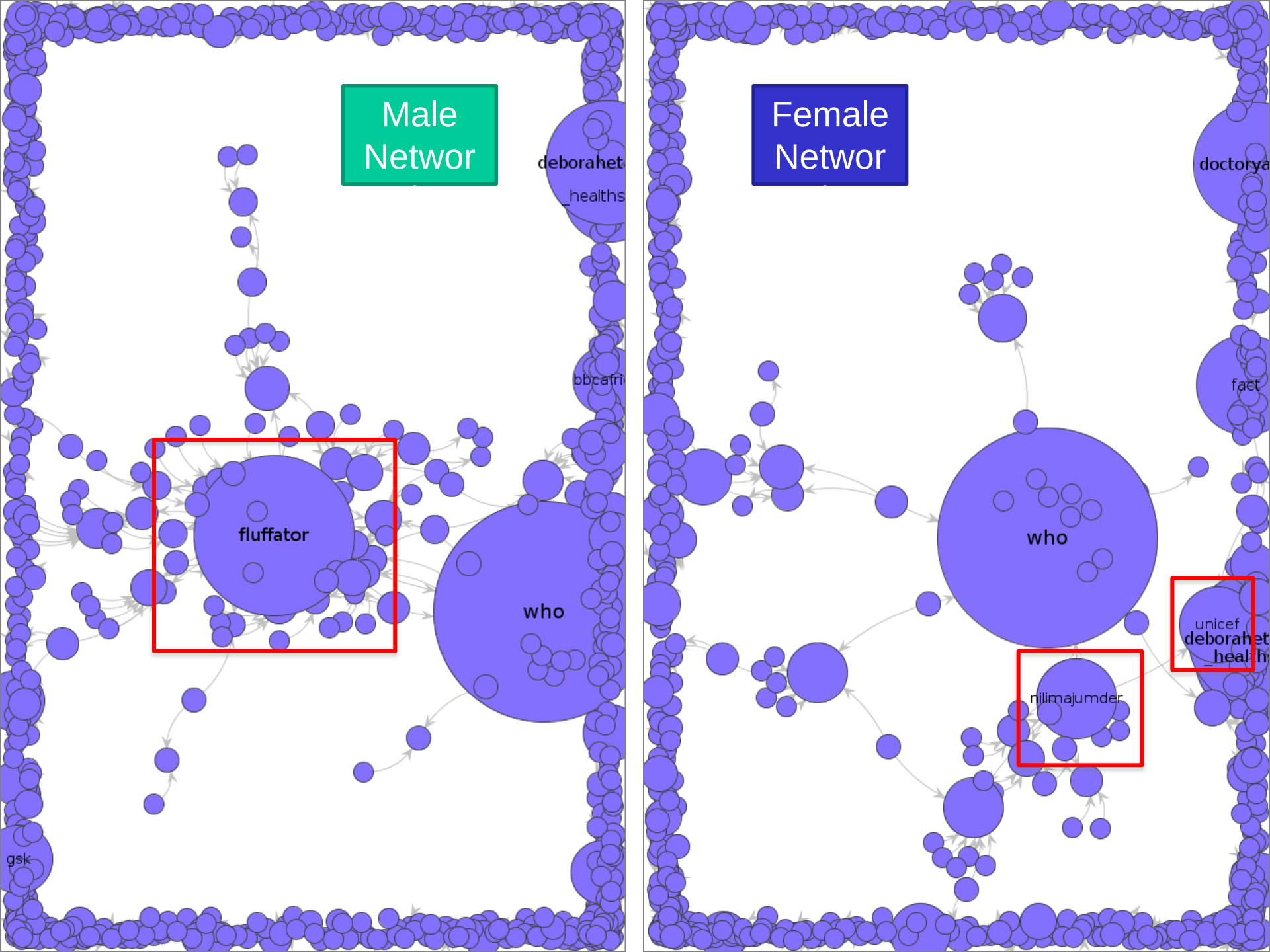
Case Study: Ebola

Gender Differences

SOCIAL DATA SCIENCE LAB.


Frequency View of Male_Only
Day Frequency
Tuesday 20 to Saturday 24 January, 2015
0
3,114
6,228
12,456
Hour Frequency
00:00 on Wednesday 21 to 00:00 on Friday 23 January, 2015
0
431
862
1,293
1,724
Minute Frequency
20:00 on Wednesday 21 to 04:00 on Thursday 22 January, 2015
0
5
10
15
20
Selection


Frequency View of Female_Only
Day Frequency
Tuesday 20 to Saturday 24 January, 2015
0
2,320
4,640
6,960
9,280
Hour Frequency
00:00 on Wednesday 21 to 00:00 on Friday 23 January, 2015
0
360
720
1,080
1,440
Minute Frequency
20:00 on Wednesday 21 to 04:00 on Thursday 22 January, 2015
0
4
8
12
16
Selection


Male
Networ
deborahet
_healths
bbcafri
fluffator
who
gsk
Female
Networ
doctorya
fact
who
unicef
deborahet
_heal
nilimajumder

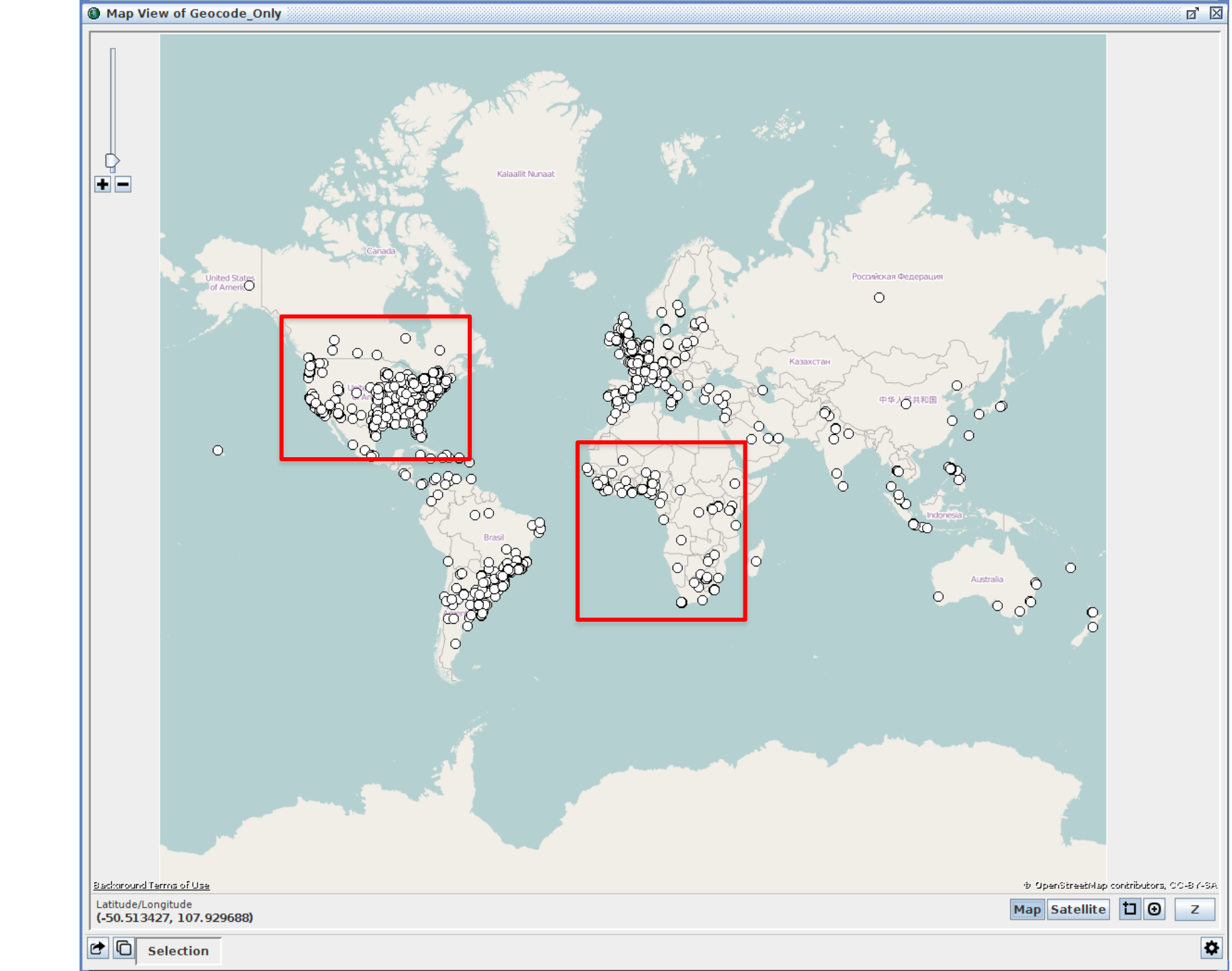

Word Cloud View of Male_Only
Tweet Text
#sotu @who africa against
arrive back batch better cases crisis end
epidemic experimental fight first
gone good gorillas health it's killed
leone liberia los
more new news now oms one out
outbreak over people que shit
sierra still third today trial
vaccine via
virus warns wash way west
world world's
Selection

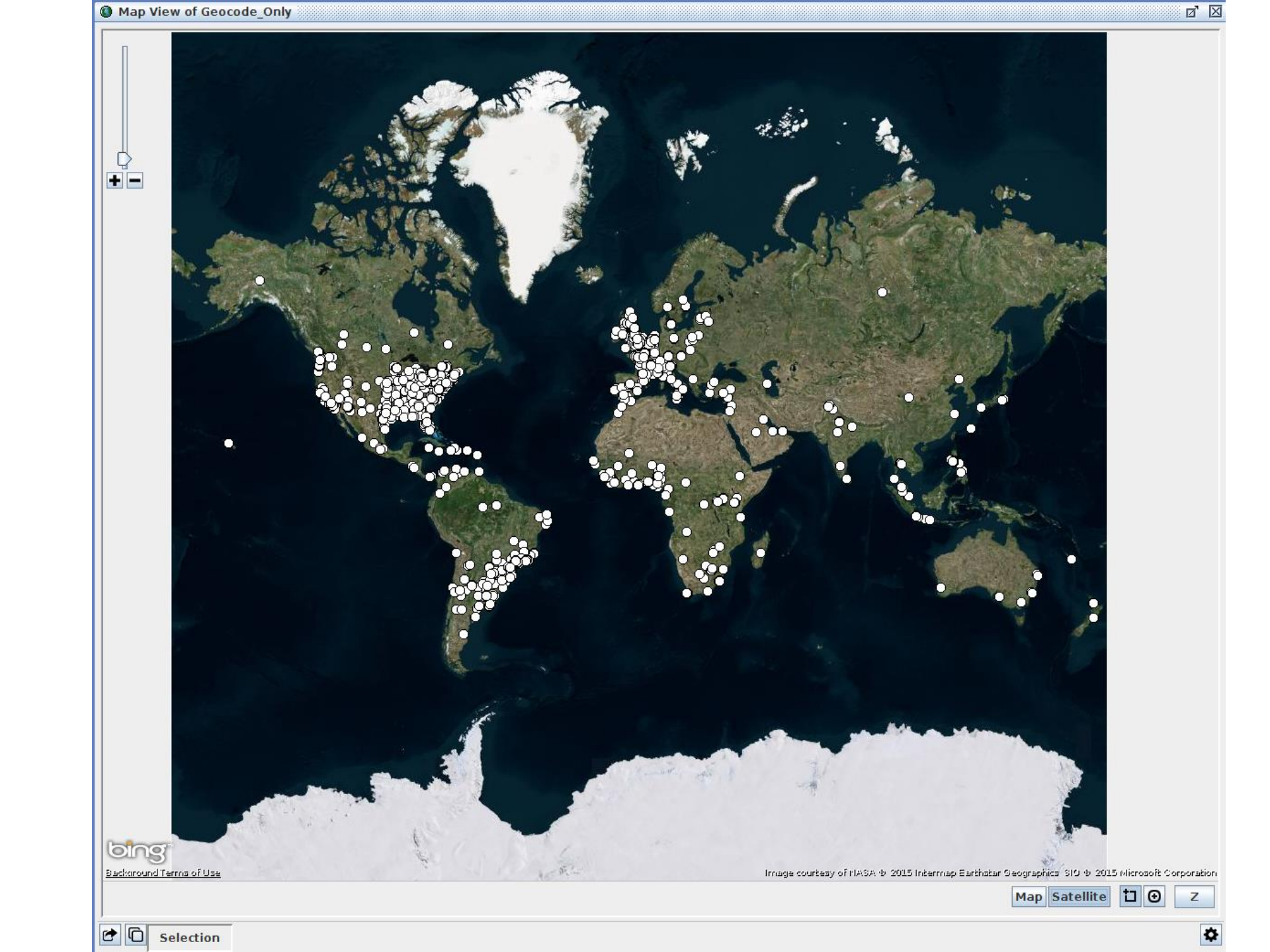

Word Cloud View of Female_Only
Tweet Text
#fightebola #sotu @who
africa against arrive back
better cases crisis cuba
experimental fight first gas
good gorillas health it's killed
leone liberia look
los more new news now obama one
out outbreak people que shit
sierra still think third today trial
vaccine via
virus wash way west
workers world world's
Selection

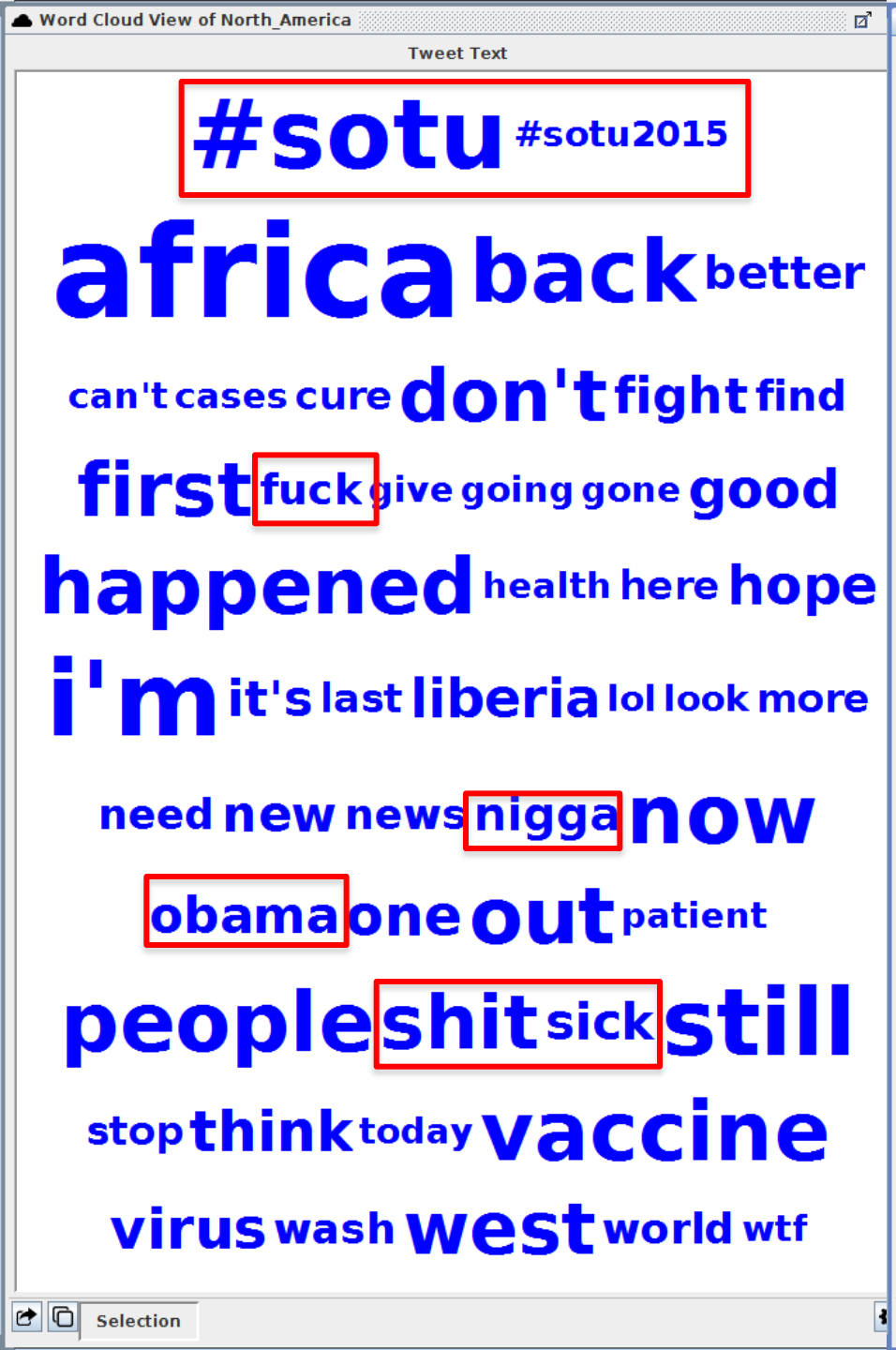
Case Study: Ebola

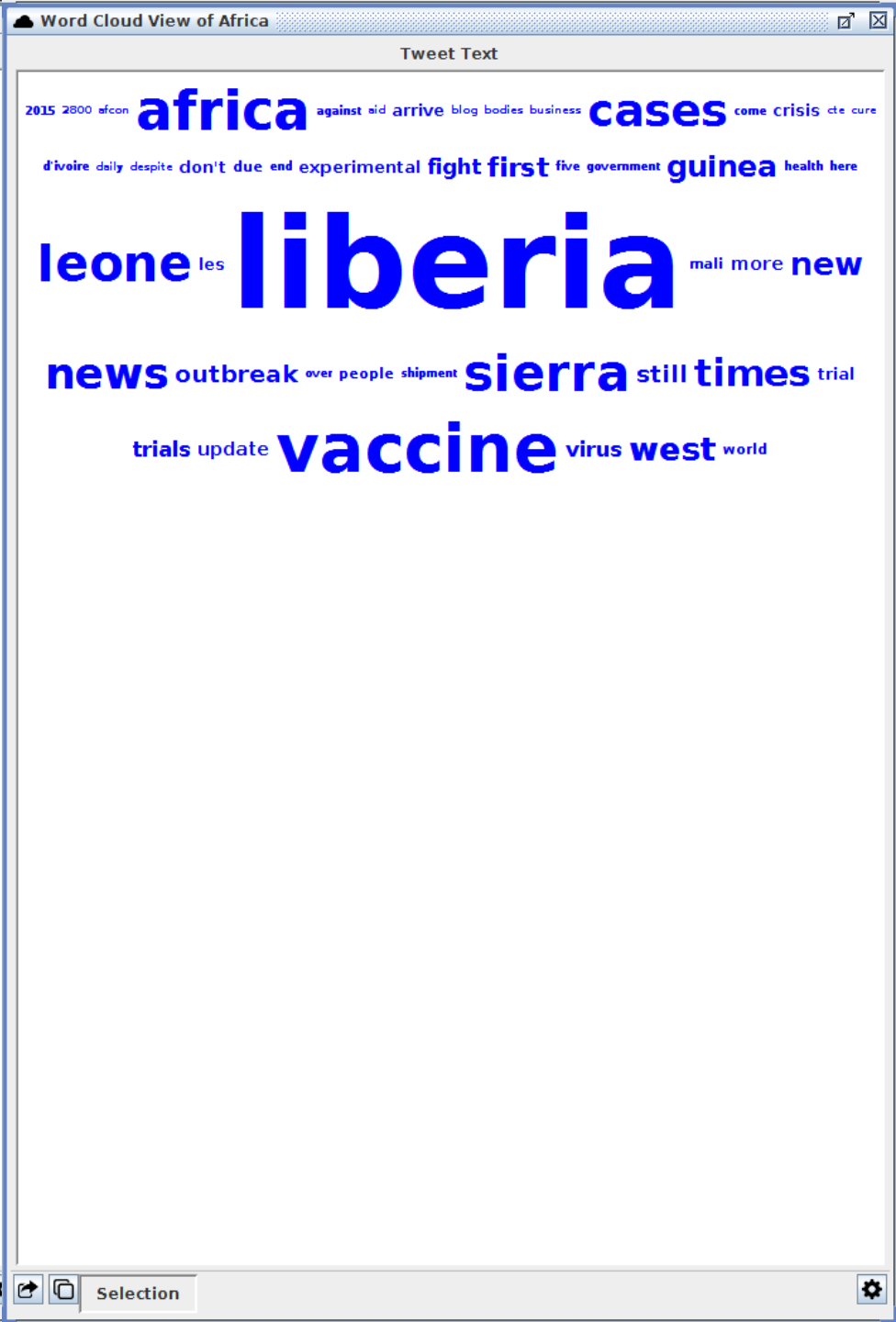
Geographical Differences

SOCIAL DATA SCIENCE LAB.


Map View of Geocode_Only
Kalaallit Nunaat
Canada
United States of America
Российская Федерация
Казахстан
中华人民共和国
Indonesia
Australia
Brasil
Background Terms of Use
© OpenStreetMap contributors, CC-BY-SA
Latitude/Longitude
(-50.513427, 107.929688)
Map
Satellite
Z
Selection


Map View of Geocode_Only
bing
Background Terms of Use
Image courtesy of NASA © 2015 Intermap Earthstar Geographics SIO © 2015 Microsoft Corporation
Map
Satellite
Z
Selection

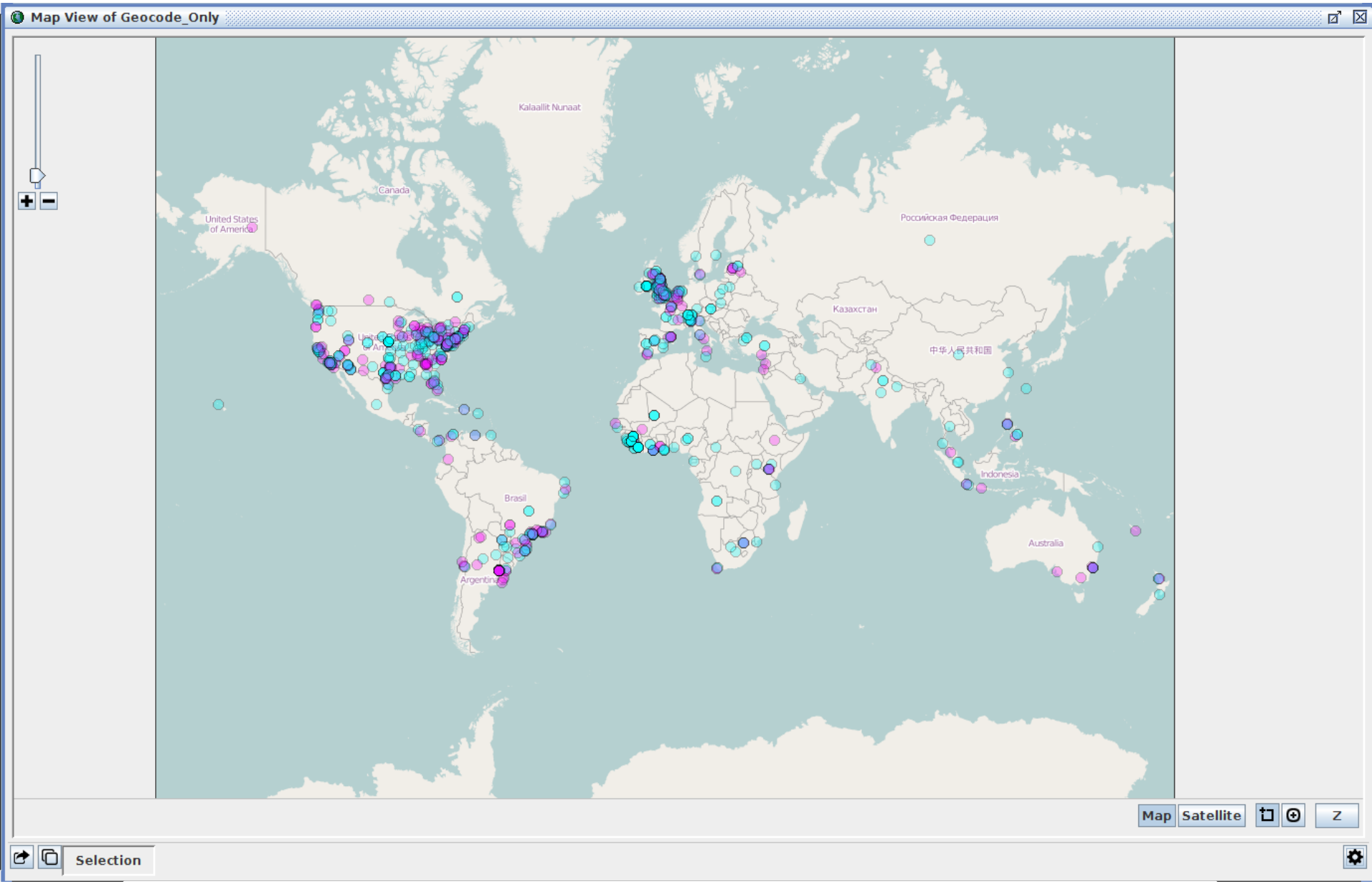

Word Cloud View of North_America
Tweet Text
#sotu #sotu2015
africa back better
can't cases cure don't fight find
first fuck give going gone good
happened health here hope
i'm it's last liberia lol look more
need new news nigga now
obama one out patient
people shit sick still
stop think today vaccine
virus wash west world wtf
Selection


Word Cloud View of Africa
Tweet Text
2015 2800 afcon africa against aid arrive blog bodies business cases come crisis cte cure
d'ivoire daily despite don't due end experimental fight first five government guinea health here
leone les liberia mali more new
news outbreak over people shipment sierra still times trial
trials update vaccine virus west world
Selection

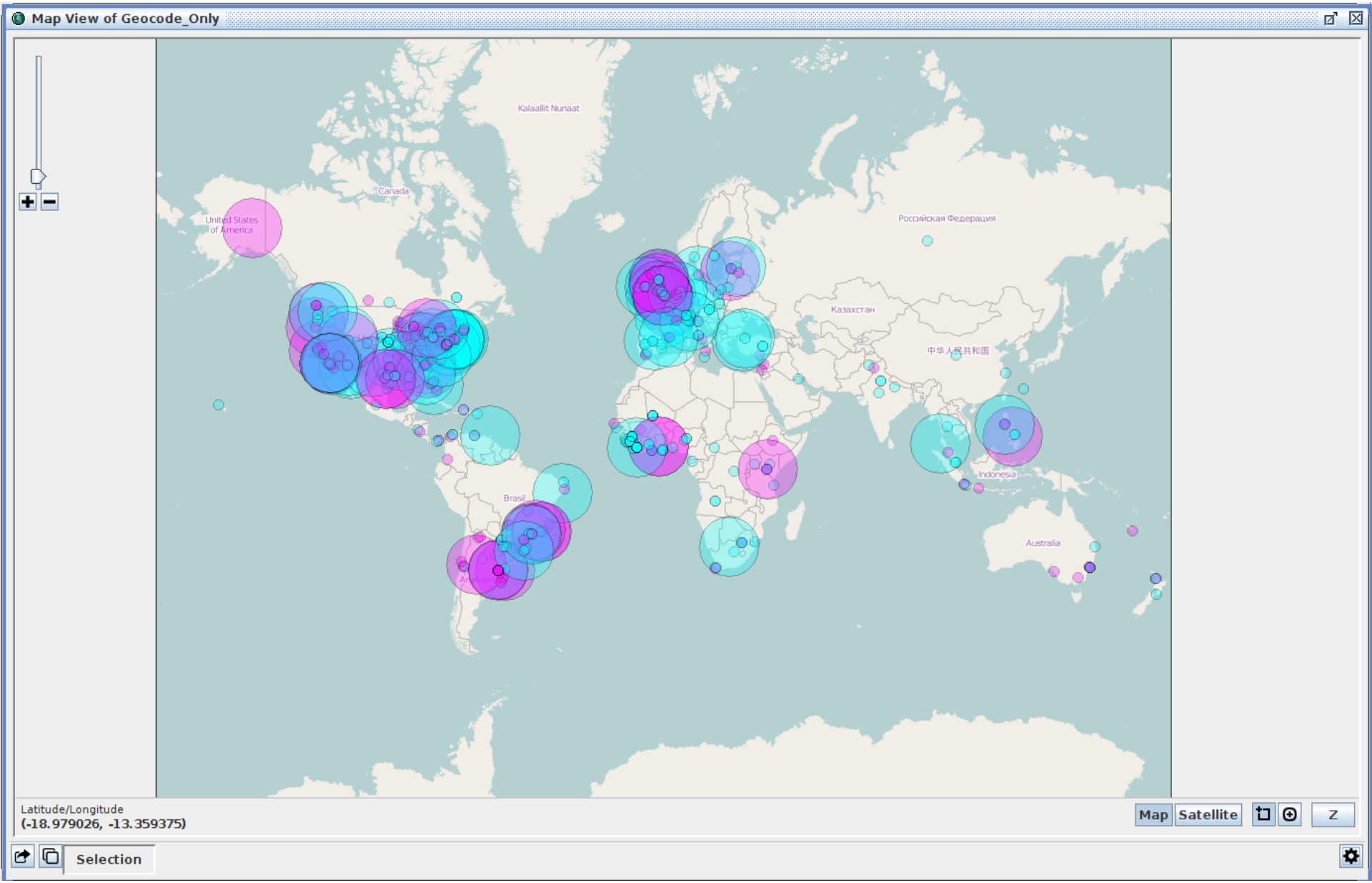
Case Study: Ebola

Intersectionality:
Gender, Geography & Sentiment (+ive)

SOCIAL DATA SCIENCE LAB.



Pink = female tweeters Blue = male tweeters



Pink = female tweeters Blue = male tweeters +ive sentiment = larger

Ebola: Summary

- Demographic characteristics seem to impact upon online behaviour (words & networks)
- All of this analysis can be done on COSMOS Desktop
- Scoping the issues, focus on more in-depth analysis
- Potential for social media analytics to provide real-time information on ebola
- Understanding demographic difference in networks and information flows enables intelligent interventions (see Sloan et al. 2014 for a food industry example)

SOCIAL DATA SCIENCE LAB.

References

Burnap, P. and Williams, M. (2015) 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', *Policy & Internet* (7:2)

Burnap, P., Williams, M.L. et al. (2014), 'Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack', *Social Network Analysis and Mining* (4:2)

Edwards et al. (2013) Computational social science and methodological innovation: surrogacy, augmentation or reorientation?, *International Journal of Social Research Methods*, 16:3

Gayo-Avello (2012) I wanted to Predict Elections with Twitter and all I got was this Lousy Paper: A Balanced Survey on Election Prediction using Twitter Data, *Department of Computer Science*, University of Oviedo Spain

Mislove et al. (2011) Understanding the demographics of Twitter users, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*

Schwartz et al. (201) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach, *PLOS ONE*, 8:9 (DOI: 10.1371/journal.pone.0073791)

Sloan et al. (2013) Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter, *Sociological Research Online*, 18:3 (http://www.socresonline.org.uk/18/3/7.html)

Sloan et al. (2014) Going Viral in Social Media – Networks and Intercepted Misinformation, *Software Sustainability Institute*, Cardiff University

Sloan et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

Williams, M. L. and Burnap, P. (2015) 'Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*

SOCIAL DATA SCIENCE LAB.

web: socialdatalab.net

Questions?

@MattLWilliams @DrLukeSloan

@socdatalab

socialdatalab.net

SOCIAL DATA SCIENCE LAB.

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD