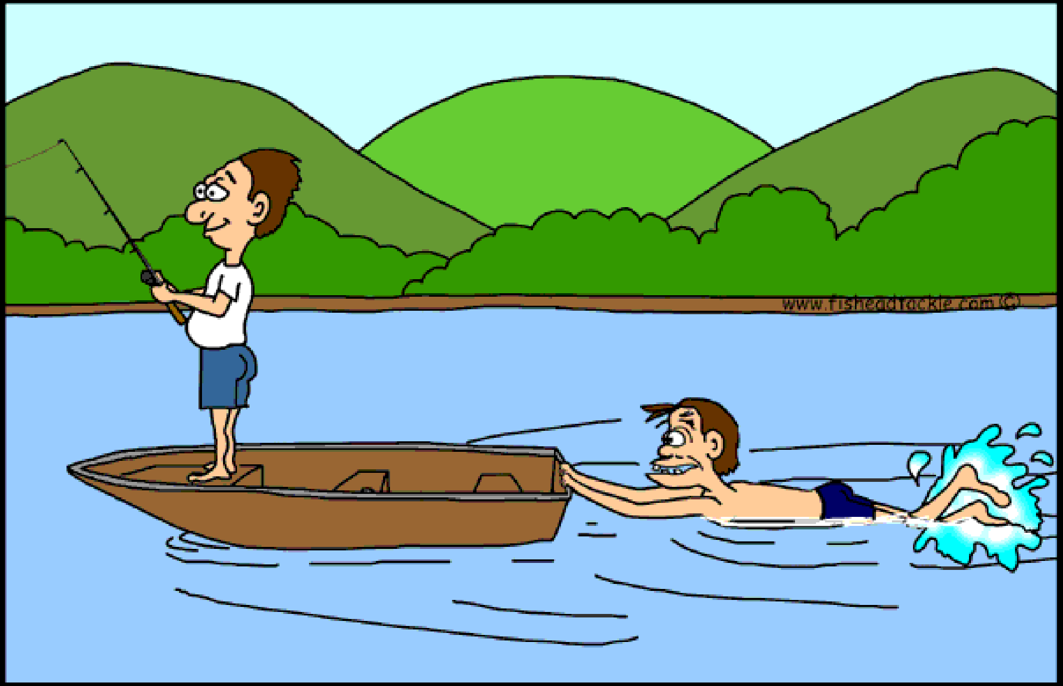
Big Data Mining

Zhifei Zhang

Outline

- Big data vs. Small data
- Software for big data
- An example

Small Data vs. Big Data



Big Data

Attributes

- Volume
- Variety
- Velocity
- Variability
- Veracity

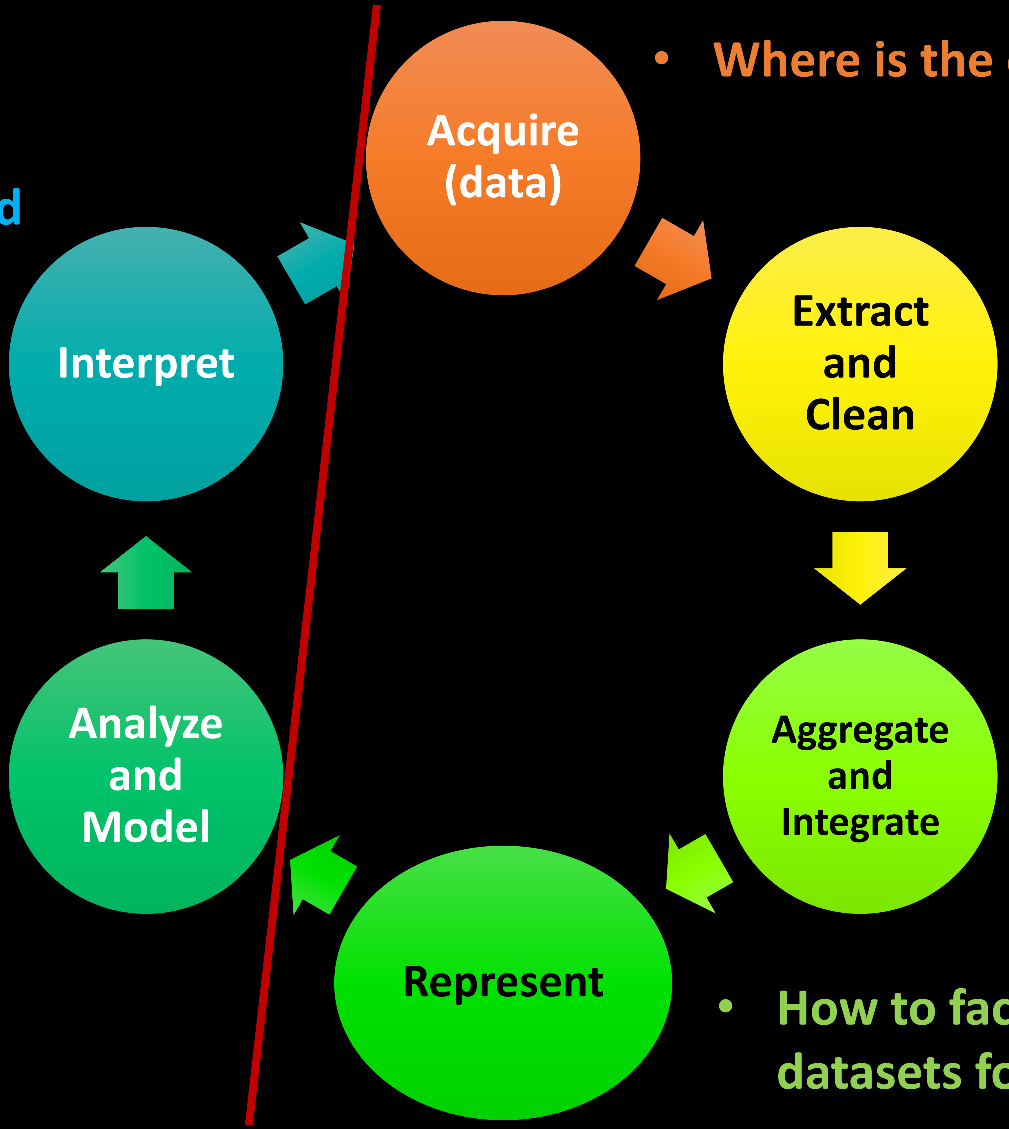
Types of Datasets

- High dimensional
- Sparse
- Graph
- Infinite/streaming
- Labeled

Compute Models

- Map Reduce
- Streams / Online
- Single machine in-memory

Big Data



- Where is the data coming from?

- How to deal with missing/incomplete /noisy data?

- How to bring datasets together?

- How to facilitate and host datasets for analysis?

- How to query, mine and model data?

- How to ensure that end users understand the data?

Algorithms for Big Data

- Classification: Bayes, clustering.
- Regression: polynomial, SVM.
- Dim. Reduction: PCA, SVD.

Example 1: Naïve Bayes

$$P(y|x) \propto p(x|y)P(y)$$

$$P(x|y) = \frac{C(X = x, Y = y)}{C(Y = y)}$$

$$P(y) = \frac{C(Y = y)}{C(Y = any)}$$

Given a data point (x, y)**:**

$$C(Y = ang) + +$$

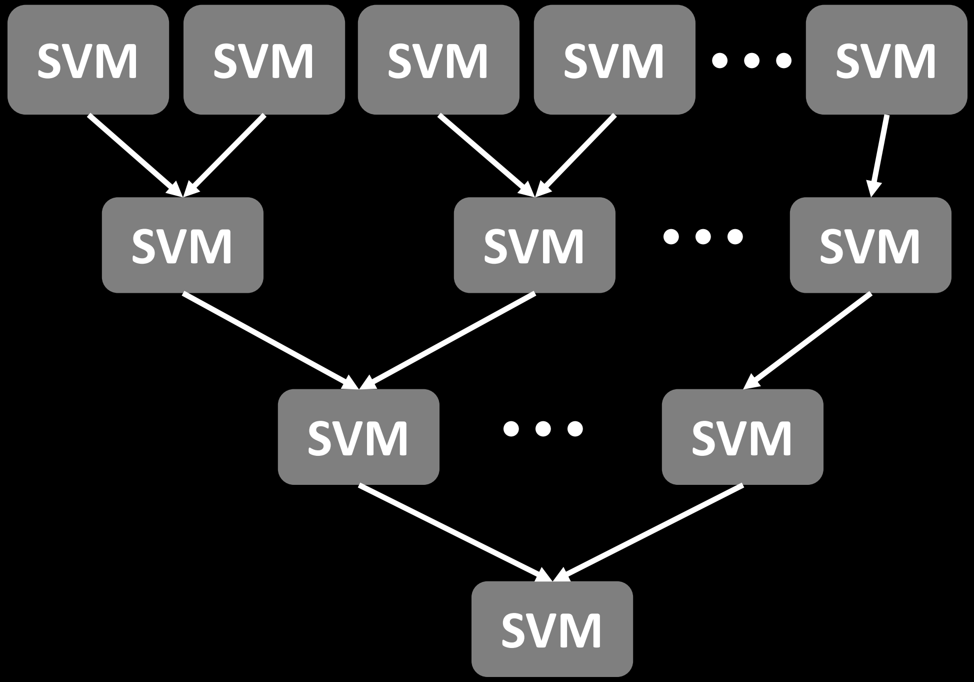
$$C(Y = y) + +$$

$$C(Y = y, X = x) + +$$

Example 2: SVM

$$\min_{\alpha}\left(\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i, x_j) - \sum_i \alpha_i\right)$$

Parallel-Hierarchical SVM



Given a partial data set:

Implement SVM

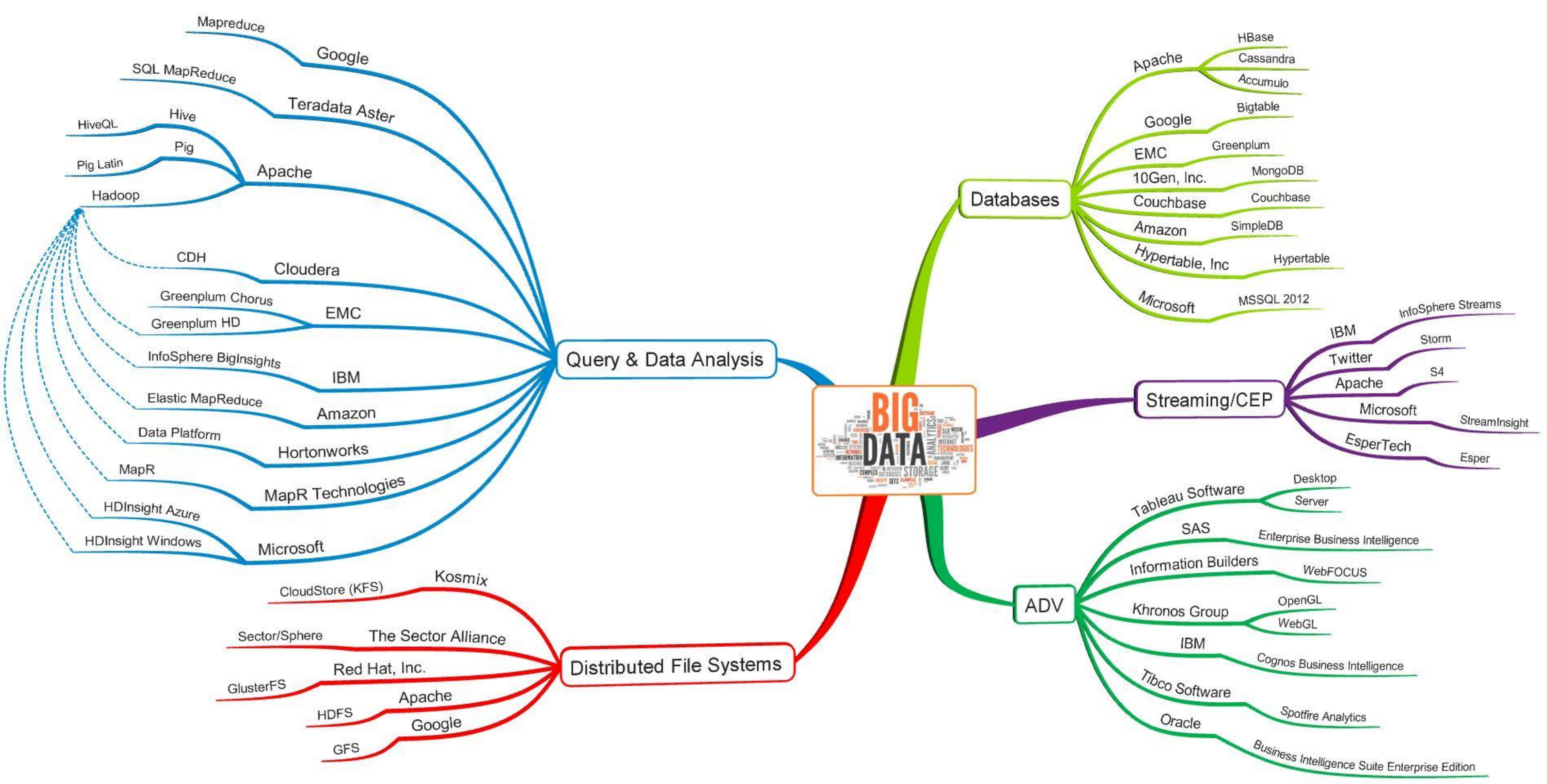
Save the local model

Combine local models

Algorithms for Big Data

- On-line version
- High efficiency
- Less iteration
- Approximation

Software for Big Data



BIG DATA

- Query & Data Analysis
  - Google
    - Mapreduce
  - Teradata Aster
    - SQL MapReduce
  - Apache
    - Hive
      - HiveQL
    - Pig
      - Pig Latin
    - Hadoop
  - Cloudera
    - CDH
  - EMC
    - Greenplum Chorus
    - Greenplum HD
  - IBM
    - InfoSphere BigInsights
  - Amazon
    - Elastic MapReduce
  - Hortonworks
    - Data Platform
  - MapR Technologies
    - MapR
  - Microsoft
    - HDInsight Azure
    - HDInsight Windows
- Distributed File Systems
  - Kosmix
    - CloudStore (KFS)
  - The Sector Alliance
    - Sector/Sphere
  - Red Hat, Inc.
    - GlusterFS
  - Apache
    - HDFS
  - Google
    - GFS
- Databases
  - Apache
    - HBase
    - Cassandra
    - Accumulo
  - Google
    - Bigtable
  - EMC
    - Greenplum
  - 10Gen, Inc.
    - MongoDB
  - Couchbase
    - Couchbase
  - Amazon
    - SimpleDB
  - Hypertable, Inc
    - Hypertable
  - Microsoft
    - MSSQL 2012
- Streaming/CEP
  - IBM
    - InfoSphere Streams
  - Twitter
    - Storm
  - Apache
    - S4
  - Microsoft
    - StreamInsight
  - EsperTech
    - Esper
- ADV
  - Tableau Software
    - Desktop
    - Server
  - SAS
    - Enterprise Business Intelligence
  - Information Builders
    - WebFOCUS
  - Khronos Group
    - OpenGL
    - WebGL
  - IBM
    - Cognos Business Intelligence
  - Tibco Software
    - Spotfire Analytics
  - Oracle
    - Business Intelligence Suite Enterprise Edition

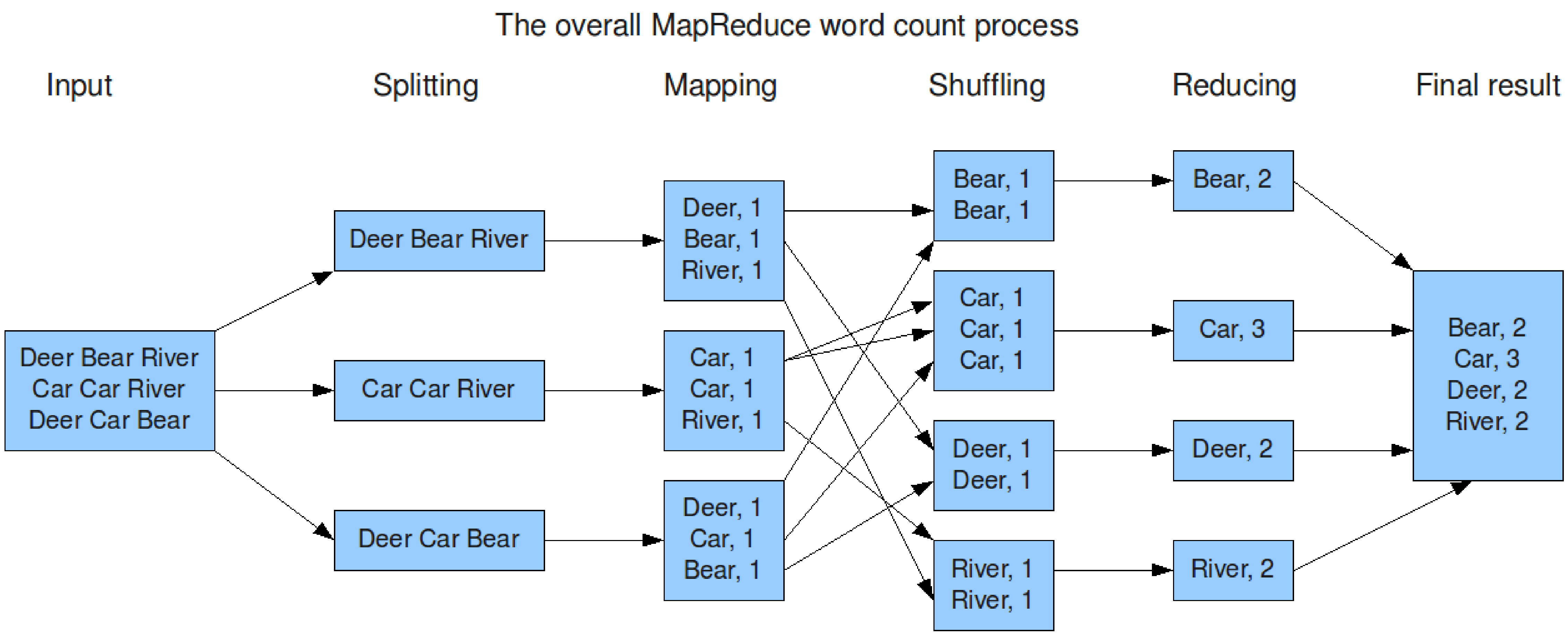
Software for Big Data

- Common Software

- Machine Learning Lib

- Faster Framework

Map & Reduce



The overall MapReduce word count process

Tracking Drinking Behavior from Twitter Data

Twitter Data

3/27/2014

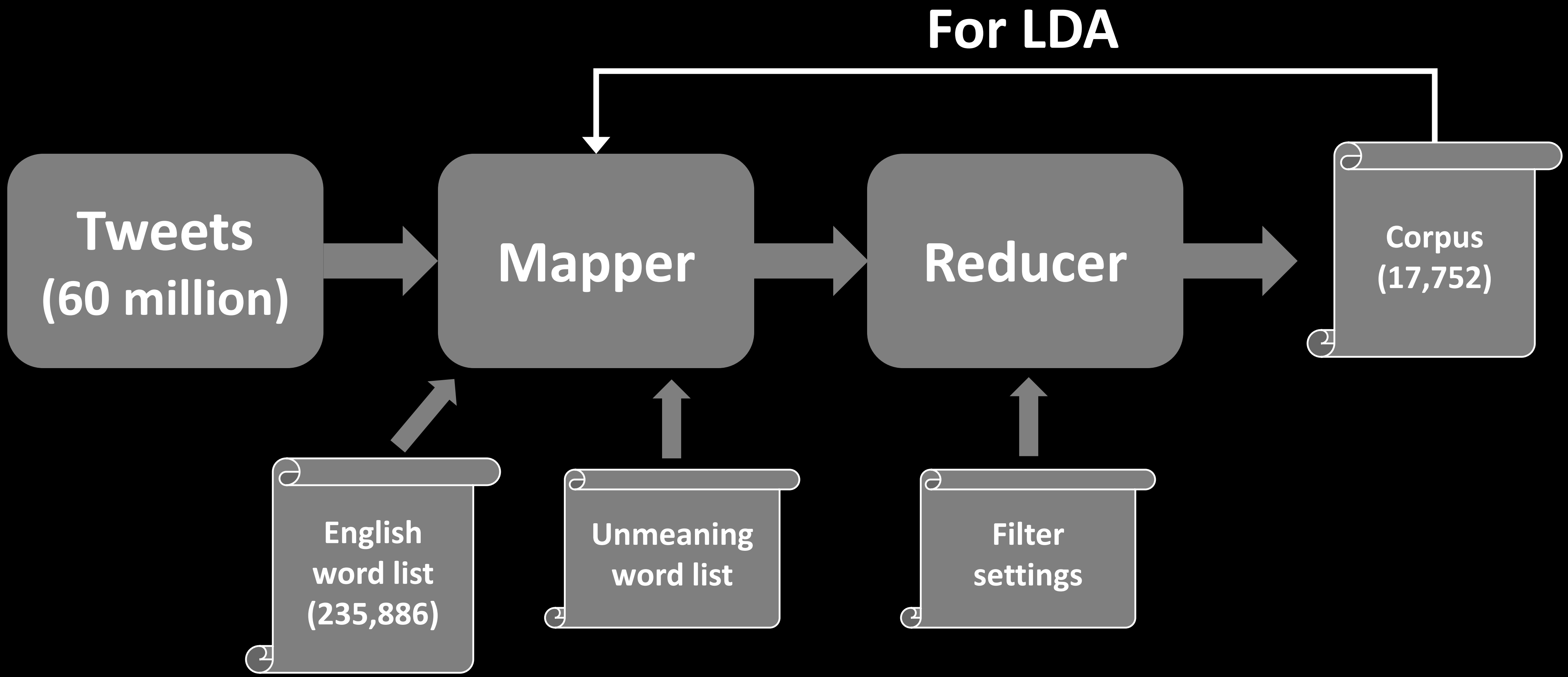
60, 000, 000 Tweets

Drinking

11,255,207 Tweets

5/1/2014

LDA + Hadoop → Drinking Words



For LDA

Tweets (60 million)

Mapper

Reducer

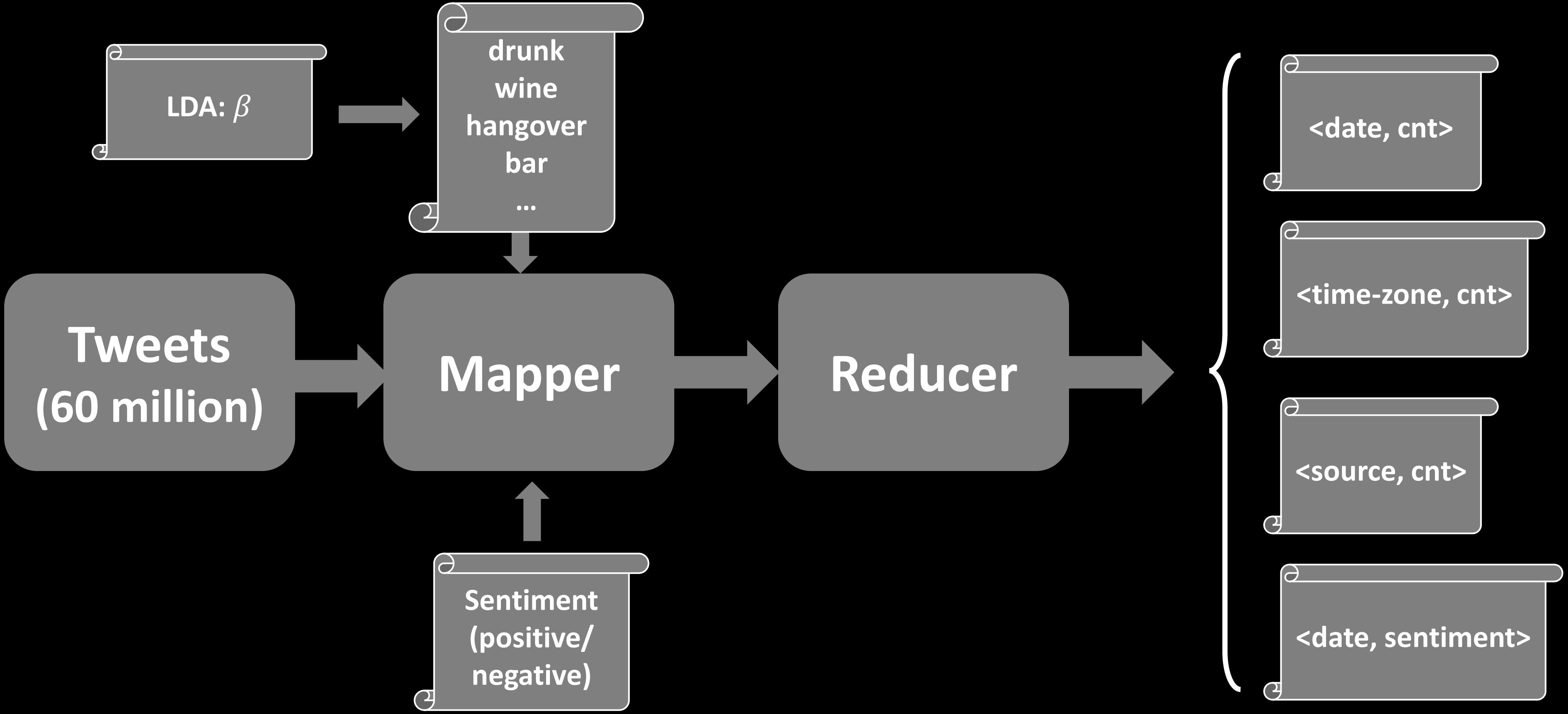
Corpus (17,752)

English word list (235,886)

Unmeaning word list

Filter settings

Hadoop → Drinking Tweets



LDA: β

drunk
wine
hangover
bar
…

Tweets
(60 million)

Mapper

Reducer

Sentiment
(positive/
negative)

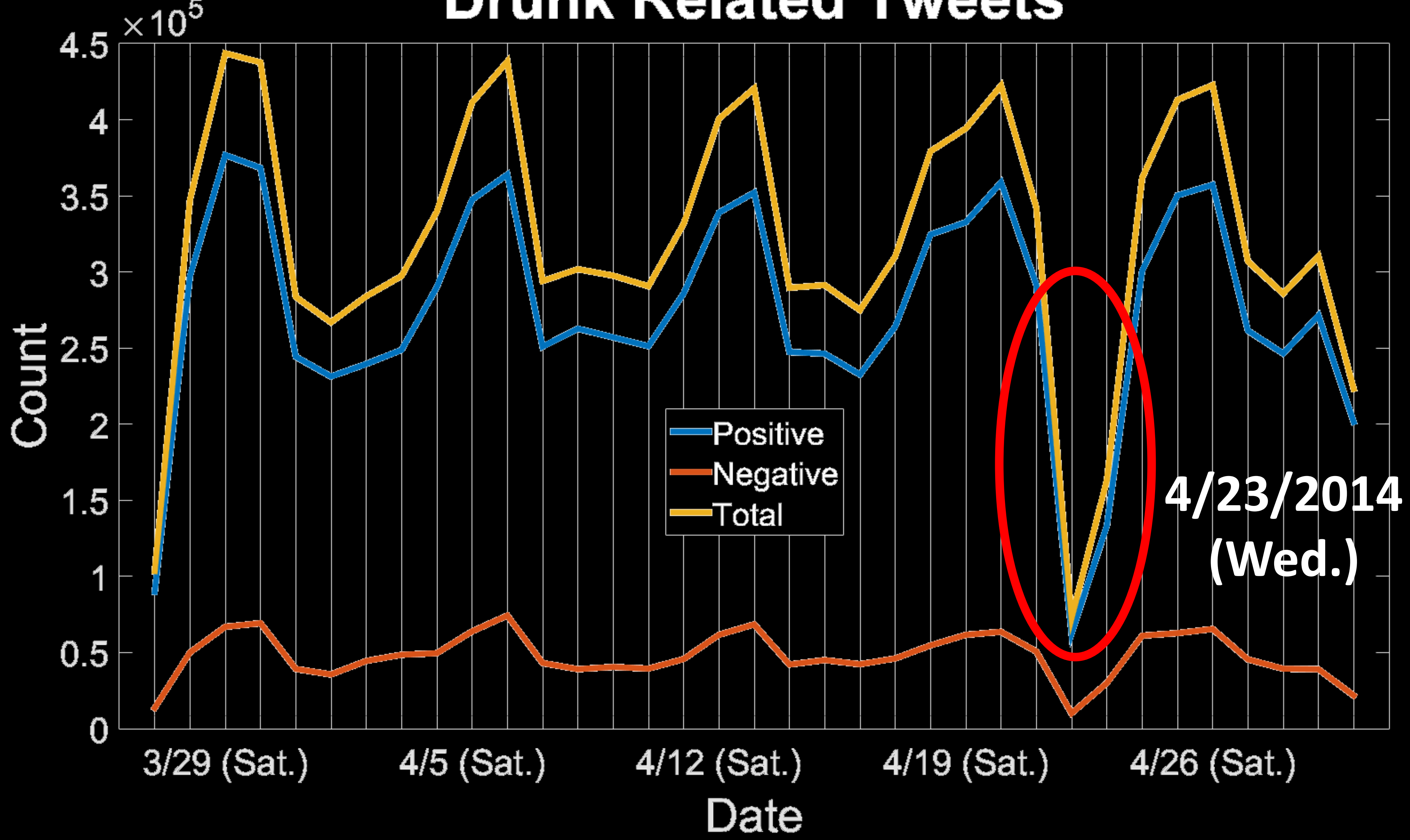
<date, cnt>

<time-zone, cnt>

<source, cnt>

<date, sentiment>

Drunk Related Tweets



4/23/2014
(Wed.)

What Happened around Apr. 23rd, 2014 ?

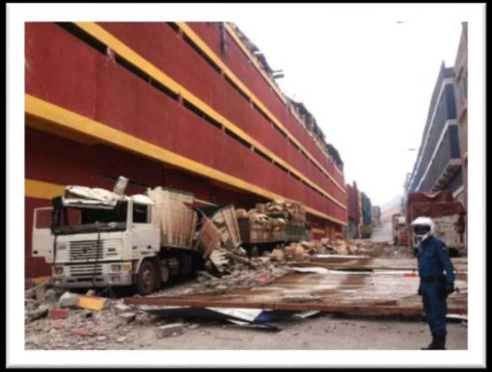
Google

Unrest in Egypt

Car bomb in Kenya, killing 4 people

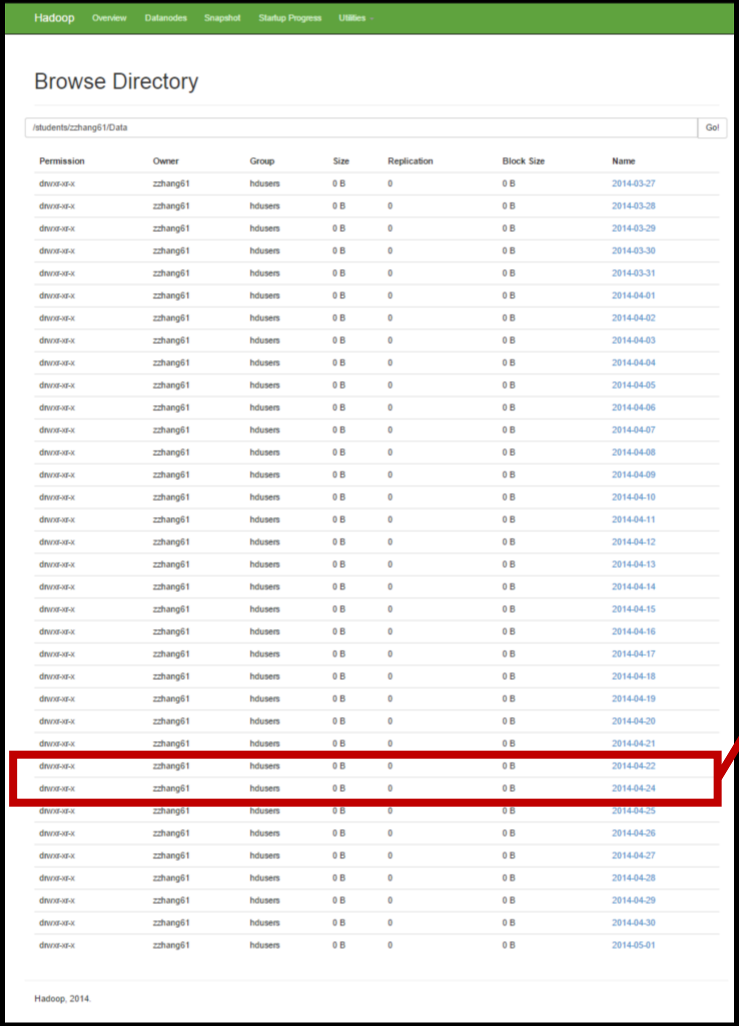
Sinking of the South Korean ferry
killed 304 passengers

Unrest in Ukraine



6.6M earthquake in Canada

The Truth is …

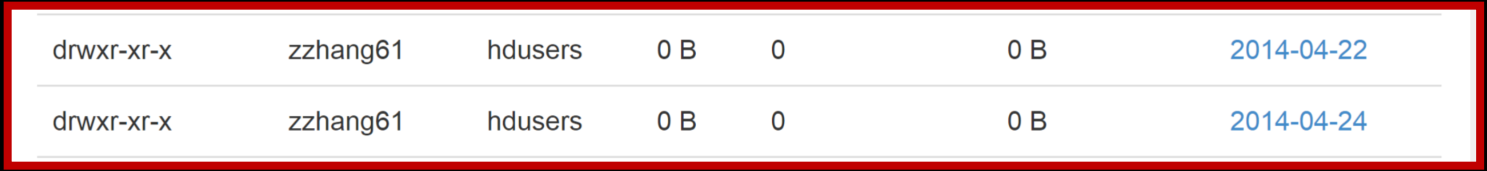


Browse Directory

/students/zzhang61/Data

Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-03-27
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-03-28
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-03-29
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-03-30
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-03-31
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-01
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-02
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-03
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-04
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-05
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-06
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-07
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-08
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-09
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-10
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-11
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-12
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-13
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-14
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-15
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-16
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-17
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-18
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-19
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-20
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-21
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-22
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-24
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-25
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-26
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-27
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-28
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-29
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-30
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-05-01

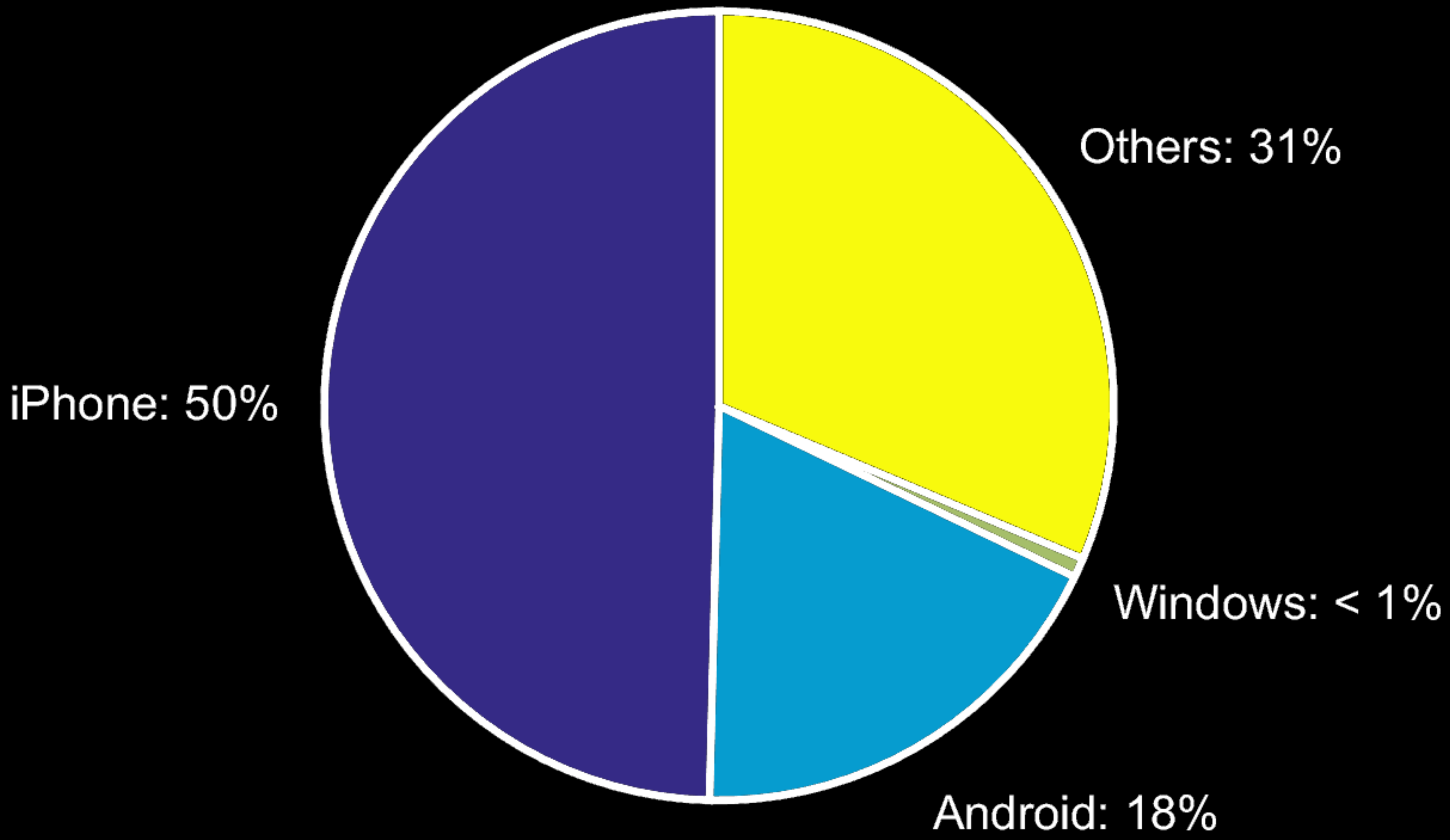
Hadoop, 2014.



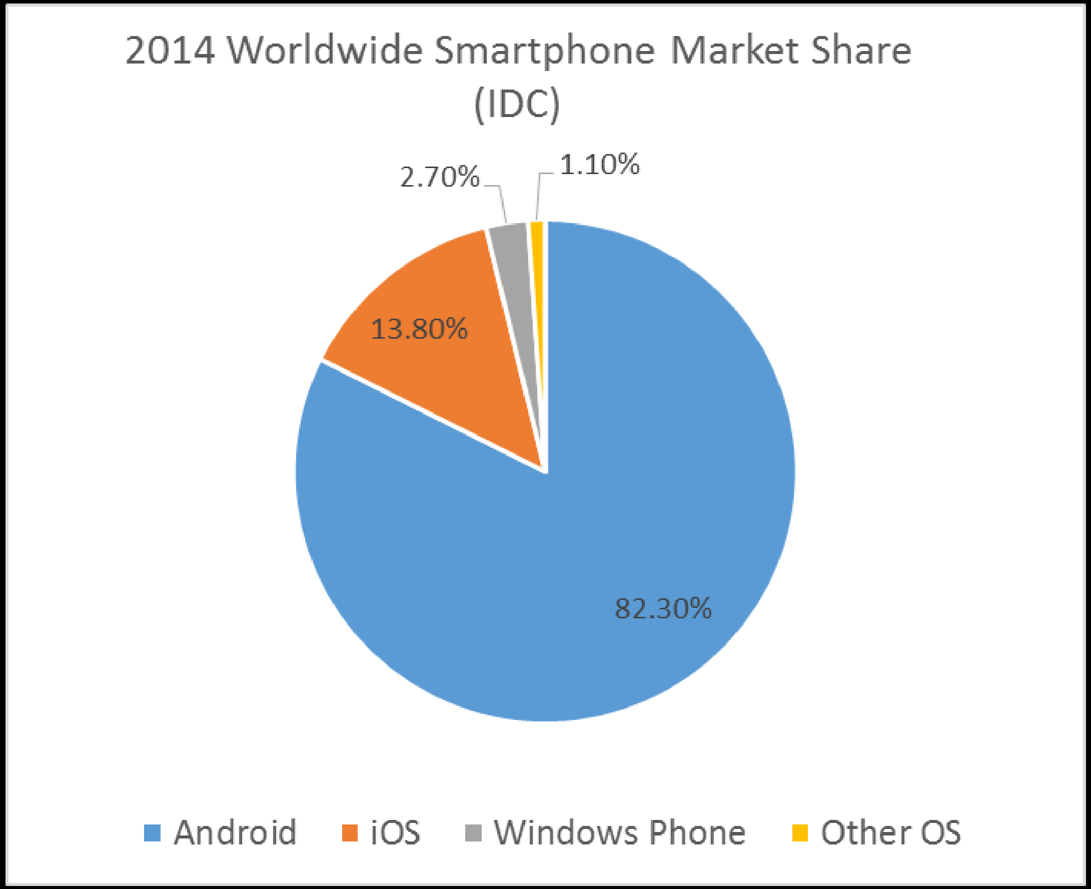
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-22
drwxr-xr-x	zzhang61	hdusers	0 B	0	0 B	2014-04-24

There is no data on 4/23

Source of Drunk Related Tweets

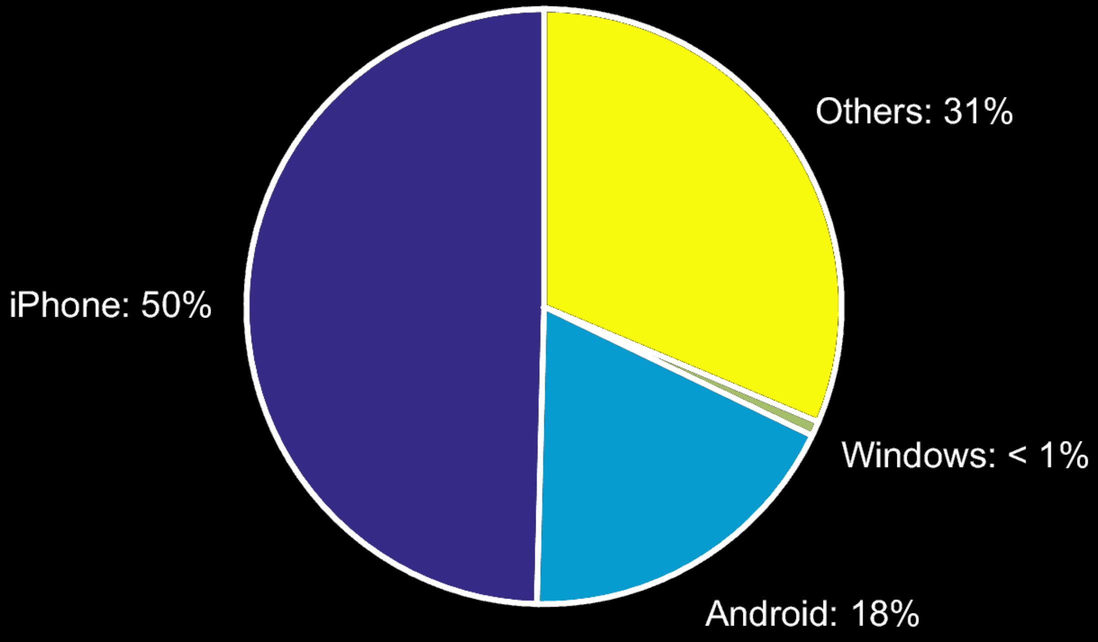


Drinkers Love iPhone More?

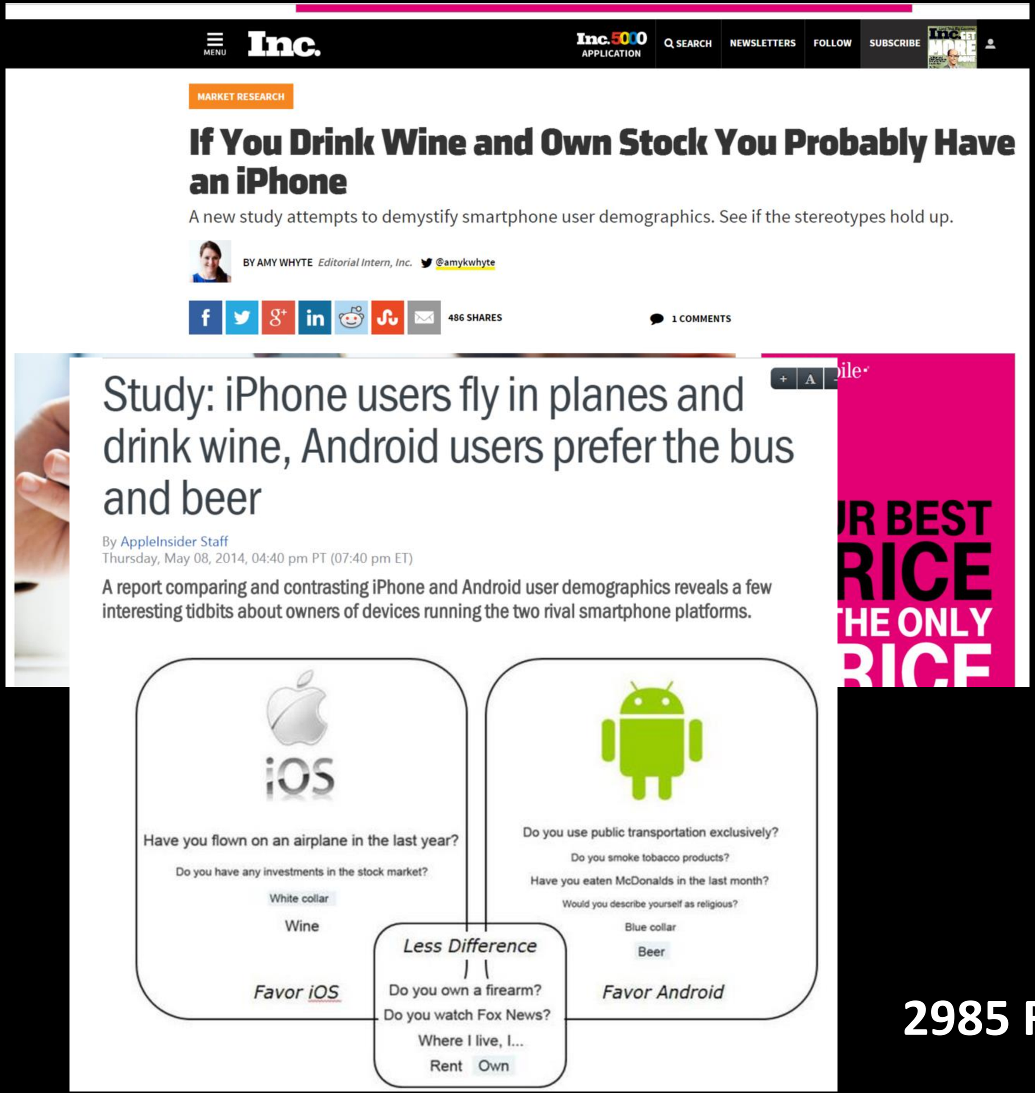


2014 Worldwide Smartphone Market Share (IDC)

Source of Drunk Related Tweets



Drinkers Love iPhone More?



MARKET RESEARCH

If You Drink Wine and Own Stock You Probably Have an iPhone

A new study attempts to demystify smartphone user demographics. See if the stereotypes hold up.

BY AMY WHYTE *Editorial Intern, Inc.* @amykwhyte

486 SHARES

1 COMMENTS

Study: iPhone users fly in planes and drink wine, Android users prefer the bus and beer

By AppleInsider Staff
Thursday, May 08, 2014, 04:40 pm PT (07:40 pm ET)

A report comparing and contrasting iPhone and Android user demographics reveals a few interesting tidbits about owners of devices running the two rival smartphone platforms.

iOS

Have you flown on an airplane in the last year?

Do you have any investments in the stock market?

White collar

Wine

Favor iOS

Do you use public transportation exclusively?

Do you smoke tobacco products?

Have you eaten McDonalds in the last month?

Would you describe yourself as religious?

Blue collar

Beer

Favor Android

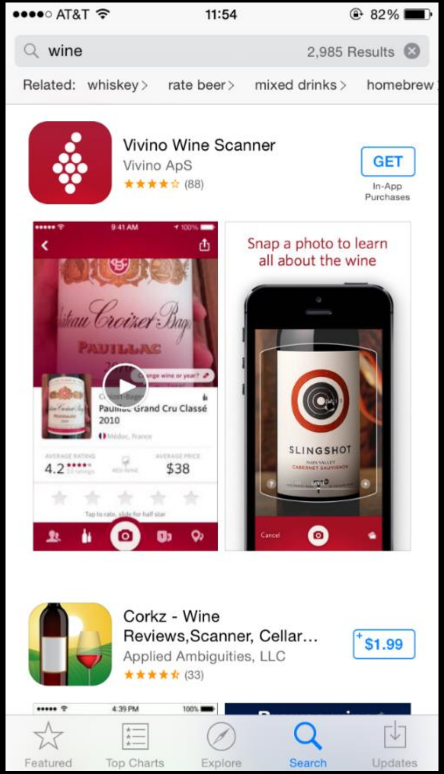
Less Difference

Do you own a firearm?

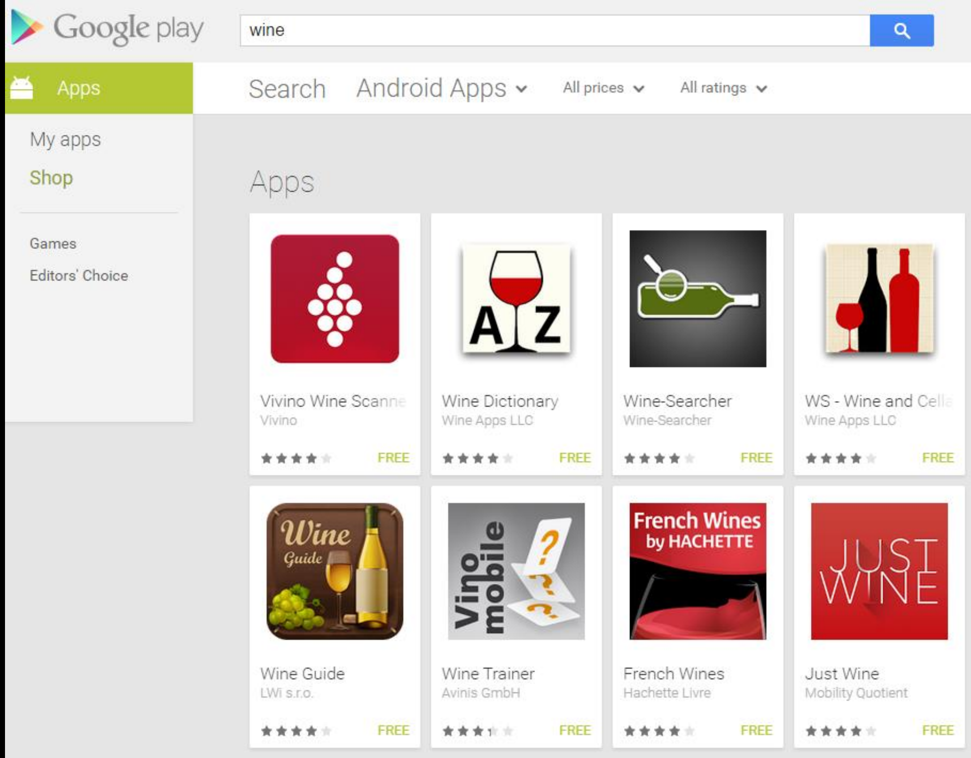
Do you watch Fox News?

Where I live, I...

Rent Own



2985 Results of “wine” in App Store

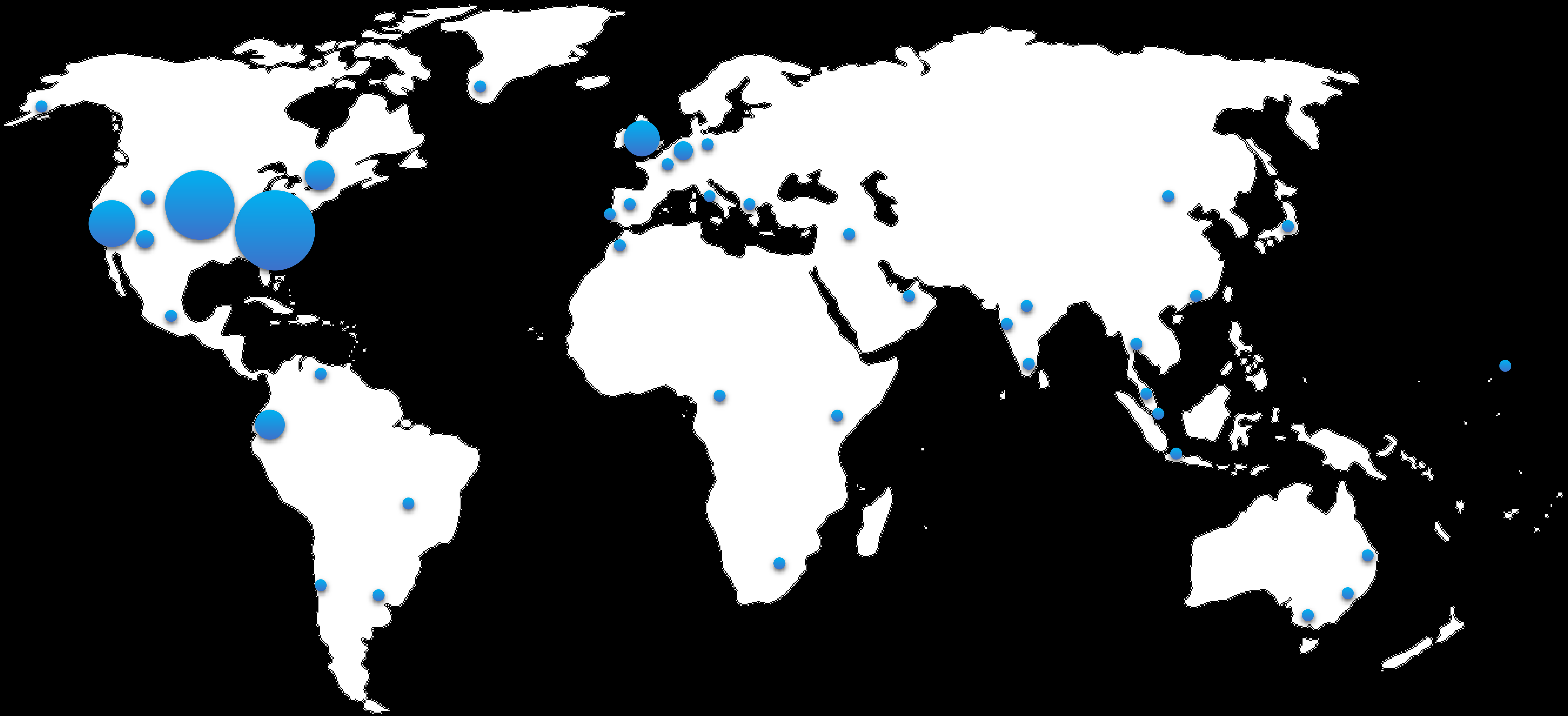


248 Results of “wine” in Google play

Location of Drinking Related Tweets

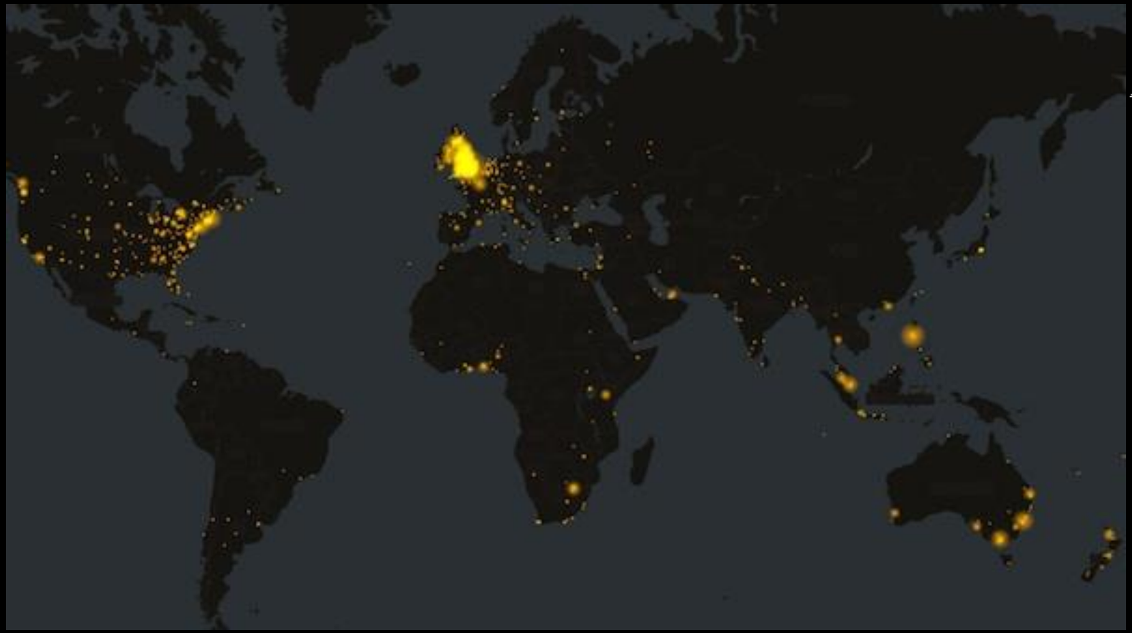
46 zones have over 10,000 drunk related tweets

'Eastern/Time/(US/&/Canada)'	1349324
'Central/Time/(US/&/Canada)'	1141305
'Pacific/Time/(US/&/Canada)'	783803
'London'	583376
'Atlantic/Time/(Canada)'	497300
'Quito'	490727
'Amsterdam'	325605
'Arizona'	291043
'Mountain/Time/(US/&/Canada)'	234895
'Hawaii'	198108
'Casablanca'	187121
'Alaska'	129952
'Athens'	102358
'Brasilia'	92126
'Beijing'	74283
'Greenland'	52823
'Chennai'	51933
'Bangkok'	49231
'Edinburgh'	44594
'Buenos/Aires'	43854
'Dublin'	38960
'Singapore'	38276
'Santiago'	37012
'Sydney'	31784
'Madrid'	31394
'Kuala/Lumpur'	29112
'Paris'	27678
'Pretoria'	26745
'Mexico/City'	26023
'Caracas'	25682

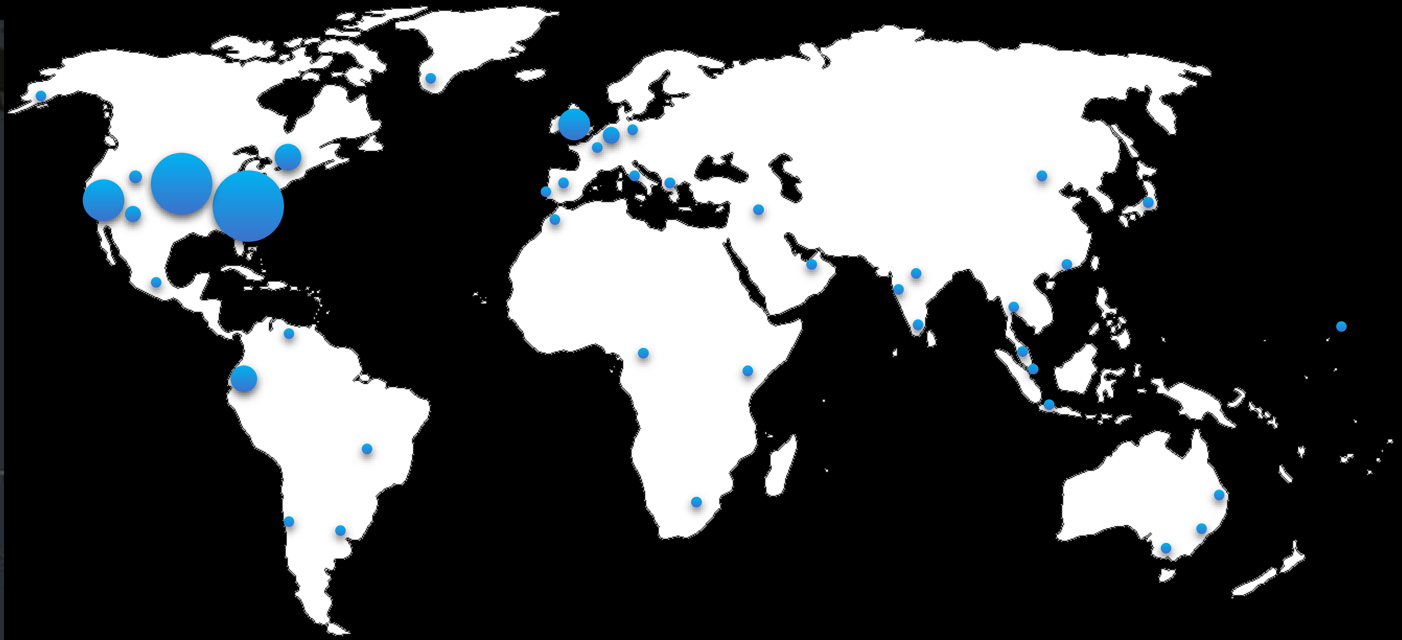


Zones with over 10,000 Drinking Related Tweets

American Drink More?



Twitter user distribution



Drinking Tweet distribution

Pros & Cons

Pros:

- Through big data mining, we may find something hard to be recognized in daily life.
- Identify effect of certain event on the public.

Cons:

- How to interpret the result? (Misleading)
- Only reflect pubic behavior but not for individuals.

THANK YOU!