

Data Scientist's Approach to Social Data

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

November 5, 2013



Social Data

Social data

created by
interactions
among people

Social data

form and content
shaped by
people's behavior

Sources

- firehoses
- APIs
- scraping

Firehose

Continuous stream
of activities
in near-real time

Activity

people interacting
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

Every day @Gnip

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

4 Billion OUT



Analysis



@DrSkippy27 @gnip

Darren Aronofsky's "Noah" Delayed Due to Flooding

Posted: October 31st, 2012 by [WorstPreviews.com Staff](#)



[SUBMIT COMMENT](#)

Darren Aronofsky ([Black Swan](#), [The Wrestler](#)) has been filming his "Noah" film, based on the Biblical tale of Noah's Ark, at Oyster Bay, NY. To make it as realistic as possible, the director built a massive ark, which measures 450 feet long, 75 feet tall and 45 feet wide.

Unfortunately, it was never meant to be sailed.



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

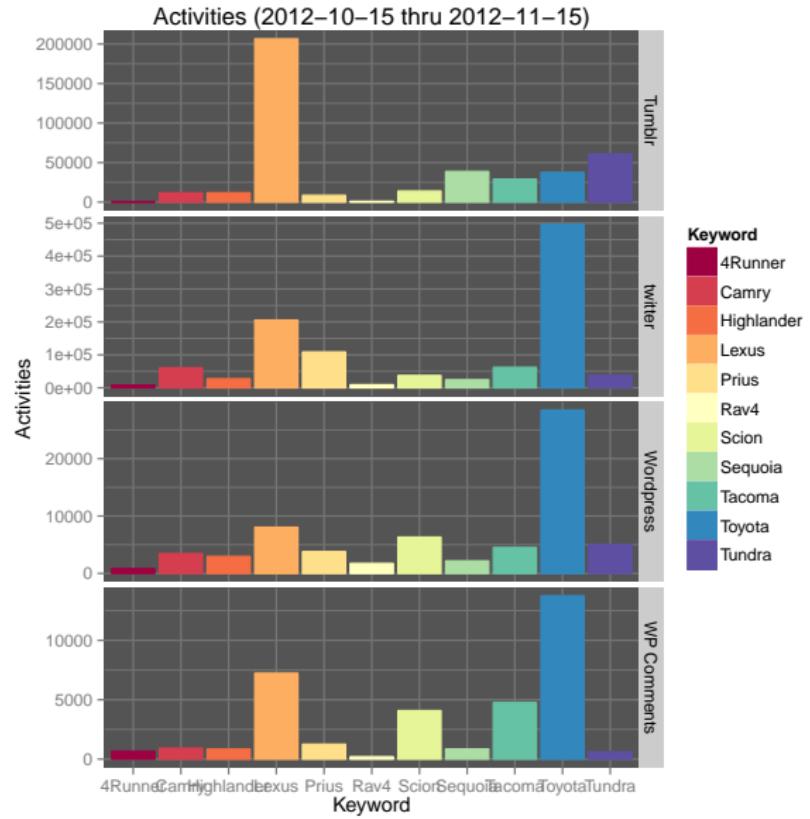
- Licensing - do you have the right to analyze, display, store data?
- Terms-of-Service Compliance - violating publishers terms of service, privacy protections?
- Cost - data collection costs? licensing costs? processing and storage costs?
- Analysis mode - batch vs. real-time? event vs. background?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

Models ...

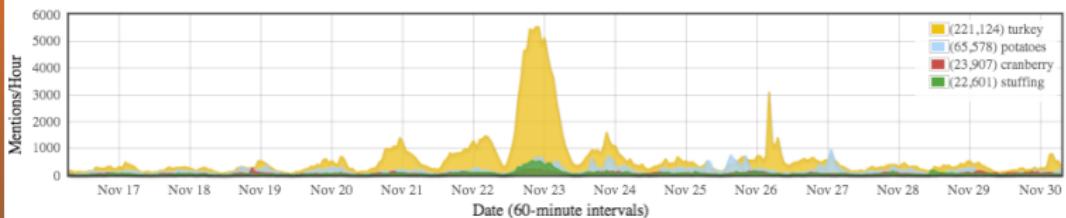
- domains - events, time series, language analysis, graph structures, etc.
- drives storage, analysis and access strategies
- analysis objectives - detection, alerting, status dashboard, discovery projects, validation, etc.



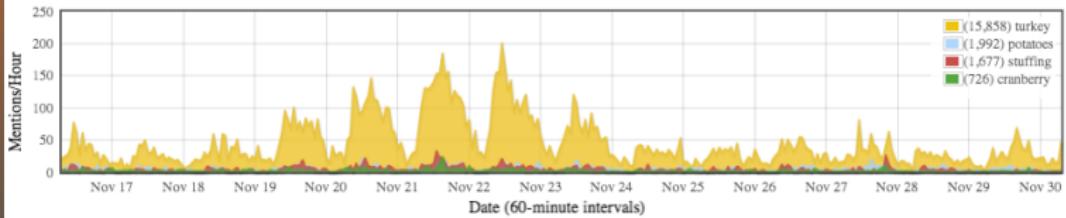
Audience



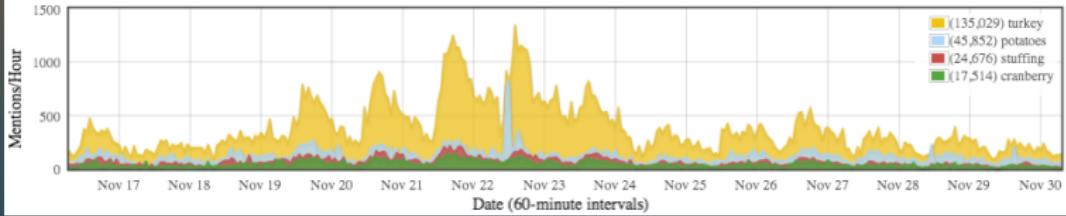
Tumblr



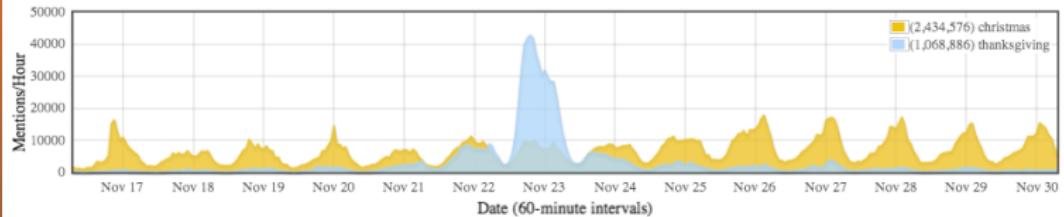
Disqus



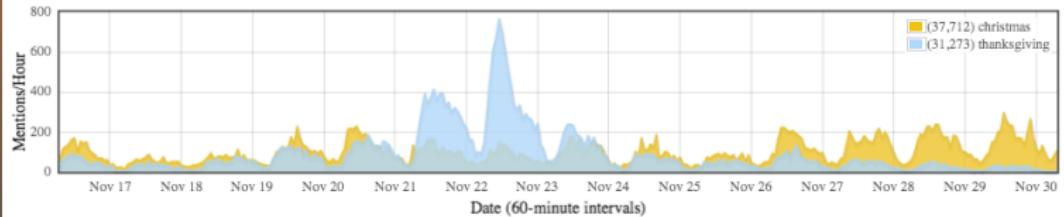
Wordpress



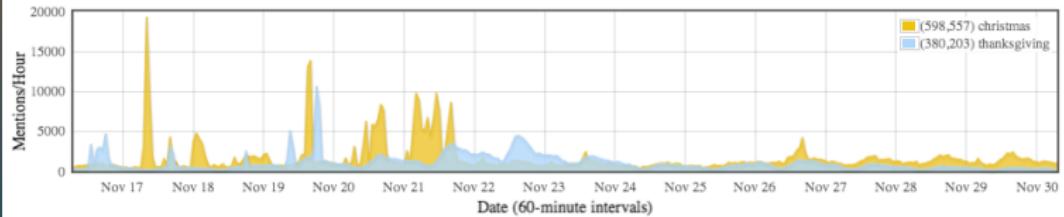
Tumblr



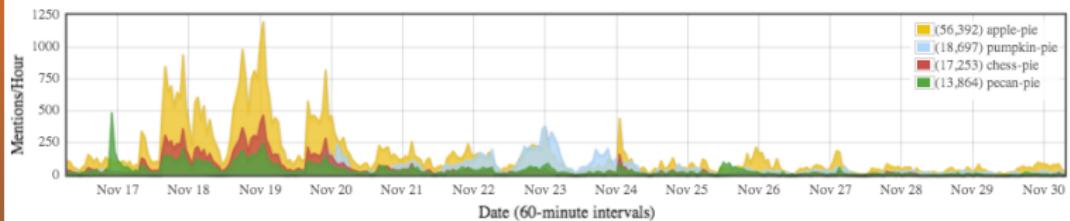
Disqus



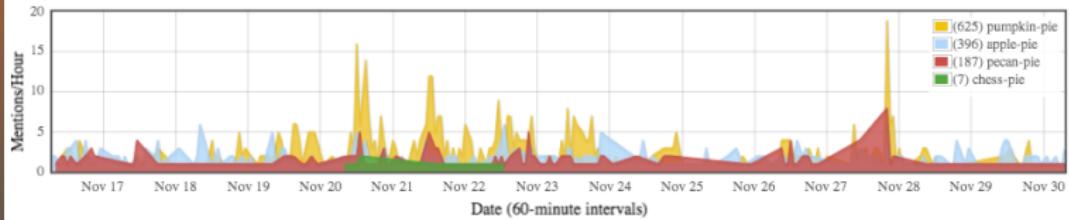
Wordpress



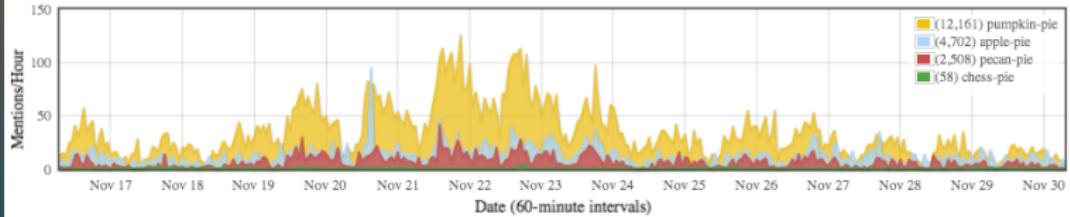
Tumblr



Disqus



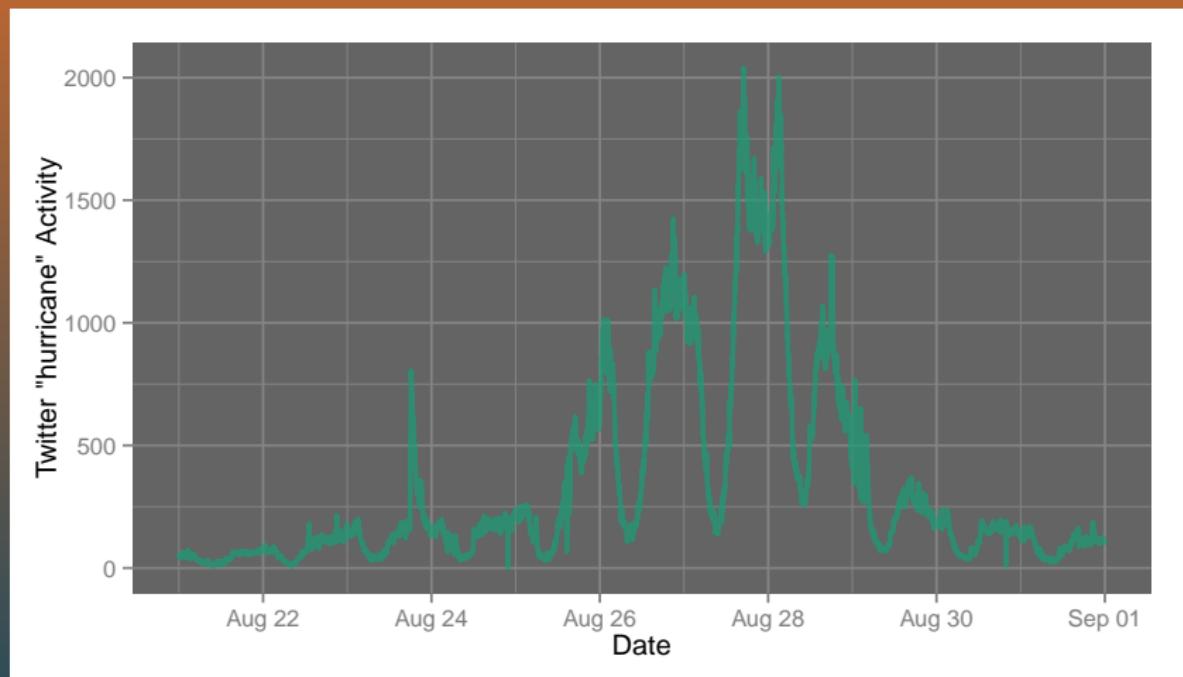
Wordpress



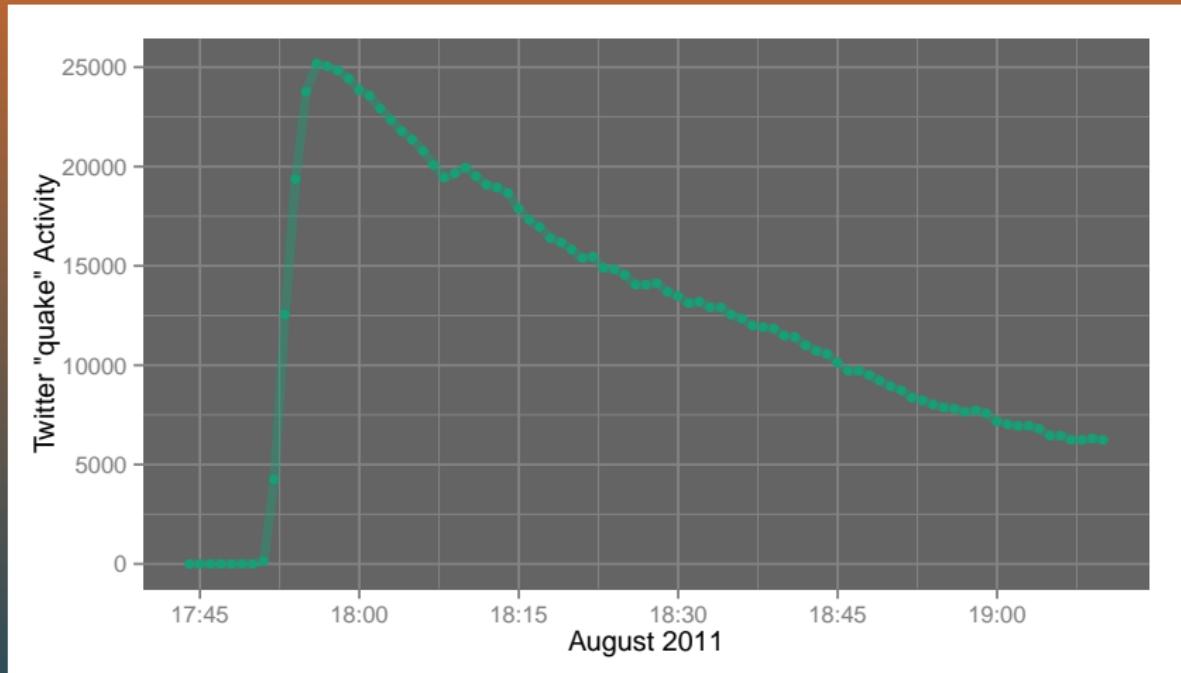


Social Medial Pulse

Expected: hurricane



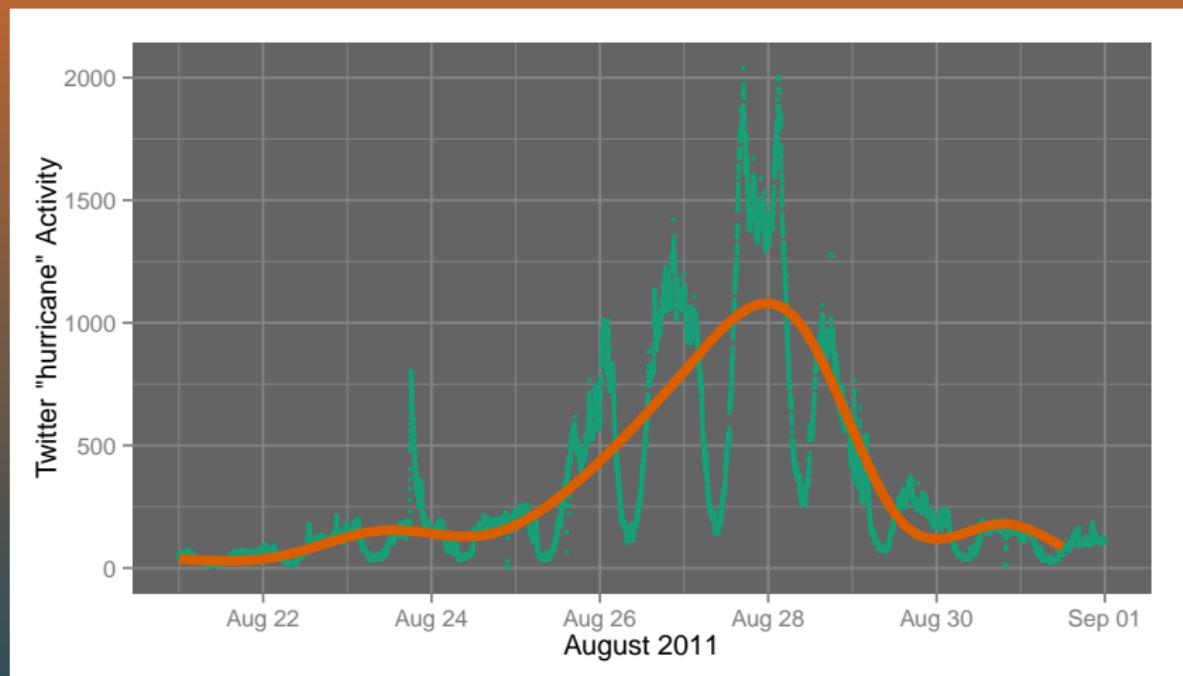
Unexpected: earthquake



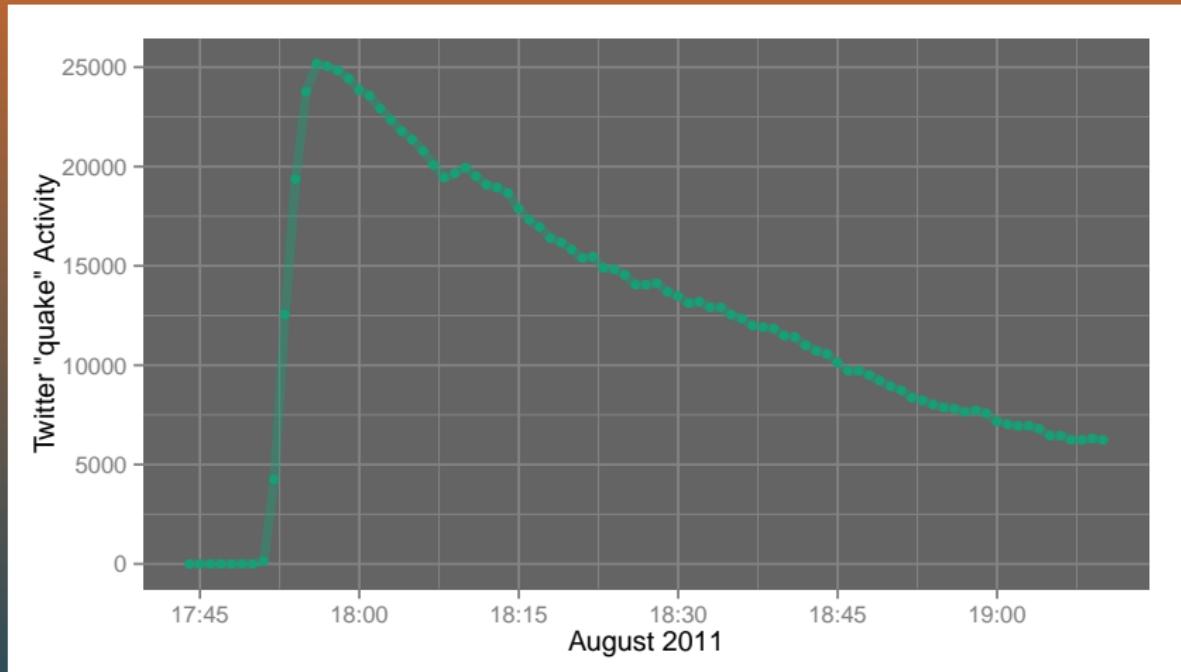
Classifying events

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

Expected: hurricane



Unexpected: earthquake



Half-life

time to observe
 $\frac{1}{2}$ of the activities
for an event

Social media pulse

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

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

Many people posting on same cue; so sum of random variables

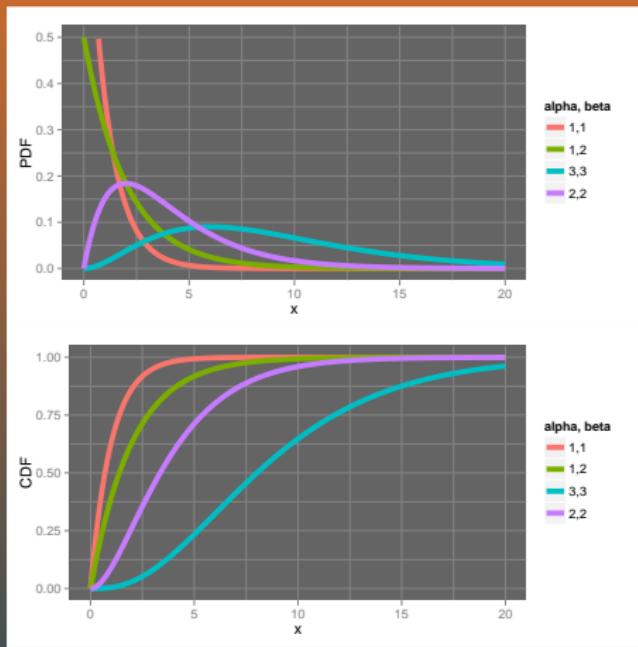
$$S = X_1 + X_2 + \dots + X_n \text{ posters}$$

Gamma probability distribution function,

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

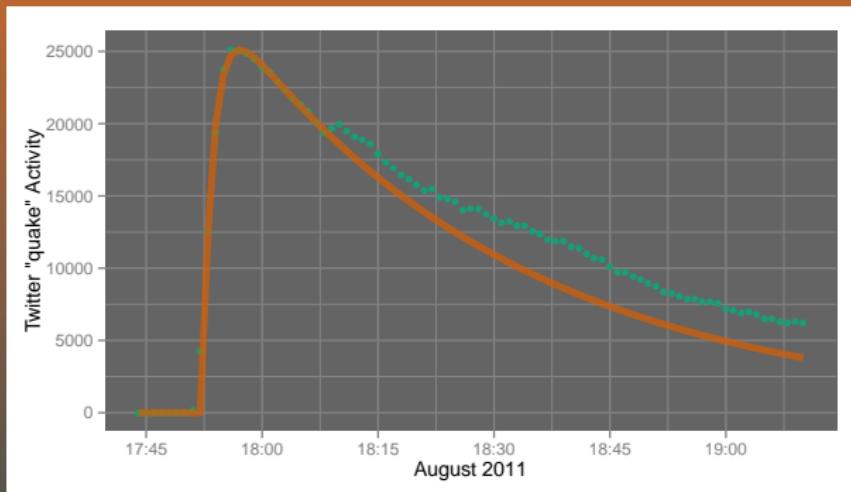
Cumulative distribution is the “generalized regularized incomplete gamma function”,

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

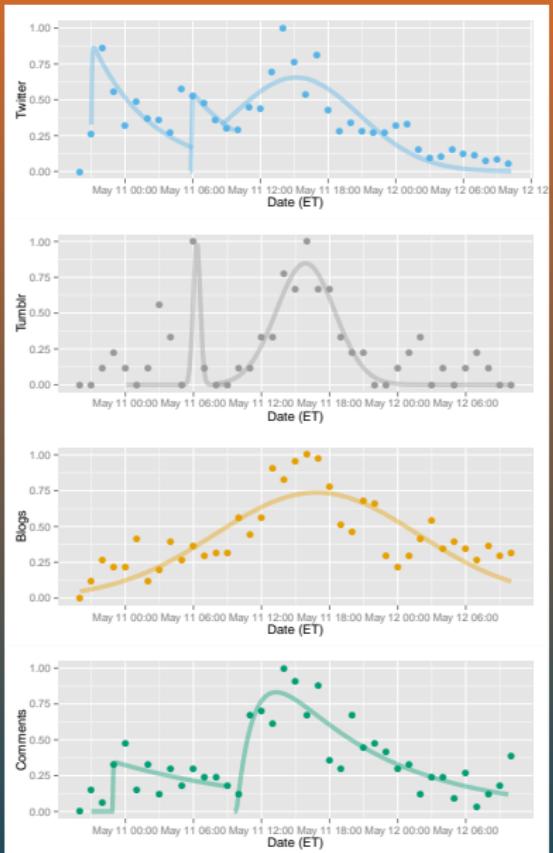


Why model half-life?

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

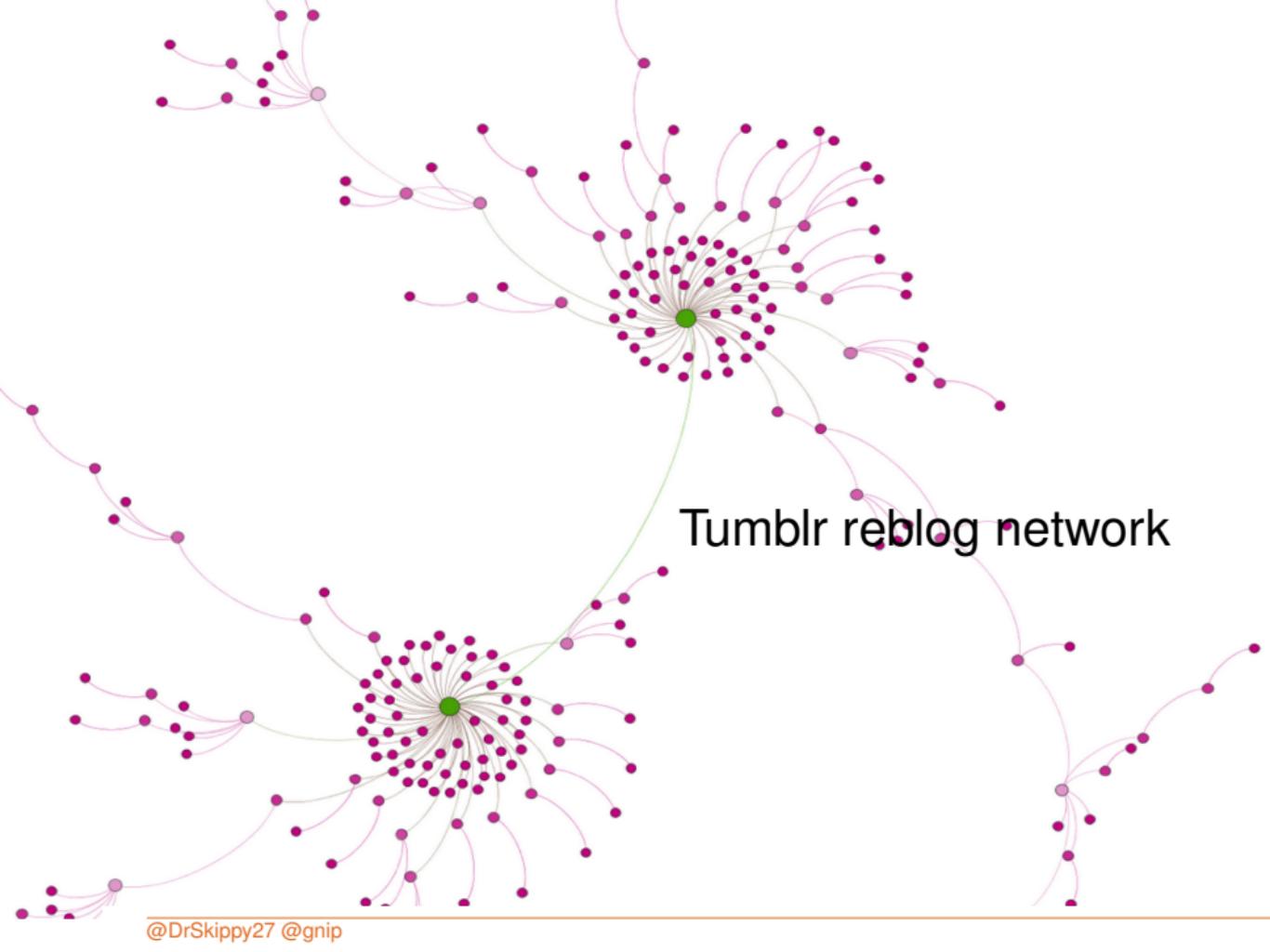


JPMorgan (-\$2B)

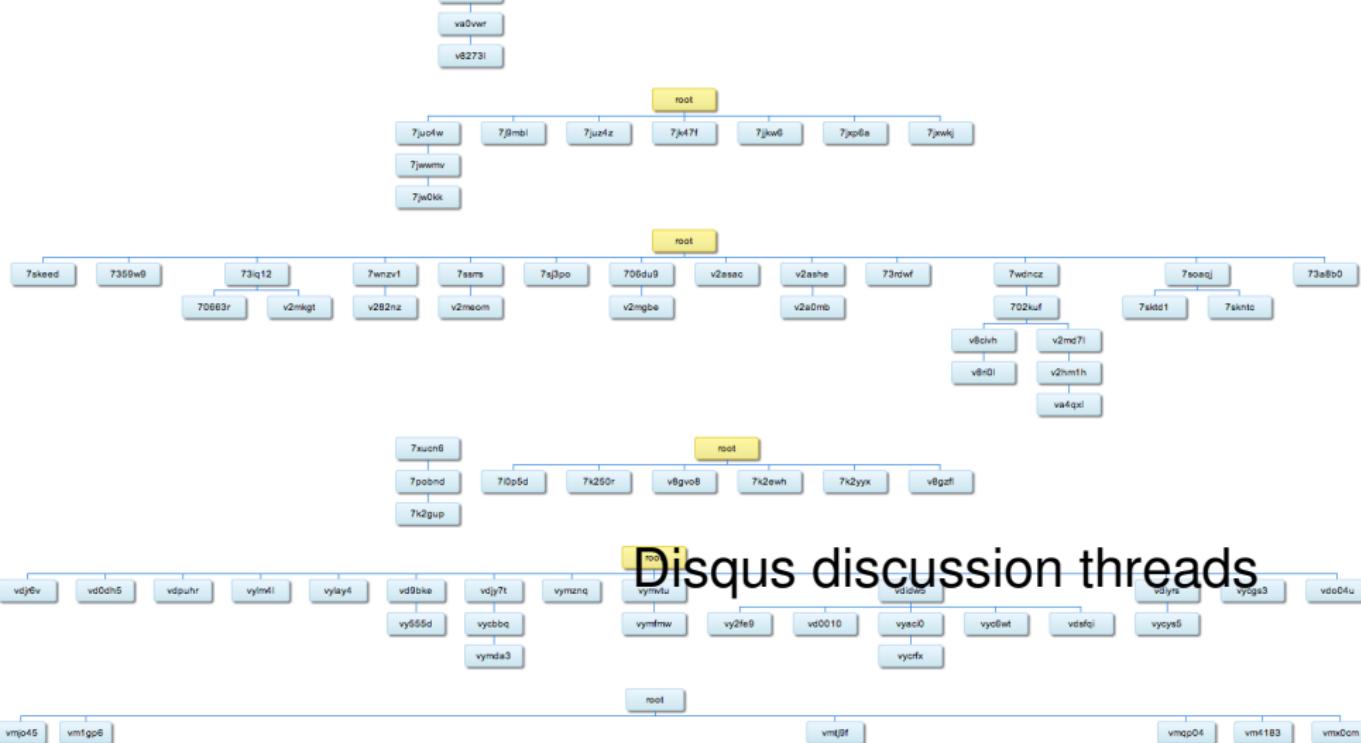




Structure



Tumblr reblog network



@DrSkippy27 @gnip



Structure and Language

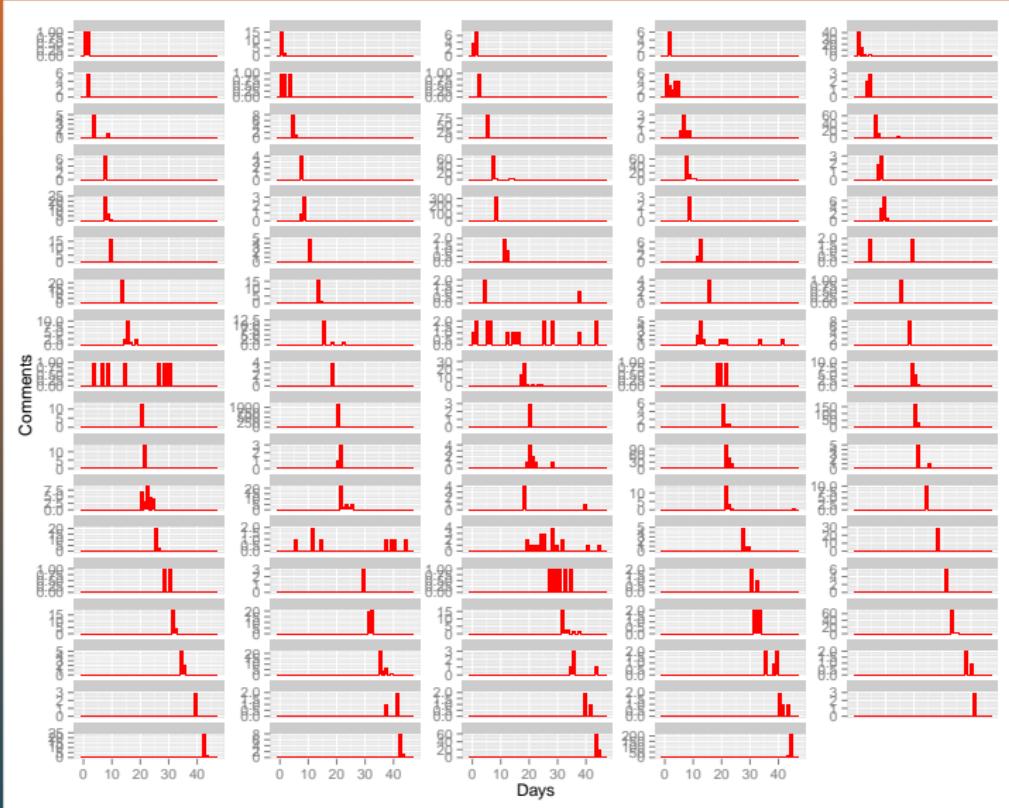
what do we talk about
when they talk about X?

Apologies: Raymond Carver

Disqus threads and topics

- 7 weeks data
- Key words: “texting,” “driving” and variants
- Select top threads based on mentions
61,406 comments from 365 threads
- Select comments based on mentions 32,856
comments from 16,886 threads
- LSI: 500 features → 80 features; 80 clusters

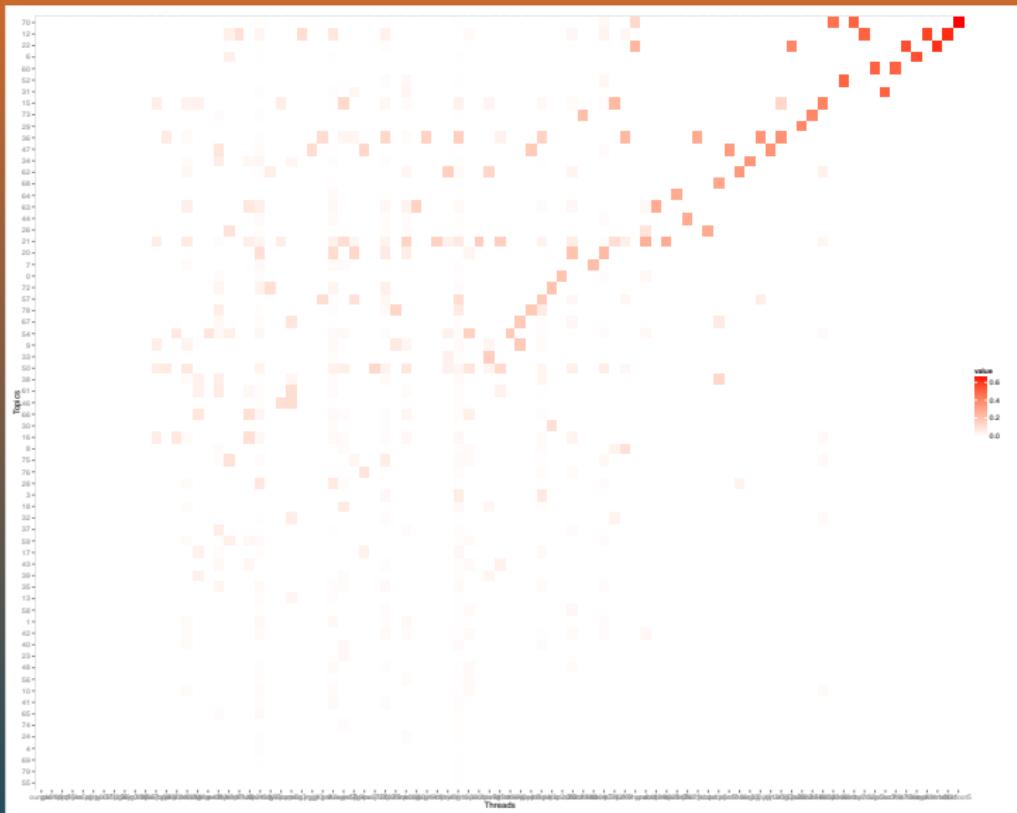
Thread activity



Topics activity



Topics × Threads



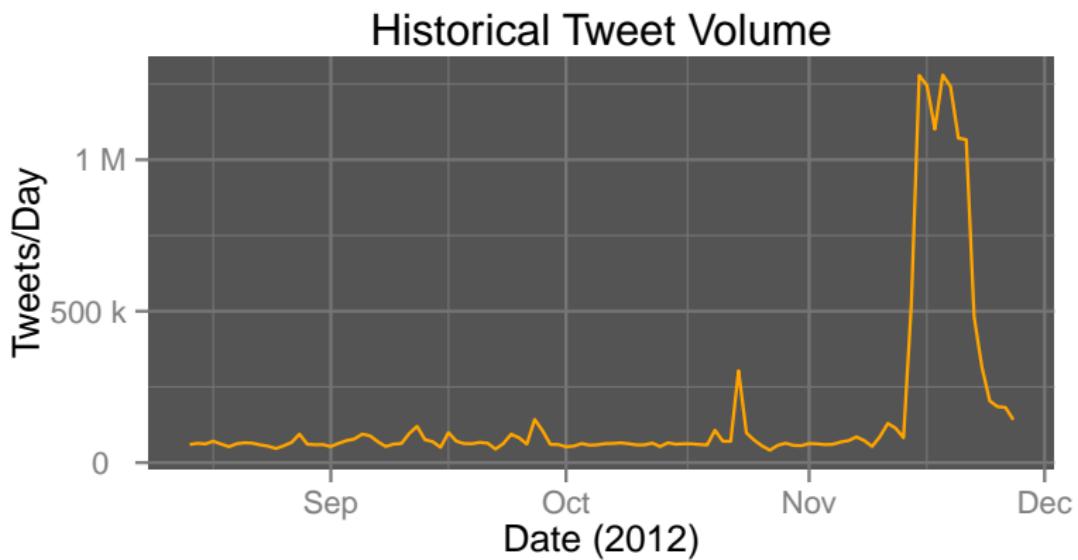
When we talk about texting and driving, we talk about ...

- 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

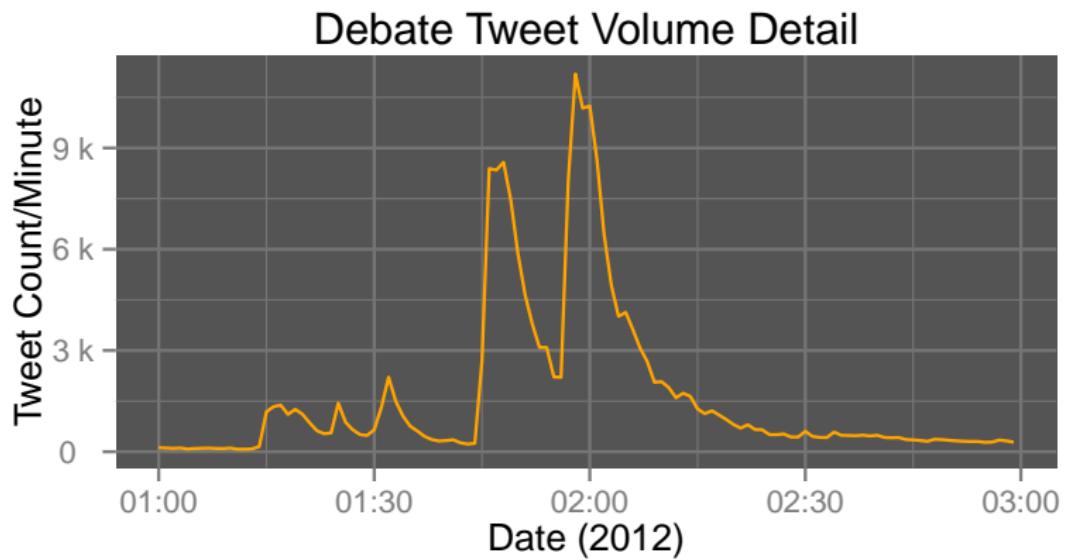
A vertical column of dark blue ink is suspended from the top center of the slide, creating a dynamic, organic shape that resembles a stylized figure or a cloud. The ink is thick at the top and tapers as it descends.

Synthesis

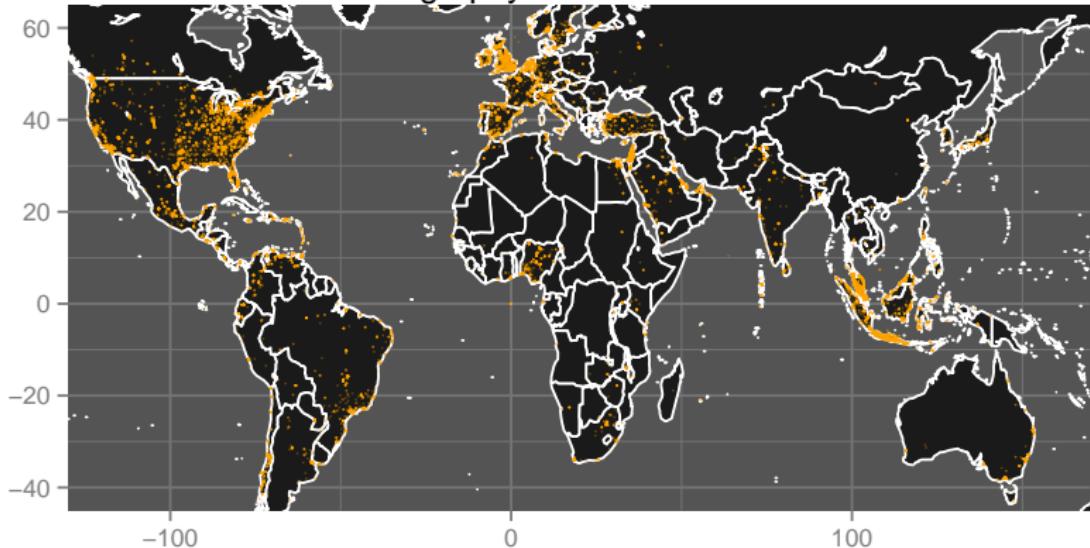
Hamas vs. Israel in Gaza 2012



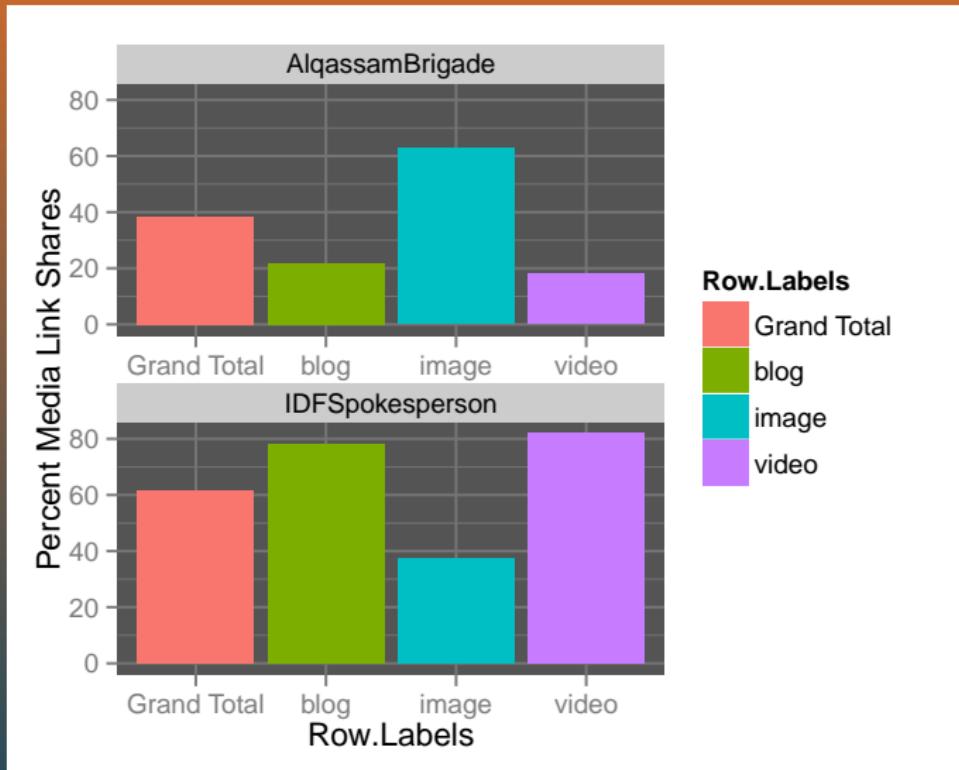
What's that leading spike?



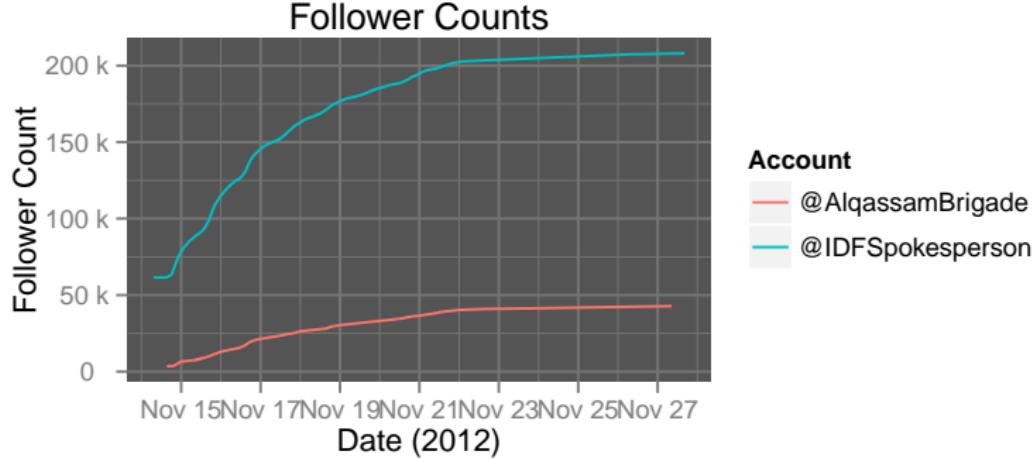
Geography of Conflict Tweets



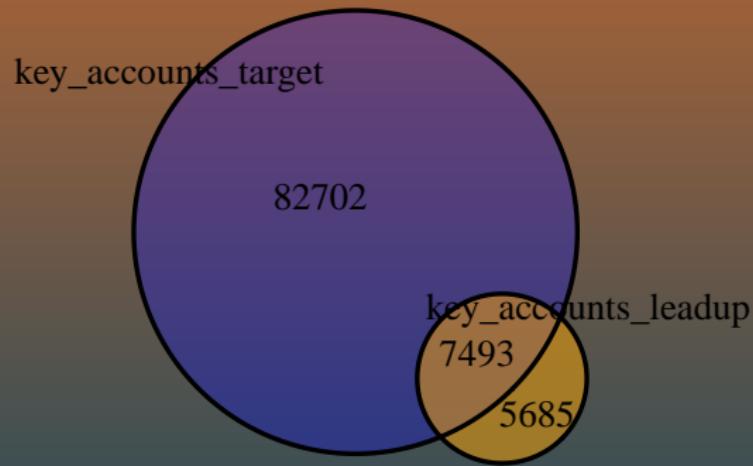
Key accounts use of media



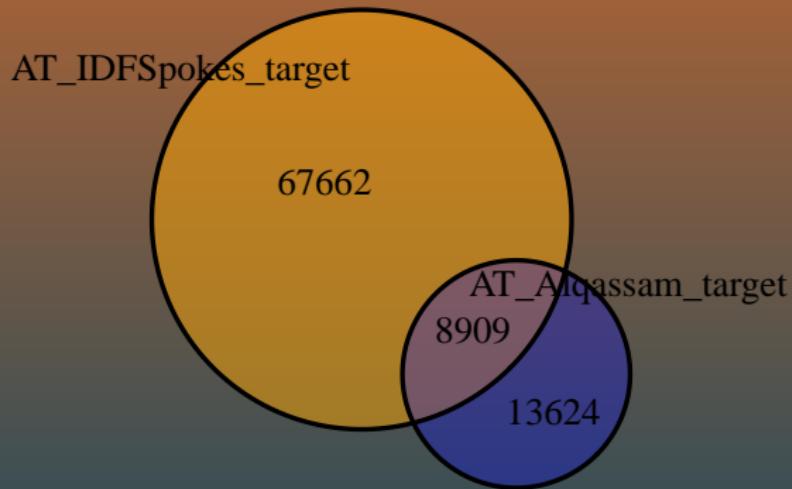
Conflict grows followers



Engagement with key accounts



Crossover engagement





Mixing it up

tumblr.



Instagram

YouTube

DISQUS



reddit

intense debate

bitly

StockTwits

flickr

vimeo

newsgator

Reaction Time



intensedebate

DISQUS



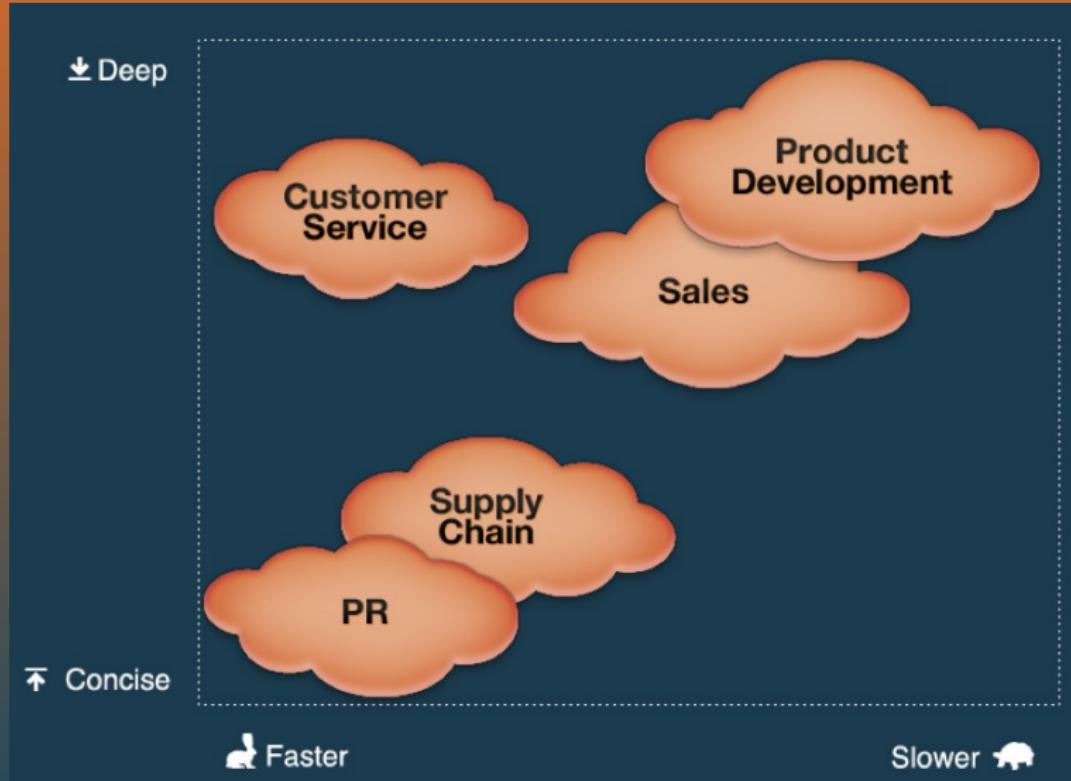
Faster

Slower

Depth



The Social Cocktail



Thank you!



- Presentation at: [http://github.com/DrSkippy27/
Approach-to-Leveraging-Social-Data_2013](http://github.com/DrSkippy27/Approach-to-Leveraging-Social-Data_2013)