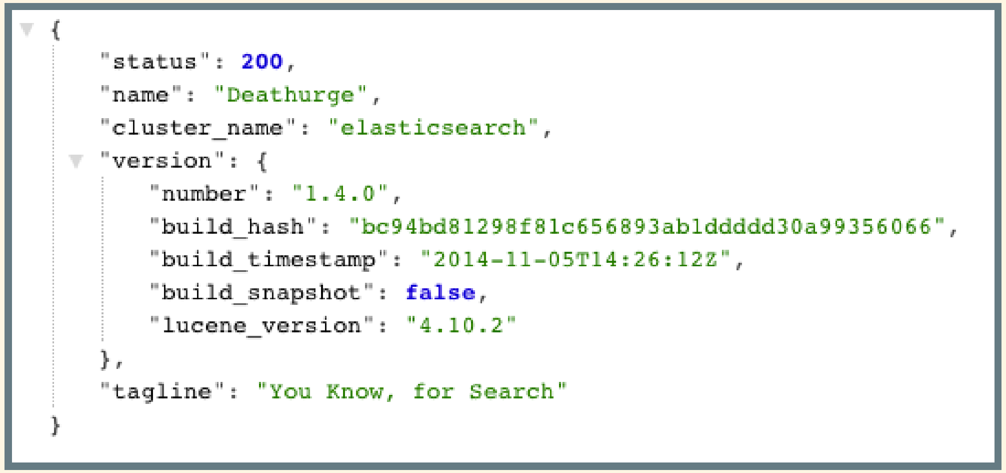
# ELASTICSEARCH, TWITTER AND SOME NEWS

## FUNCTION SCORES AND AGGREGATIONS WITH ES

Daniel Trümper / @truemped

# ABOUT ME

# ELASTICSEARCH



```
{
    "status": 200,
    "name": "Deathurge",
    "cluster_name": "elasticsearch",
    "version": {
        "number": "1.4.0",
        "build_hash": "bc94bd81298f81c656893ab1ddddd30a99356066",
        "build_timestamp": "2014-11-05T14:26:12Z",
        "build_snapshot": false,
        "lucene_version": "4.10.2"
    },
    "tagline": "You Know, for Search"
}
```

distributed, RESTful, JSON, Lucene

# THIS TALK

Create a simple news aggregator using Twitter

Rank news by retweets, favorites and time

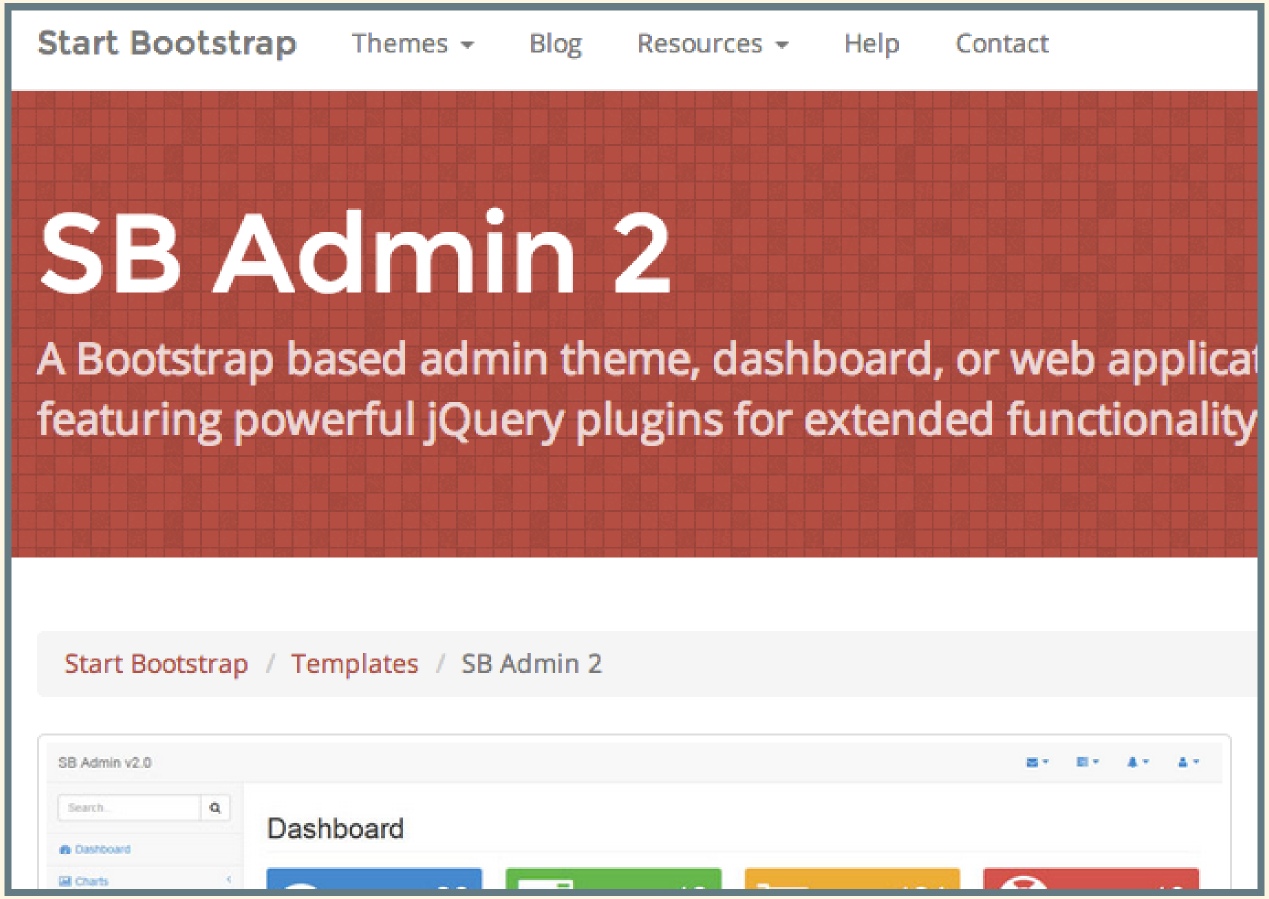
Insights into data

Filter news

# AGENDA

# FRONTEND

## Start Bootstrap Admin 2

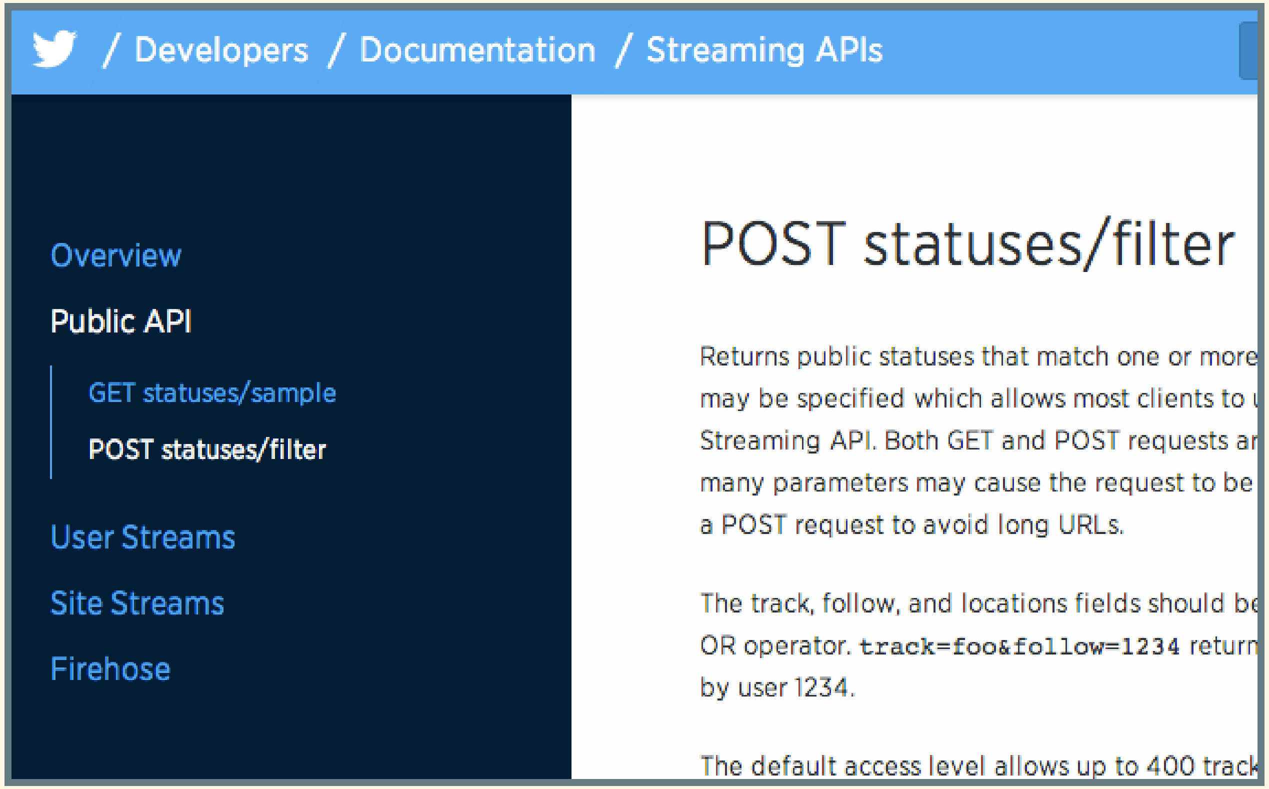


SB Admin 2

A Bootstrap based admin theme, dashboard, or web applicat
featuring powerful jQuery plugins for extended functionality

# DATA

Twitter Streaming API



POST statuses/filter

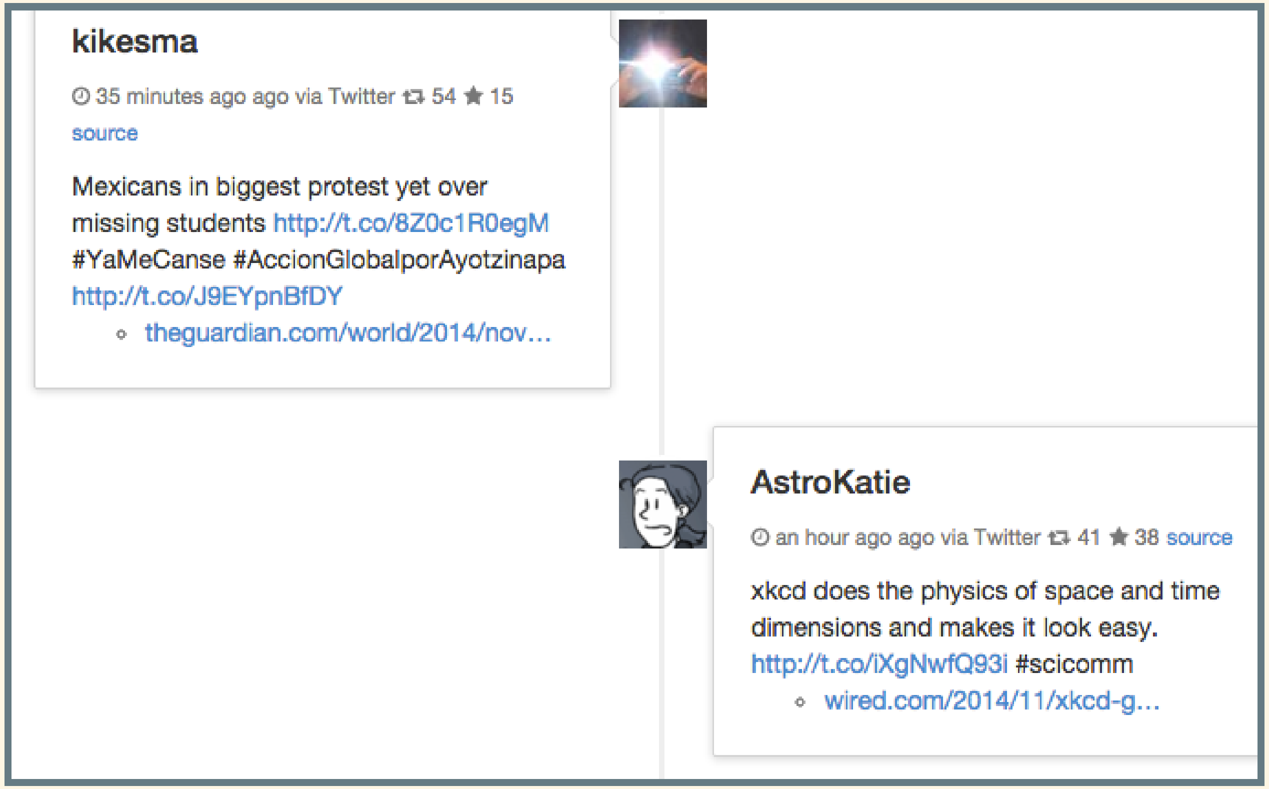
Returns public statuses that match one or more
may be specified which allows most clients to
Streaming API. Both GET and POST requests ar
many parameters may cause the request to be
a POST request to avoid long URLs.

The track, follow, and locations fields should be
OR operator. `track=foo&follow=1234` return
by user 1234.

The default access level allows up to 400 track

# FUNCTION SCORE

## Hacker News/Reddit style relevance



**kikesma**

35 minutes ago ago via Twitter 54 ★ 15

source

Mexicans in biggest protest yet over missing students http://t.co/8Z0c1R0egM #YaMeCanse #AccionGlobalporAyotzinapa http://t.co/J9EYpnBfDY

- theguardian.com/world/2014/nov…

**AstroKatie**

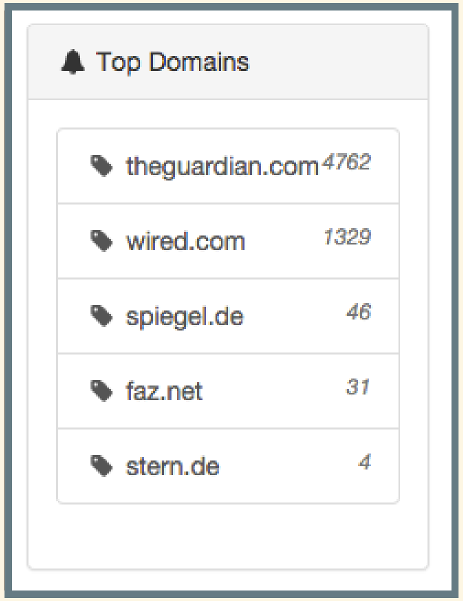
an hour ago ago via Twitter 41 ★ 38 source

xkcd does the physics of space and time dimensions and makes it look easy. http://t.co/iXgNwfQ93i #scicomm

- wired.com/2014/11/xkcd-g…

# AGGREGATIONS I

Simple aggregrations on domains

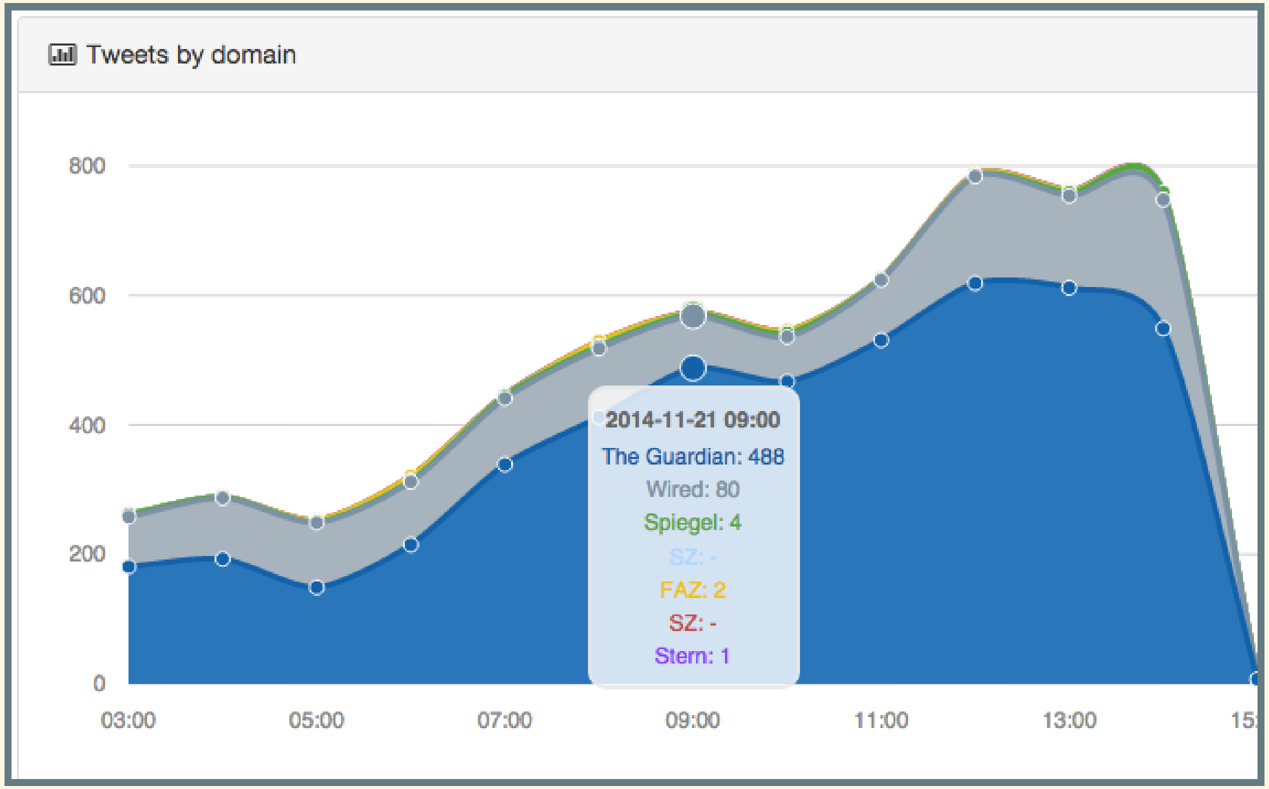


Top Domains

| | |
|---|---|
| theguardian.com | 4762 |
| wired.com | 1329 |
| spiegel.de | 46 |
| faz.net | 31 |
| stern.de | 4 |

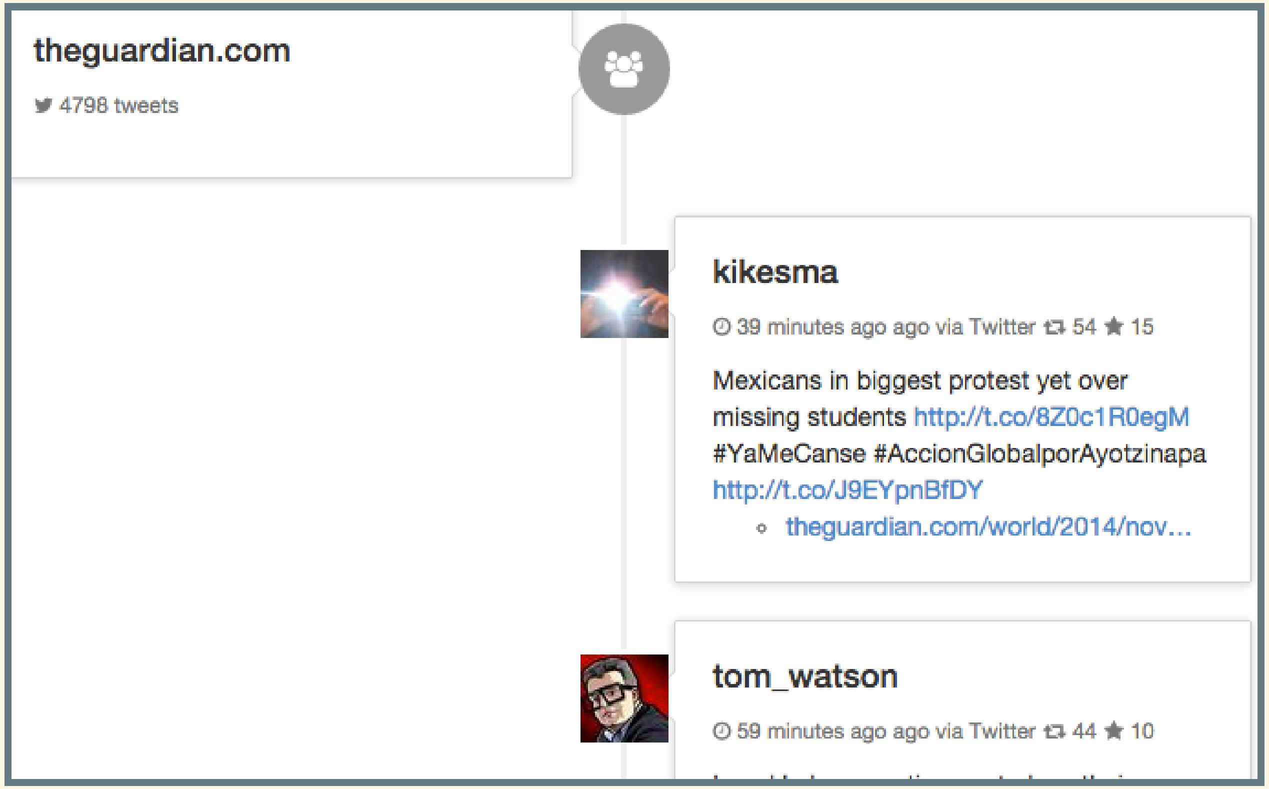
# AGGREGATIONS II

## Histogram Aggregations



# AGGREGATIONS III

## Top Hits Aggregations for grouping



**theguardian.com**

4798 tweets

**kikesma**

🕒 39 minutes ago ago via Twitter 🔁 54 ★ 15

Mexicans in biggest protest yet over missing students http://t.co/8Z0c1R0egM #YaMeCanse #AccionGlobalporAyotzinapa http://t.co/J9EYpnBfDY

- theguardian.com/world/2014/nov…

**tom_watson**

🕒 59 minutes ago ago via Twitter 🔁 44 ★ 10

# TWITTER STREAMING API

- follow - a list of user ids
- track - keywords
- locations - geo bounding boxes

https://dev.twitter.com/streaming/reference/post/statuses/filter

# TRACKING URLS

**track=theguardian com** matches tweets with links to *theguardian.com*

# A TWEET WITH URLS

## also contains the expanded full url

```
"urls": [{
    "url": "http://t.co/9bsuTIFUWs",
    "indices": [29, 51],
    "expanded_url": "http://www.theguardian.com/teacher-network/teacher-blog/2
    "display_url": "theguardian.com/teacher-networ…"
}]
```

# IN PYTHON

## Using twython library

```python
class Streamer(TwythonStreamer):

    def __init__(self, es):
        self._es = es

    def on_success(self, tweet):
        self._es.index('news-tweets', 'tweet',
                       tweet['id_str'], json.dumps(tweet))

Streamer(Elasticsearch()).statuses.filter(
    track='theguardian com', language='de,en')
```

# RETWEETS

- contain the original tweet
- update retweets
- update favorites
- only the newest copy of a tweet indexed

# UPDATE THE STREAMER

```python
def on_success(self, tweet):
    if 'retweeted_status' in tweet:
        tweet = tweet['retweeted_status']

    self._es.index('news-tweets', 'tweet',
                   tweet['id_str'], json.dumps(tweet))
```

# DELETIONS

Twitter notifies about deleted tweets

```
def on_success(self, tweet):
    if 'delete' in tweet:
        status_id = tweet['delete']['status']['id']
        self._es.delete('news-tweets', 'tweet', status_id)
        return

    if 'retweeted_status' in tweet:
        tweet = tweet['retweeted_status']

    self._es.index('news-tweets', 'tweet',
                   tweet['id_str'], json.dumps(tweet))
```

# DATA ENHANCING

Add **domain** to the urls

This allows for a simple filtering per news paper

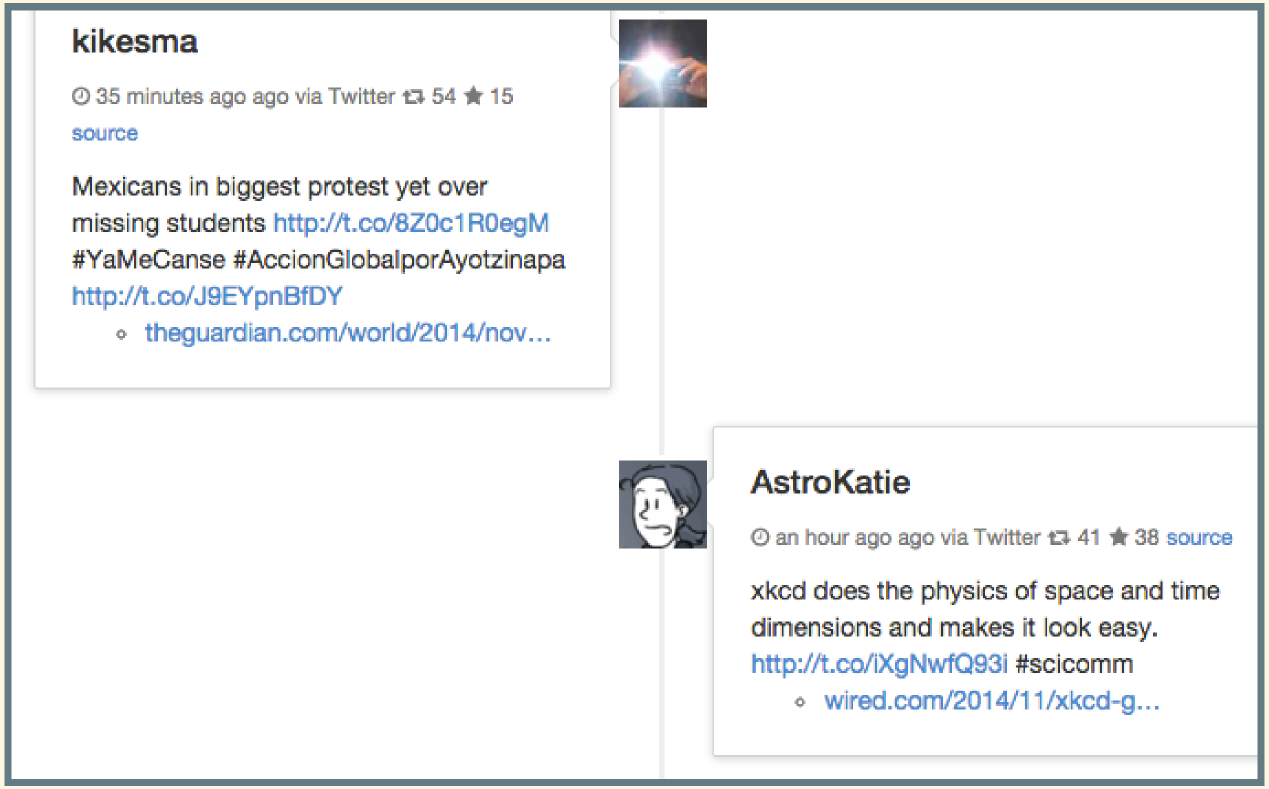
```
"urls": [{
    "url": "http://t.co/9bsuTIFUWs",
    "indices": [29, 51],
    "expanded_url": "http://www.theguardian.com/teacher-network/teacher-blog/2
    "display_url": "theguardian.com/teacher-networ…",
    "domain": "theguardian.com"
}]
```

# FINAL INDEXING METHOD

```python
def on_success(self, tweet):
    if 'delete' in tweet:
        status_id = tweet['delete']['status']['id']
        self._es.delete('news-tweets', 'tweet', status_id)
        return

    if 'retweeted_status' in tweet:
        tweet = tweet['retweeted_status']

    for url in tweet['entities']['urls']:
        if 'theguardian.com' in url['expanded_url']:
            url['domain'] = 'theguardian.com'

    self._es.index('news-tweets', 'tweet',
                   tweet['id_str'], json.dumps(tweet))
```

# TIMELINE

## Ranking news by **importance**



kikesma

35 minutes ago ago via Twitter 54 ★ 15

source

Mexicans in biggest protest yet over missing students http://t.co/8Z0c1R0egM #YaMeCanse #AccionGlobalporAyotzinapa http://t.co/J9EYpnBfDY

- theguardian.com/world/2014/nov...

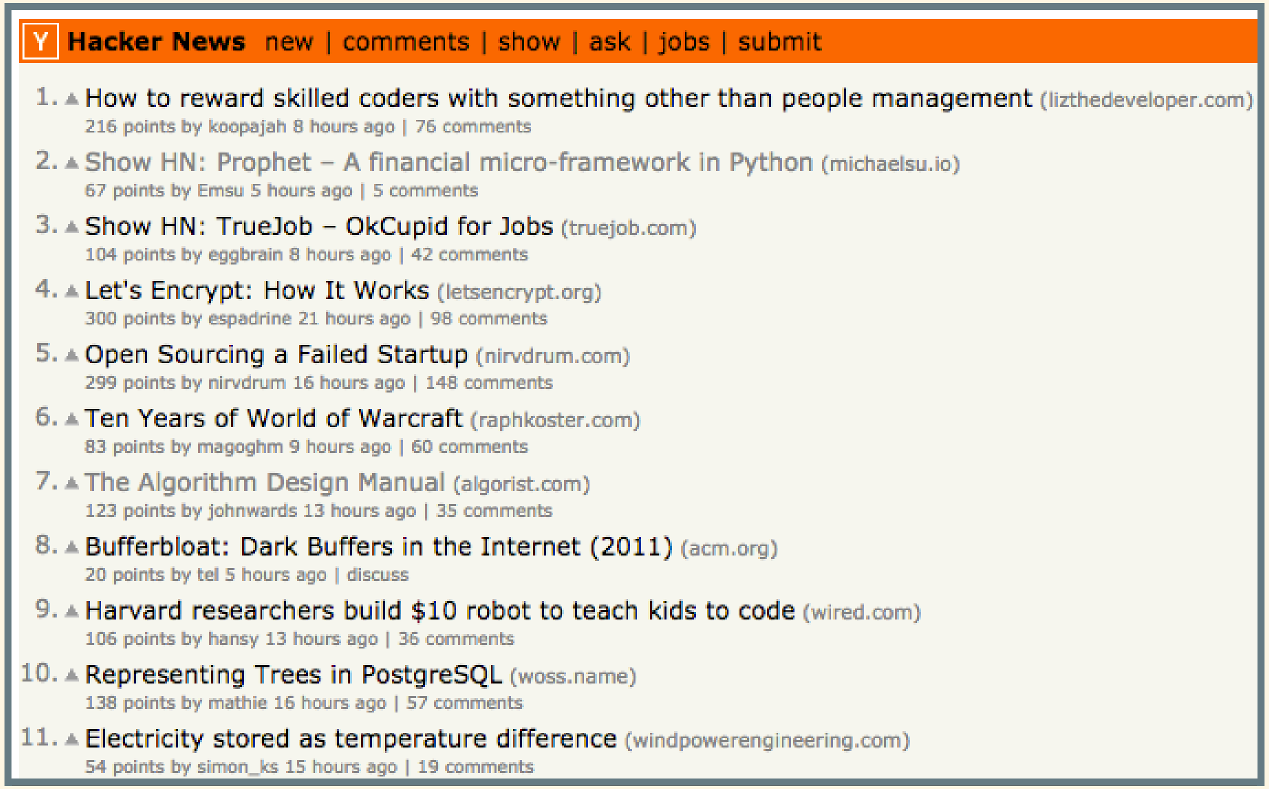
AstroKatie

an hour ago ago via Twitter 41 ★ 38 source

xkcd does the physics of space and time dimensions and makes it look easy. http://t.co/iXgNwfQ93i #scicomm

- wired.com/2014/11/xkcd-g...

# HACKER NEWS



Y **Hacker News** new | comments | show | ask | jobs | submit

1. ▲ How to reward skilled coders with something other than people management (lizthedeveloper.com)
   216 points by koopajah 8 hours ago | 76 comments
2. ▲ Show HN: Prophet – A financial micro-framework in Python (michaelsu.io)
   67 points by Emsu 5 hours ago | 5 comments
3. ▲ Show HN: TrueJob – OkCupid for Jobs (truejob.com)
   104 points by eggbrain 8 hours ago | 42 comments
4. ▲ Let's Encrypt: How It Works (letsencrypt.org)
   300 points by espadrine 21 hours ago | 98 comments
5. ▲ Open Sourcing a Failed Startup (nirvdrum.com)
   299 points by nirvdrum 16 hours ago | 148 comments
6. ▲ Ten Years of World of Warcraft (raphkoster.com)
   83 points by magoghm 9 hours ago | 60 comments
7. ▲ The Algorithm Design Manual (algorist.com)
   123 points by johnwards 13 hours ago | 35 comments
8. ▲ Bufferbloat: Dark Buffers in the Internet (2011) (acm.org)
   20 points by tel 5 hours ago | discuss
9. ▲ Harvard researchers build $10 robot to teach kids to code (wired.com)
   106 points by hansy 13 hours ago | 36 comments
10. ▲ Representing Trees in PostgreSQL (woss.name)
    138 points by mathie 16 hours ago | 57 comments
11. ▲ Electricity stored as temperature difference (windpowerengineering.com)
    54 points by simon_ks 15 hours ago | 19 comments

# RANKING

Two components: votes and time since published

$$score = votes/(t + 2)^g$$

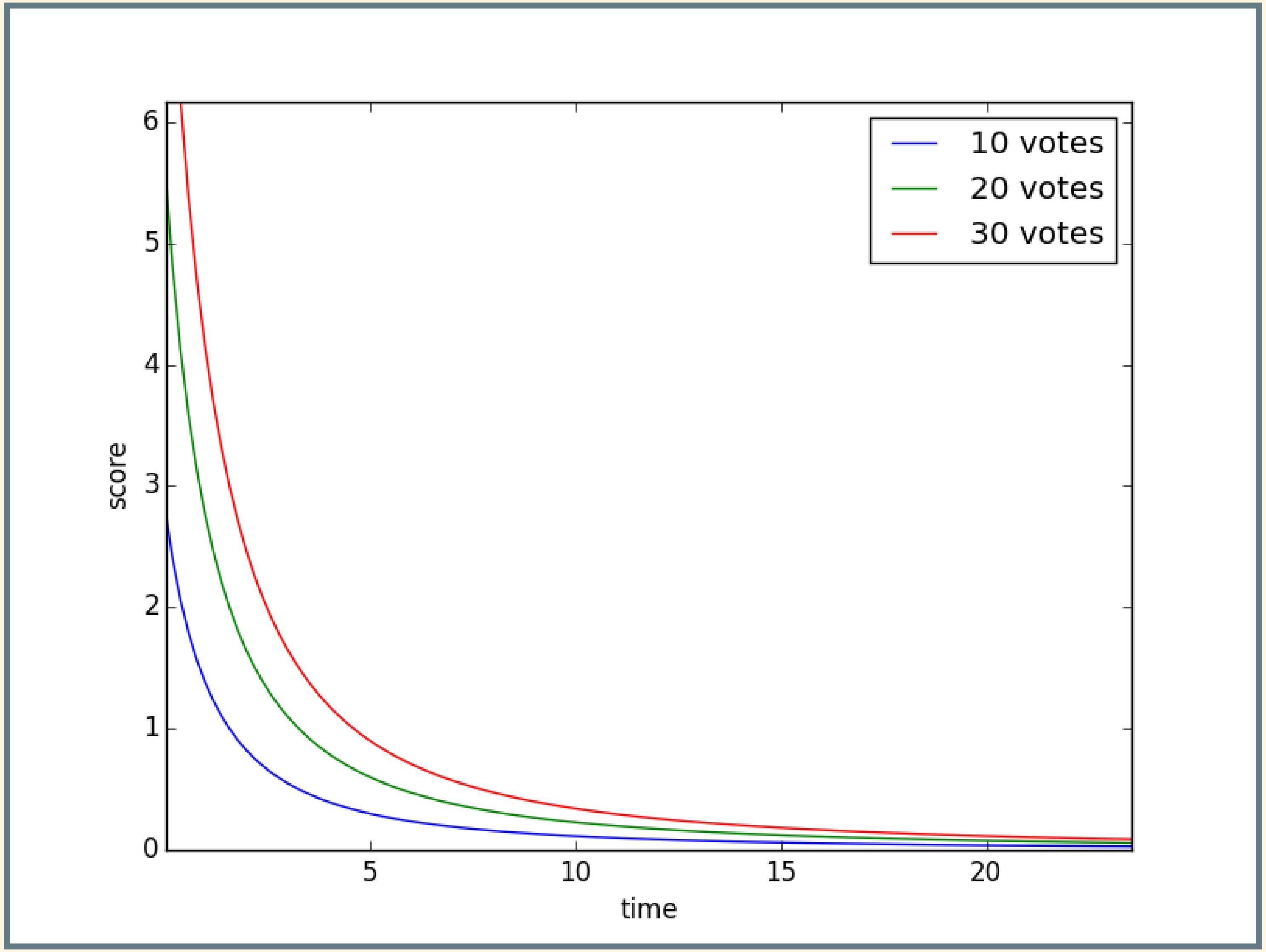
Nice blog post about HN/Reddit scoring: http://amix.dk/blog/post/19574

# GRAVITY

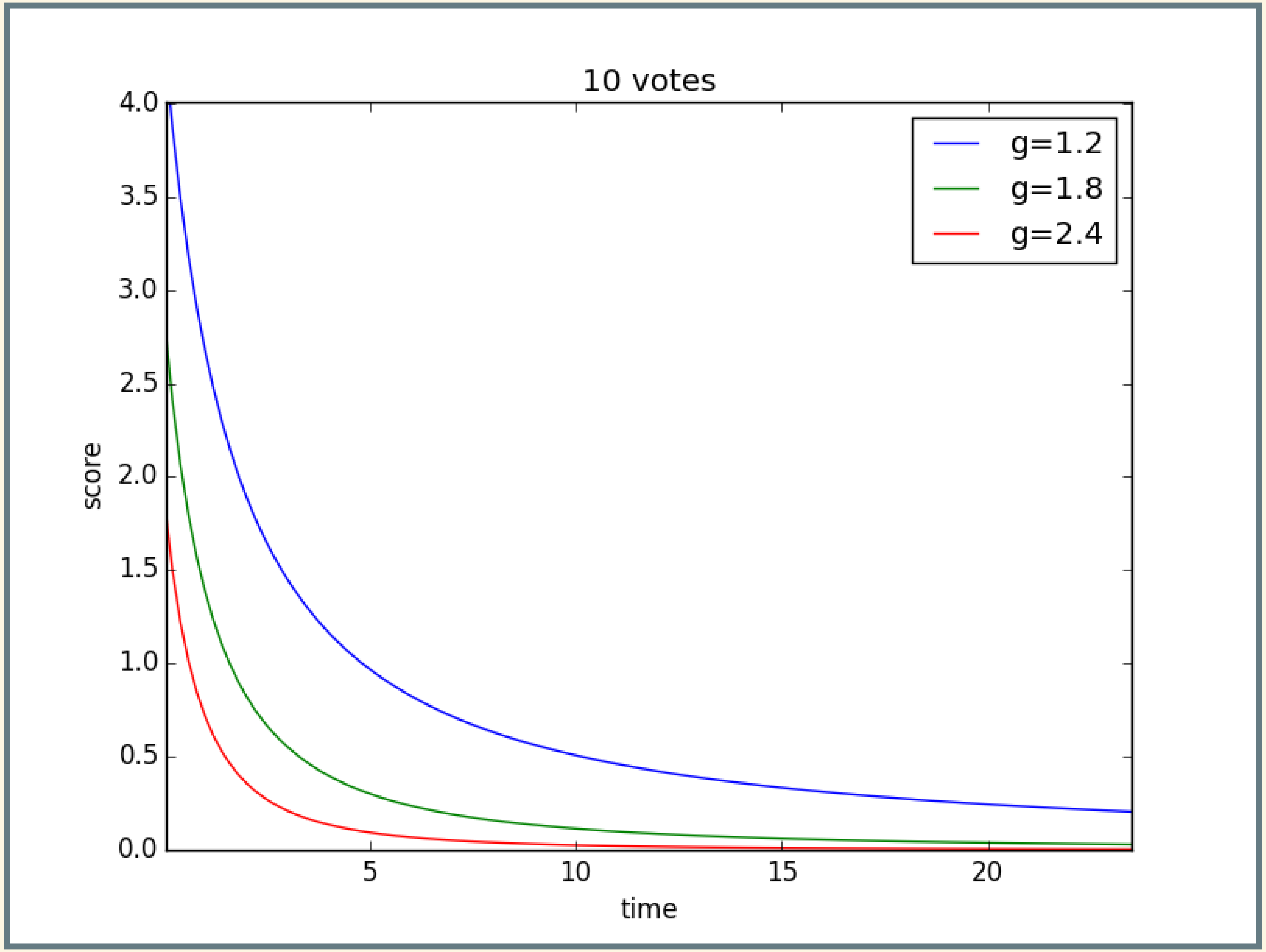
**g** determines the importance of time

smaller **g** means more influence of votes

# EXAMPLE G=1.8

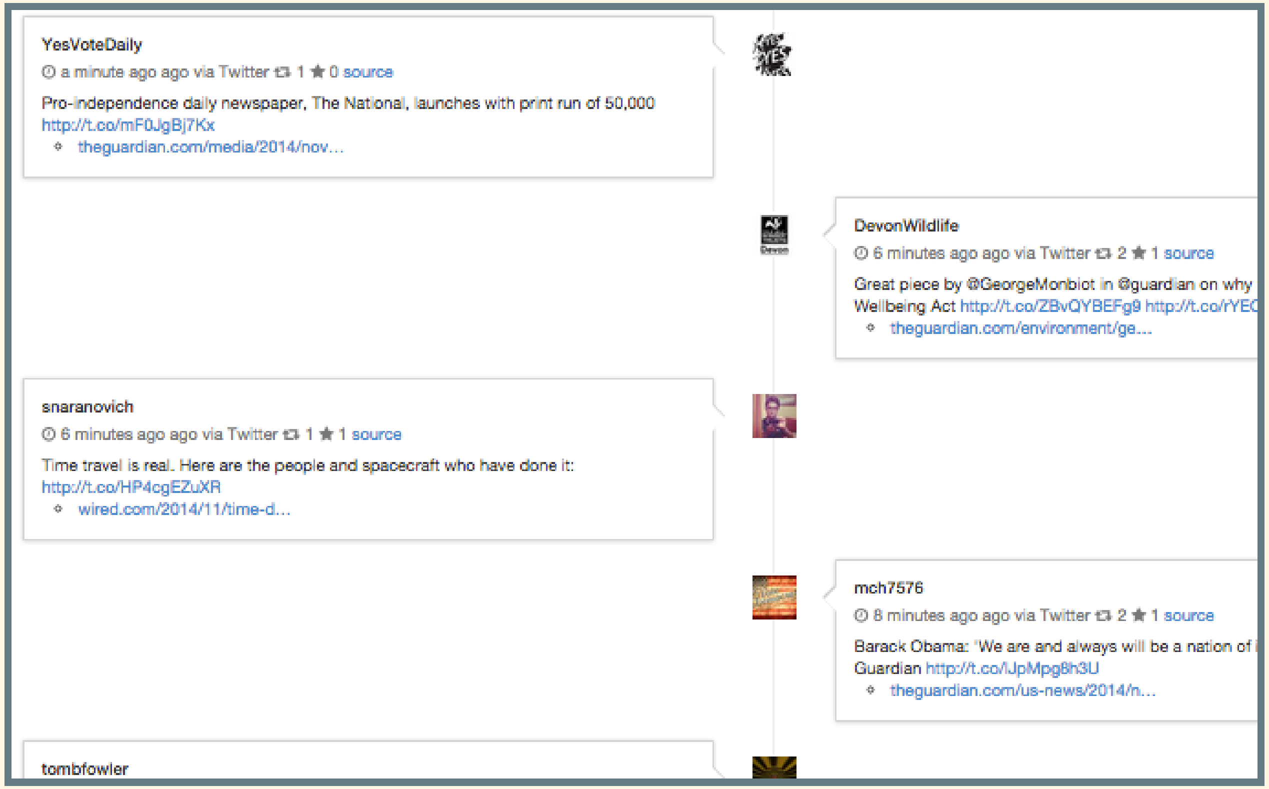


# G-EFFECT


10 votes
score
time
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5
10
15
20
g=1.2
g=1.8
g=2.4

# FUNCTION SCORE IN ES

```
def scoring(now, g=1.8):
    return {"function_score": {
            "query": {"match_all": {}},
            "script_score": {
                "params": {
                    "g": g,
                    "now": now,
                },
                "script": ("(doc['favorite_count'].value + " +
                           "doc['retweet_count'].value) / " +
                           "pow((now - doc['created_at'].value + 2), g)")
        }}}
```



YesVoteDaily

a minute ago ago via Twitter 1 ★ 0 source

Pro-independence daily newspaper, The National, launches with print run of 50,000 http://t.co/mF0JgBj7Kx

- theguardian.com/media/2014/nov…

DevonWildlife

6 minutes ago ago via Twitter 2 ★ 1 source

Great piece by @GeorgeMonbiot in @guardian on why Wellbeing Act http://t.co/ZBvQYBEFg9 http://t.co/rYEC

- theguardian.com/environment/ge…

snaranovich

6 minutes ago ago via Twitter 1 ★ 1 source

Time travel is real. Here are the people and spacecraft who have done it: http://t.co/HP4cgEZuXR

- wired.com/2014/11/time-d…

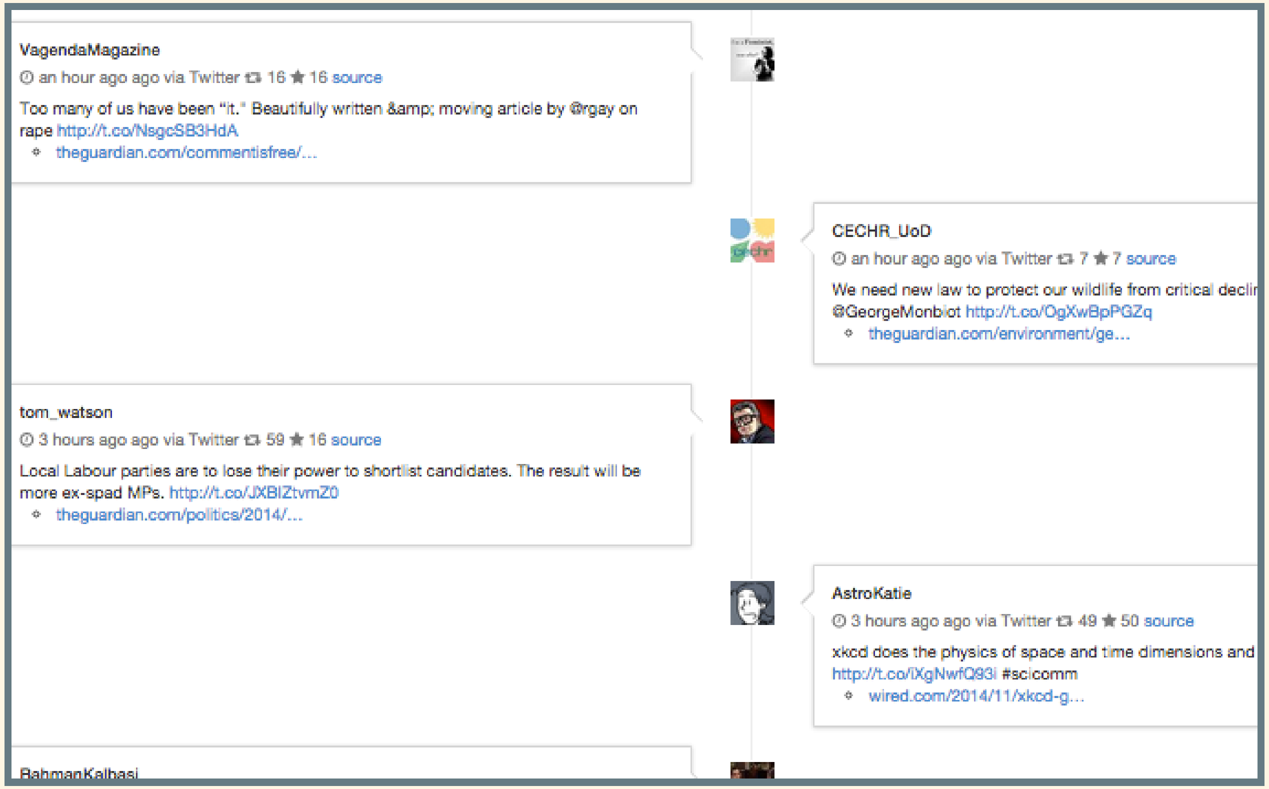
mch7576

8 minutes ago ago via Twitter 2 ★ 1 source

Barack Obama: 'We are and always will be a nation of i Guardian http://t.co/lJpMpg8h3U

- theguardian.com/us-news/2014/n…

tombfowler



VagendaMagazine

🕒 an hour ago ago via Twitter 🔁 16 ★ 16 source

Too many of us have been "it." Beautifully written &amp; moving article by @rgay on rape http://t.co/NsgcSB3HdA

- theguardian.com/commentisfree/...

CECHR_UoD

🕒 an hour ago ago via Twitter 🔁 7 ★ 7 source

We need new law to protect our wildlife from critical declir
@GeorgeMonbiot http://t.co/OgXwBpPGZq

- theguardian.com/environment/ge...

tom_watson

🕒 3 hours ago ago via Twitter 🔁 59 ★ 16 source

Local Labour parties are to lose their power to shortlist candidates. The result will be more ex-spad MPs. http://t.co/JXBIZtvmZ0

- theguardian.com/politics/2014/...

AstroKatie

🕒 3 hours ago ago via Twitter 🔁 49 ★ 50 source

xkcd does the physics of space and time dimensions and
http://t.co/iXgNwfQ93i #scicomm

- wired.com/2014/11/xkcd-g...

BahmanKalbasi

# SIMPLE AGGREGATIONS

Retrieve the top domains and the top hashtags

# DOMAIN

```
def domain():
    return {'aggregations': {'domains': {
        'terms': {'field': 'entities.urls.domain'}}}}
```

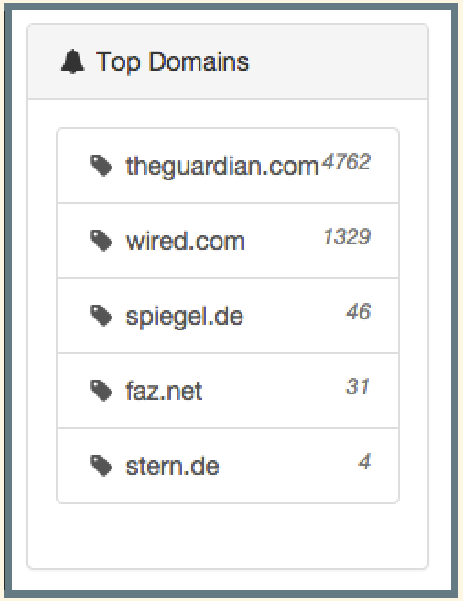
# HASHTAGS

```
def hashtags():
    return {'aggregations': {'hashtags': {
        'terms': {'field': 'entities.hashtags.text'}}}}
```

# ES RESPONSE

```
{"aggregations": {
    "domains": {
        "doc_count_error_upper_bound": 0,
        "sum_other_doc_count": 0,
        "buckets": [
            {
                "key": "theguardian.com",
                "doc_count": 5682
            },
            {
                "key": "wired.com",
                "doc_count": 1665
            }
        ]
    }
}}
```

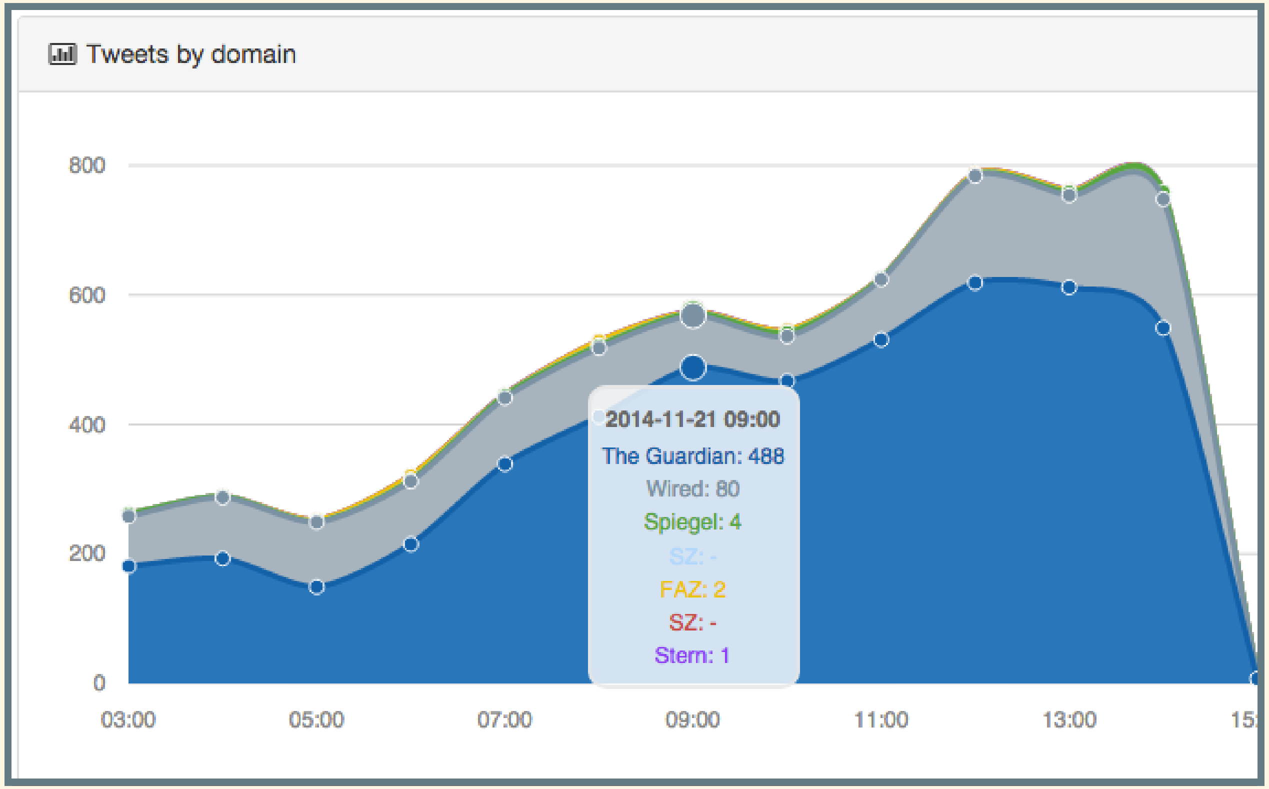
# FRONTEND



Top Domains

| | |
|---|---|
| theguardian.com | 4762 |
| wired.com | 1329 |
| spiegel.de | 46 |
| faz.net | 31 |
| stern.de | 4 |

# HISTOGRAM OF LINKS PER DOMAIN



# ES AGGREGATION BASICS

- create buckets
- count documents per bucket
- sub-aggregations on each bucket

# DATE HISTOGRAM

```
def domain_histogram():
    return {'aggregations': {'date_hist': {
        'date_histogram': {
            'field': 'created_at',
            'interval': '1h'
        }}}}
```

# ALMOST THERE

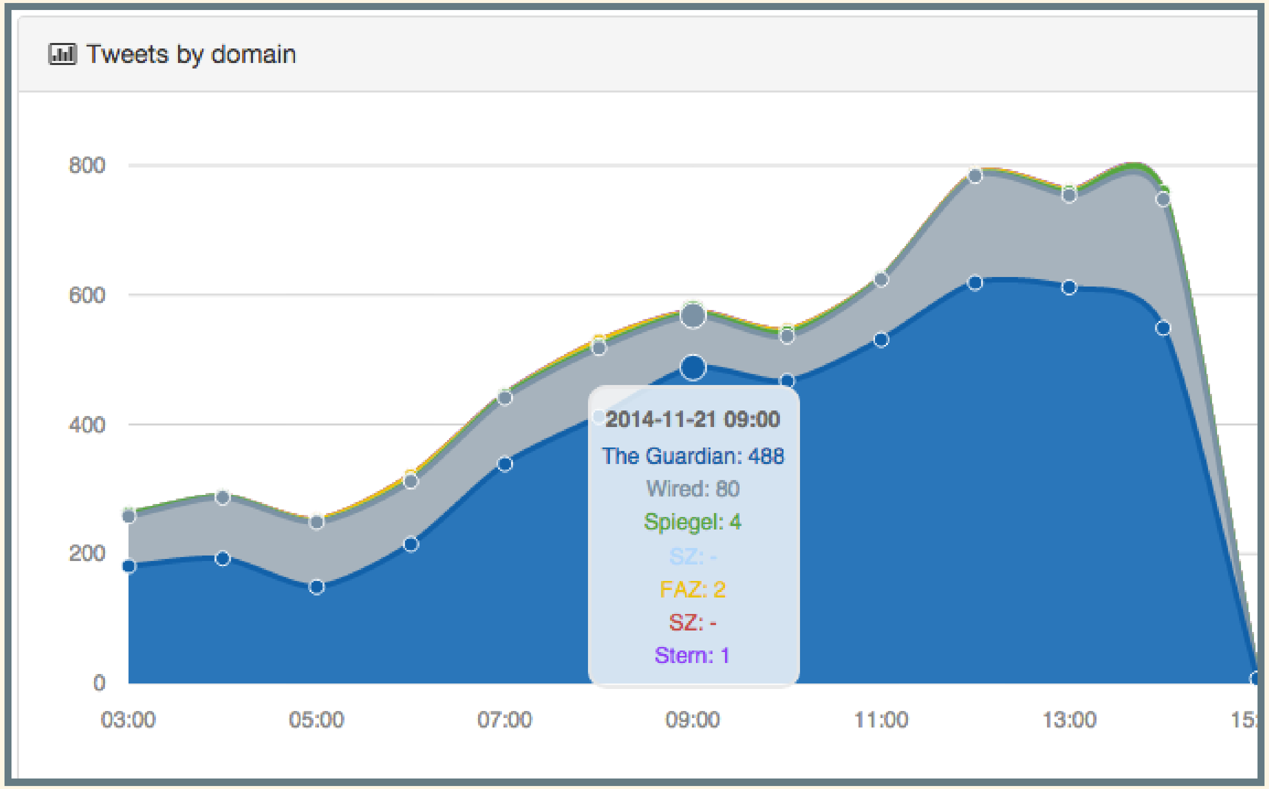
```
{"aggregations": {
    "date_hist": {
        "buckets": [
            {
                "key_as_string": "Fri Nov 21 05:00:00 +0000 2014",
                "key": 1416546000000,
                "doc_count": 253
            },
            {
                "key_as_string": "Fri Nov 21 06:00:00 +0000 2014",
                "key": 1416549600000,
                "doc_count": 321
            }
        ]
    }}}
```

# SUB-AGGREGRATIONS

```
def domain_histogram():
    return {'aggregations': {
        'date_hist': {
            'date_histogram': {
                'field': 'created_at',
                'interval': '1h'
            },
            'aggregations': {
                'domains': {
                    'terms': {
                        'field': 'entities.urls.domain'
    }}}}}}
```

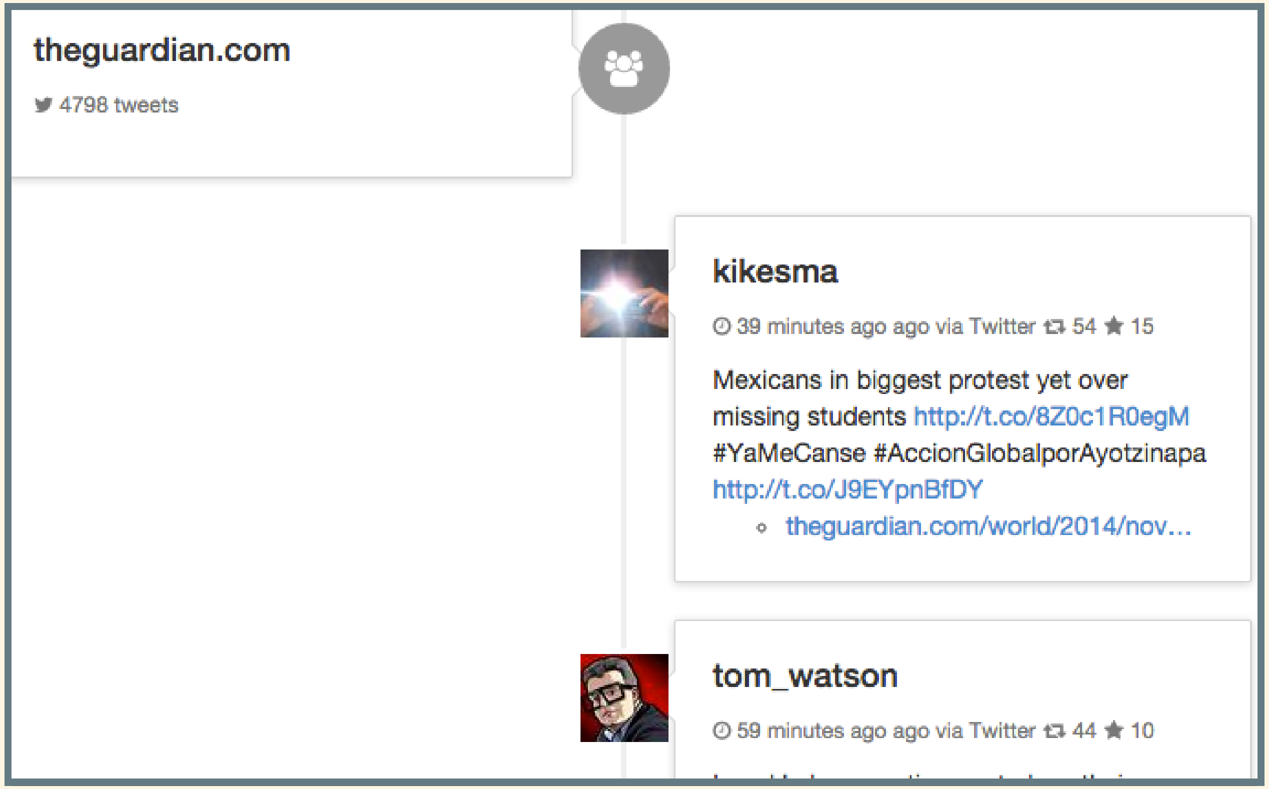
# ES RESPONSE

```
{"aggregations": {
    "date_hist": {
        "buckets": [
            {
                "key_as_string": "Fri Nov 21 05:00:00 +0000 2014",
                "key": 1416546000000,
                "doc_count": 253,
                "domain": {
                    "doc_count_error_upper_bound": 0,
                    "sum_other_doc_count": 0,
                    "buckets": [
                        {
                            "key": "theguardian.com",
                            "doc_count": 149
                        },
                        {
                            "key": "wired.com"
```



2014-11-21 09:00
The Guardian: 488
Wired: 80
Spiegel: 4
SZ: -
FAZ: 2
SZ: -
Stern: 1

# TOP TWEETS PER DOMAIN



theguardian.com

4798 tweets

kikesma

39 minutes ago ago via Twitter 54 ★ 15

Mexicans in biggest protest yet over missing students http://t.co/8Z0c1R0egM #YaMeCanse #AccionGlobalporAyotzinapa http://t.co/J9EYpnBfDY

- theguardian.com/world/2014/nov...

tom_watson

59 minutes ago ago via Twitter 44 ★ 10

# AGGREGATE DOMAINS

```python
def topdomains():
    return {'aggregations': {'top_tweets': {
        'terms': {'field': 'entities.urls.domain'}}
        }}
```

# GET THE TOP HITS PER DOMAIN

```
def topdomains():
    return {'aggregations': {'top_tweets': {
        'terms': {'field': 'entities.urls.domain'}},
        'aggregations': {'top_domain_hits': {
            'top_hits': {}
        }}}
```

# ES RESULT

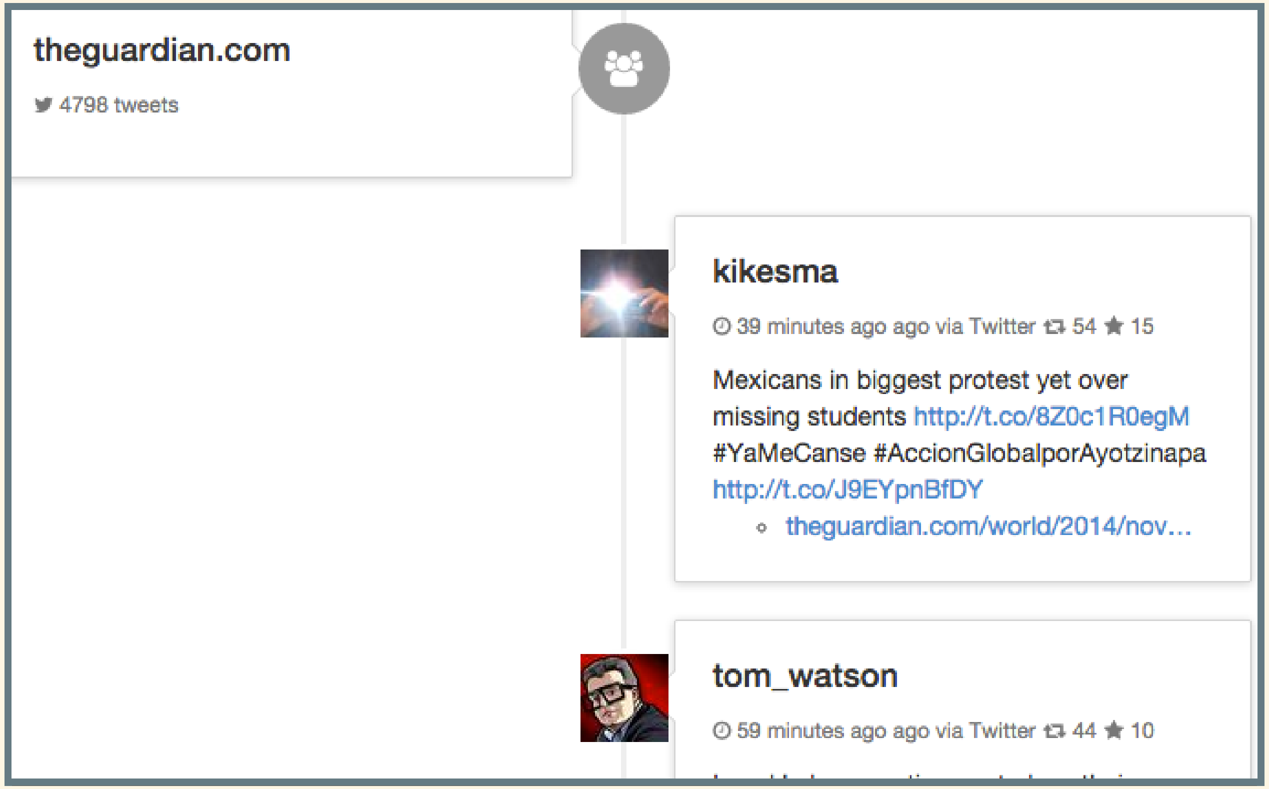
```
...
"buckets": [{"key": "theguardian.com",
    "doc_count": 6361,
    "top_domain_hits": {
        "hits": {
            "total": 6361,
            "max_score": 1.2440135e-14,
            "hits": [{
                    "_index": "news-tweets",
                    "_type": "tweet",
                    "_id": "535810990223687680",
                    "_score": 1.2440135e-14,
                    "_source": {...}
                },{
                    "_index": "news-tweets",
                    "_type": "tweet",
                    "_id": "535815107364651008"
```

# SORTING THE BUCKETS

Currently extra aggregator is necessary to ensure sorting of **terms** buckets

```
def topdomains():
    return {'aggregations': {'top_tweets': {
        'terms': {'field': 'entities.urls.domain',
                  'order': {'top_term': 'desc'}}},
        'aggregations': {
            'top_domain_hits': {'top_hits': {}},
            'top_term': {'max': {'script': '_score'}}
        }}}
```

# RESULT



theguardian.com

4798 tweets

kikesma

39 minutes ago ago via Twitter 54 ★ 15

Mexicans in biggest protest yet over missing students http://t.co/8Z0c1R0egM #YaMeCanse #AccionGlobalporAyotzinapa http://t.co/J9EYpnBfDY

- theguardian.com/world/2014/nov...

tom_watson

59 minutes ago ago via Twitter 44 ★ 10

# CRAZY IDEAS

- tweets per url
- download and index news
- extract more metadata (category, author...)
- classify news I read

# THE END

SLIDES AND CODE ON HTTPS://GITHUB.COM/TRUEMPED/ES-TWITTER-AND-NEWS