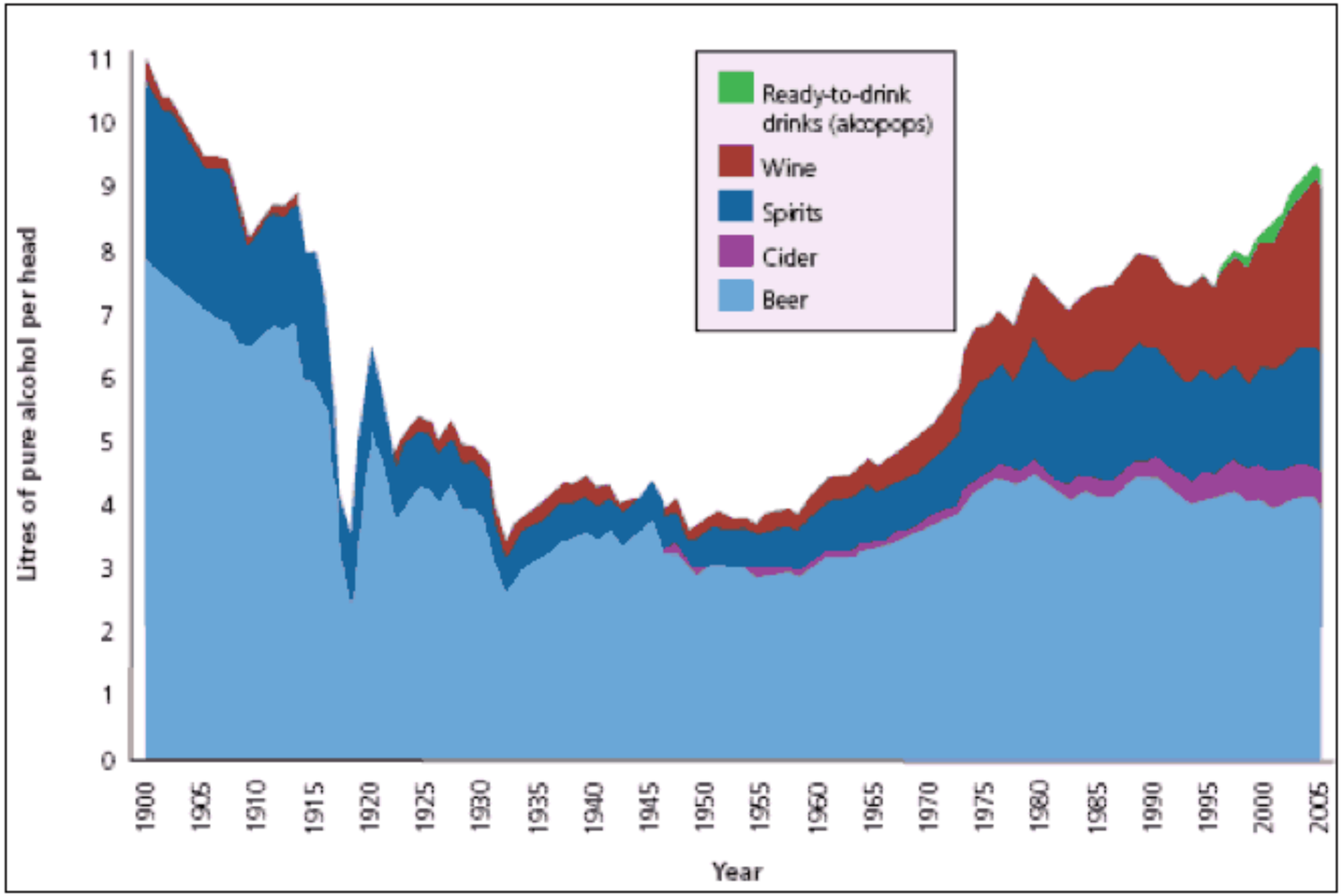
Monitoring Regional Alcohol Consumption through Social Media

Daniel Kershaw
HighWire DTC

@danjamker

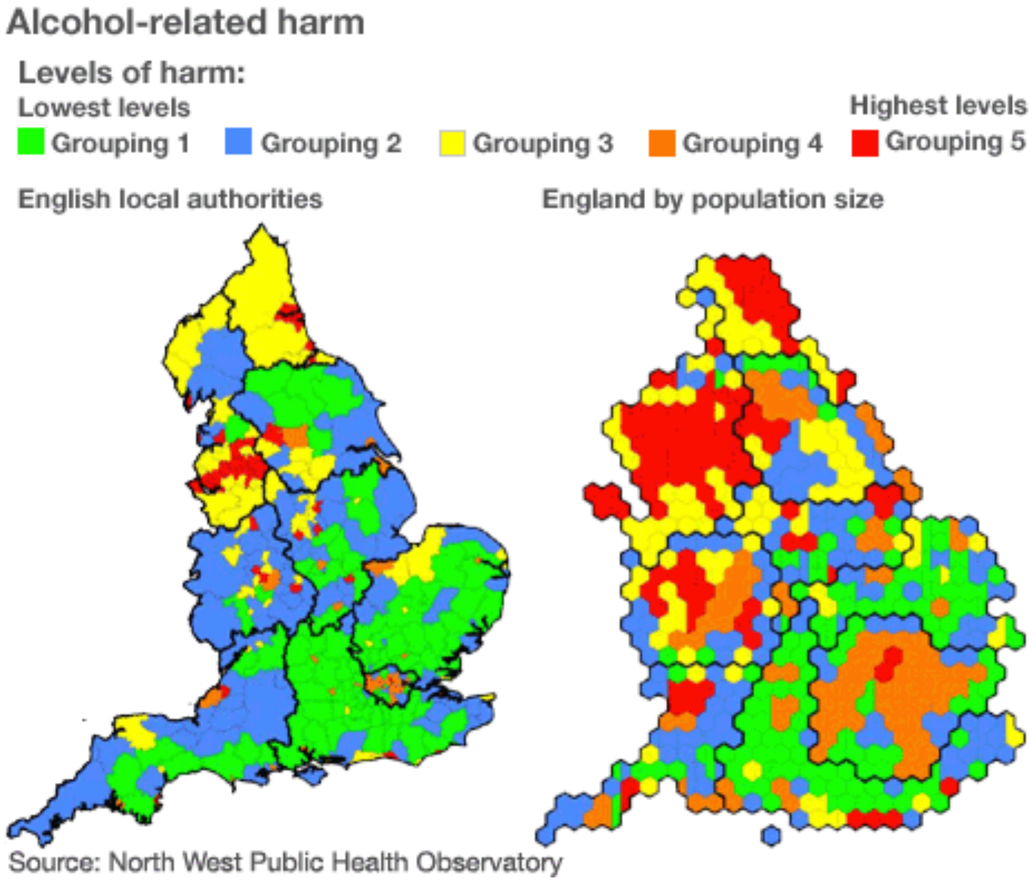
People like to drink

Alcohol Consumption form 1950's



source: scotland.gov.uk

Varying rates of harm



source: BBC News

Collecting Current Data

- Quantity Frequency Questionnaires (QF)
- Time Line Method (TL)
- Both talk a while to perform
- Expensive
- Data is only a snapshot of the past

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002)
Babor, T. F., Stephens, R. S., & Marlatt, G. A. (1987)

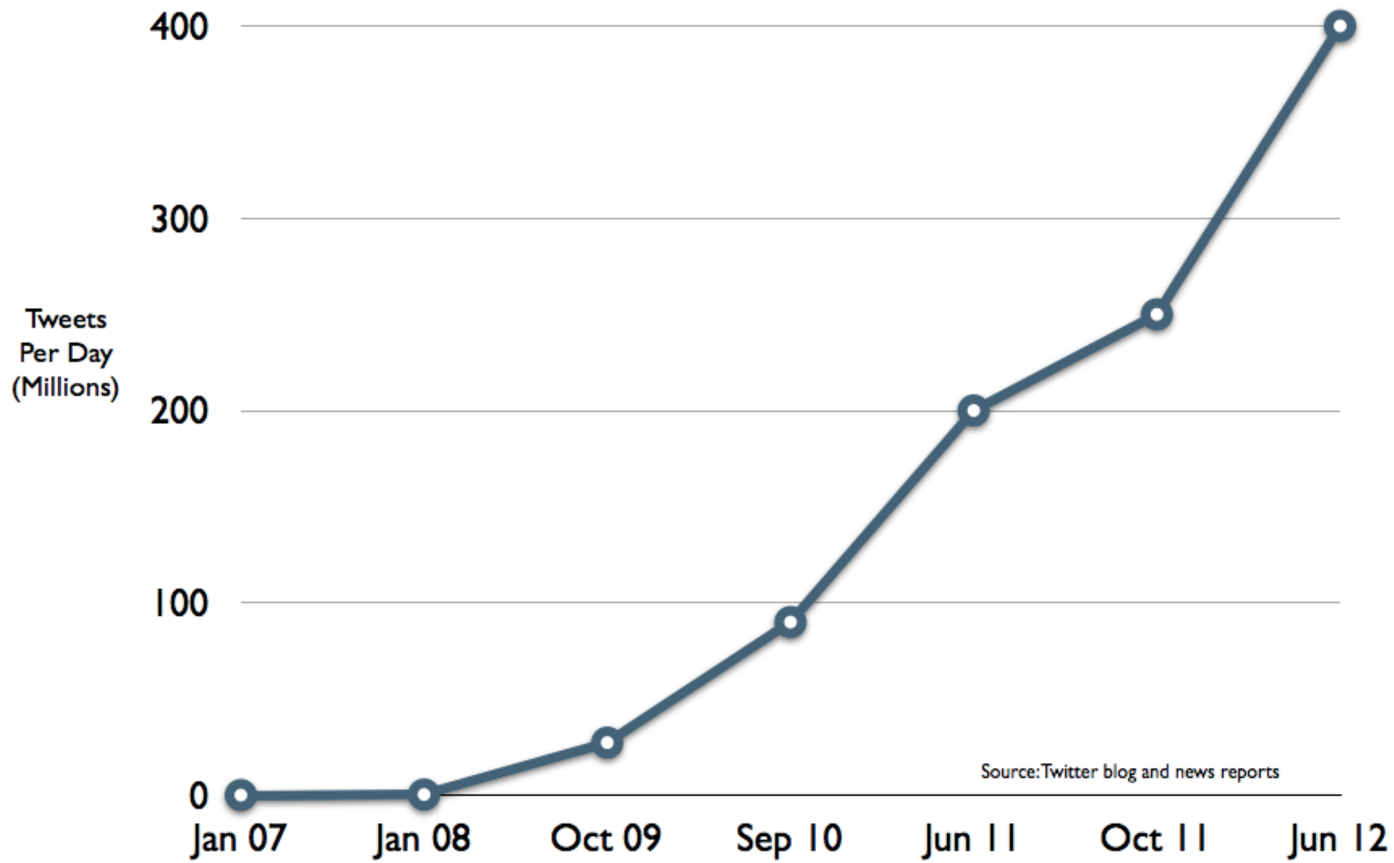
Data Collection Errors

- Selective reporting
- Recall bias
- Accidental under-estimation
- Study level under estimation by up to 40%

Boniface, S., & Shelton, N. (2013)
Del Boca, F. K., & Darkes, J. (2003)

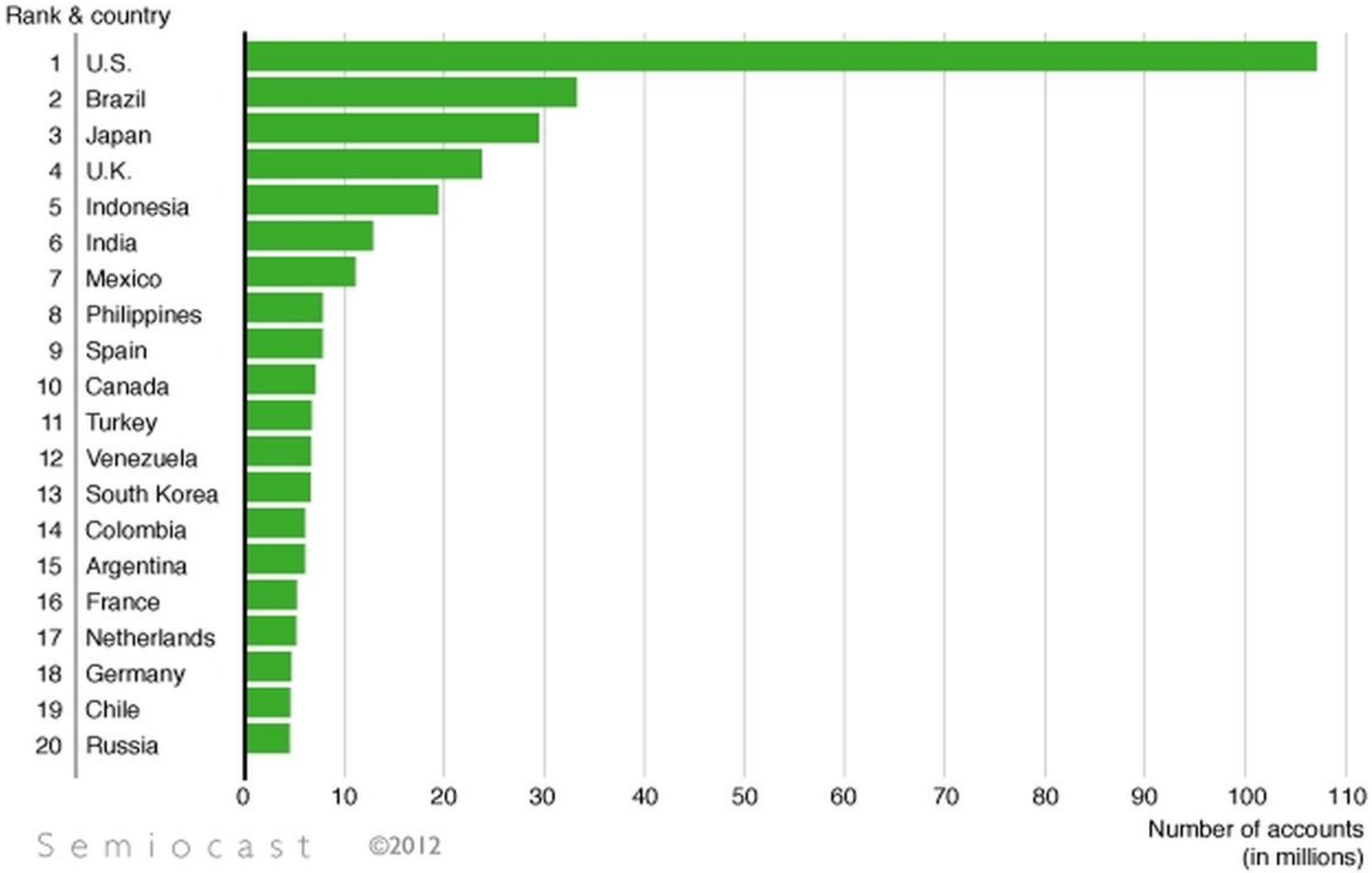
People like to Tweet

Growth of Twitter



source: forbes.com

Twitter users by county



source: indoboom.com

Reasons people use Twitter

- Minimal Effort
- Mobile and pervasive
- People-based RSS feeds
- Broadcast Nature of Twitter
- Keeping in touch with friends and family / Raising visibility
- Gathering information / Seeking help / Releasing emotional stress

Twitter as a Spatio-Temporal Sense Network

Do people Tweet when they are drinking Alcohol?

Do people Tweet that they are drinking Alcohol?

Is it possible to track drinking habits on Twitter?

Research Question

Is it possible to characterise and model UK alcohol consumption patterns of alcohol on social media data such as Twitter, and if so is there a variation across geographical location in drinking patterns and terminology usage?

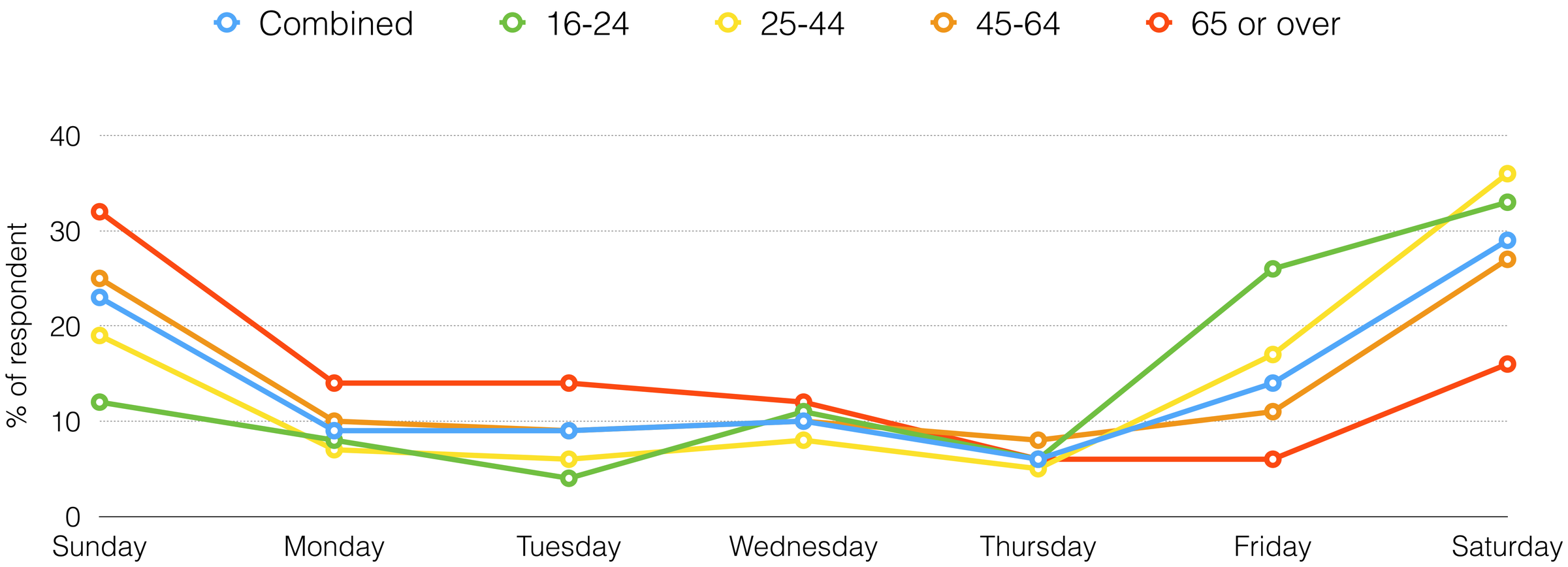
Previous Work

- Monitoring Flu spreading though twitter - Culotta, A. (2010)
- Social media to track depression on a global scale - De Choudhury, M., Counts, S., & Horvitz, E. (2013)
- Stock market prediction through sentiment analysis - Bollen, Mao, Zeng. (2011)
- Detecting earthquakes through peoples tweets - Sakaki, T., Okazaki, M., & Matsuo, Y. (2010)

Finding the Grounded Truth

- Health and Social Care Information Centre (HSCIC)
- Statistics on Alcohol Report
- Looking for daily granularity

Grounded Truth

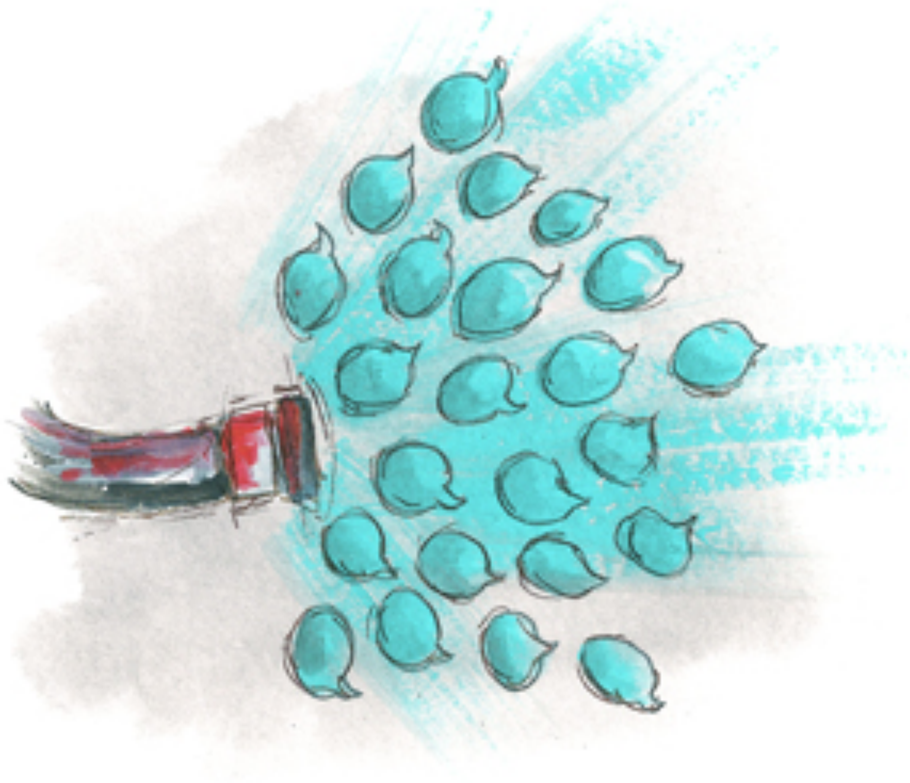


Health and Social Care Information Centre, L. S. (2012, May 31)

I need Tweets

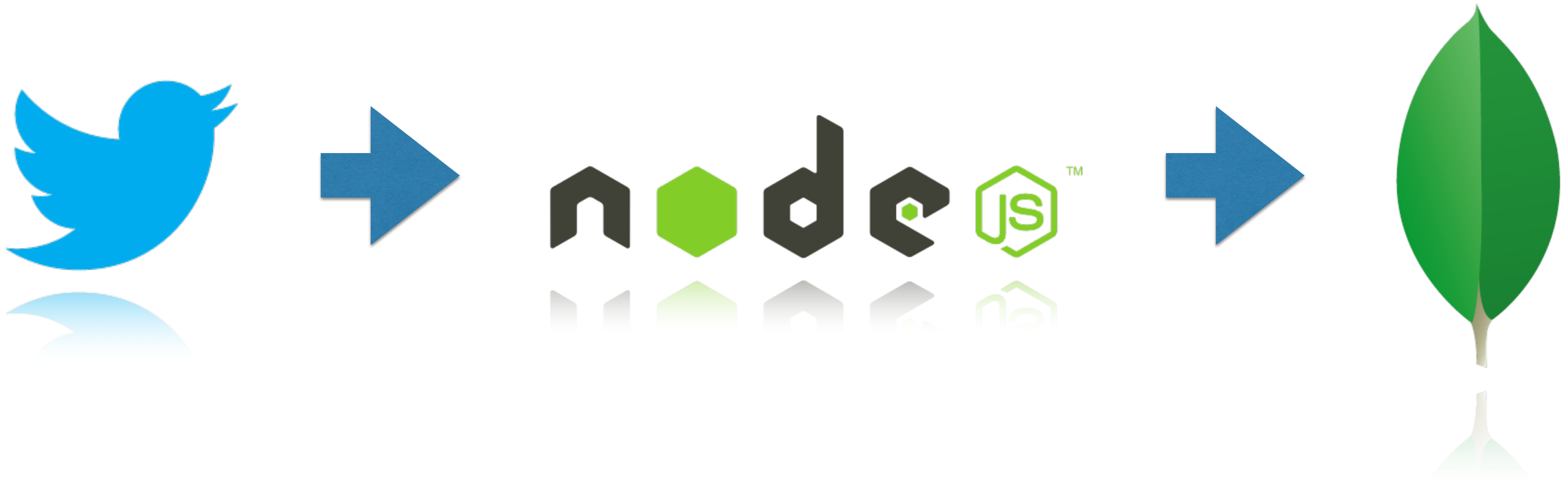
A lot of Tweets

Twitter Streaming API



Twitter Streaming API

- 23rd November 2013 - 5th January 2013
- Bounding Box applied limiting results
- Only GeoTagged to the UK
- JSON Object for each returned



The Numbers

- 31.6 million Tweets over 6 week period
- 700,000 tweets/daily
- 500 tweets/minute
- 8 tweets/second
- 40Gb of Data to process

Method

- Each tweet given an ‘Alcohol Score’
- Based on the number of keywords per tweet
- Key terms chosen to indicate alcohol consumption
- Closed language method

Key-terms

drunk	wine	wasted
pissed	hungover	hangover
vodka		

Alex Williams
@alexwilliams_xx

Follow

Really wanna get drunk tonight but I'm a loner with no drink left☐

Reply Retweet Favorite ••• More

5:30 PM - 2 Jan 14 from Bromsgrove, Worcestershire

Reply to @alexwilliams_xx

Matt Fothers @MattFothers 2m
@alexwilliams_xx get drunk on your own? #Sad
Details Reply Retweet Favorite ••• More

Alex Williams @alexwilliams_xx 10s
@MattFothers clearly not that's why I said 'but'
Details Reply Retweet Favorite ••• More

Sue Griffiths @SueGriffiths19 13m

Thanks for the donations, they definitely kept me on track for #soberoctober and I raised £170 for Macmillan. What shall I do at 00.01?

Details

Louise Daniel @LDanielPlym 3m

@SueGriffiths19 Whine? ;-)

Details

Sue Griffiths
@SueGriffiths19

Follow

@LDanielPlym I won't be whining at midnight but might've with a hangover tomorrow after several crisp dry whites!

Reply Retweet Favorite More

11:30 AM - 31 Oct 13 from Teignbridge, Devon

Reply to @SueGriffiths19 @LDanielPlym

Louise Daniel @LDanielPlym 59s

@SueGriffiths19 be rude not to - well done btw

Details


b.lo'
@BeelowKnows
Follow
Now I'm pissed that I thought I was too old for Halloween. Lol. I feel like throwing a costume party next year.
Reply
Retweet
Favorite
More
11:37 AM - 31 Oct 13
Reply to @BeelowKnows
© 2013 Twitter
About
Help
Ads

The Maths

$$\mathrm{SMAI}(T, M) = \frac{\sum_{t \in T} s(t)}{|T|}$$

$$s(t, M) = \frac{\sum_{m \in M} t(m)}{|tokens(t)|}$$

$$t(m) = \sum_{w \in tockens(t)} f(w, m)$$

$$f(w, m_j) = \begin{cases} 1, & \text{if } m_j = w \\ 0, & \text{otherwise} \end{cases}$$

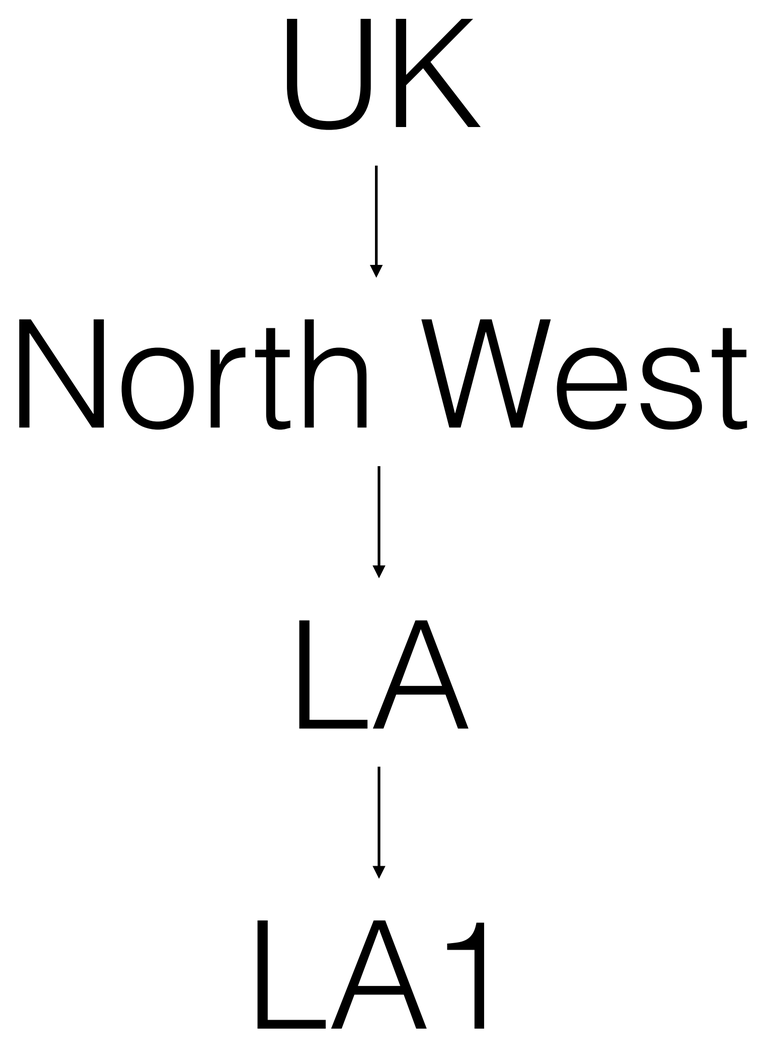
Groupings

	National	Regonal	Post Code Regon	Local Post Code
Daily				
Hourly				

31.6 million tweets becomes 252.8 million data points • 320 Gb to process

Geographical Groupings

- Each Tweet has Longitude and Latitude
- Shortest distance to the nearest centre of postcode
- Postcode Districts are from the initial characters of the alphanumeric UK postcode.
- Regions are groups of postcode regions

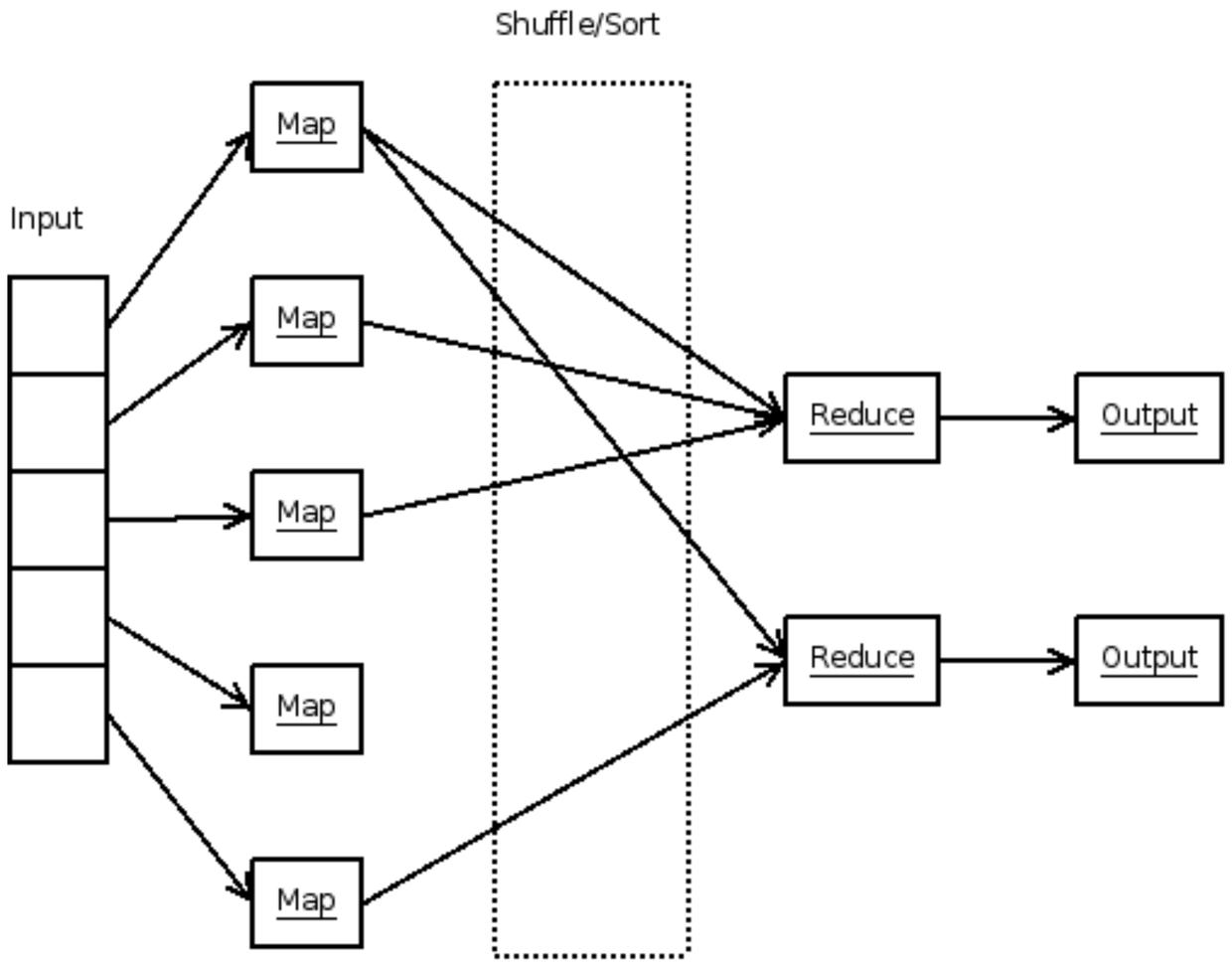

UK
North West
LA
LA1

Processing Data

- Data to big to process on one machine
- Task can be spread across a number of machines
- Can be implemented in a map reduce style

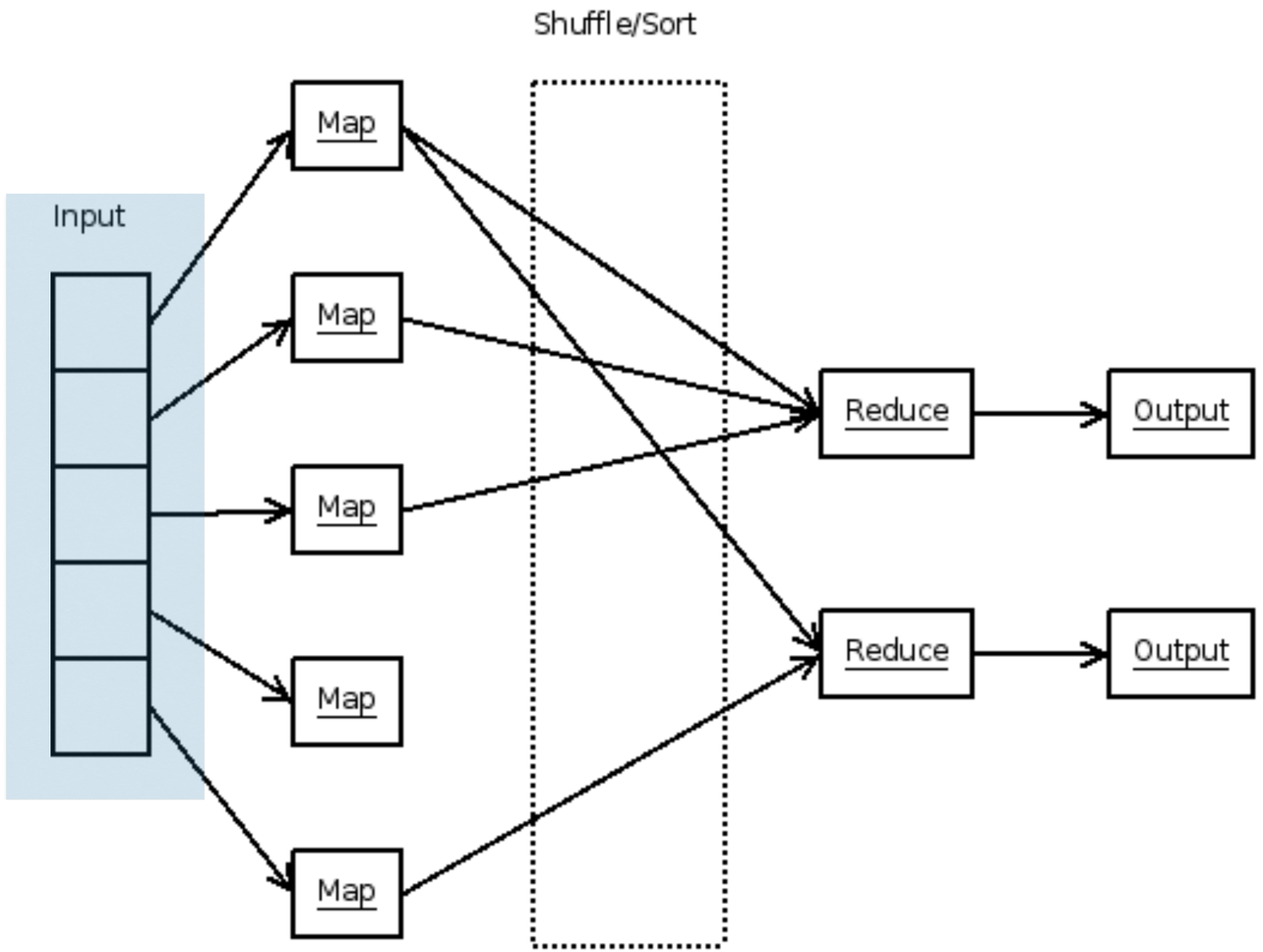
Map Reduce Explained

- **Challenge**: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user_id, value=1
- Shuffle: sort by user_id
- Reduce: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



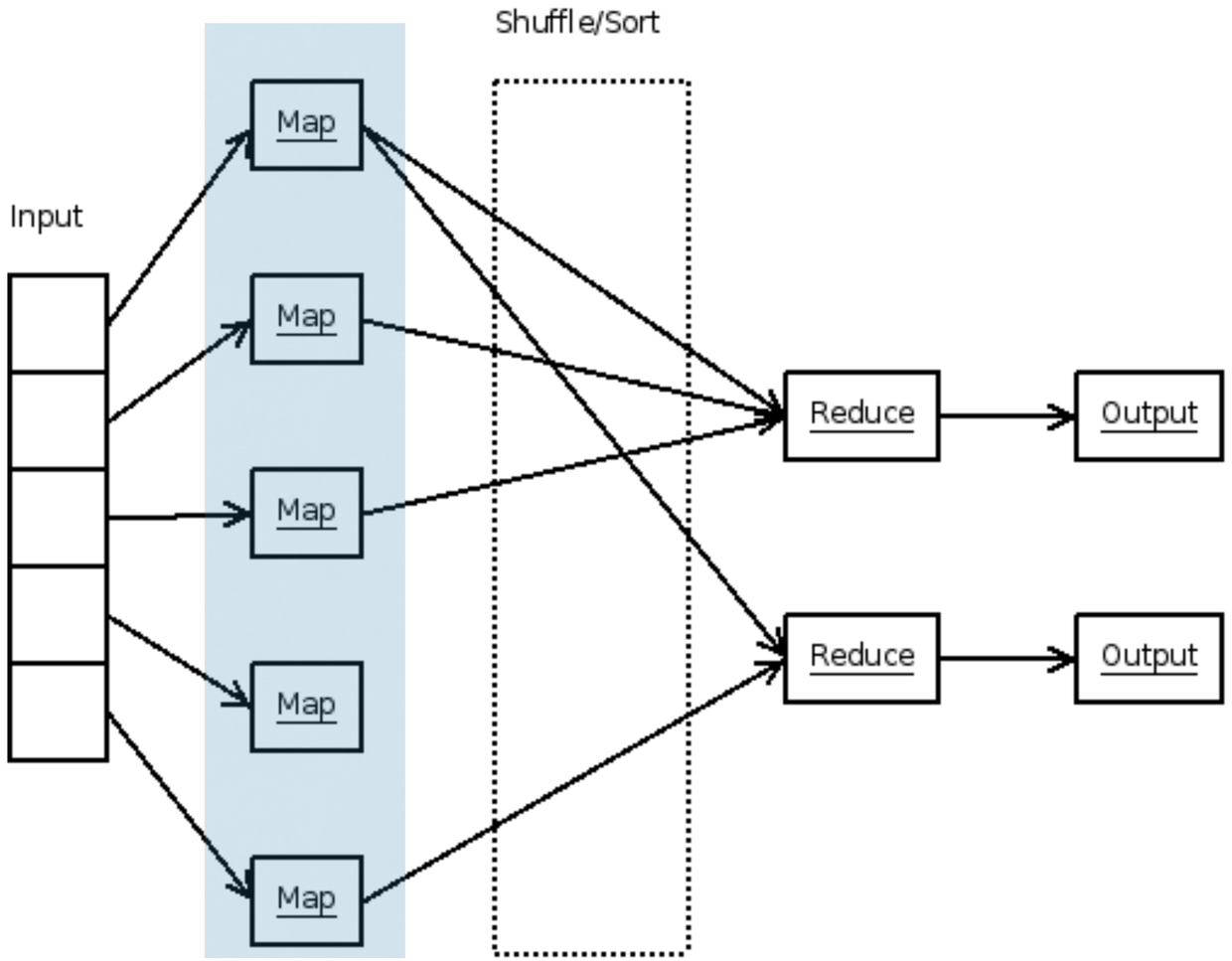
Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- **Input**: key=row, value=tweet info
- Map: output key=user_id, value=1
- Shuffle: sort by user_id
- Reduce: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



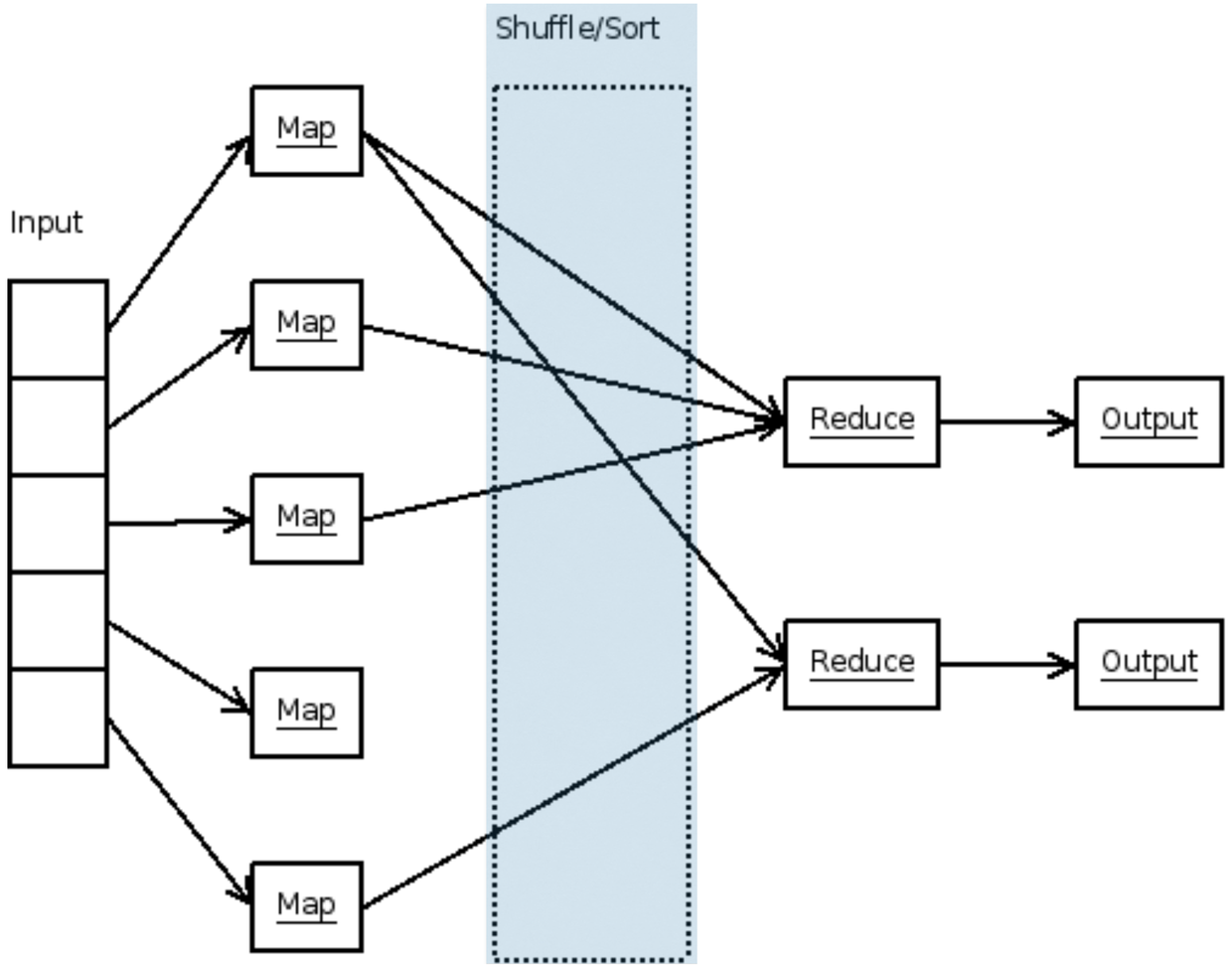
Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- **Map**: output key=user_id, value=1
- Shuffle: sort by user_id
- Reduce: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



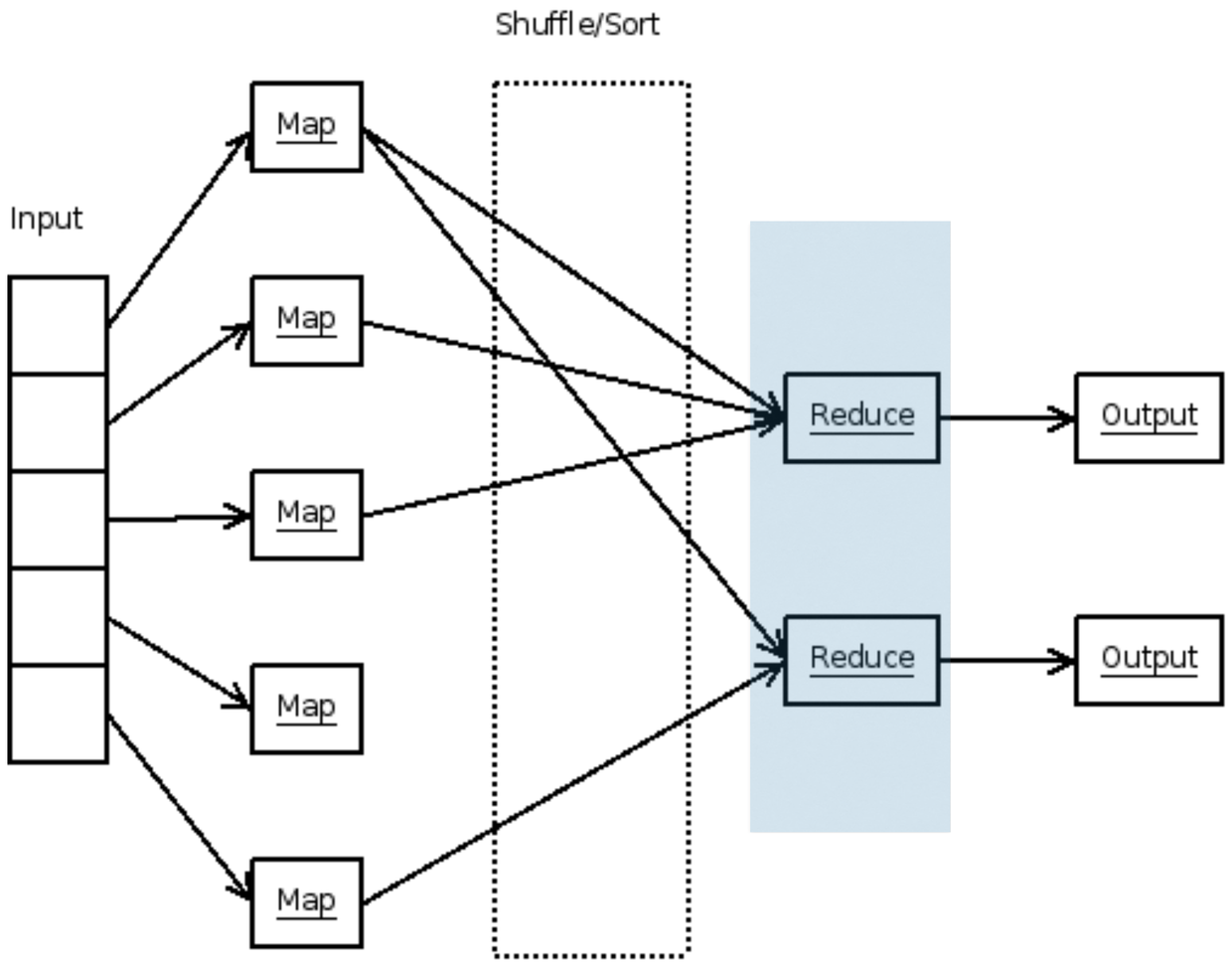
Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user_id, value=1
- **Shuffle**: sort by user_id
- Reduce: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



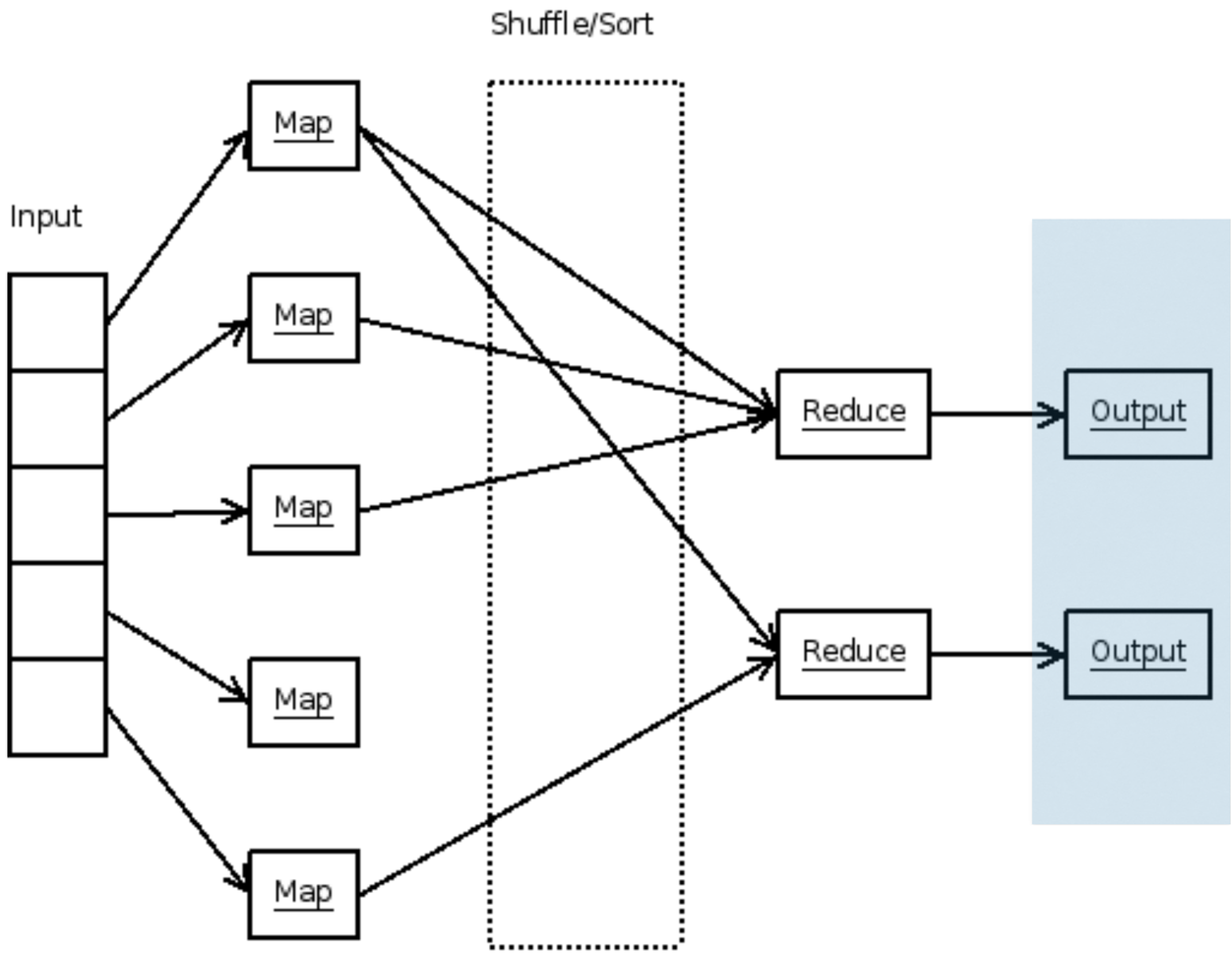
Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user_id, value=1
- Shuffle: sort by user_id
- **Reduce**: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



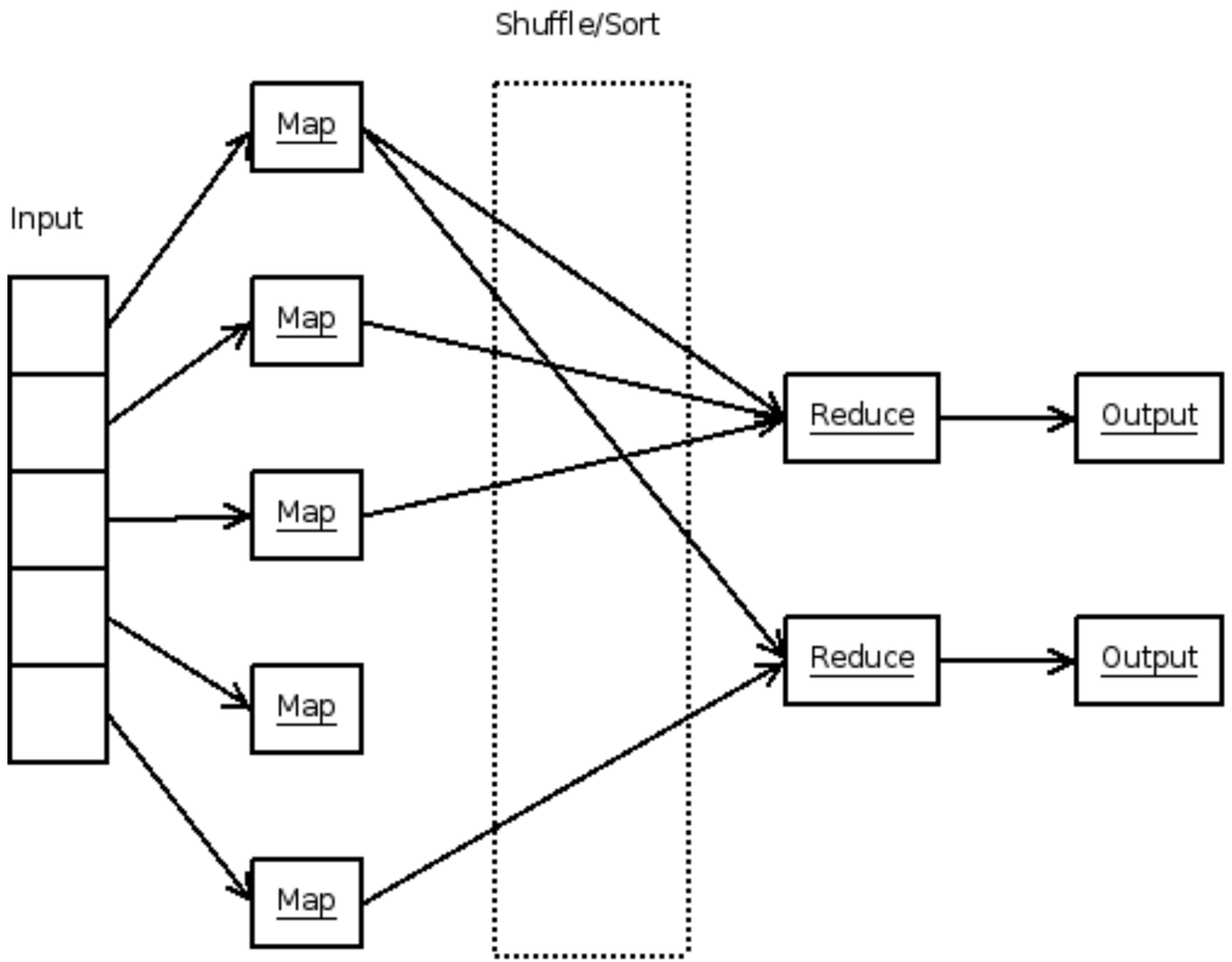
Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user_id, value=1
- Shuffle: sort by user_id
- Reduce: for each user_id, sum
- **Output**: user_id, tweet count
- With 2x machines, runs close to 2x faster.



Map Reduce Explained

- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user_id, value=1
- Shuffle: sort by user_id
- Reduce: for each user_id, sum
- Output: user_id, tweet count
- With 2x machines, runs close to 2x faster.



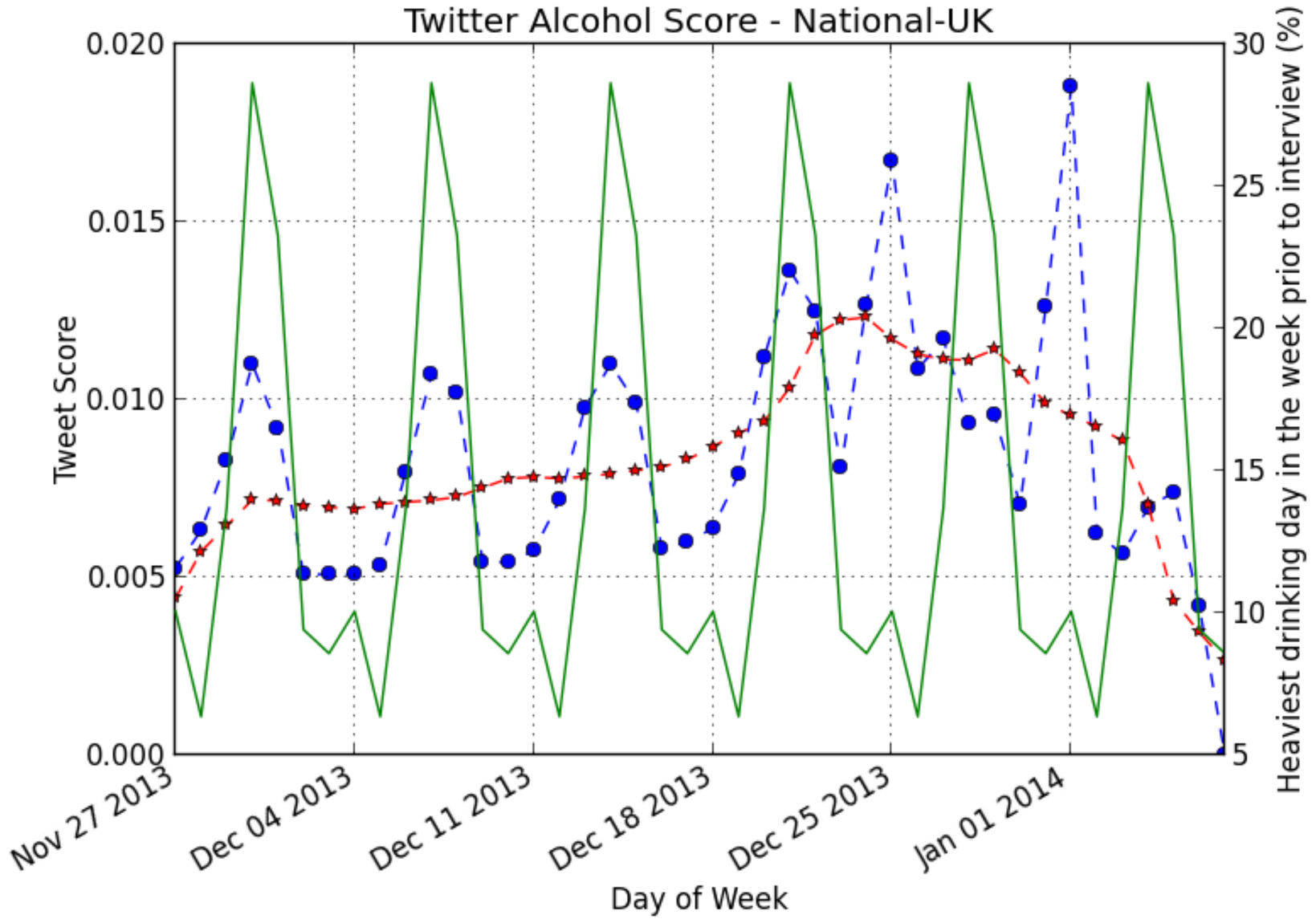
Map Reduce

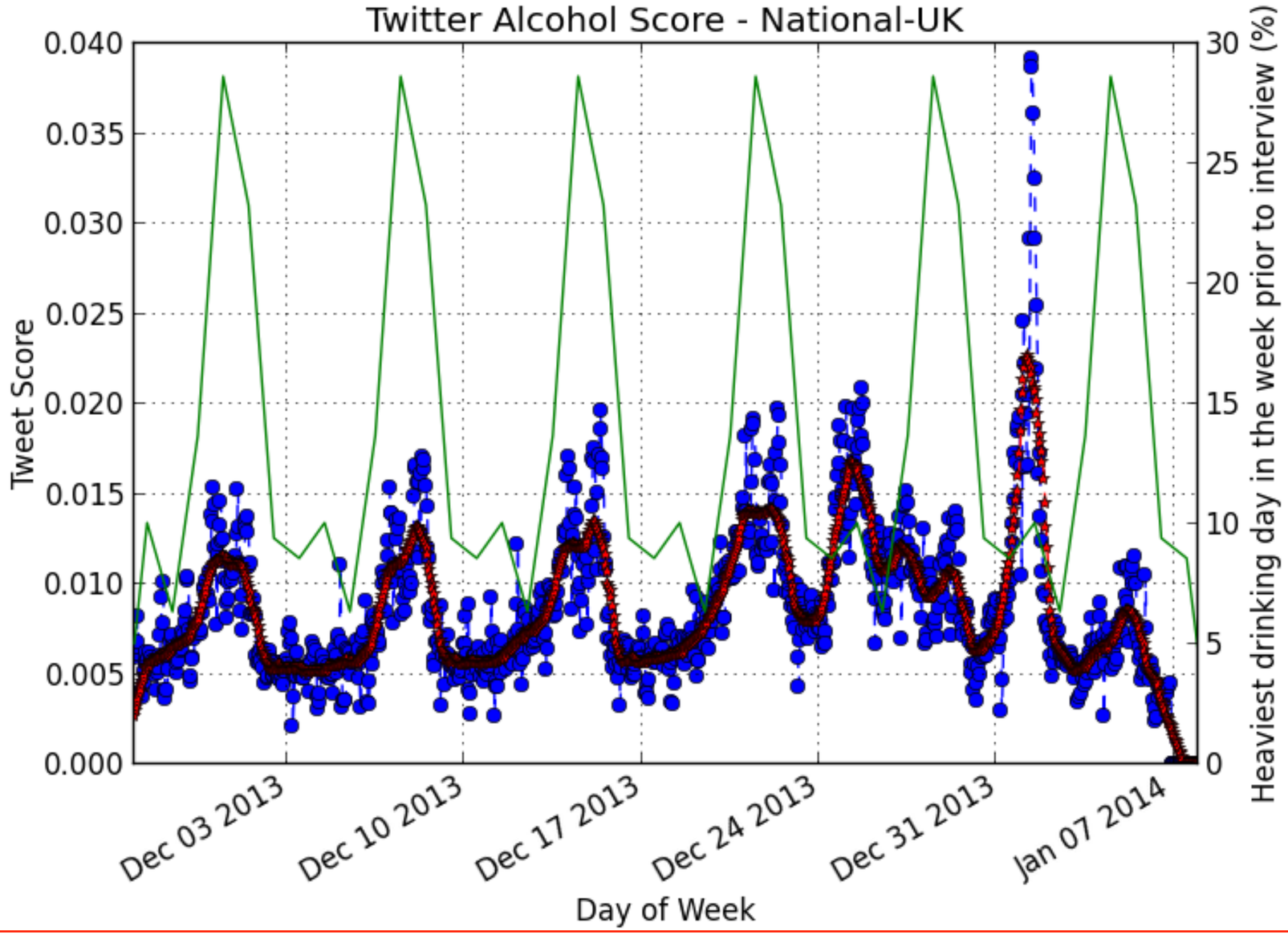
- **Input**: key=null, value=tweet info
- **Map**: output key=location+time, value=alcohol twitter score
- **Shuffle**: sort by location+time
- **Reduce**: for each location+time, average(alcohol twitter score), collocation for each each set
- **Output**: location+time, average(alcohol twitter score), collocation

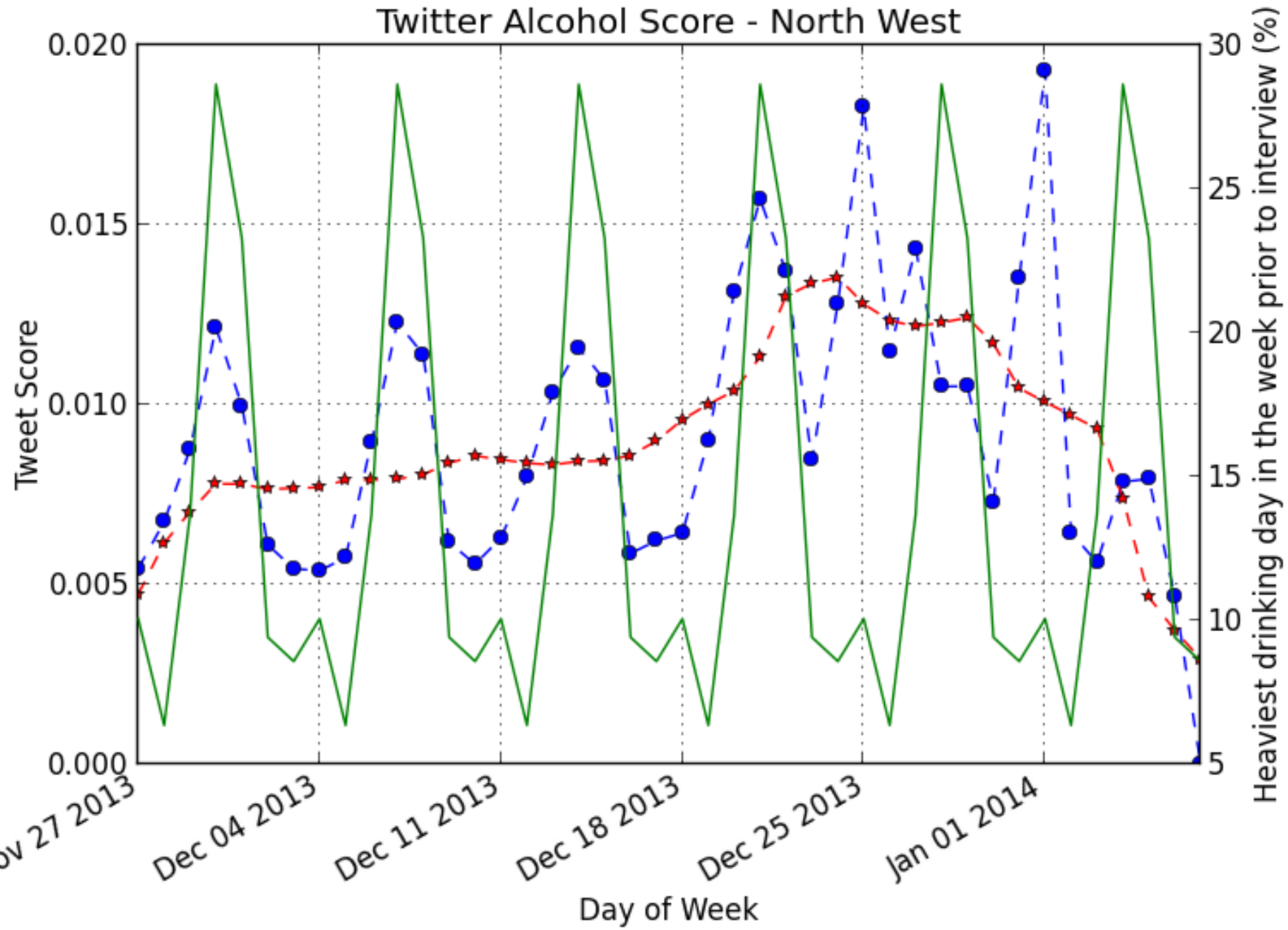
A day of processing later

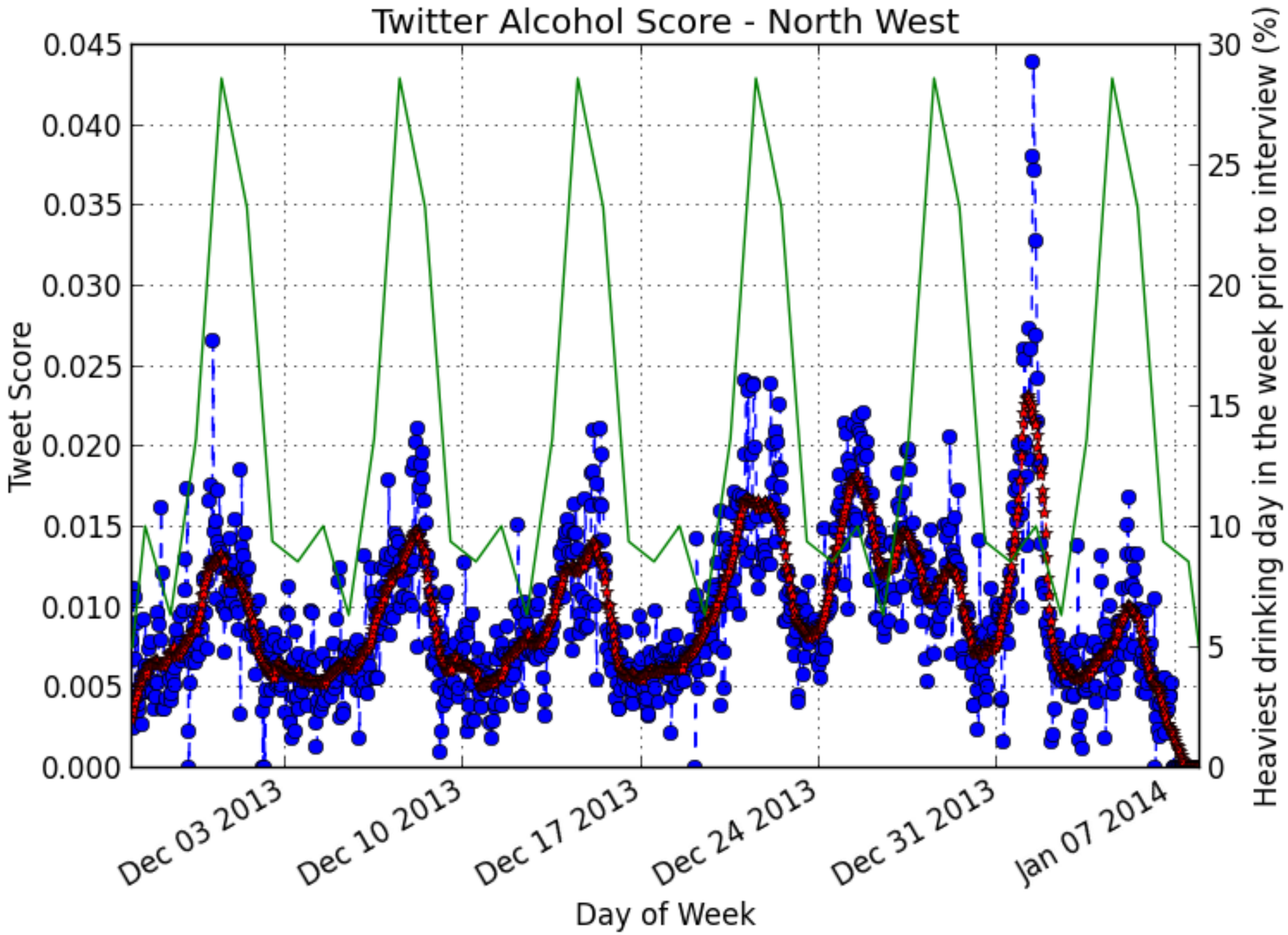
Lets see the data

I'll drink to that

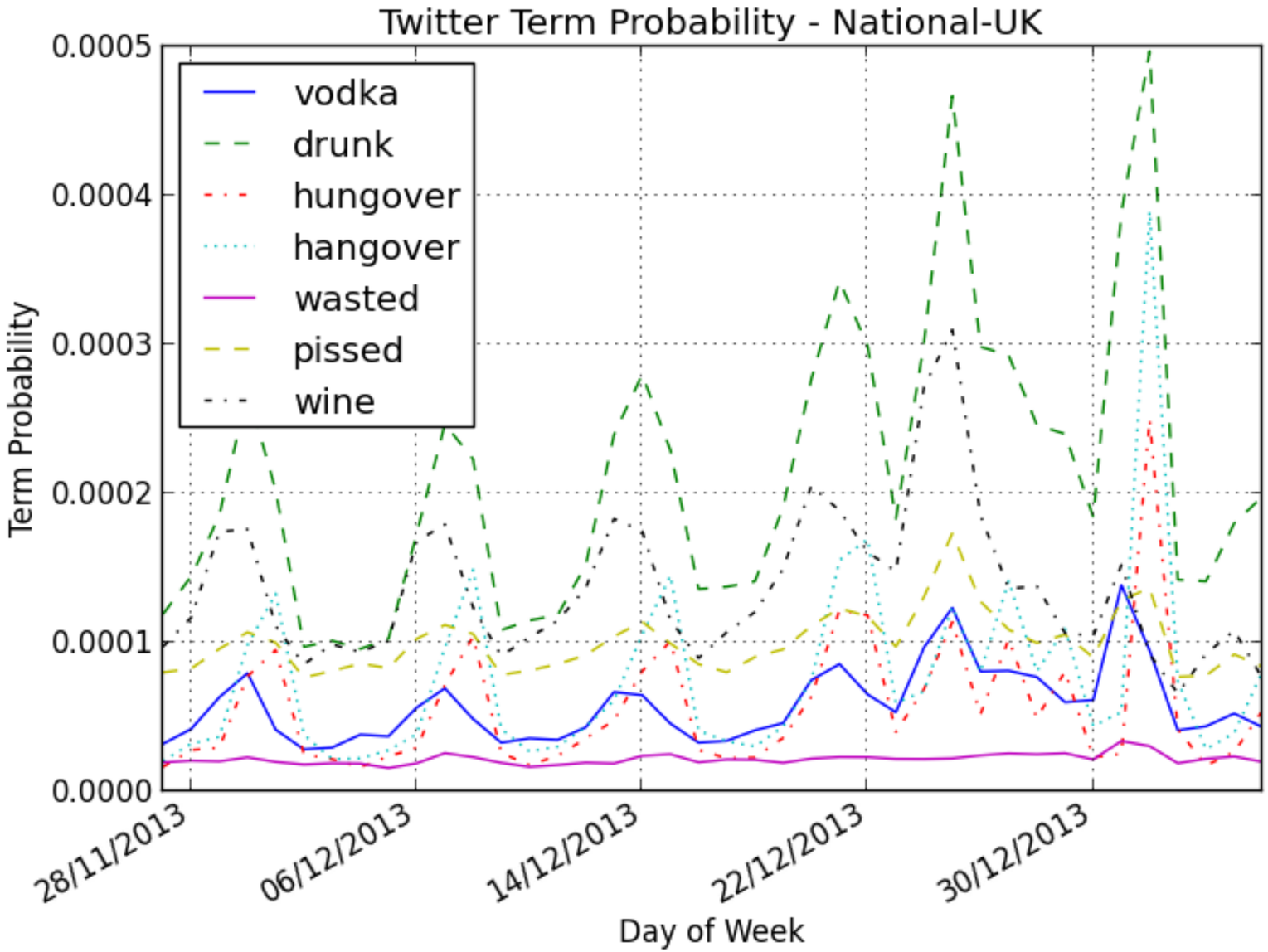

Twitter Alcohol Score - National-UK
Tweet Score
Heaviest drinking day in the week prior to interview (%)
Day of Week
0.020
0.015
0.010
0.005
0.000
30
25
20
15
10
5
Nov 27 2013
Dec 04 2013
Dec 11 2013
Dec 18 2013
Dec 25 2013
Jan 01 2014

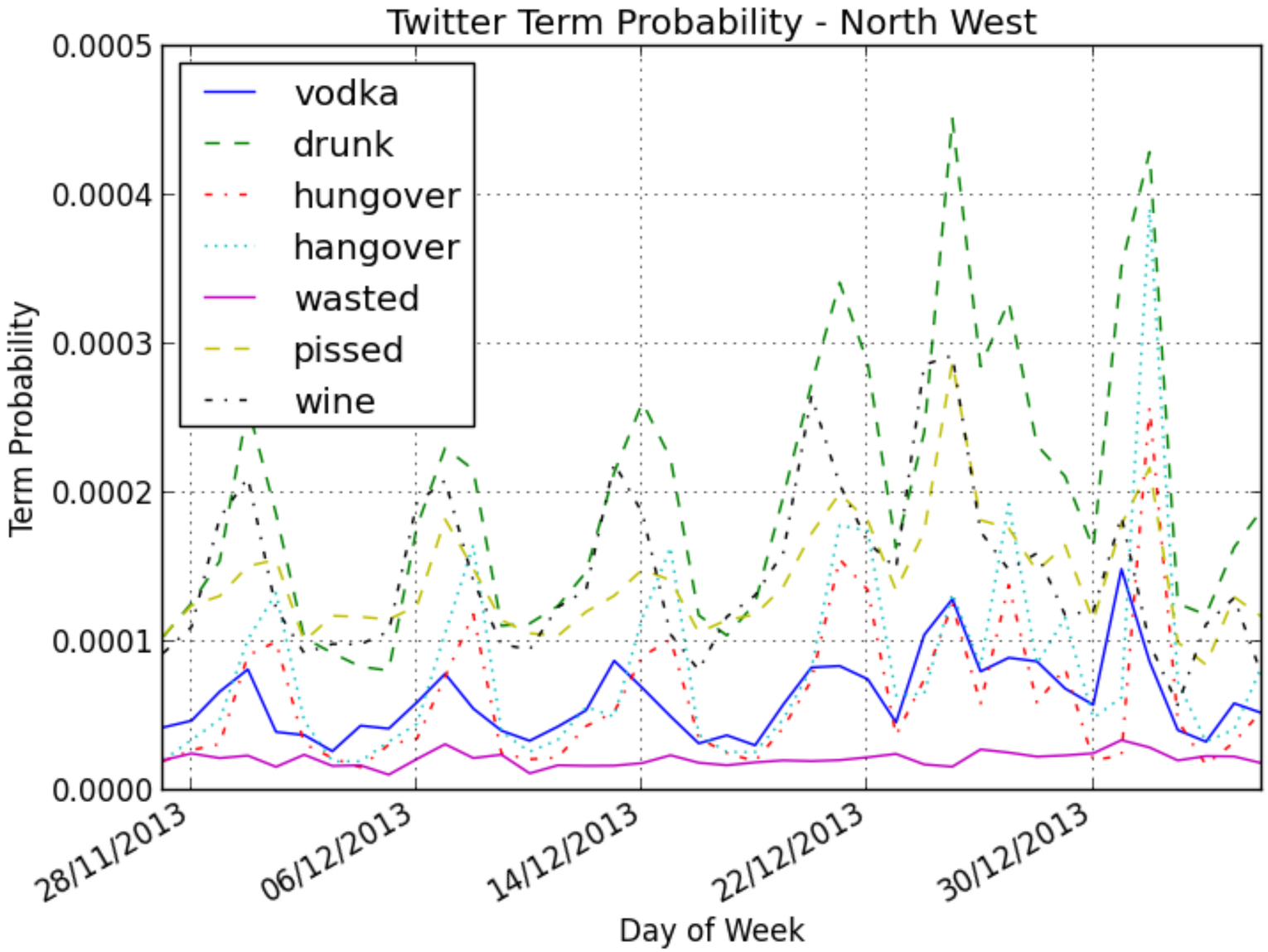

Twitter Alcohol Score - National-UK
Tweet Score
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Heaviest drinking day in the week prior to interview (%)
30
25
20
15
10
5
0
Dec 03 2013
Dec 10 2013
Dec 17 2013
Dec 24 2013
Dec 31 2013
Jan 07 2014
Day of Week

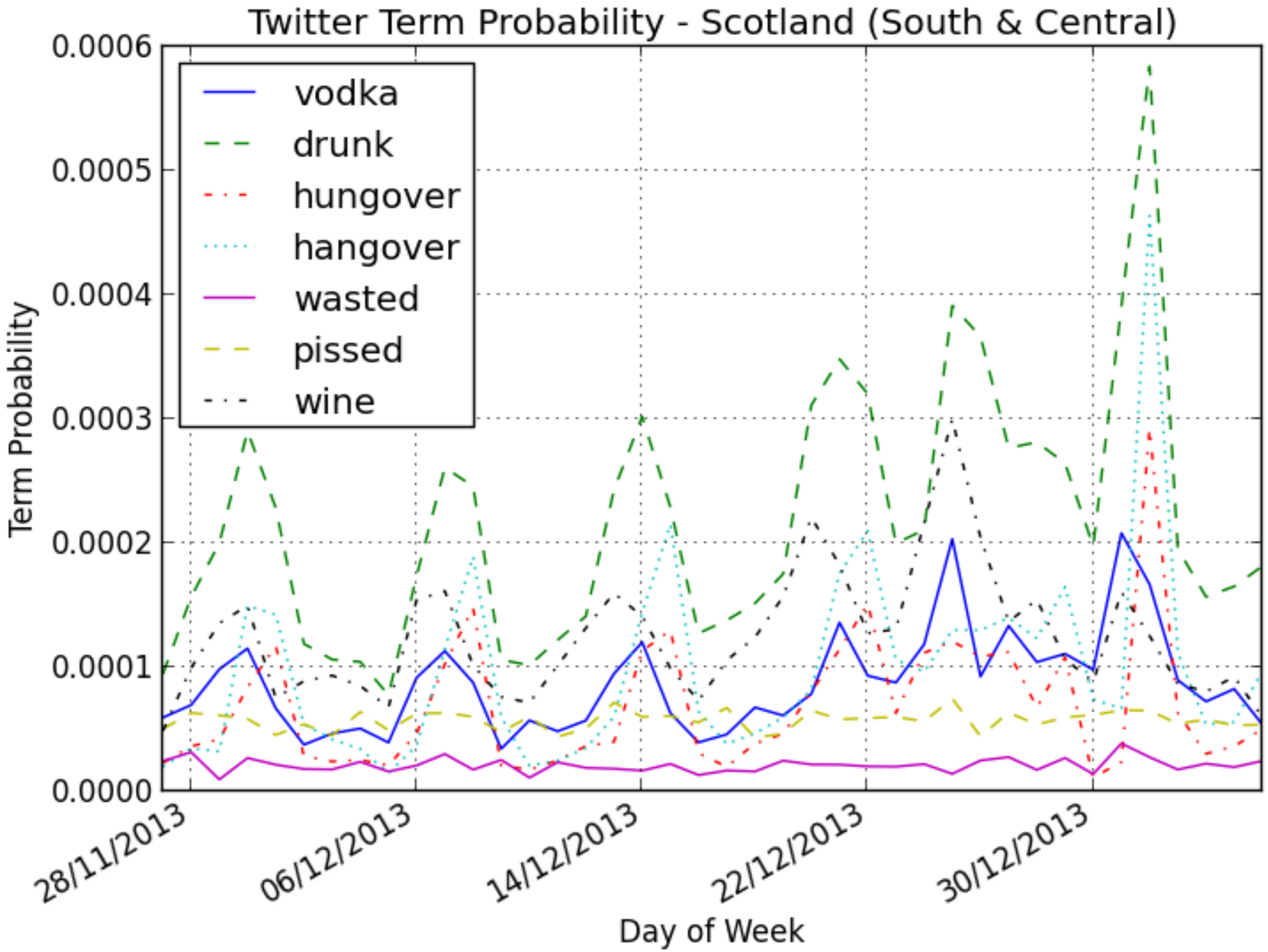

Twitter Alcohol Score - North West
Tweet Score
Heaviest drinking day in the week prior to interview (%)
Day of Week
0.020
0.015
0.010
0.005
0.000
30
25
20
15
10
5
Nov 27 2013
Dec 04 2013
Dec 11 2013
Dec 18 2013
Dec 25 2013
Jan 01 2014

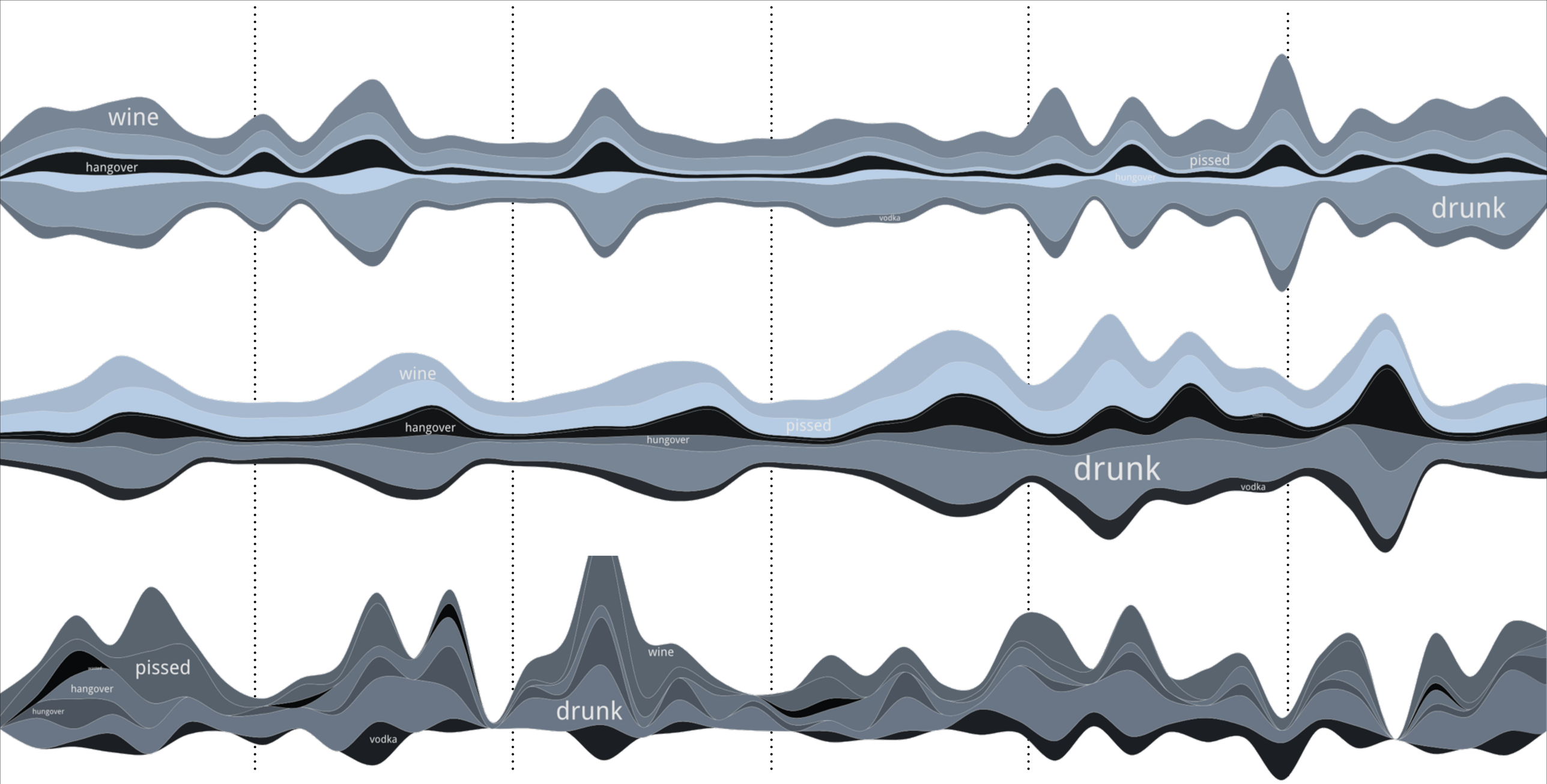

Twitter Alcohol Score - North West
Tweet Score
Heaviest drinking day in the week prior to interview (%)
Day of Week
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
30
25
20
15
10
5
0
Dec 03 2013
Dec 10 2013
Dec 17 2013
Dec 24 2013
Dec 31 2013
Jan 07 2014

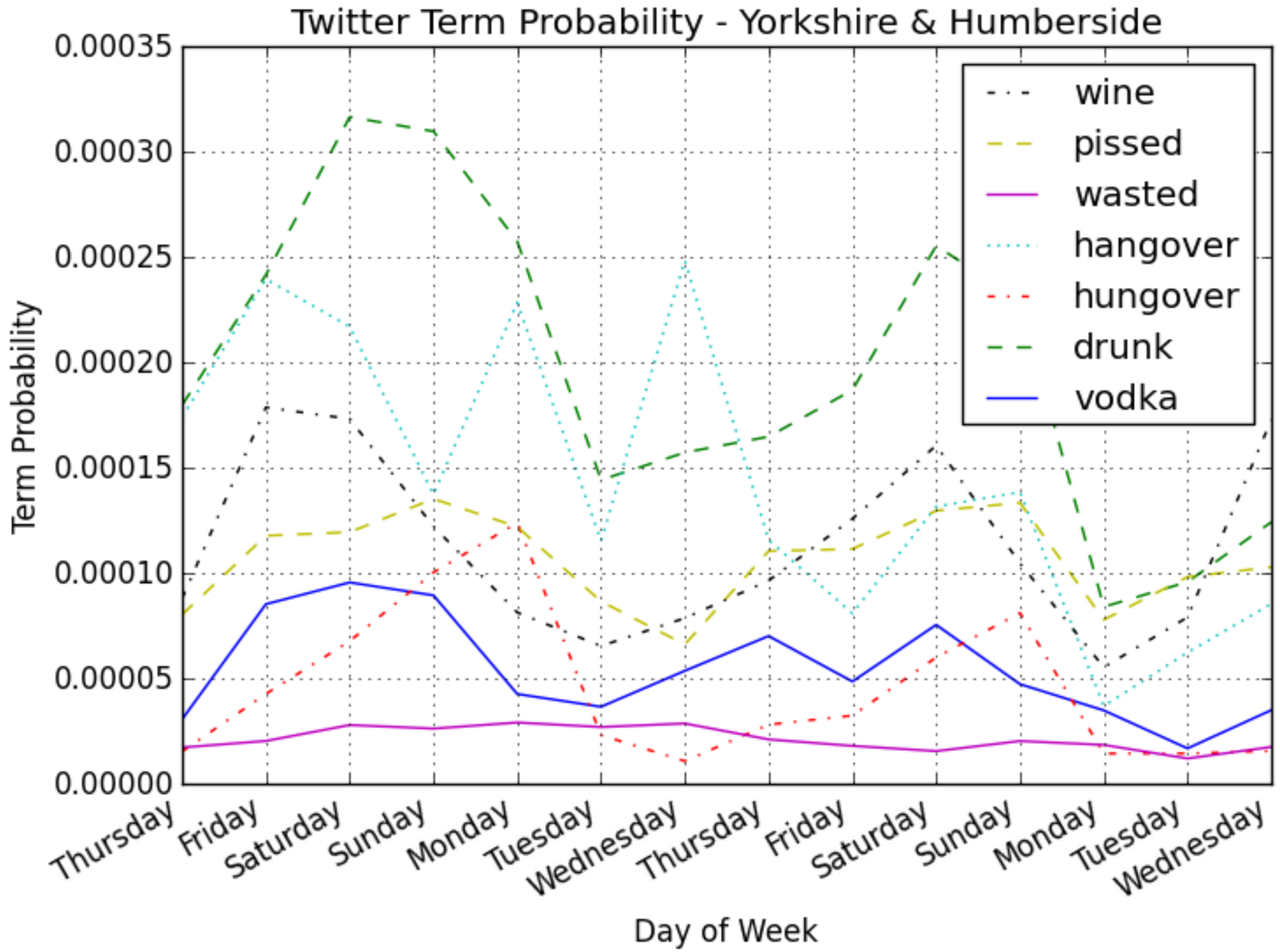
	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
National UK	0.93 ***	0.96 ***	0.86 **	0.74 **	-0.27 *	0.05 *
North West	0.92 ***	0.97 ***	0.84 **	0.76 **	-0.22 *	0.11 *
Yorkshire & Humberside	0.93 ***	0.96 ***	0.79 ***	0.71 **	-0.41 *	0.00 *
Greater London	0.86 **	0.93 ***	0.80 **	0.67 **	-0.27 *	0.06 *
South West	0.94 ***	0.94 ***	0.81 ***	0.66 *	-0.33 *	0.05 *
South East	0.91 ***	0.96 ***	0.87 ***	0.58 **	-0.29	0.06 *
Northern Ireland	0.91 ***	0.89 ***	0.80 **	0.57 **	-0.12 *	0.17 *
West Midlands	0.88 ***	0.96 ***	0.84 **	0.59 *	-0.26 *	0.09 *
Channel Islands	0.00 *	0.00 *	-0.30*	-0.35 *	-0.37 *	-0.22 *
Home Counties	0.91 ***	0.95 ***	0.90 ***	0.78 **	-0.24 *	0.06 *
Scotland (North)	0.93 ***	0.96 ***	0.88 ***	0.95 ***	-0.08 *	-0.06 *
East England	0.94 ***	0.97 ***	0.85 ***	0.72 **	-0.20 *	0.052 *
Scotland (South & Central)	0.89 ***	0.97 ***	0.93 ***	0.88 ***	-0.16 *	-0.08 *
Wales (South)	0.97 ***	0.90 ***	0.89 ***	0.78 **	-0.27 *	-0.04 *
Wales (North)	0.96 ***	0.98 ***	0.93 ***	0.76 **	-0.33 *	0.19 *
East Midlands	0.90 ***	0.90 ***	0.69 **	0.69 **	-0.19 *	0.09 *
North East	0.94 ***	0.91 ***	0.79 **	0.81 **	-0.27 *	0.04 *
Avrage	0.86	0.88	0.76	0.66	-0.23	0.03

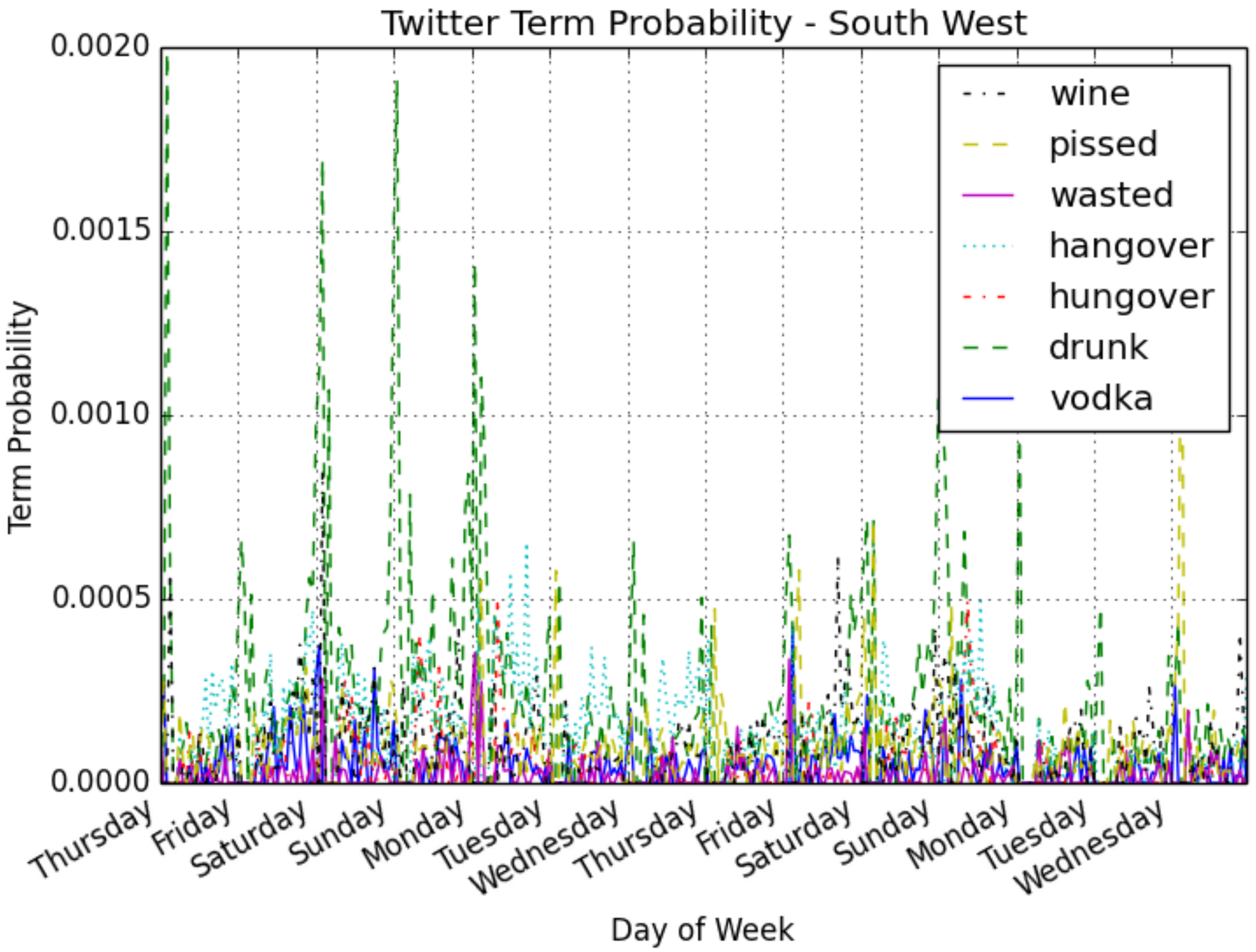

Twitter Term Probability - National-UK
vodka
drunk
hungover
hangover
wasted
pissed
wine
Term Probability
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
28/11/2013
06/12/2013
14/12/2013
22/12/2013
30/12/2013
Day of Week


Twitter Term Probability - North West
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
Term Probability
vodka
drunk
hungover
hangover
wasted
pissed
wine
28/11/2013
06/12/2013
14/12/2013
22/12/2013
30/12/2013
Day of Week


Twitter Term Probability - Scotland (South & Central)
Term Probability
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
vodka
drunk
hungover
hangover
wasted
pissed
wine
28/11/2013
06/12/2013
14/12/2013
22/12/2013
30/12/2013
Day of Week


wine
hangover
pissed
hungover
drunk
vodka
wine
hangover
pissed
hungover
drunk
vodka
pissed
hangover
hungover
vodka
drunk
wine


Twitter Term Probability - Yorkshire & Humberside
Term Probability
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
wine
pissed
wasted
hangover
hungover
drunk
vodka
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Day of Week


Twitter Term Probability - South West
Term Probability
0.0020
0.0015
0.0010
0.0005
0.0000
wine
pissed
wasted
hangover
hungover
drunk
vodka
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Day of Week

	North West	Yorkshire & Humberside	Greater London	South West	South East	Northern Ireland	West Midlands	Channel Islands	Home Counties	Scotland (North)	East England	Scotland (South & Central)	Wales (South)	Wales (North)	East Midlands	North East
North West	1															
Yorkshire & Humberside	0.96	1														
Greater London	0.92	0.94	1													
South West	0.94	0.96	0.94	1												
South East	0.92	0.95	0.95	0.97	1											
Northern Ireland	0.86	0.88	0.83	0.91	0.89	1										
West Midlands	0.96	0.96	0.93	0.96	0.95	0.88	1									
Channel Islands	0.01	0.02	0.02	0	0	0	0.01	1								
Home Counties	0.93	0.96	0.95	0.97	0.97	0.9	0.96	0.01	1							
Scotland (North)	0.85	0.89	0.86	0.91	0.92	0.9	0.88	0	0.91	1						
East England	0.93	0.96	0.94	0.97	0.98	0.89	0.96	-0.01	0.97	0.91	1					
Scotland (South & Central)	0.86	0.9	0.87	0.93	0.93	0.93	0.89	0.01	0.92	0.95	0.91	1				
Wales (South)	0.91	0.92	0.88	0.91	0.91	0.9	0.89	0	0.91	0.9	0.9	0.9	1			
Wales (North)	0.9	0.89	0.87	0.9	0.88	0.83	0.89	-0.03	0.88	0.84	0.88	0.85	0.89	1		
East Midlands	0.94	0.96	0.94	0.97	0.97	0.88	0.97	-0.01	0.96	0.9	0.97	0.9	0.91	0.89	1	
North East	0.93	0.93	0.89	0.93	0.91	0.89	0.91	0	0.91	0.9	0.92	0.9	0.94	0.9	0.92	1

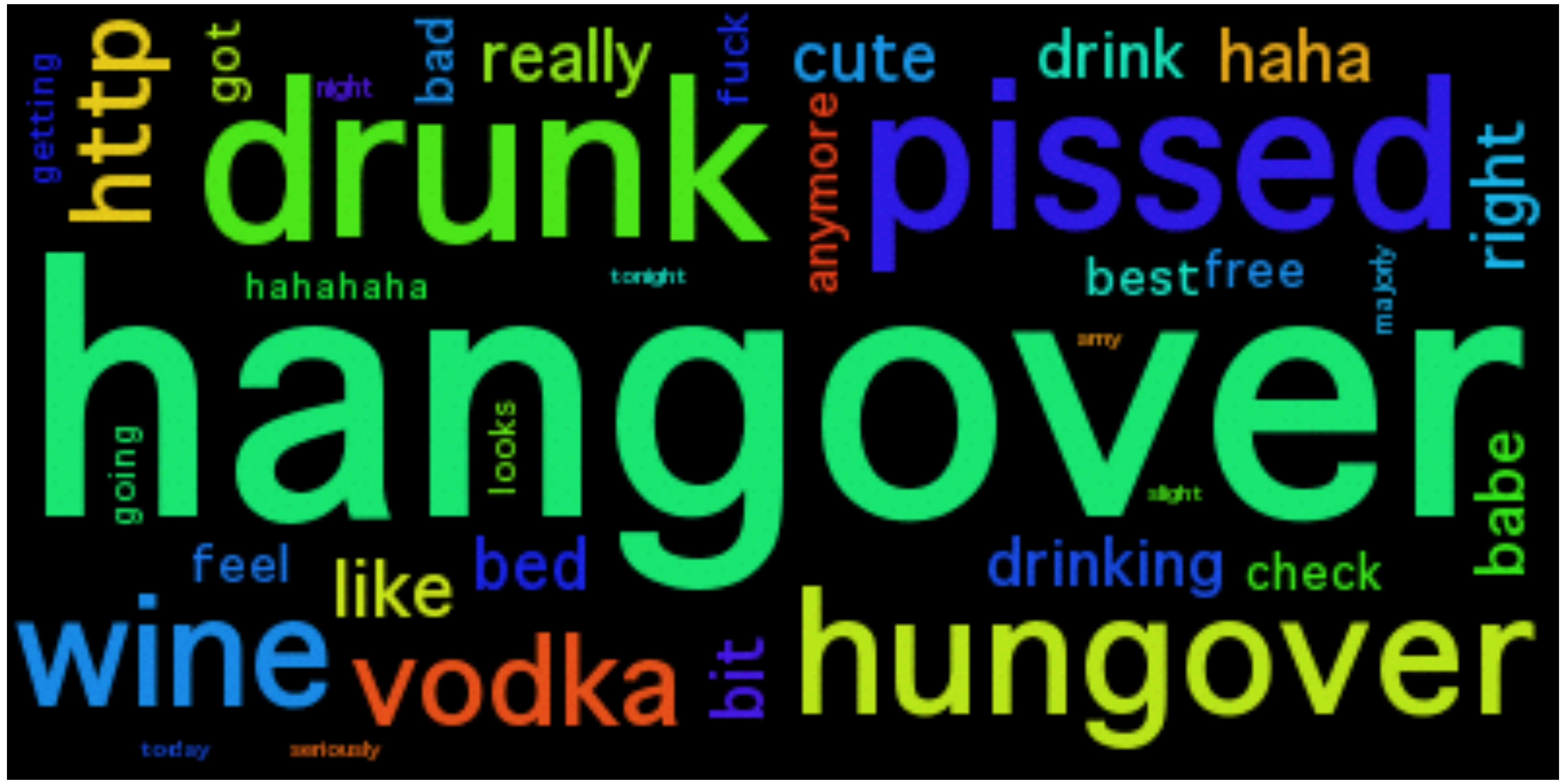
Collocations

Thursday

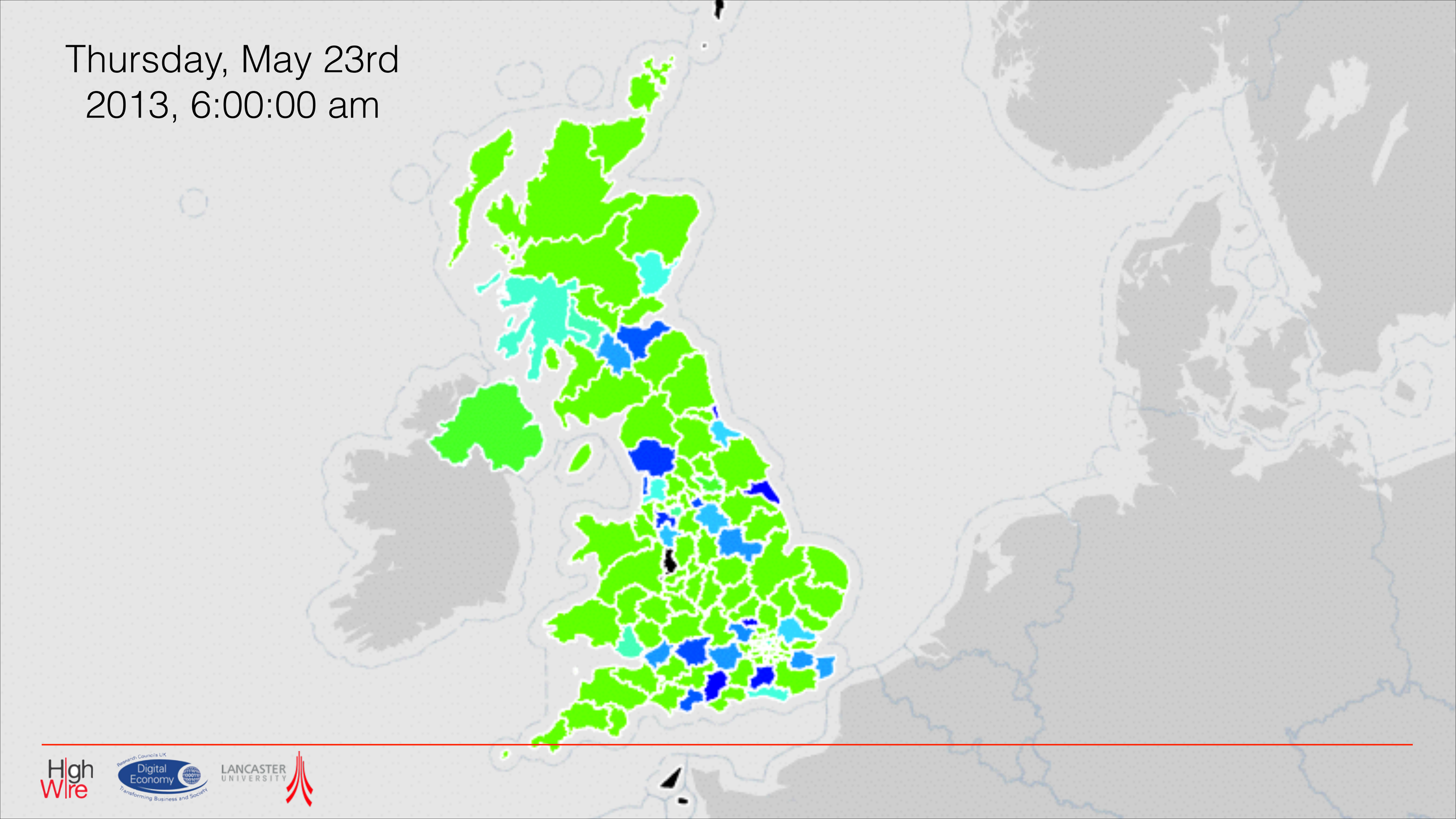
Friday

Saturday

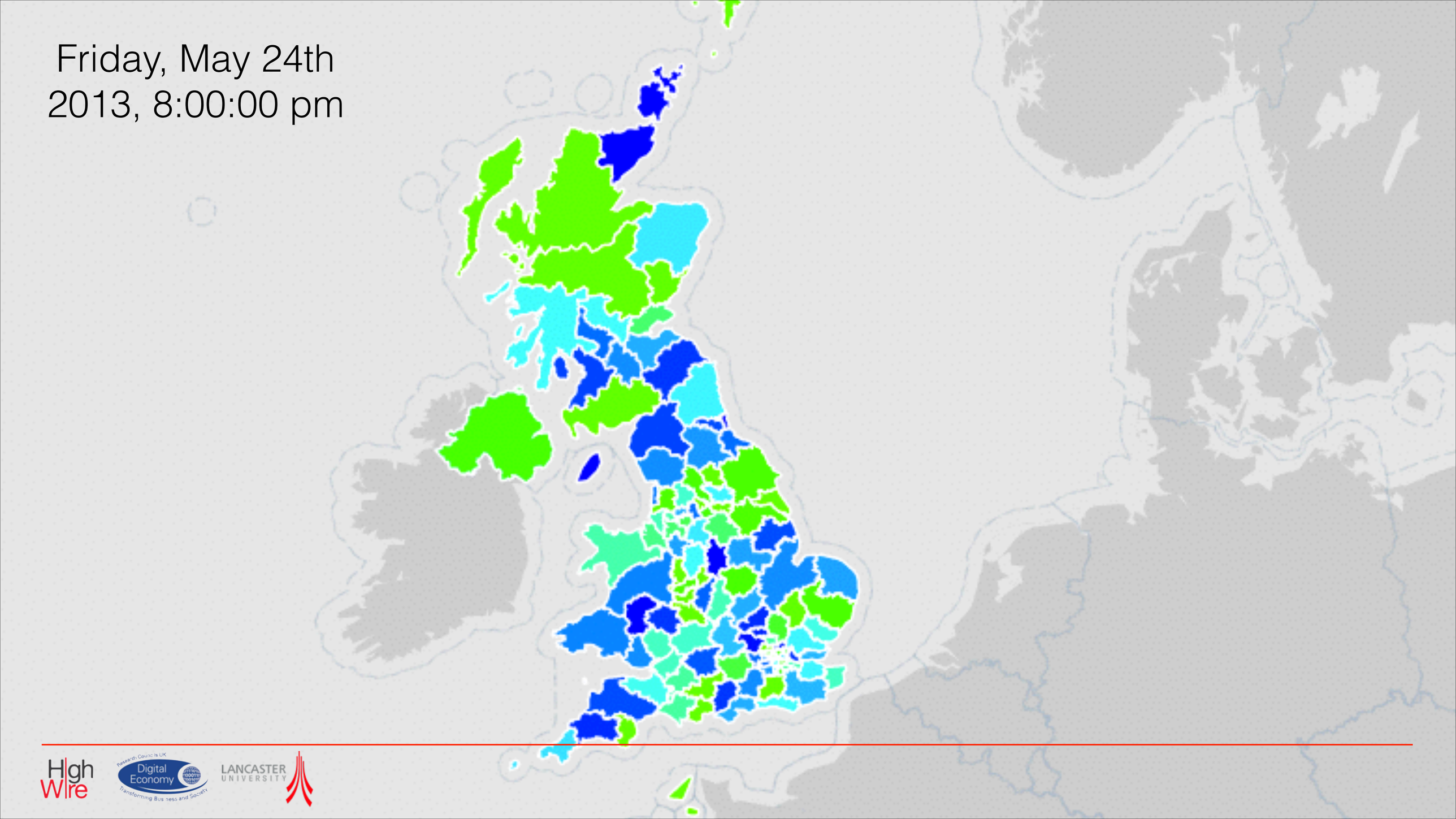
Yorkshire and Humber


getting
http
got
night
bad
really
fuck
cute
drink
haha
drunk
pissed
anymore
right
hahahaha
tonight
best
free
majority
army
hangover
going
looks
slight
babe
feel
like
bed
drinking
check
wine
vodka
bit
hungover
today
seriously


man
double
lol
tonight
free
drink
food
really
tomorrow
people
getting
like
good
cure
today
work
extremely
little
away
worst
big
feel
need
tweets
slightly
half
home
fairly
waster
haha
watching
severely
nearly
quite
bad
movies
love
guy
drinking
day
right
got
bit
tak
pretty
red
http
new
bar
time
hate
kicking
maybe
fucking
bottle
white
feeling
hahaha


Thursday, May 23rd
2013, 6:00:00 am

Friday, May 24th
2013, 8:00:00 pm



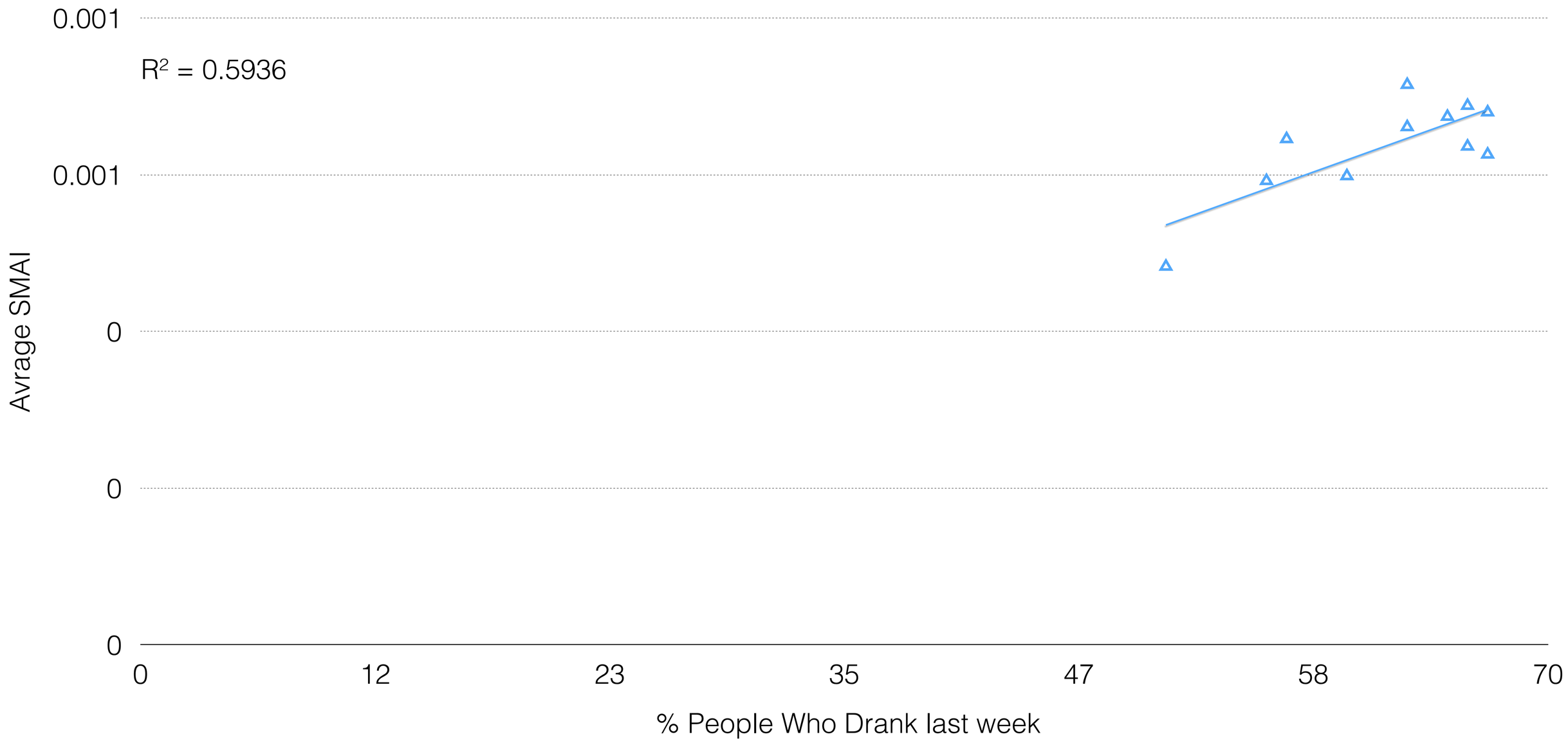
High Wire

Research Councils UK Digital Economy Transforming Business and Society

LANCASTER UNIVERSITY

Some new tables

Region	Drank last week	Avrage SMAI
North East	63	0.000715033
North West	63	0.00066108
Yorkshire and the Humber	67	0.000679924
East Midlands	65	0.000674305
West Midlands	60	0.000598496
East of England	66	0.000636277
London	51	0.000482961
South East	67	0.000626092
South West	66	0.000688288
Wales	57	0.000645785
Scotland	56	0.00059233


0.001
R^2 = 0.5936
0.001
Avrage SMAI
0
0
0
0
12
23
35
47
58
70
% People Who Drank last week

Limitations

- Sample not representative of general population
    - 70% of Americans online but only 18% report using twitter
- Users are younger
- Some ethnic groups are under represented
- Small number of key terms measured

Brenner, J. (2013, December 31)

Further Work

- Expanded keywords list
- Open Langauge method
- Apply machine learning techniques
- Make it real time
- Explore using Twitter Store vs Apache Hadoop - http://storm-project.net/

References

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. Presented at the SOMA '10: Proceedings of the First Workshop on Social Media Analytics, ACM Request Permissions. doi:10.1145/1964858.1964874

Bollen, Mao, Zeng. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 8–8. doi:10.1016/j.jocs.2010.12.007

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Presented at the WWW '10: Proceedings of the 19th international conference on World wide web, ACM. doi:10.1145/1772690.1772777

Health and Social Care Information Centre, L. S. (2012, May 31). Statistics on Alcohol: England, 2012. *Catalogue.Ic.Nhs.Uk*. Retrieved May 19, 2013, from https://catalogue.ic.nhs.uk/publications/public-health/alcohol/alco-eng-2012/alco-eng-2012-rep.pdf

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations (pp. 47–56). Presented at the WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference, New York, New York, USA: ACM Request Permissions. doi:10.1145/2464464.2464480

Strunin, L. (2001). Assessing alcohol consumption: developments from qualitative research methods. *Social Science & Medicine*. doi:10.1016/S0277-9536(00)00332-4

Health and Social Care Information Centre, L. S. (2012, May 31). Statistics on Alcohol: England, 2012. *Catalogue.Ic.Nhs.Uk*. Retrieved May 19, 2013, from https://catalogue.ic.nhs.uk/publications/public-health/alcohol/alco-eng-2012/alco-eng-2012-rep.pdf

Del Boca, F. K., & Darkes, J. (2003). The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction*, *98*(s2), 1–12. doi:10.1046/j.1359-6357.2003.00586.x

Babor, T. F., Stephens, R. S., & Marlatt, G. A. (1987). Verbal report methods in clinical research on alcoholism: Response bias and its minimization. *Journal of Studies on Alcohol and …*.

Penny, G. N., & Armstrong-Hallam, S. (2010). Student Choices and Alcohol Matters (SCAM): A multi-level analysis of student alcohol (mis) use and its implications for policy and prevention strategies within universities, cognate educational establishments and the wider community.

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002). Test-retest reliability of alcohol measures: Is there a difference between Internet-based assessment and traditional methods?, *16*(1), 56–63. doi:10.1037/0893-164X.16.1.56

Boniface, S., & Shelton, N. (2013). How is alcohol consumption affected if we account for under-reporting? A hypothetical scenario. *The European Journal of Public Health*, *23*(6), ckt016–1081. doi:10.1093/eurpub/ckt016

Brenner, J. (2013, December 31). Pew Internet: Social Networking (full detail) | Pew Research Center's Internet & American Life Project. Pew Internet. Retrieved February 6, 2014, from http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx

Thank you

http://alcohol-twitter-map.herokuapp.com/