Gnip Twitter Firehose and PowerTrack

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

March 12, 2013

Gnip firehose

Continuous stream

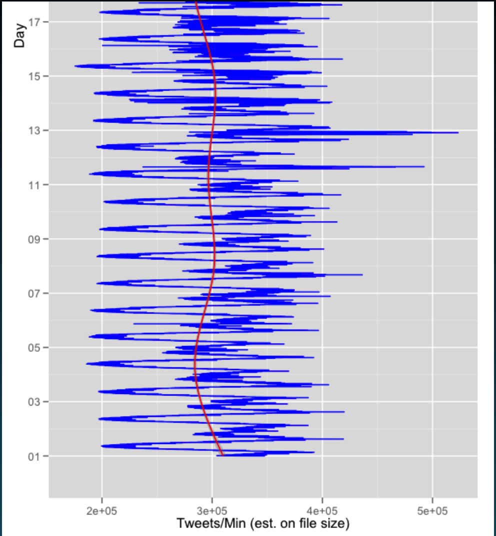
of JSON tweets

in near-real time

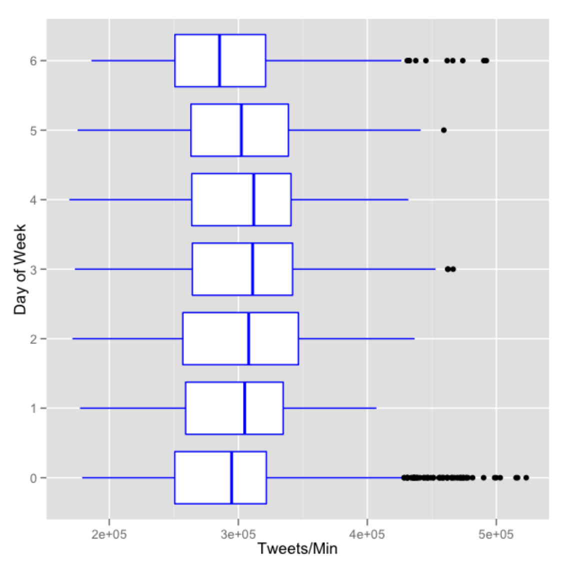
Example firehose volumes

Publisher	Daily Activity
Twitter	400M
Tumblr	75M
Wordpress Posts	615k
Wordpress Comments	1.1M
Disqus	1.3M
Engagement (likes, votes)	2.4M

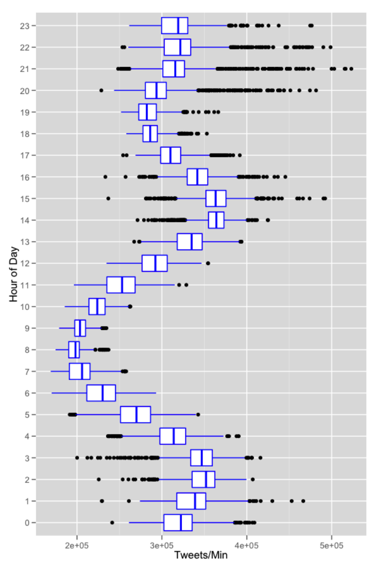
Twitter volumes – two weeks



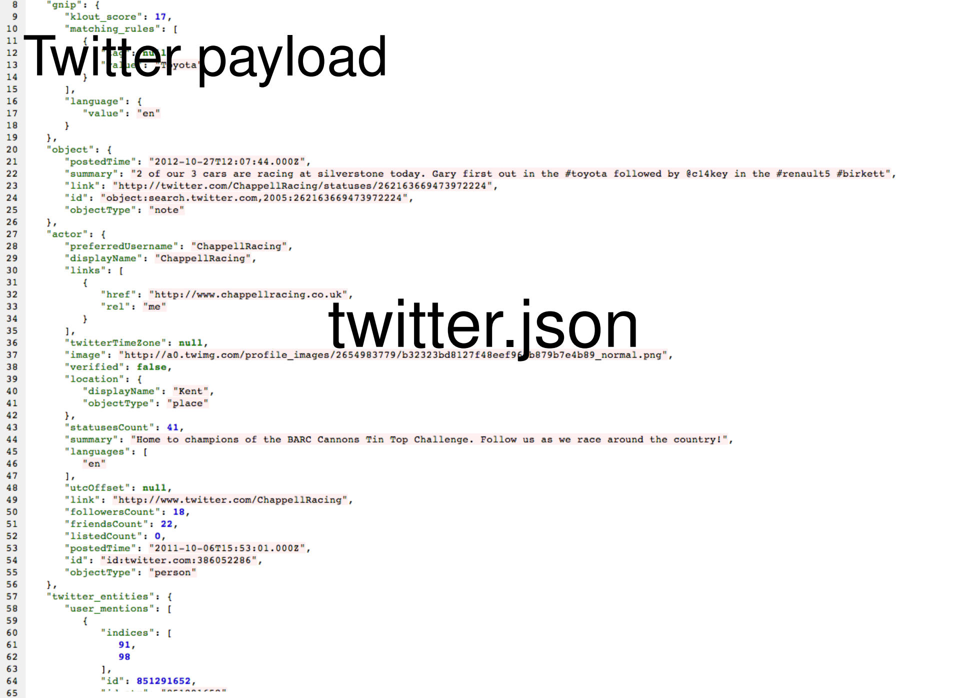
Twitter volumes - day of week



Twitter volumes - hour of day



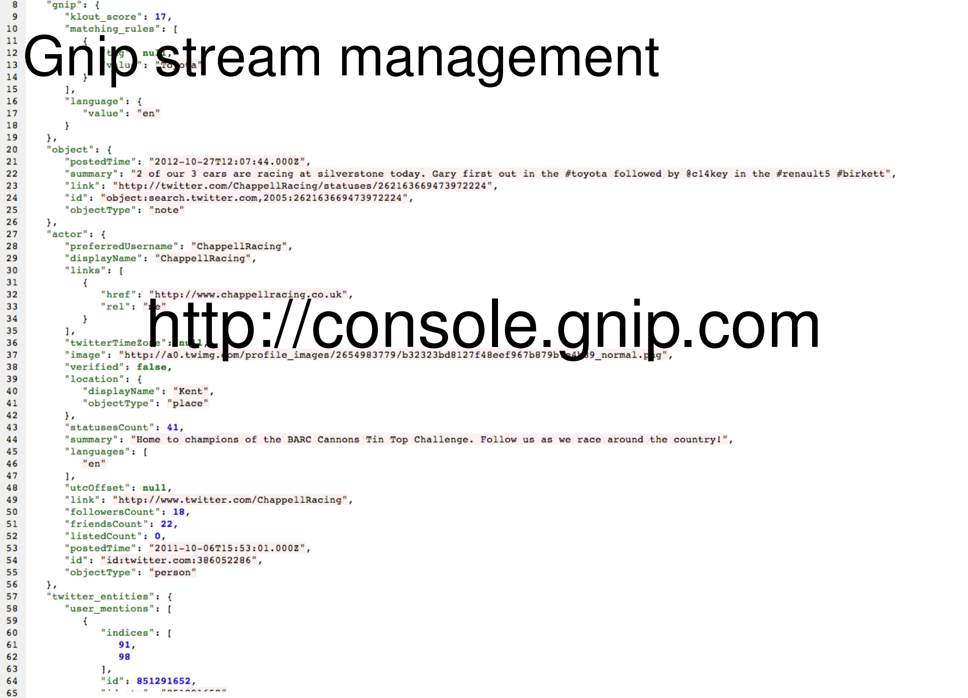
Twitter payload



twitter.json

```
   "gnip": {
      "klout_score": 17,
      "matching_rules": [
         {
            "tag": null,
            "value": "Toyota"
         }
      ],
      "language": {
         "value": "en"
      }
   },
   "object": {
      "postedTime": "2012-10-27T12:07:44.000Z",
      "summary": "2 of our 3 cars are racing at silverstone today. Gary first out in the #toyota followed by @cl4key in the #renault5 #birkett",
      "link": "http://twitter.com/ChappellRacing/statuses/262163669473972224",
      "id": "object:search.twitter.com,2005:262163669473972224",
      "objectType": "note"
   },
   "actor": {
      "preferredUsername": "ChappellRacing",
      "displayName": "ChappellRacing",
      "links": [
         {
            "href": "http://www.chappellracing.co.uk",
            "rel": "me"
         }
      ],
      "twitterTimeZone": null,
      "image": "http://a0.twimg.com/profile_images/2654983779/b32323bd8127f48eef96 b879b7e4b89_normal.png",
      "verified": false,
      "location": {
         "displayName": "Kent",
         "objectType": "place"
      },
      "statusesCount": 41,
      "summary": "Home to champions of the BARC Cannons Tin Top Challenge. Follow us as we race around the country!",
      "languages": [
         "en"
      ],
      "utcOffset": null,
      "link": "http://www.twitter.com/ChappellRacing",
      "followersCount": 18,
      "friendsCount": 22,
      "listedCount": 0,
      "postedTime": "2011-10-06T15:53:01.000Z",
      "id": "id:twitter.com:386052286",
      "objectType": "person"
   },
   "twitter_entities": {
      "user_mentions": [
         {
            "indices": [
               91,
               98
            ],
            "id": 851291652,
```

Gnip stream management



```
   "gnip": {
      "klout_score": 17,
      "matching_rules": [
         {
            "tag": null,
            "value": "Toyota"
         }
      ],
      "language": {
         "value": "en"
      }
   },
   "object": {
      "postedTime": "2012-10-27T12:07:44.000Z",
      "summary": "2 of our 3 cars are racing at silverstone today. Gary first out in the #toyota followed by @cl4key in the #renault5 #birkett",
      "link": "http://twitter.com/ChappellRacing/statuses/262163669473972224",
      "id": "object:search.twitter.com,2005:262163669473972224",
      "objectType": "note"
   },
   "actor": {
      "preferredUsername": "ChappellRacing",
      "displayName": "ChappellRacing",
      "links": [
         {
            "href": "http://www.chappellracing.co.uk",
            "rel": "me"
         }
      ],
      "twitterTimeZone": null,
      "image": "http://a0.twimg.com/profile_images/2654983779/b32323bd8127f48eef967b879b7e4b89_normal.png",
      "verified": false,
      "location": {
         "displayName": "Kent",
         "objectType": "place"
      },
      "statusesCount": 41,
      "summary": "Home to champions of the BARC Cannons Tin Top Challenge. Follow us as we race around the country!",
      "languages": [
         "en"
      ],
      "utcOffset": null,
      "link": "http://www.twitter.com/ChappellRacing",
      "followersCount": 18,
      "friendsCount": 22,
      "listedCount": 0,
      "postedTime": "2011-10-06T15:53:01.000Z",
      "id": "id:twitter.com:386052286",
      "objectType": "person"
   },
   "twitter_entities": {
      "user_mentions": [
         {
            "indices": [
               91,
               98
            ],
            "id": 851291652,
            "id_str": "851291652"
```

http://console.gnip.com

Streaming data from the firehose

Curl the firehose

```
curl --compressed -v \
-ushendrickson@gnip.com  \
"https://stream.gnip.com:443/accounts/shendrickson/
    publishers/twitter/streams/track/track2.json"
```

More curling the firehose

```
curl --compressed -s \
-ushendrickson@gnip.com:<password>  \
"https://stream.gnip.com:443/accounts/shendrickson/
   publishers/twitter/streams/track/track2.json" \
-o outfile.json
```

PowerTrack: filter and shape

- core idea: exact token matches (e.g. "obama", "beer" …)
- non-token matches: “happy birthday” and “contains:dog”
- meta-data operators: geo, language, bios, …
- shaping operators: (e.g. “sample:10” gives 10%)
- operators: (by publisher) user, hashtag, language,…
- filter on 100% of the firehose

PowerTrack: combining rules

$$\begin{aligned} newline &= OR \\ space &= AND \\ “OR'' &= OR \\ “-'' &= NOT \\ “(\ldots)'' &= grouping \end{aligned}$$

PowerTrack: rule limits

- A single PowerTrack rule may contain up to 10 positive clauses, and up to 50 negative clauses
- A single PowerTrack rule may not have more than 1024 characters, including OR operators and parentheses
- Max 250K rules

Example Toyota PowerTrack rules

```
(sexy OR speed OR speeding OR \"sport utility\" OR ...
    suv OR toyota) (infiniti OR infinitis OR #infiniti OR @infiniti) -job
    -\"<money>\" -\"<phone>\" -jobs -deal -review -#jobs -tattoo
    -giveaway -deals -discount -reviews -#job -jewelry -jewelry
@toyotacanada sample:40
lang:en toyota recall
lang:it toyota window
lang:fr toyota recall
lang:en toyota auris -crime -lease -sells -thief -police -robbed -robber
lang:ru toyota dyna -lkw -aqua -bail -died -film -toka -camry
```

Enterprise PowerTrack features

- Update individual rules without disconnect ($\lesssim 1s$ update time for 100s of rules)
- Rule tagging
- Keep alive - signal that connection is live, even when no data is coming (30 s)
- Low latency: avg 1s Twitter raw; 10s Twitter enriched
- Redundancy - multiple simultaneous connections available
- Backfill - buffer data and fill in if short term disconnect
- PowerTrack Replay - connect with start and end dates to stream past time periods (<5 days)
- Historical PowerTrack - Twitter historical filtering for any time period

Rule JSON

```
{
  "rules": [
    {
      "tag": "presidents",
      "value": "obama"
    },
    {
      "tag": "musicians",
      "value": "gaga"
    },
    {
      "tag": "musicians",
      "value": "bieber"
    }
  ]
}
```

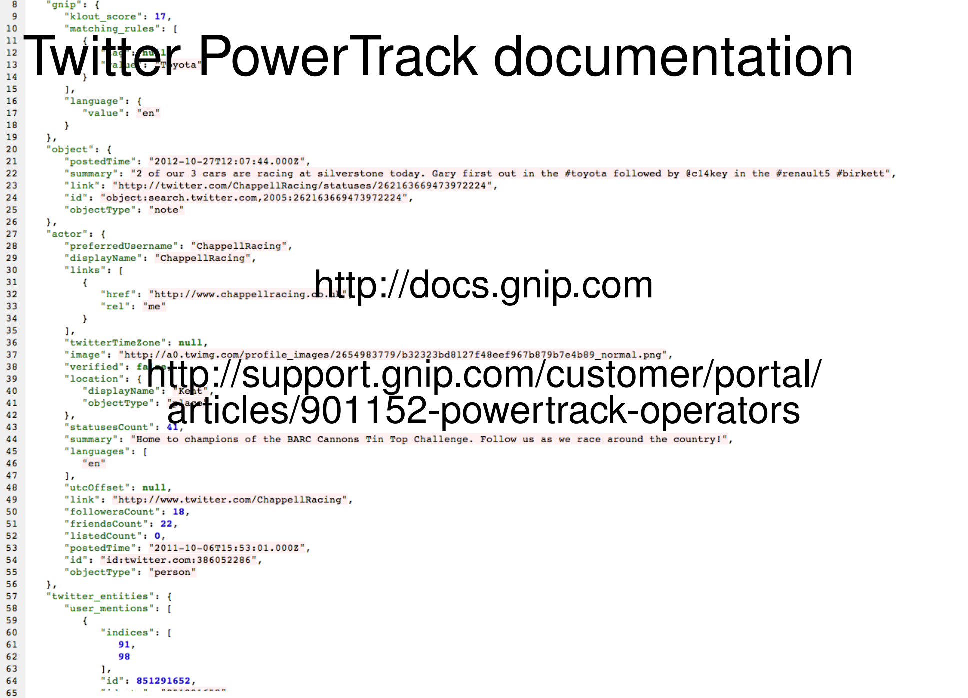
Rules REST API

- POST (add rules)
- DELETE (rule match by value)
- GET (rule list)
- UPDATE pattern: GET, (alter rule), ADD, DELETE
- https://api.gnip.com:443/accounts/shendrickson/publishers/twitter/streams/track/track2/rules.json

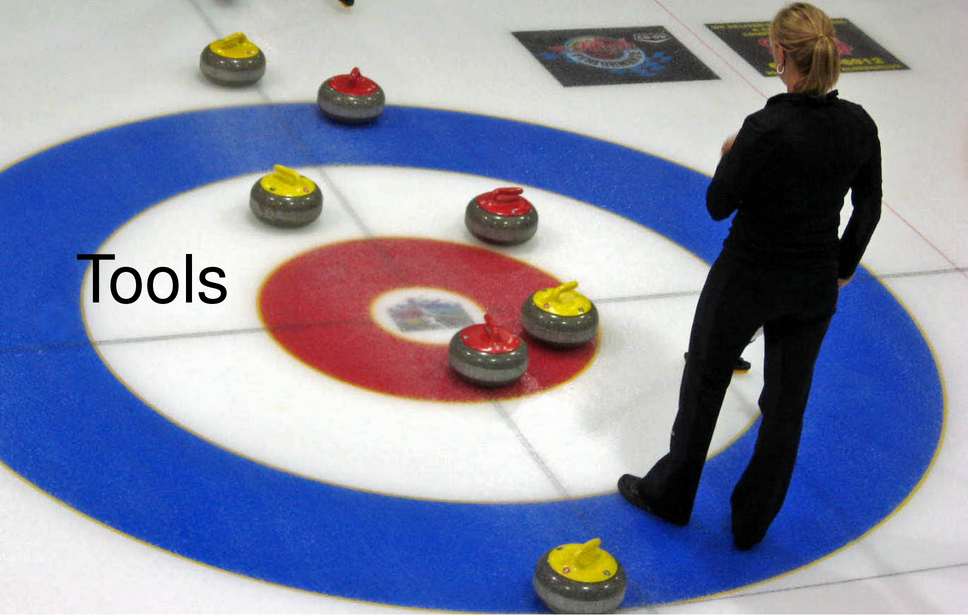
Twitter PowerTrack documentation

http://docs.gnip.com

http://support.gnip.com/customer/portal/articles/901152-powertrack-operators



```
   "gnip": {
      "klout_score": 17,
      "matching_rules": [
         {
            "tag": null,
            "value": "Toyota"
         }
      ],
      "language": {
         "value": "en"
      }
   },
   "object": {
      "postedTime": "2012-10-27T12:07:44.000Z",
      "summary": "2 of our 3 cars are racing at silverstone today. Gary first out in the #toyota followed by @cl4key in the #renault5 #birkett",
      "link": "http://twitter.com/ChappellRacing/statuses/262163669473972224",
      "id": "object:search.twitter.com,2005:262163669473972224",
      "objectType": "note"
   },
   "actor": {
      "preferredUsername": "ChappellRacing",
      "displayName": "ChappellRacing",
      "links": [
         {
            "href": "http://www.chappellracing.co.uk",
            "rel": "me"
         }
      ],
      "twitterTimeZone": null,
      "image": "http://a0.twimg.com/profile_images/2654983779/b32323bd8127f48eef967b879b7e4b89_normal.png",
      "verified": false,
      "location": {
         "displayName": "Kent",
         "objectType": "place"
      },
      "statusesCount": 41,
      "summary": "Home to champions of the BARC Cannons Tin Top Challenge. Follow us as we race around the country!",
      "languages": [
         "en"
      ],
      "utcOffset": null,
      "link": "http://www.twitter.com/ChappellRacing",
      "followersCount": 18,
      "friendsCount": 22,
      "listedCount": 0,
      "postedTime": "2011-10-06T15:53:01.000Z",
      "id": "id:twitter.com:386052286",
      "objectType": "person"
   },
   "twitter_entities": {
      "user_mentions": [
         {
            "indices": [
               91,
               98
            ],
            "id": 851291652,
            "id_str": "851291652"
```



Tools

Simple parser: TWitterACTivitieS

- core idea: use twacs to parse common twitter elements to pipe-delimited (flat) structure
- requires: Python
- github: https://github.com/DrSkippy27/Twacs
- From PyPi: sudo pip install twacs

twacs.py examples - prettifier

```
> gzip -cd twitter_oneDay_onePercent.json.gz | twacs-prettifier.py
{
  "body": "Giving them 1 inch and they take a mile",
  "retweetCount": 0,
  "generator": {
    "link": "http://twitter.com/download/iphone",
    "displayName": "Twitter for iPhone"
  },
  "gnip": {
    "klout_score": 47,
    "language": {
      "value": "en"
    }
  },

…
```

twacs.py examples - basic parse

```
> gzip -cd twitter_oneDay_onePercent.json.gz | twacs.py
tag:search.twitter.com,2005:309063808016584704|
      2013-03-05T22:12:08.000Z|
      Giving them 1 inch and they take a mile
tag:search.twitter.com,2005:309063808041771008|
      2013-03-05T22:12:08.000Z|
    @luizaaguiarb sÃ³ se for o teu filho! O meu vai ser super higienizado e cheiros
tag:search.twitter.com,2005:309063808331153409|
      2013-03-05T22:12:08.000Z|
      @SlyOuu Mdrrr le negro s'emballe les coquilles
tag:search.twitter.com,2005:309063808427622402|
      2013-03-05T22:12:08.000Z|
      RT @NotARapistHere: My favorite pickup line: Get in the van.
 Audsbgivasiugbasdpiub

…
```

twacs.py examples - help

```
>twacs.py -h
Usage: twacs.py [options]

Options:
  -h, --help       show this help message and exit
  -g, --geo        Include geo fields
  -u, --user       Include user fields
  -r, --rules      Include rules fields
  -s, --urls       Include urls fields
  -l, --lang       Include language fields
  -p, --pretty     Pretty JSON output of full records
  -c, --csv        Comma-delimited output (default is | without quotes)
 -x, --explain     Show field names in output for for sample input records
  -i, --influence  Show user’s influence metrics
```

Curling and parsing the firehose

```
curl --compressed -s \
-ushendrickson@gnip.com:<password>  \
"https://stream.gnip.com:443/accounts/shendrickson/
    publishers/twitter/streams/track/track2.json" | twacs.py
```

Rules management

- core idea: use to list, delete, add and update rules
- requires: Python
- library and command line utilities
- github: https://github.com/DrSkippy27/Gnip-Python-PowerTrack-Rules

Twitter stream attributes

inReplyTo data element

```
{
 "body": "@rachelschadd @kylefraley3 @toritabin don’t read, just tweet!",
  "inReplyTo": {
  "link": "http://twitter.com/rachelschadd/statuses/309064691186020352"
  }

{
  "body": "@stonesy10 clearly but Madrid can!! Arsenal won’t have to
          worry bout that next season though",
  "inReplyTo": {
  "link": "http://twitter.com/stonesy10/statuses/309063725309104128"
  }
```

Retweets

- about 17% of twitter activities are retweets
- convention “RT …” added by many clients to text
- unattributed quoting

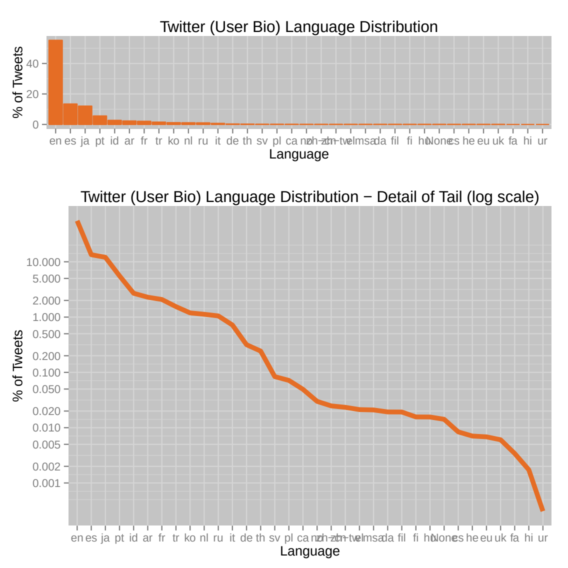
```
{
  "body": "RT @UberBullshit: Snoop Dogg changed his name to Snoop
        Lion after losing a bet in which he was out-smoked by Justin Bieber.",
   "retweetCount": 1979,
...
```

Retweets link

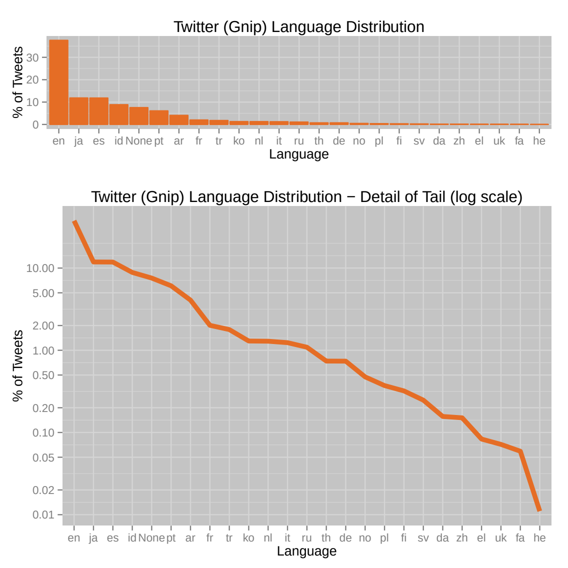
```
   "object": {
     "postedTime": "2013-03-05T22:08:54.000Z",
     "summary": "Fergie, that was for Jonjo Shelvey.",
  "link": "http://twitter.com/KopiteKru/statuses/309062995366010882"
...
```

http://twitter.com/KopiteKru/statuses/309062995366010882

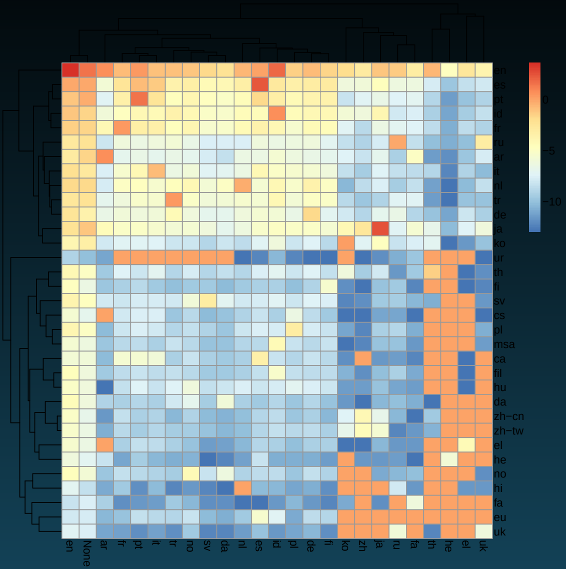
Twitter bio langauges



Twitter tweet (Gnip) languages



Twitter bio vs. tweet languages

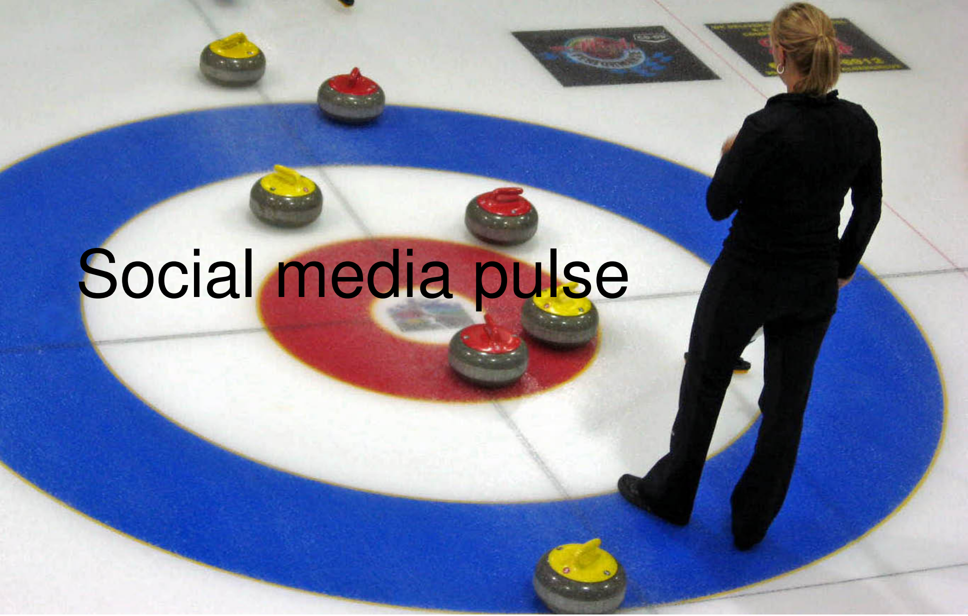


Geo information in tweets

```
[ tweet location] | Tweet location place id’d by twitter | User bio location
--------------------------------------------------------------------------------
[’41.02117698’, ’-73.8731331’]|Mercy College, Dobbs Ferry|US|NYC
[’-7.54556’, ’110.82484’]|Banjarsari, Surakarta|ID|Indonesia
[’51.7541896’, ’-0.34086304’]|Saint Albans, Hertfordshire|GB|St Albans
[’51.8446547’, ’4.3364468’]|Spijkenisse|NL|DEDICATED FOR LIFE
[’18.22484423’, ’-65.9027102’]|Ceiba Norte, PR|US|Juncos
[’40.21630994’, ’28.96884114’]|TÃ¼rkiye|TR|Erdek /Bursa
[’36.89167243’, ’30.67495879’]|TÃ¼rkiye|TR|big drummer
[’-6.2590775’, ’106.868624’]|Kramat Jati, Jakarta Timur|ID|Random
```

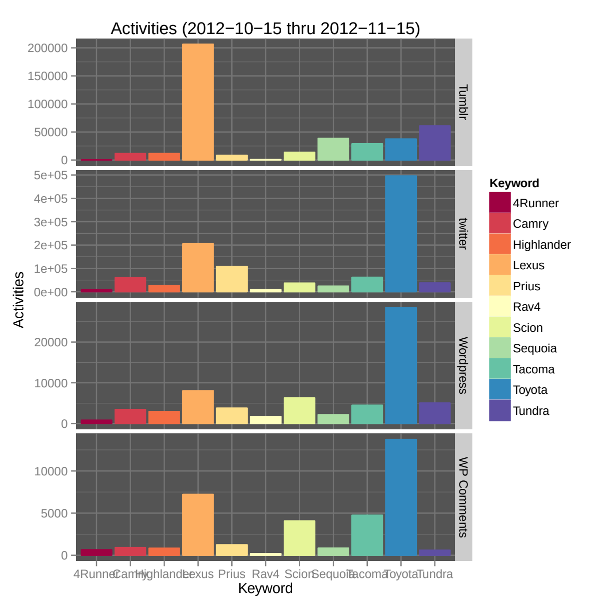
Geo location of Tweets

Type	Precision	Frequency
Geo Tagged: (Lat, Long) "Point"	High	1.235%
Geo Tagged: (Lat, Long) points "Polygon"	Medium-Low	1.418%
–	With either Point, Polygon or Both	1.596%
Country Code	Medium	1.43%
User Bio Place	High (long, lat)-Low (gibberish)	57.67%
Timezone Offset	Medium-Low	73.6%



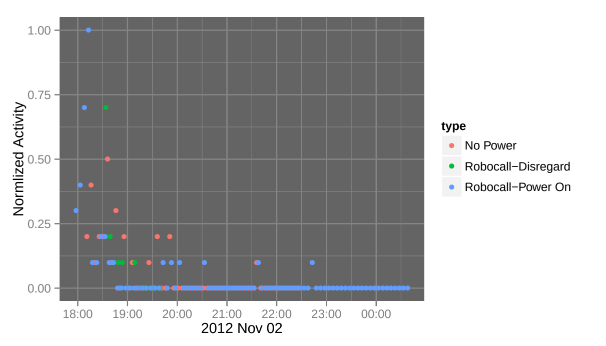
Social media pulse

Audience & perspective, Timing

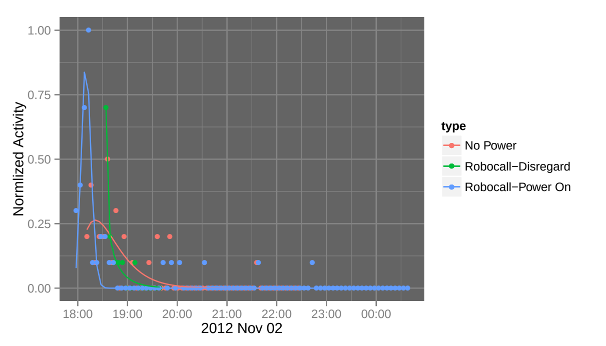

Activities (2012−10−15 thru 2012−11−15)
Activities
0
50000
100000
150000
200000
Tumblr
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
twitter
0
10000
20000
Wordpress
0
5000
10000
WP Comments
4Runner
Camry
Highlander
Lexus
Prius
Rav4
Scion
Sequoia
Tacoma
Toyota
Tundra
Keyword
Keyword
4Runner
Camry
Highlander
Lexus
Prius
Rav4
Scion
Sequoia
Tacoma
Toyota
Tundra

noise or signal?

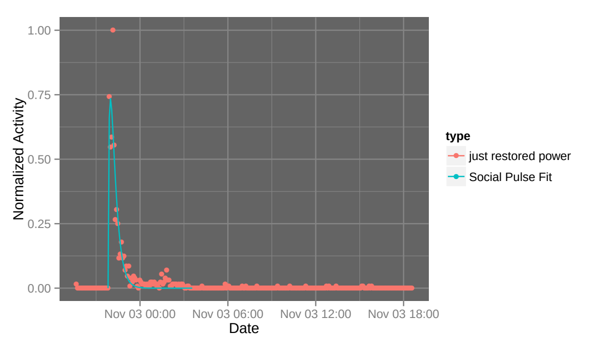
Sandy – Chelsea Power Outage



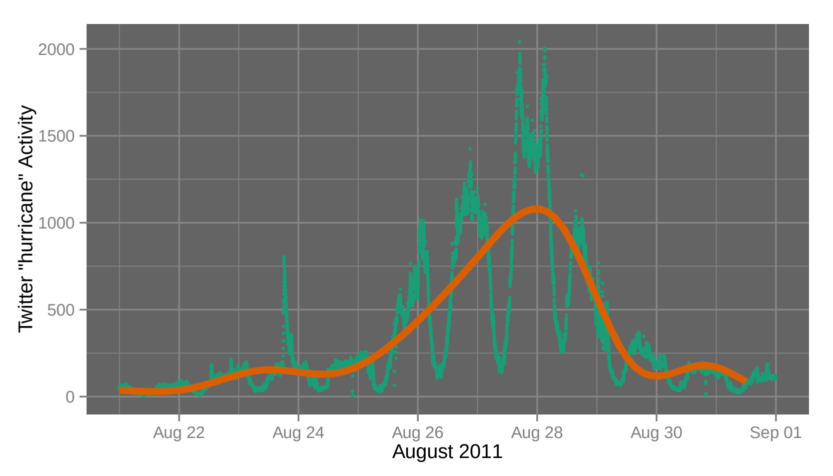
Better Statistical Model



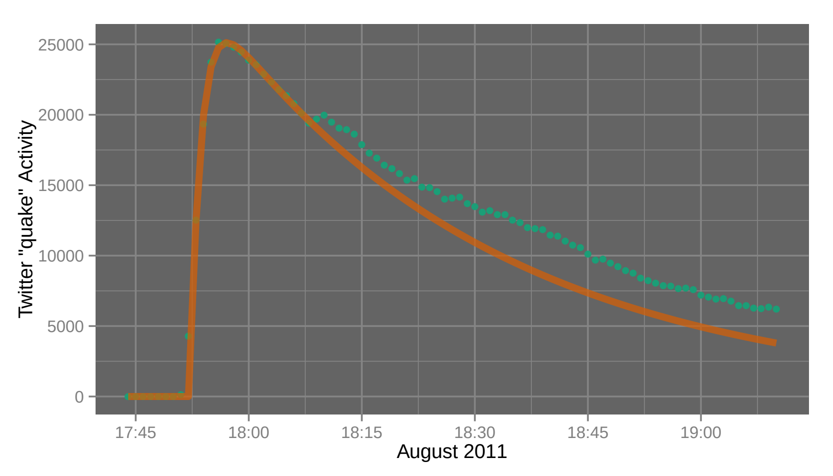
Real event has much higher volume



Expected: Hurricane



Unexpected: Earthquake



Classifying Events

Type	Response	Examples
Expected	Approx. Symmetric	Hurricane Sandy Olympics
Unexpected (many obs.)	Social Media Pulse	Beyoncé VMAs Mexico earthquake Steve Jobs
Unexpected (spread)	Network Models	Osama bin Laden Whitney Houston Syrian dissidents

Half-life

time to observe

$\frac{1}{2}$ of the activities

for an event

Social media pulse

Given an event, the probability of a activity from one person,

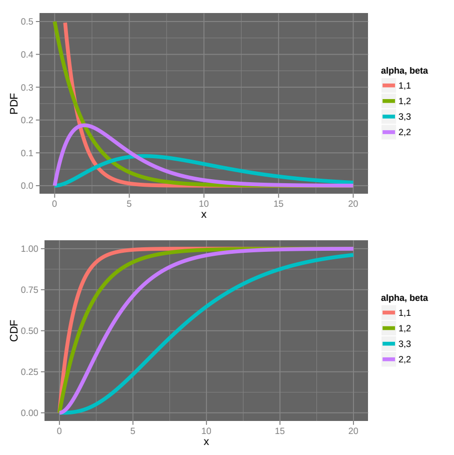
$$f(t) = \lambda \exp(-\lambda t), \text{ for } t \geq 0.$$

Many people posting, so sum of random variables $S = X_1 + X_2 + \ldots + X_{n\text{ posters}}$.
Probability distribution function,

$$f_S(t) = \frac{\beta^{-\alpha} t^{\alpha-1} \exp(\frac{-t}{\beta})}{\Gamma(\alpha)}$$

Cumulative distribution is the “generalized regularized incomplete gamma function”,

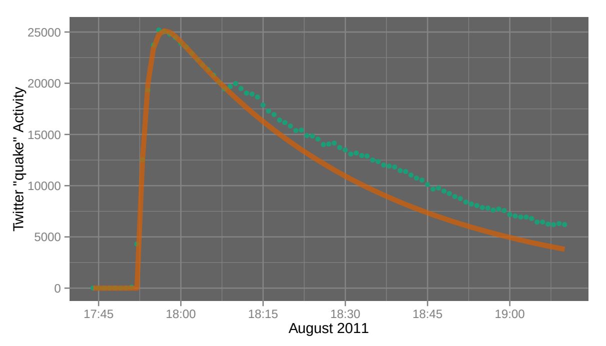
$$F_S(t) = Q(\alpha, 0, \frac{t}{\beta})$$

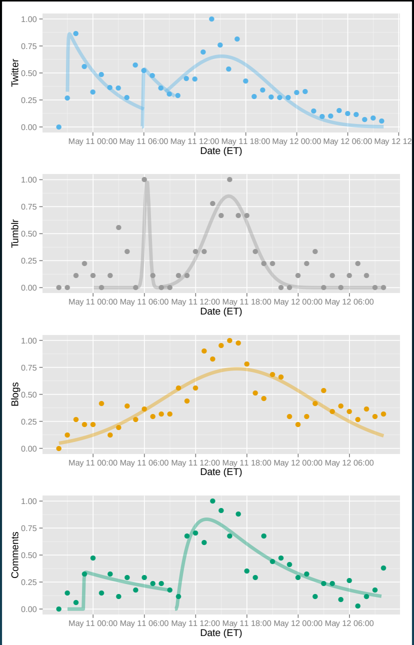


0.5
0.4
0.3
0.2
0.1
0.0
PDF
0
5
10
15
20
x
alpha, beta
1,1
1,2
3,3
2,2
1.00
0.75
0.50
0.25
0.00
CDF
0
5
10
15
20
x
alpha, beta
1,1
1,2
3,3
2,2

Why model half-life?

- predict total story volume
- compare half-lives
- anomalous story evolution


25000
20000
15000
10000
5000
0
Twitter "quake" Activity
17:45
18:00
18:15
18:30
18:45
19:00
August 2011


Twitter
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
May 12 12
Date (ET)
Tumblr
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
Date (ET)
Blogs
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
Date (ET)
Comments
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
Date (ET)

Thank you!

- Presentation, data, code at:
  http://github.com/DrSkippy27/GreenPlum2013