Data Scientist's Approach to Social Data

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

November 5, 2013

Social Data

created by
interactions
among people

Social Data

form and content

shaped by

people

Sources of Social Data

- Firehoses
- APIs
- Scraping

Firehose

Continuous stream
of activities
in near-real time

Social Data Activity

People interact
on social media platform

Firehose volumes

Publisher	Daily Activity
Twitter	500M
Tumblr	105M
Foursquare	4.3M
GetGlue	430k
Wordpress Posts	919k
Wordpress Comments	1.7M
Disqus	1.9M
Engagement (likes, votes)	59M

Daily @Gnip

$\frac{3}{4}$ Billion IN

4 Billion OUT

Analysis Considerations

- Technology - interfaces, tools, infrastructure for accessing
- Latency - how soon after activity as created?
- Uniformity - how hard/costly to normalize data formats?
- Coverage - do you need it all? a defined sample?
- Meta-data - how much and what kind of data about the data?

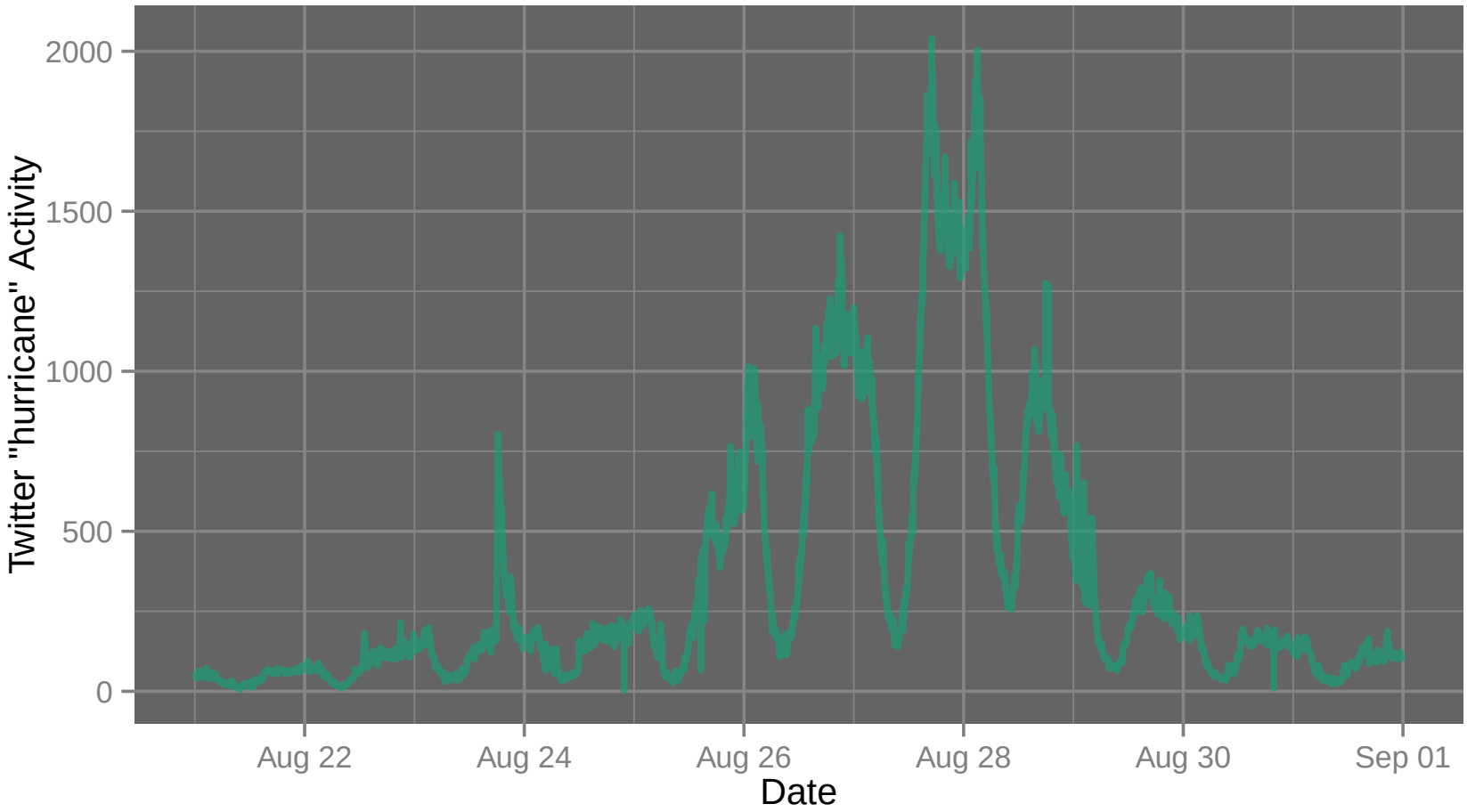
Business Considerations

- Licensing - do you have the right to analyze, display, store data?
- Terms-of-Service Compliance - violating publishers terms of service, privacy protections?
- Cost - data collection costs? licensing costs? processing and storage costs?
- Analysis mode - batch vs. real-time? event vs. background? time, structure, language, people?

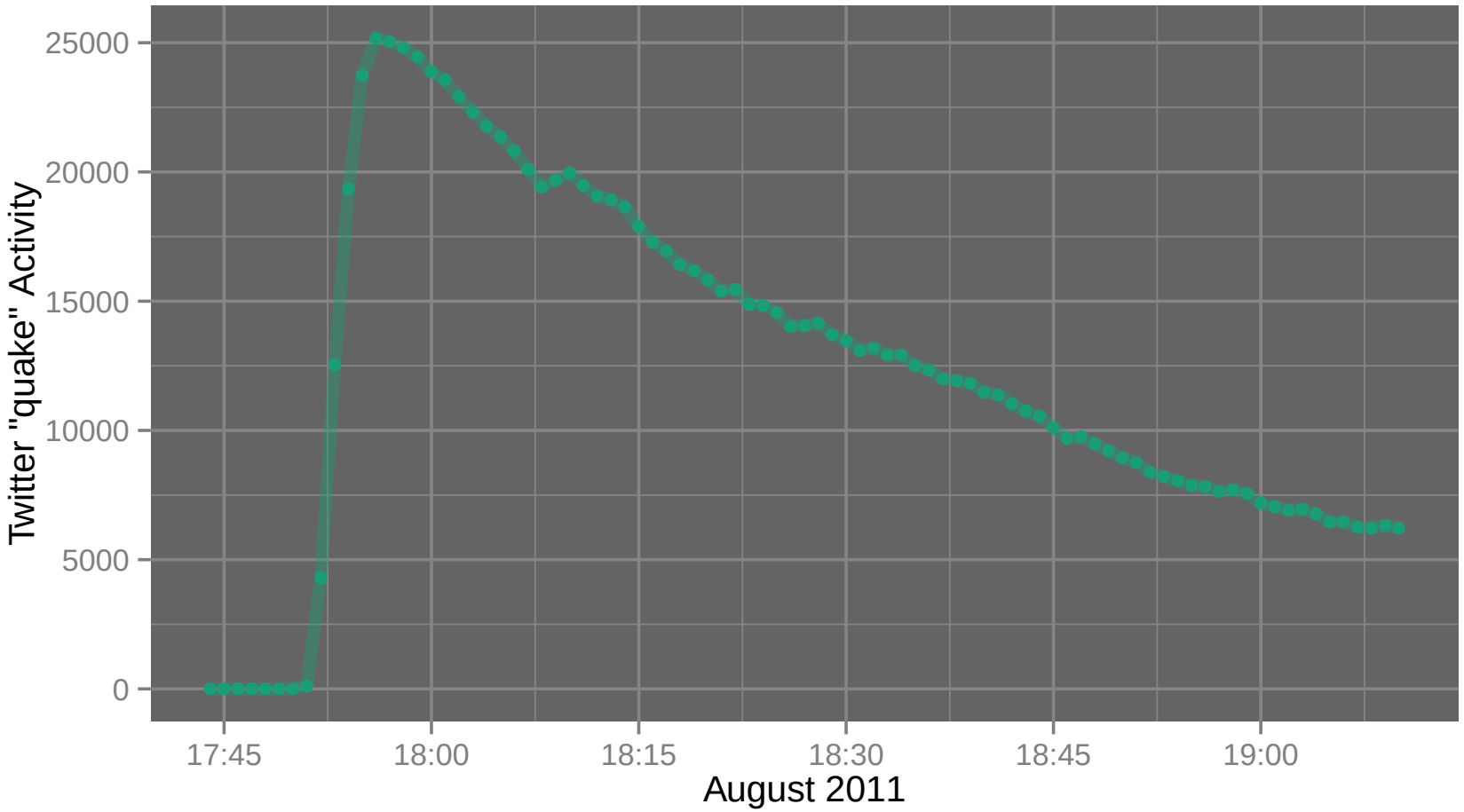
Models

- Model domain - timeseries, language analysis, network structures
- Model domain drives storage/access strategy - test files, spreadsheets, relational dbs, no-sql dbs...
- Analysis - dashboard vs. discovery projects

Expected: Hurricane



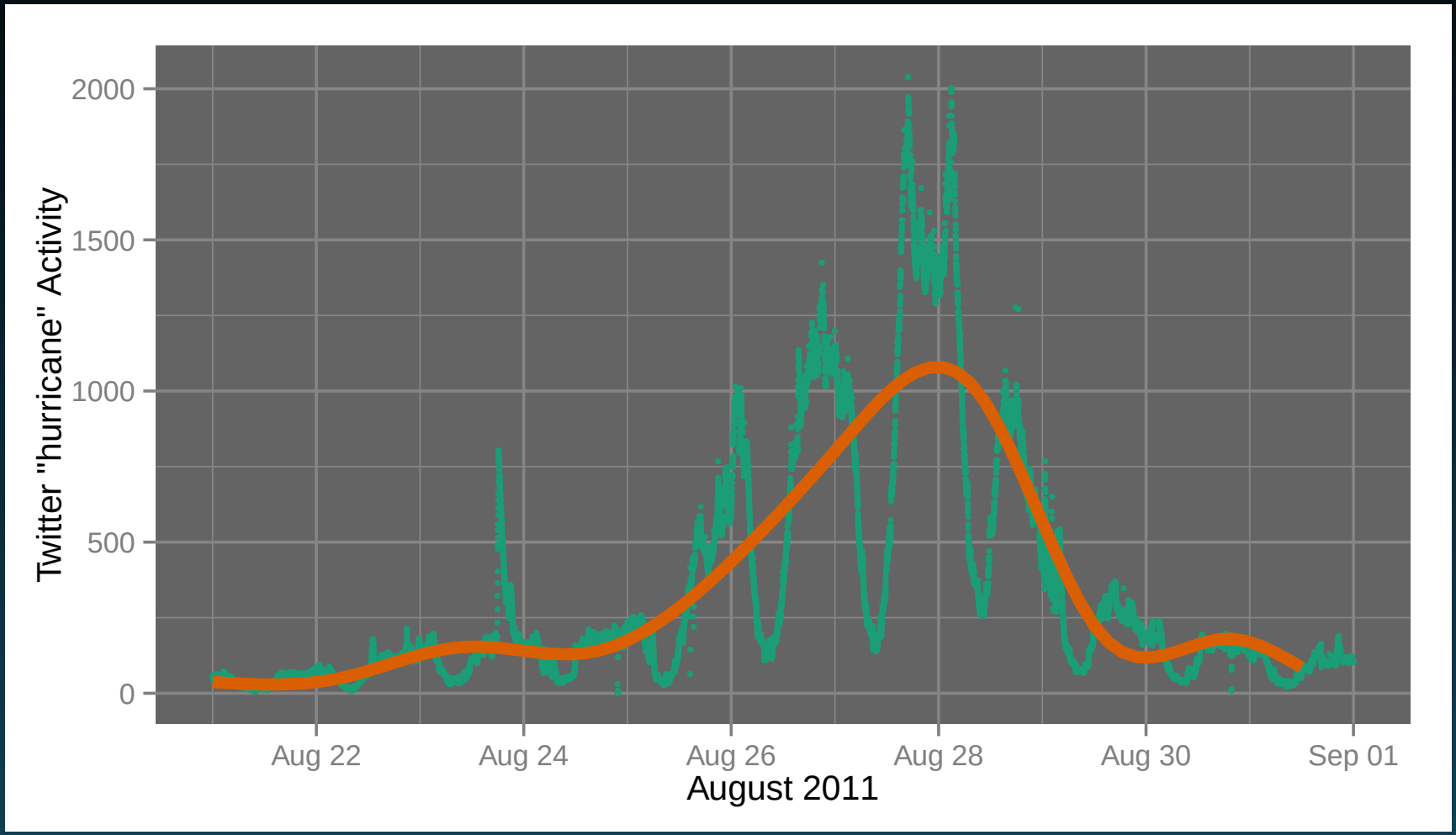
Unexpected: Earthquake



Classifying Events

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

Expected: Hurricane



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,

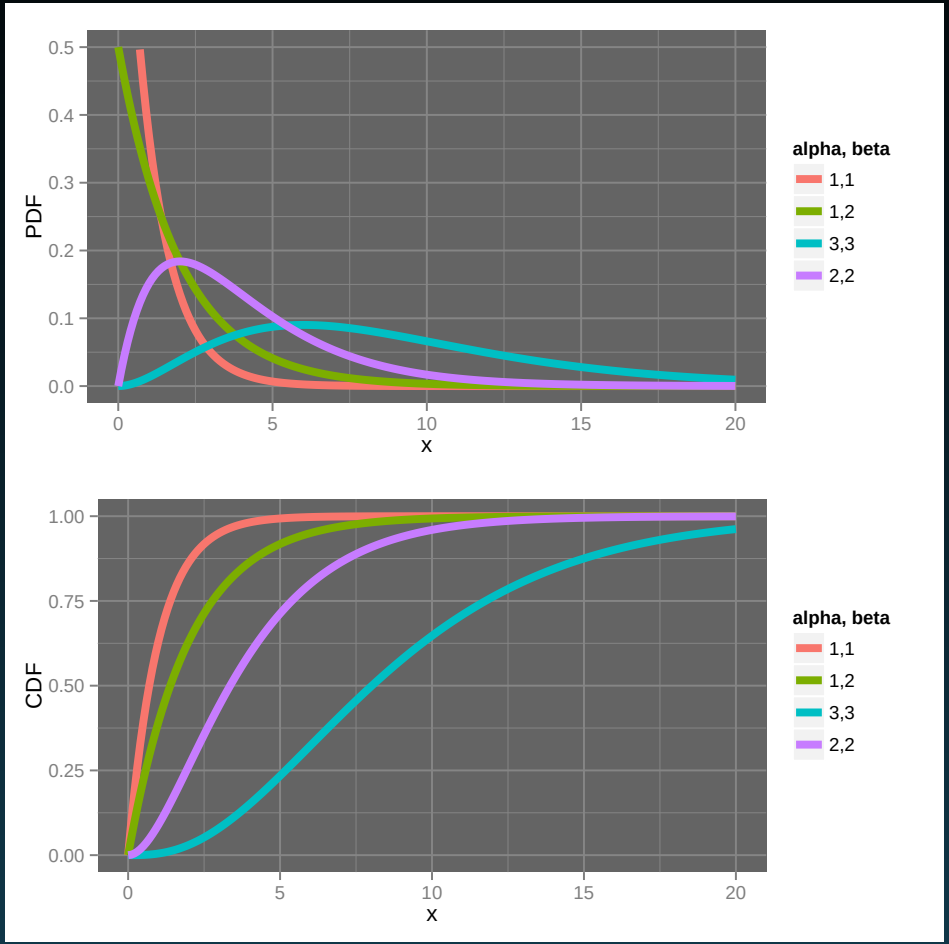
$$f(t) = \lambda \exp(-\lambda t), \text{ for } t \geq 0.$$

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

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

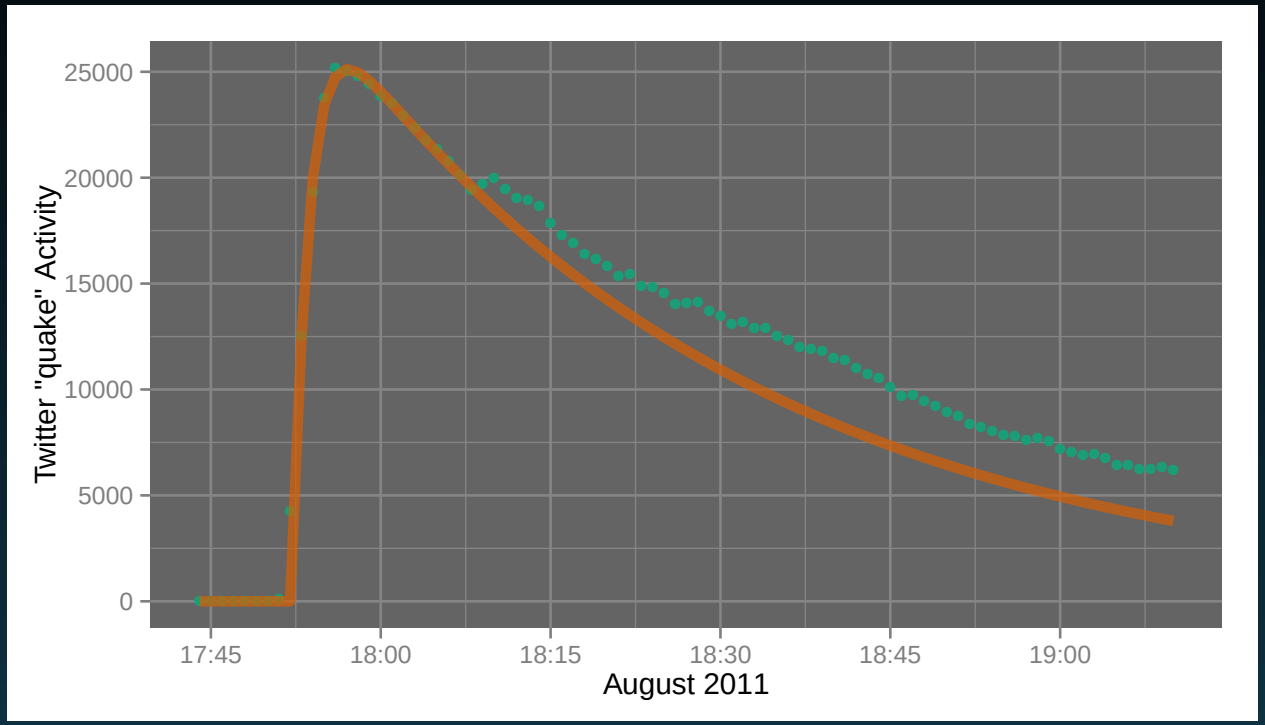
Cumulative distribution is the “generalized regularized incomplete gamma function”,

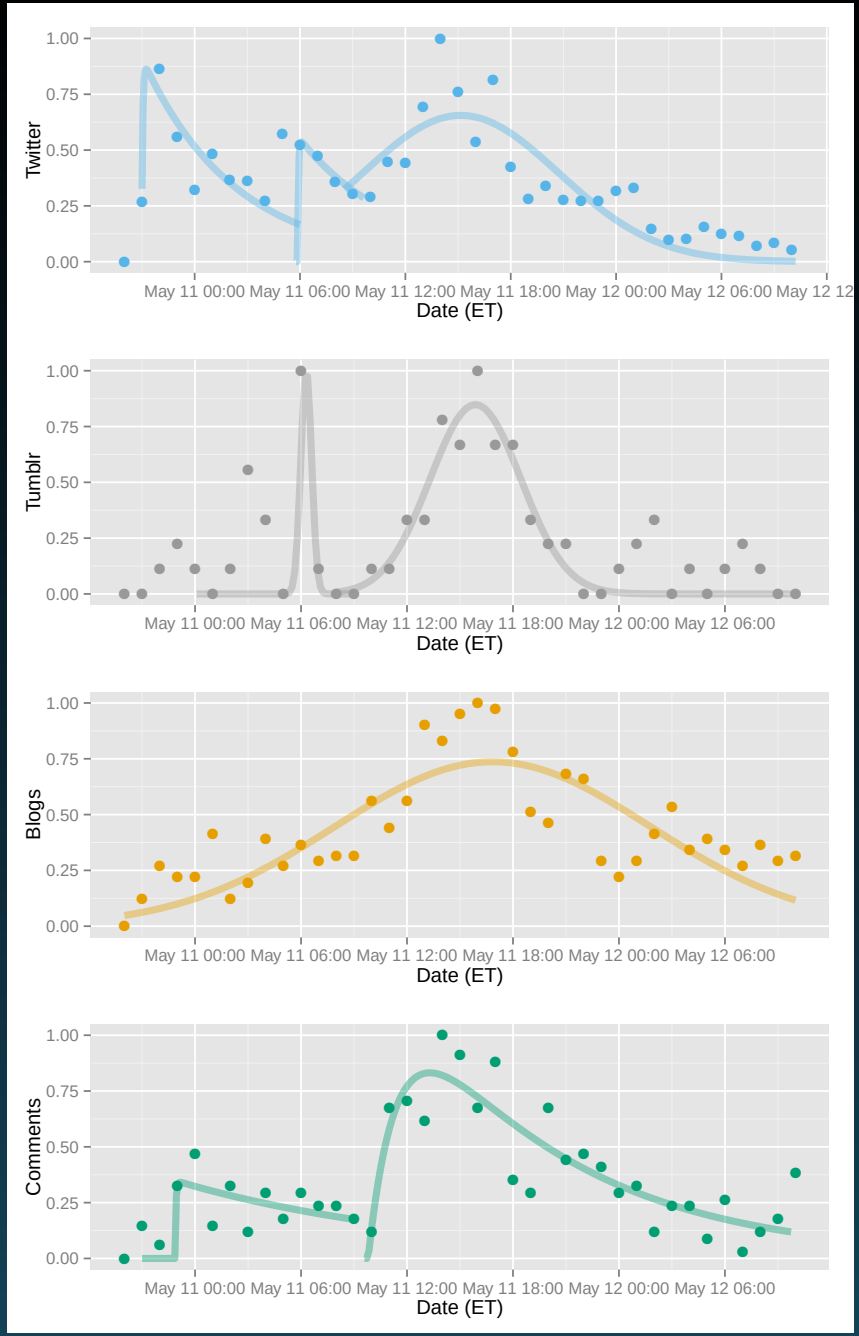
$$F_S(t) = Q(\alpha, 0, \frac{t}{\beta})$$


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

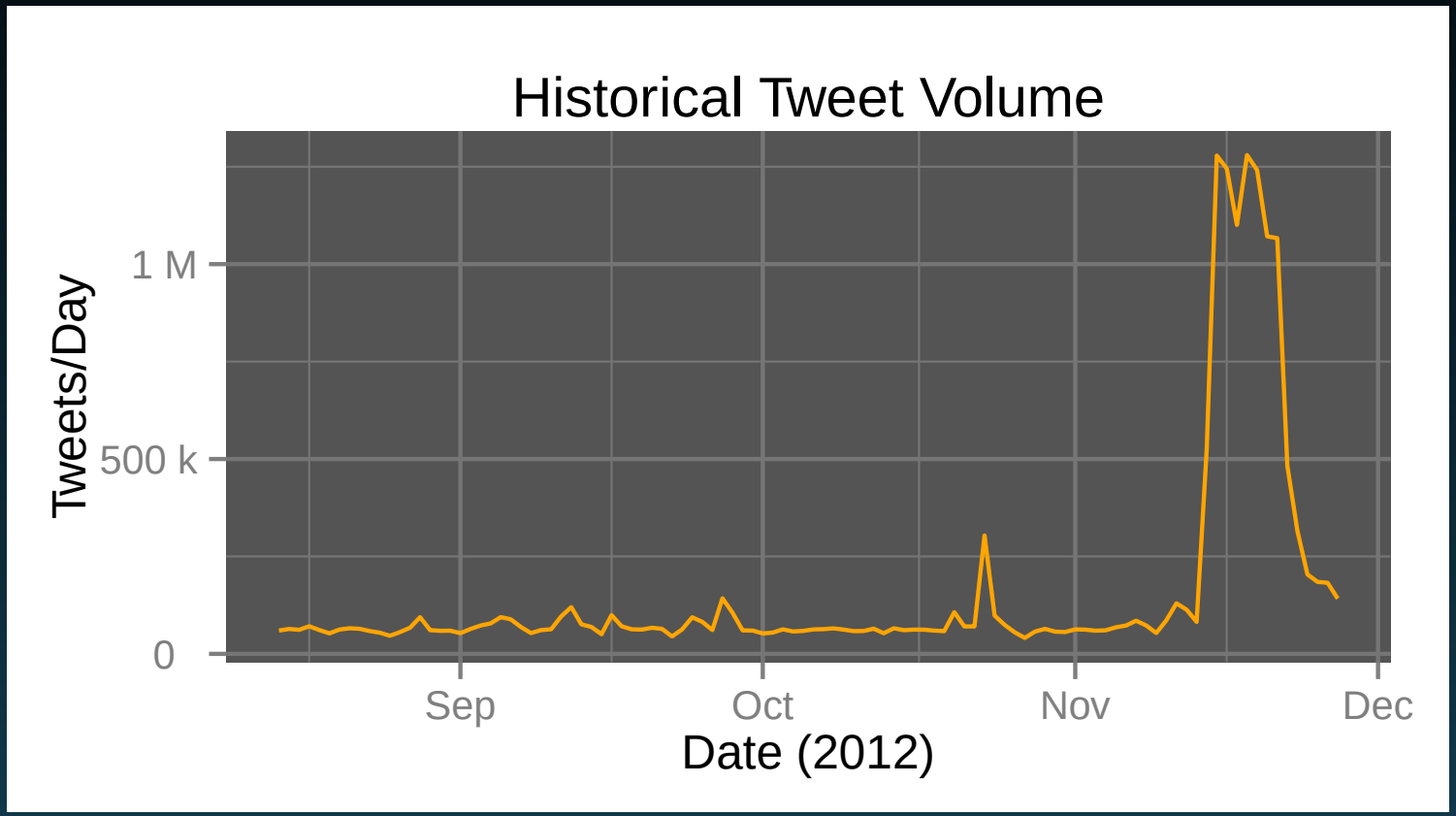
Why model half-life?

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

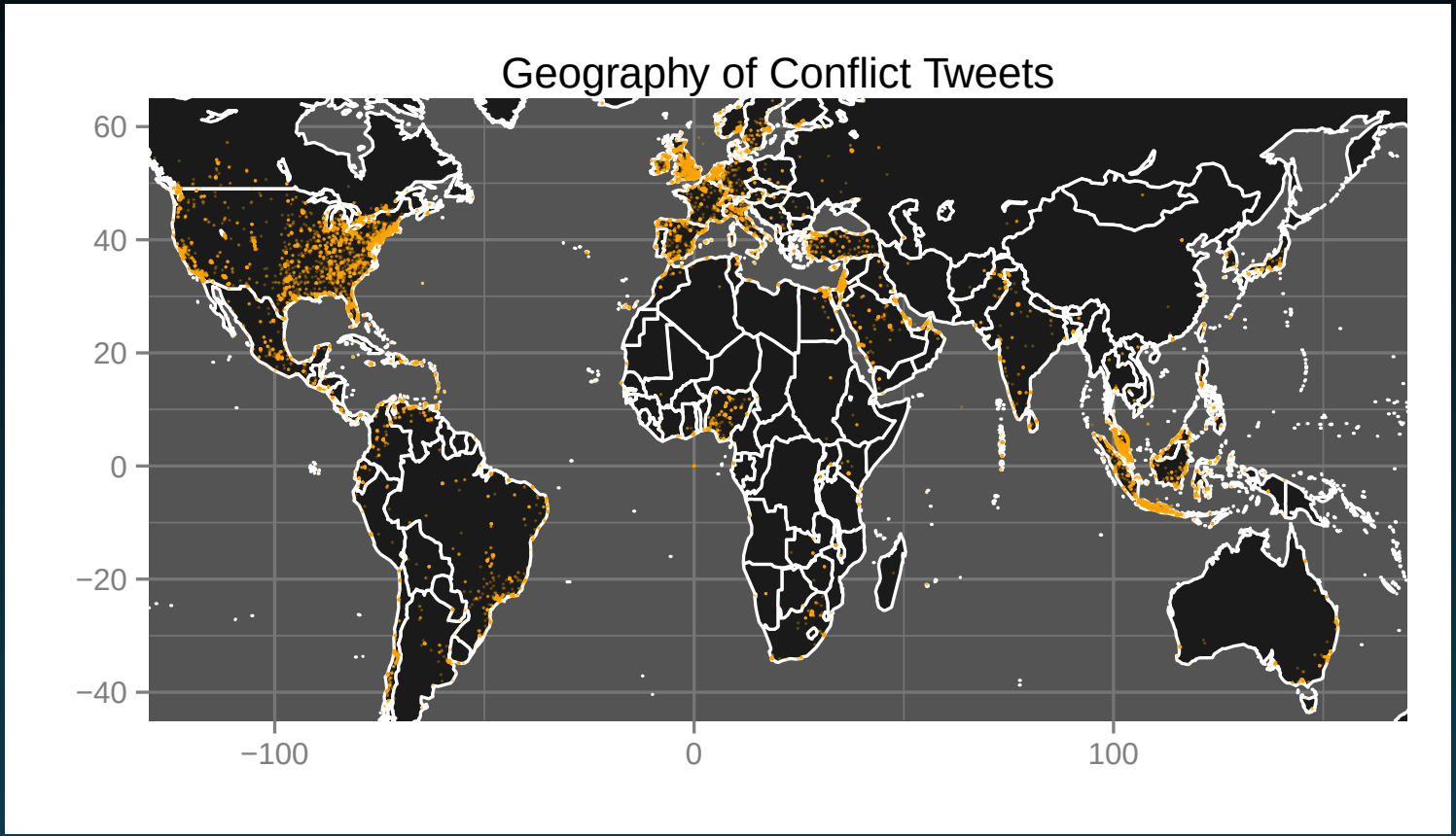

25000
20000
15000
10000
5000
0
Twitter "quake" Activity
17:45
18:00
18:15
18:30
18:45
19:00
August 2011


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)

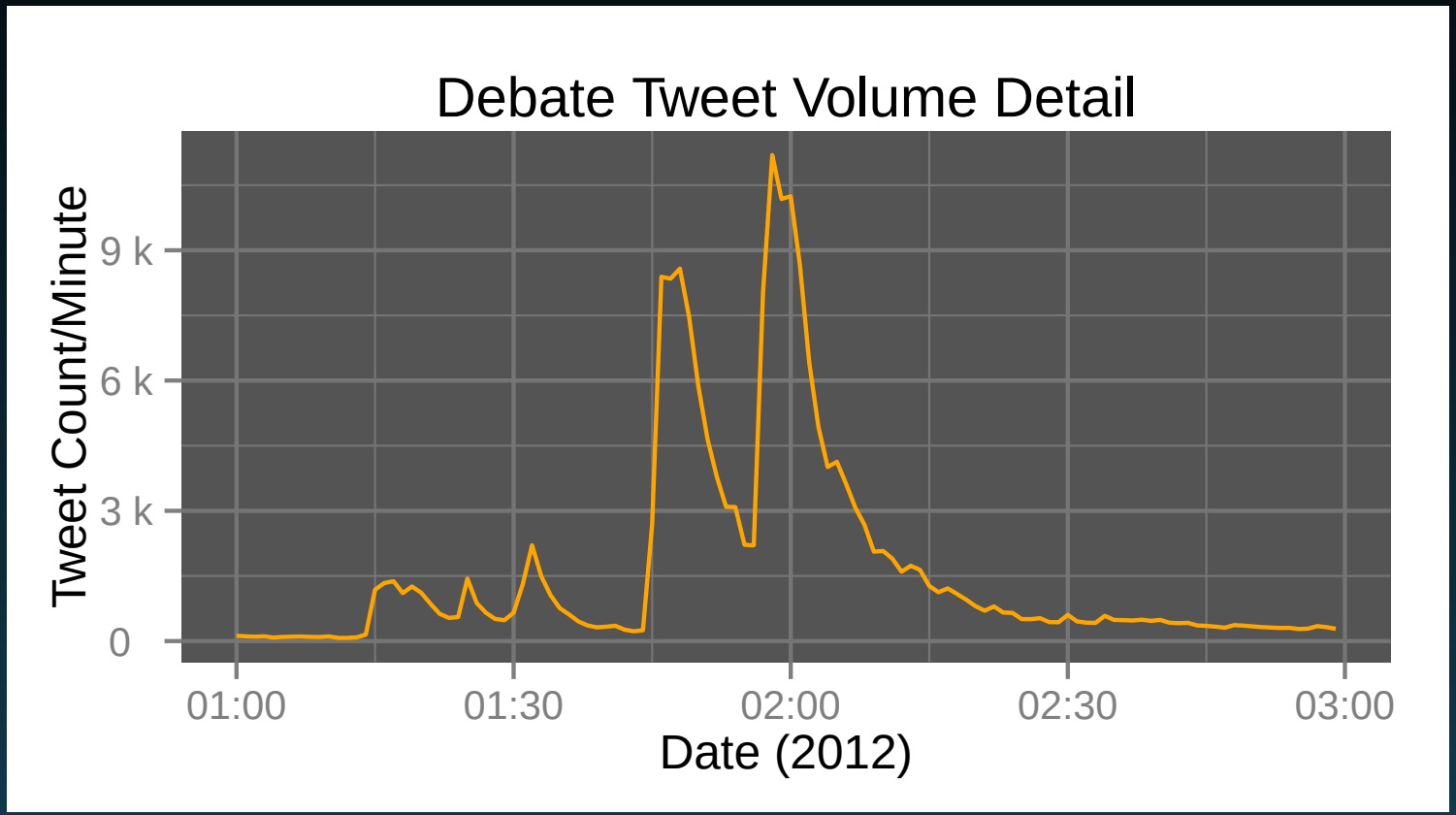
Comments with key words over time



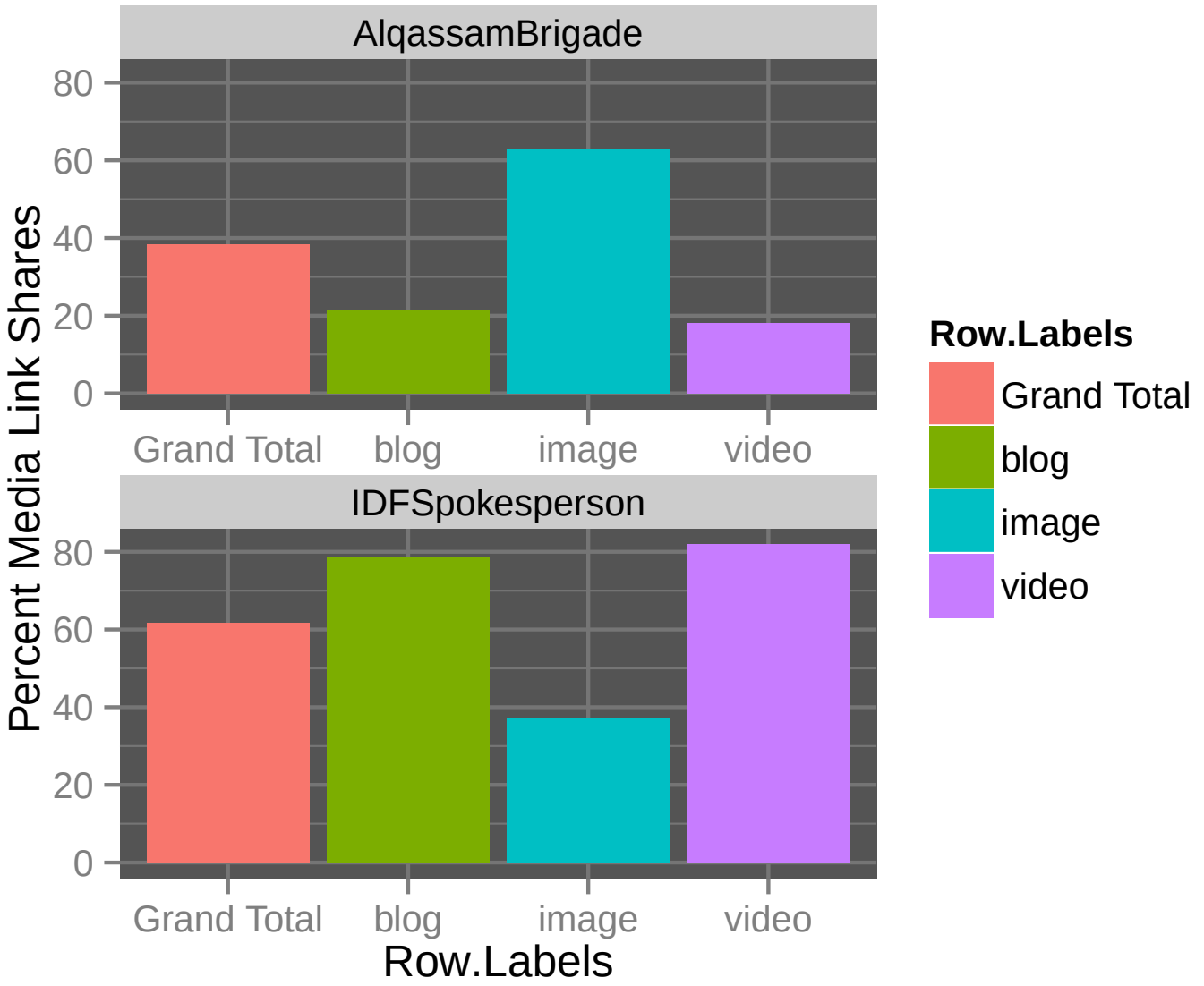
Comments with key words over time



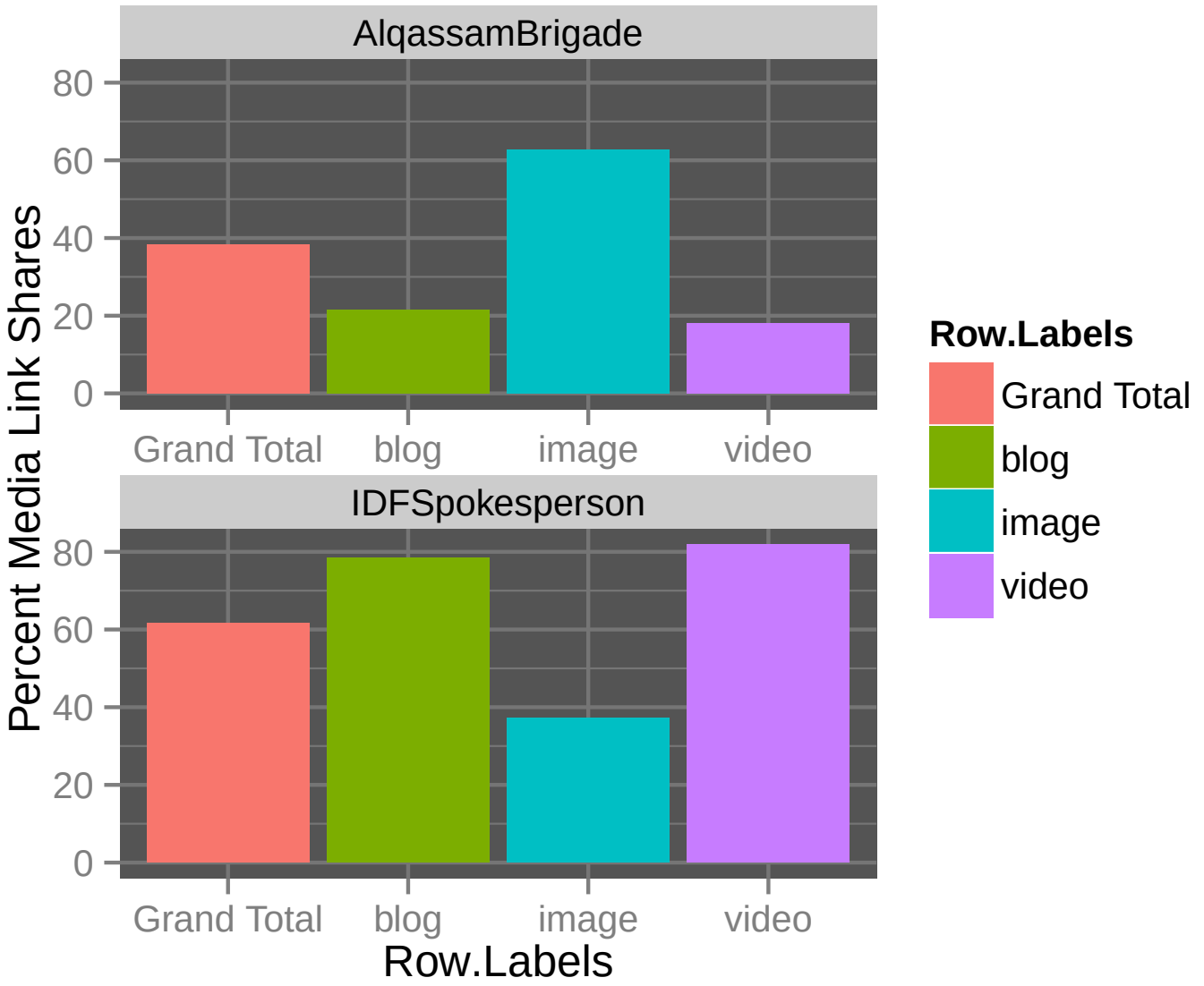
Comments with key words over time



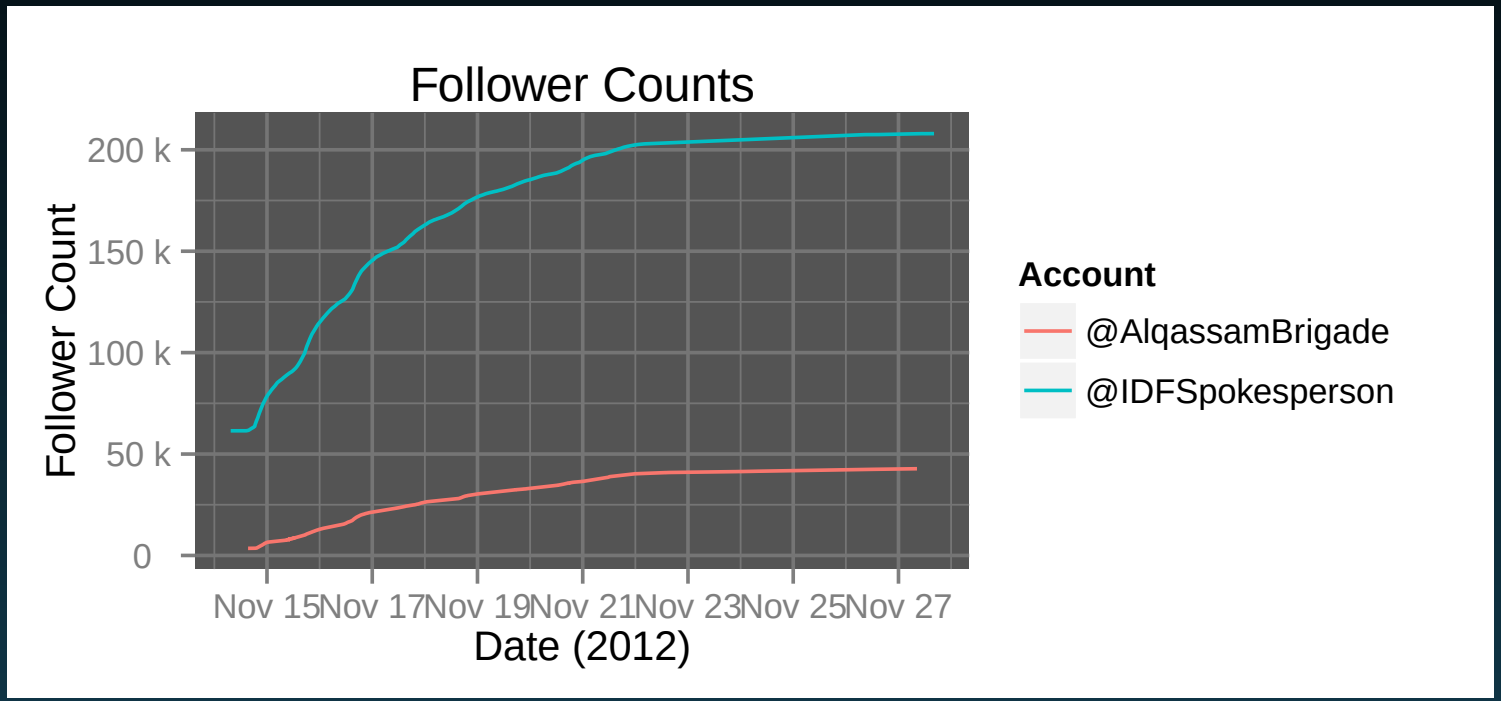
Comments with key words over time



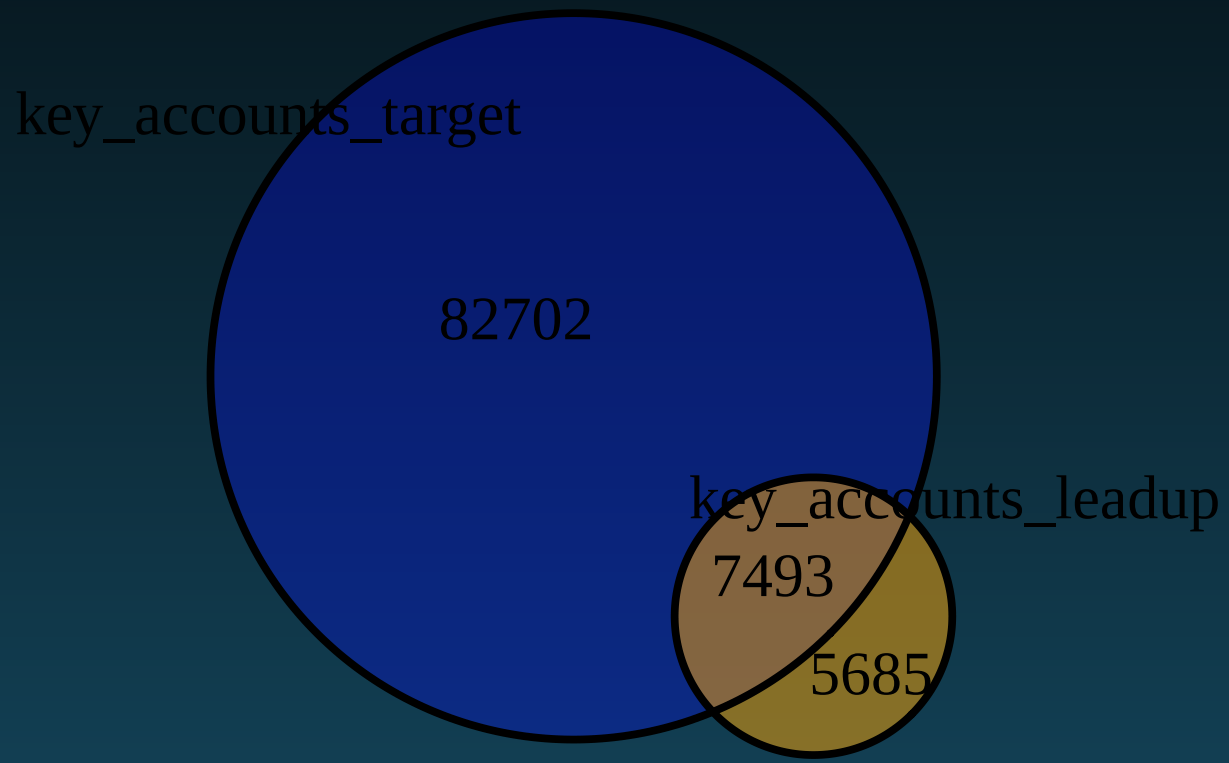
Comments with key words over time



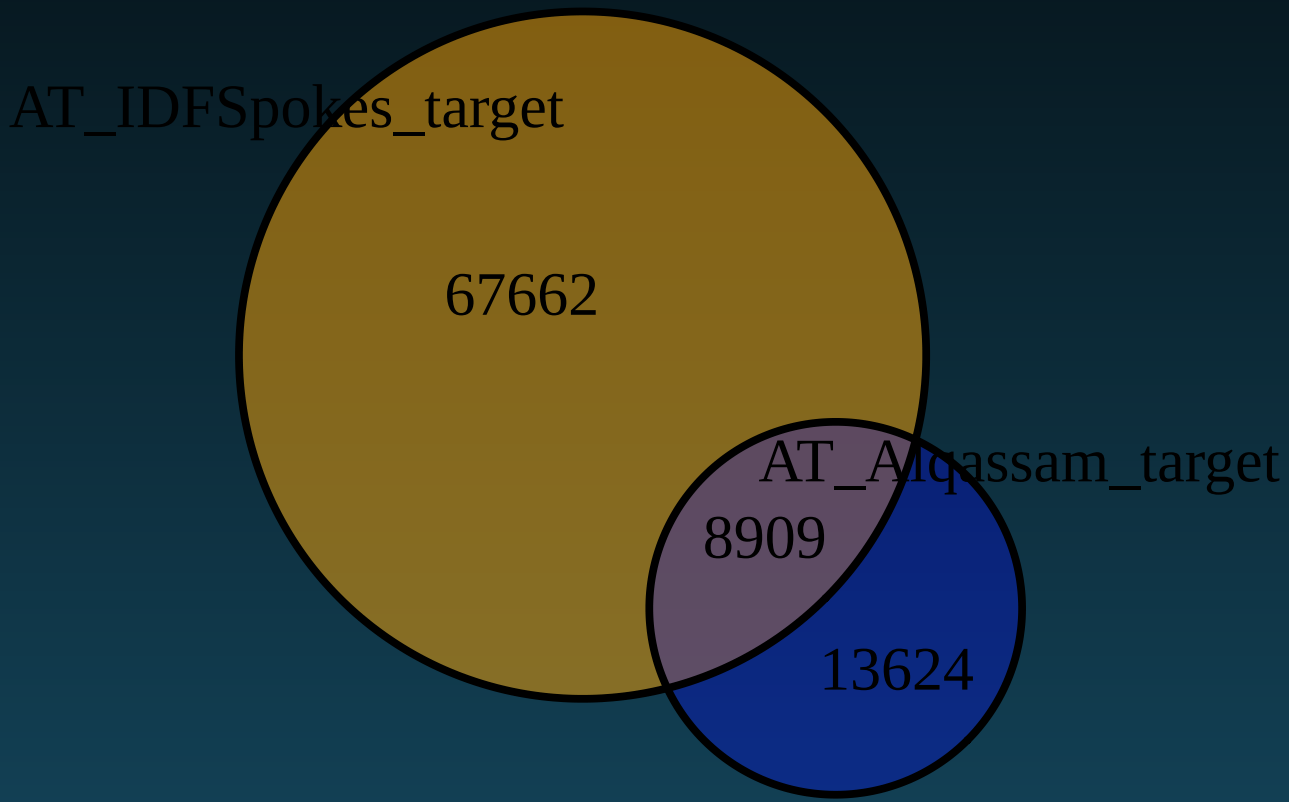
Comments with key words over time

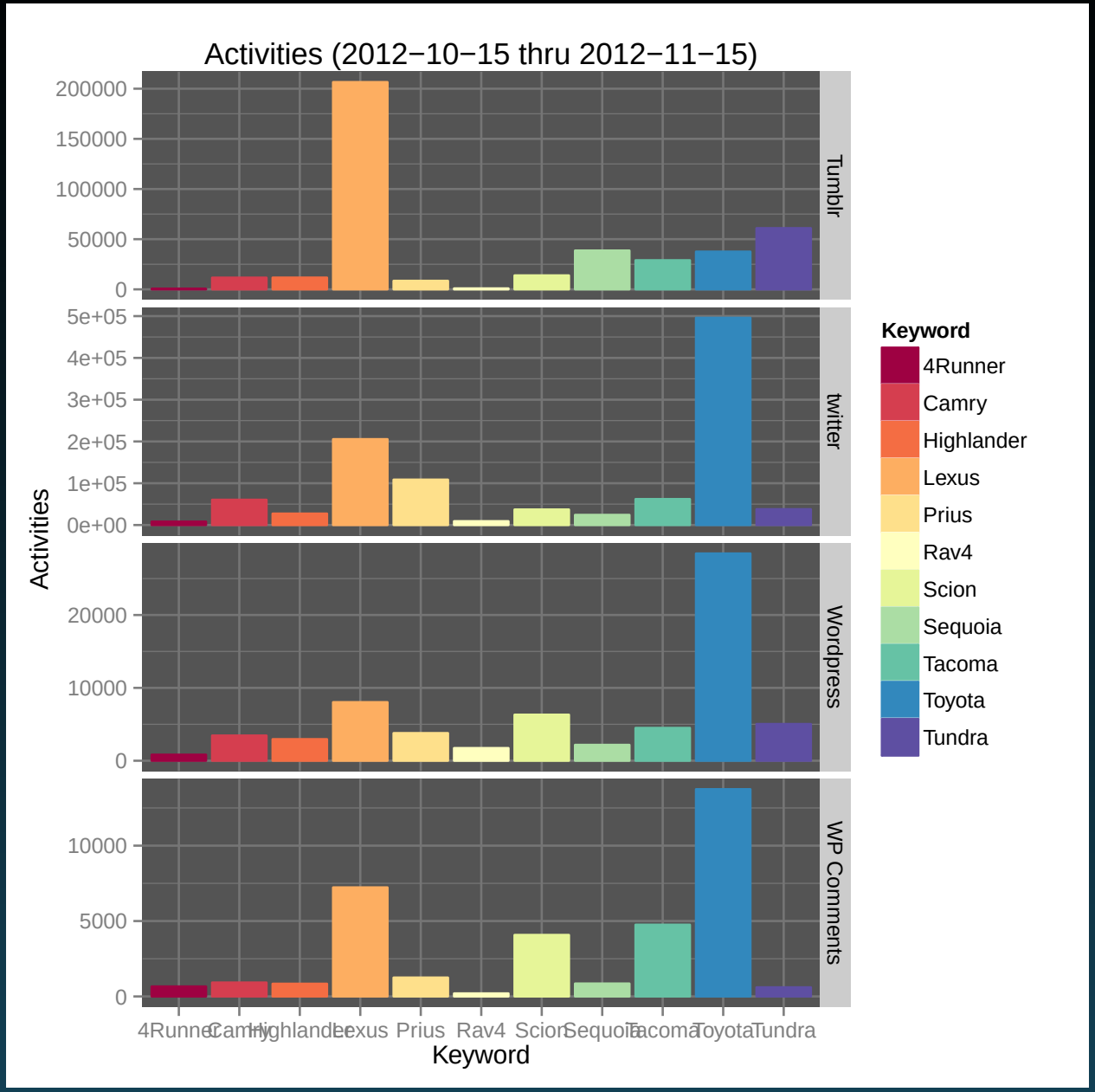


Comments with key words over time



Comments with key words over time




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

What do we talk about when they talk about X?

Apologies: Raymond Carver

Disqus Tree Structures

articles $\leftarrow$ comments

comments $\leftarrow$ comments

Disqus Threads

- 7 weeks
- Key words: “texting,” “driving” and variants
- Select top threads based on mentions
- 61,406 comments from 365 threads

Disqus Topic Model Approach

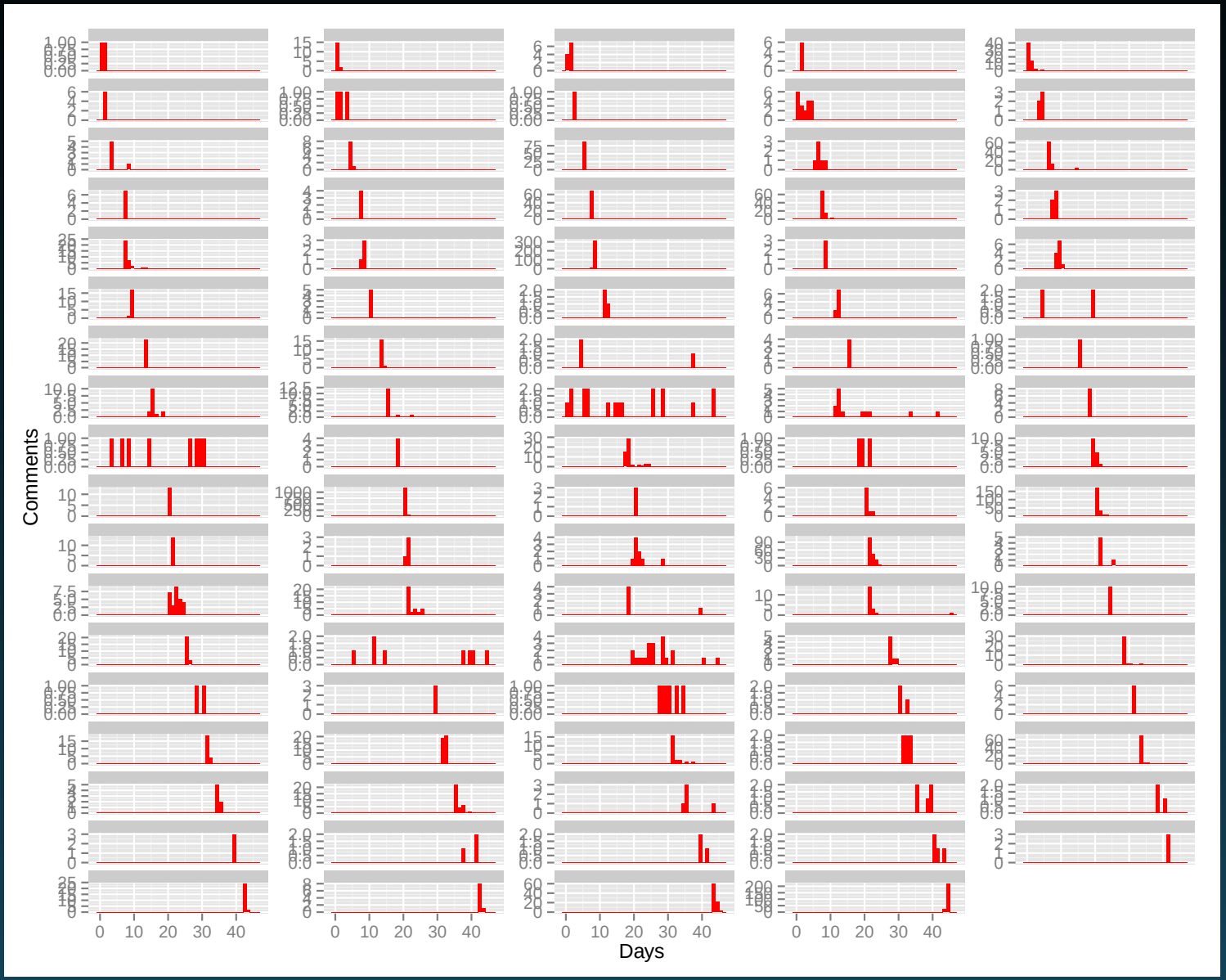
- Find comments that mention key words
- Corpus of comments (across many threads)
- tf-idf matrix: terms $\times$ comments
- LSI (rotate space to align with “important” dimensions, cut dimensions)
- K-means (quick-and-dirty clustering in reduced dimensional space)
- …rinse and repeat (looking for distinction and cohesion)

Disqus Topic Model

- Same 7 weeks; same keywords
- 32,856 comments from 16,886 threads
- LSI: 500 features $\rightarrow$ 80 features
- K-means: 80 clusters as topics (?!)

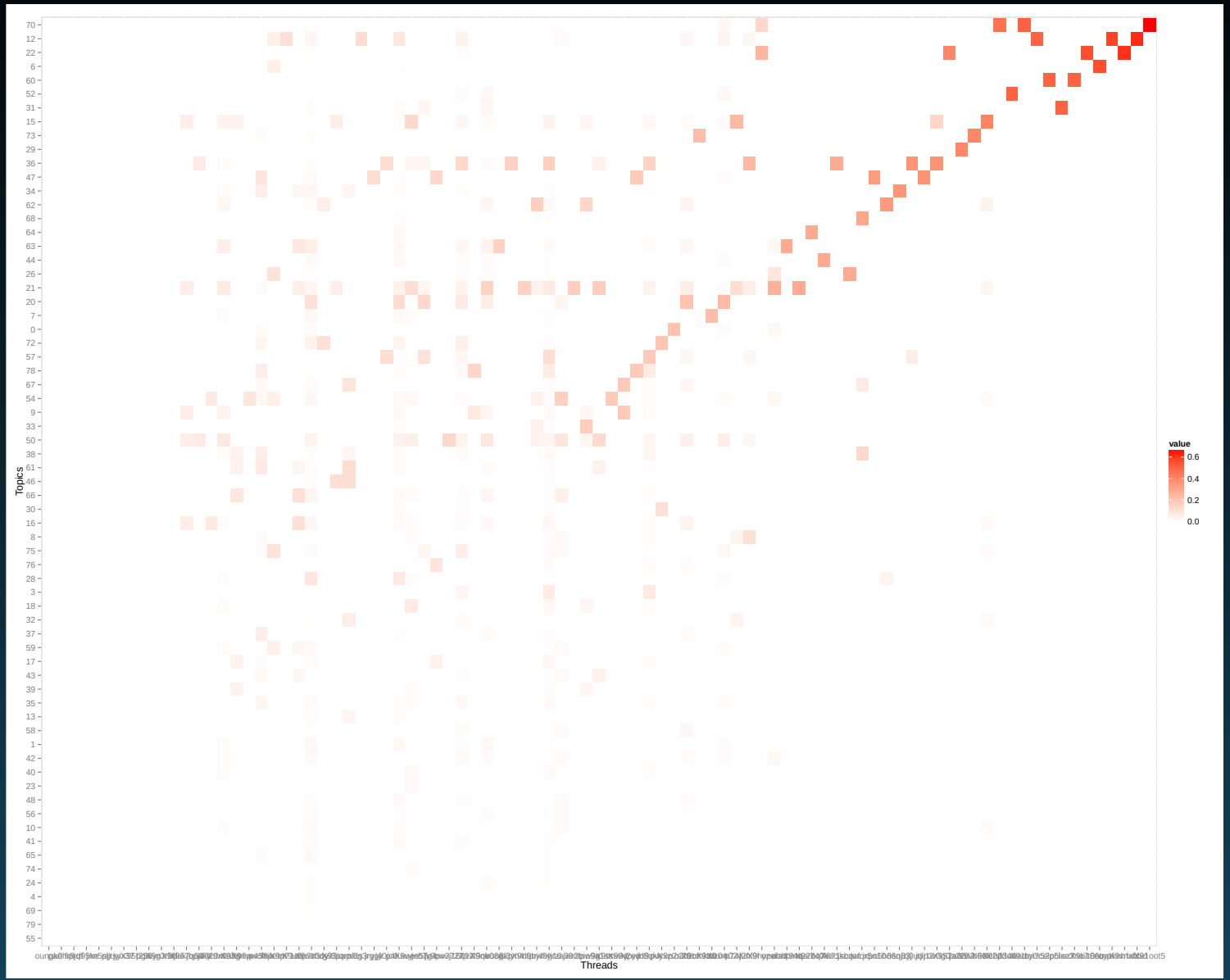
Focus on the intersection of Thread and Topic models

Disqus Thread Activity over Time



Disqus Topics Activity over Time

Dominant Topics × Threads?



When we talk about texting and driving, we talk about …

- Topic 12: poor graphic design
- Topic 50: fake ids and fake drivers licenses
- Topic 58: health/accident insurance
- Topic 62: drunk drivers
- Topic 64: buses and bus drivers
- Topic 67: bikes, bike lanes
- Topic 68: trucks and truck drivers

Thank you!

- Presentation, data, vis. code at: http://github.com/DrSkippy27/Approach-to-Leveraging-Social-Data_2013