Gnip, with examples…

Data Science/Business Analytics Meetup - Denver

Scott Hendrickson
Data Scientist
@DrSkippy27

December 6, 2012

We believe social data has unlimited value and near limitless application

…this refrain helps keep us focused on our ultimate goal: to be the source of record for all public social conversation.

Providing Social Media Data to the Enterprise

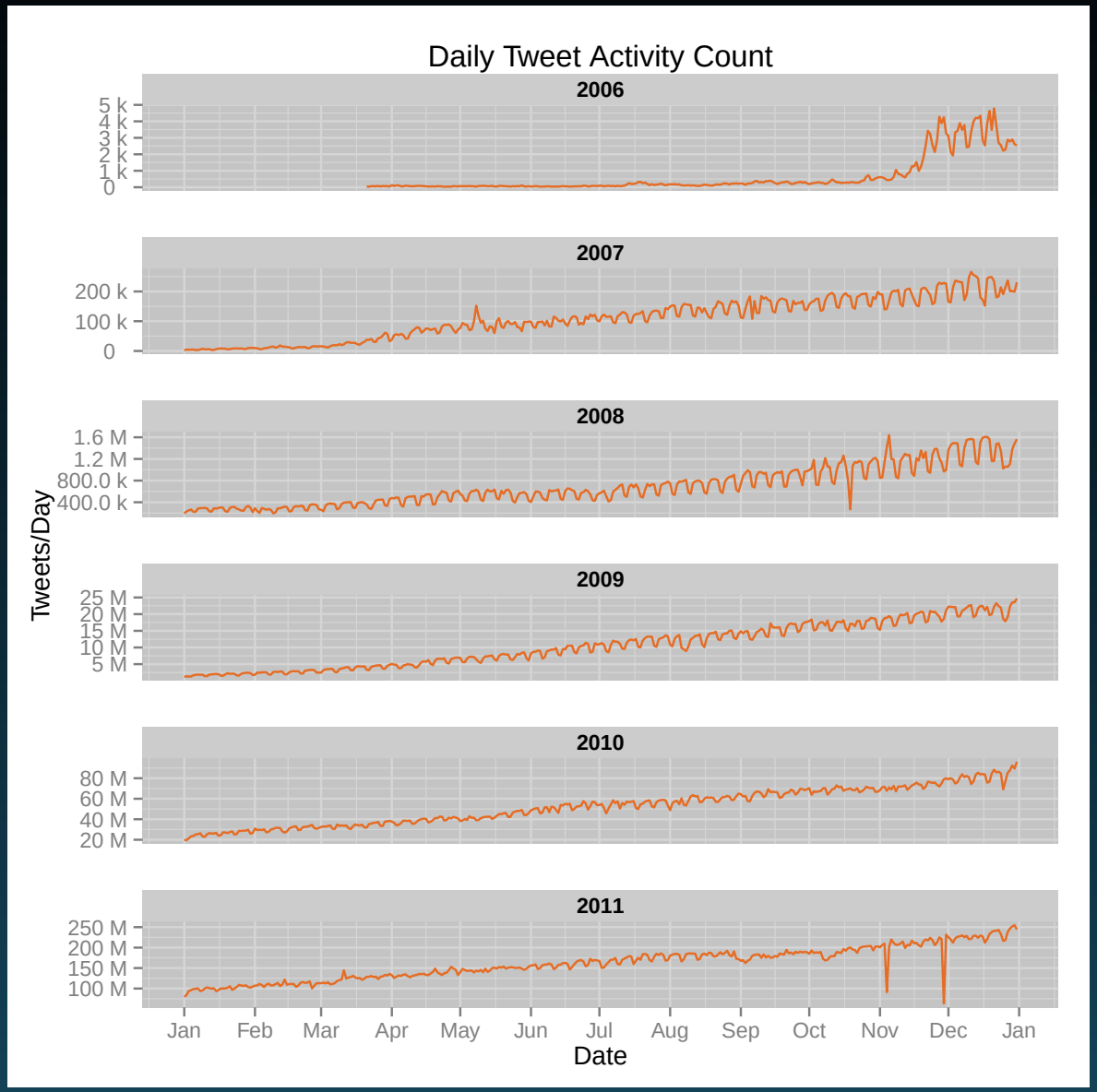
- Firehoses -
  Twitter, Tumblr, Wordpress, Disqus …
- Data Collectors -
  API access (Gnip !Ping) to Facebook, YouTube Comments, Reddit, G+, Flickr …

Firehose

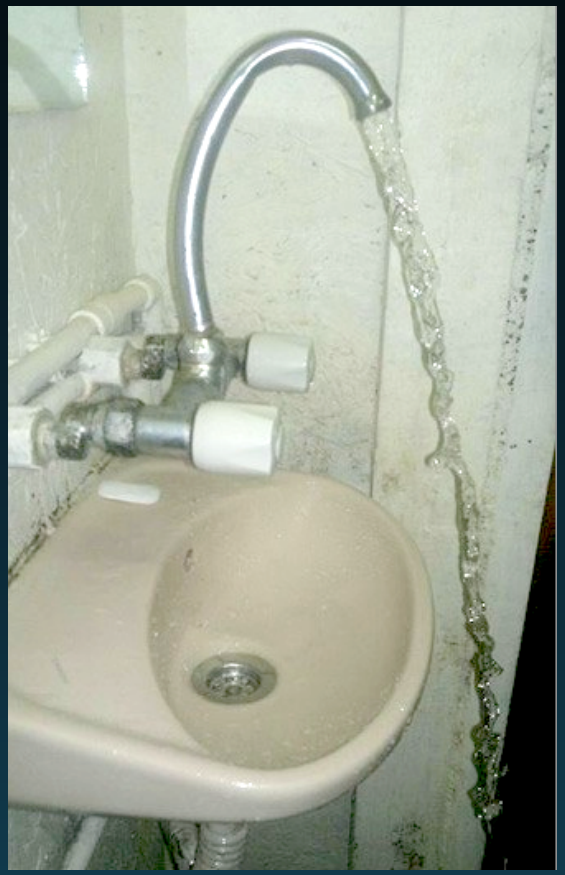
Continuous stream of flexibly structured social media activities in near-real time.

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

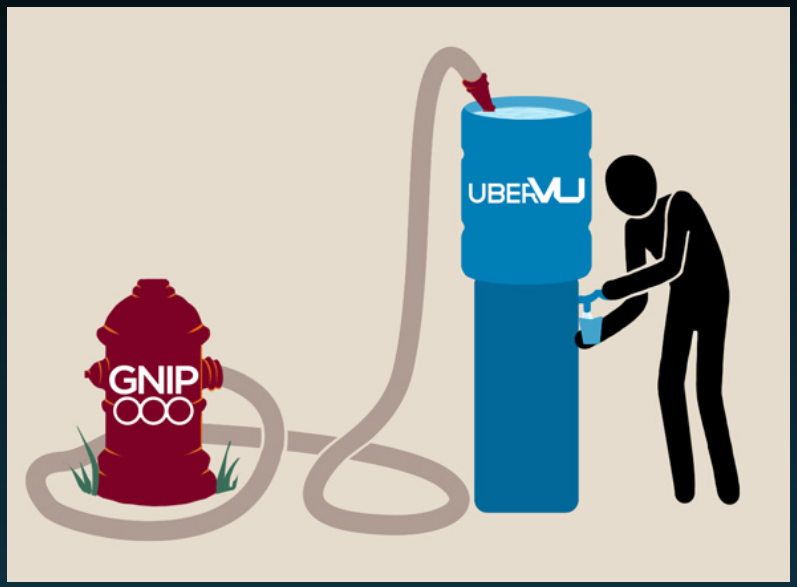
Twitter Growth (2006 - 2010)…



Without Gnip…

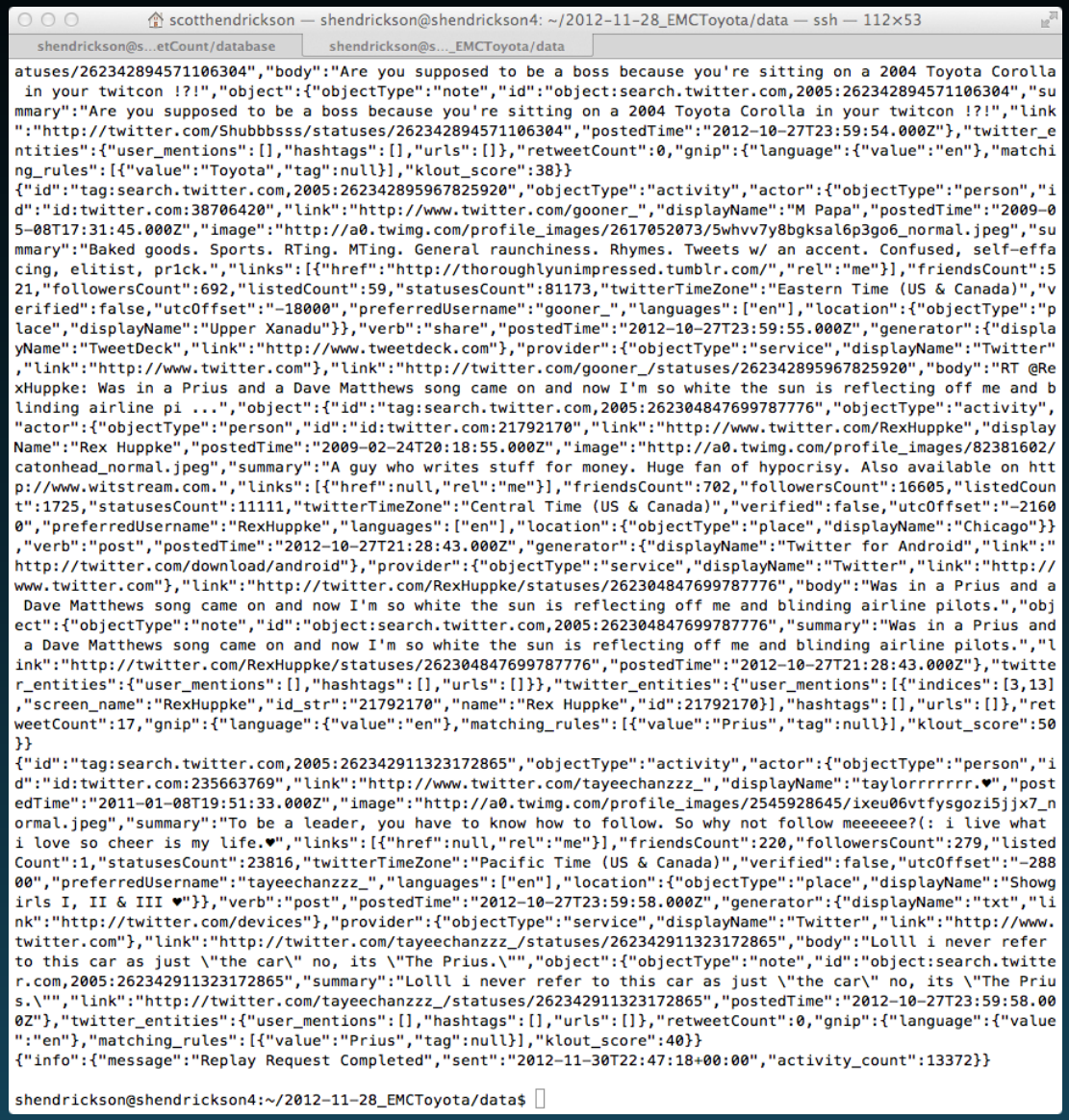


With Gnip…



http://blog.ubervu.com/ubervu-is-now-plugged-in-to-gnip.html

Streaming firehose with cURL …



Example Activity: 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/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,
```

PowerTrack: Shape and Filter

- core idea: token filter (e.g. "obama", "beer" …)
- operators: meta-data (geo, language, bios, …)
- operators: shaping (e.g. sample:10
- operators: (by publisher) user, hashtag, language,…
- combinations:

$$
\begin{aligned}
\textit{newline} &= \textit{OR} \\
\textit{space} &= \textit{AND} \\
\text{“}\textit{AND}\text{”} &= \textit{AND} \\
\text{“}\textit{OR}\text{”} &= \textit{OR} \\
\text{“}\_\text{”} &= \textit{NOT} \\
\text{“}(\ldots)\text{”} &= \textit{grouping} \\
\text{“}\ldots\text{”} &= \textit{literal}
\end{aligned}
$$

Example Toyota PowerTrack Rules

-evt10023885212 (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

lang:en toyota recall

lang:ru 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

lang:en toyota yaris -lkw -aqua -bail -died -film -toka -camry

Enterprise Features

- Rules: 250K rules
- Update individual rules without disconnect, $\lesssim$ 1s update time (100s of rules)
- Rule tagging
- Keep alive - signal that connection is live, even when no data is coming
- Low latency: avg 1s Twitter; 10s Twitter enriched …
- Redundancy - multiple simultaneous connections
- Backfill - buffer data and fill in if short term disconnect
- Replay - connect with start and end dates to stream past time periods (<5 days)
- Historical PowerTrack - Twitter historical filtering for any time period

Social media pulse: The shape of breaking news on social media

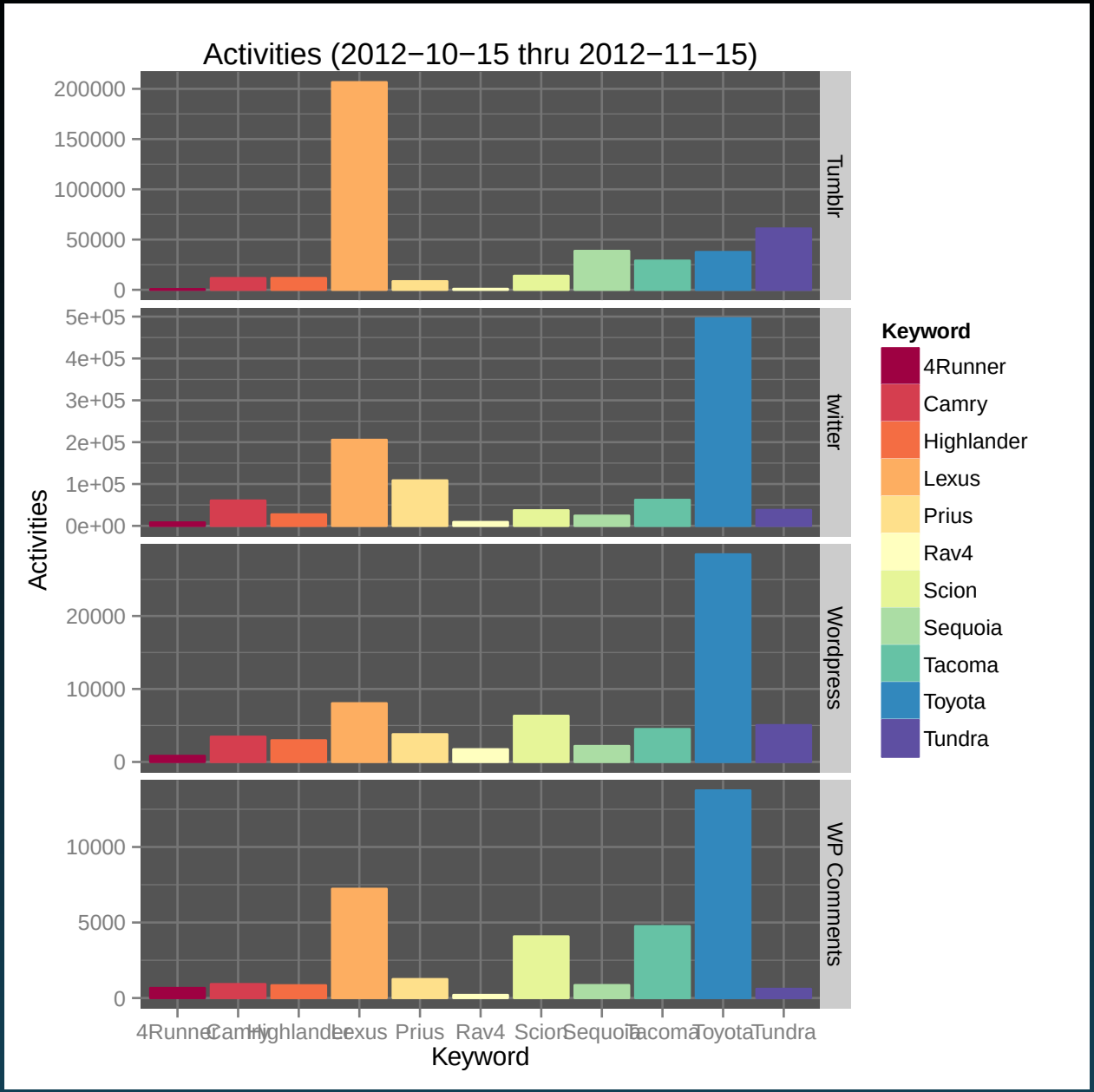
why social media for breaking stories?

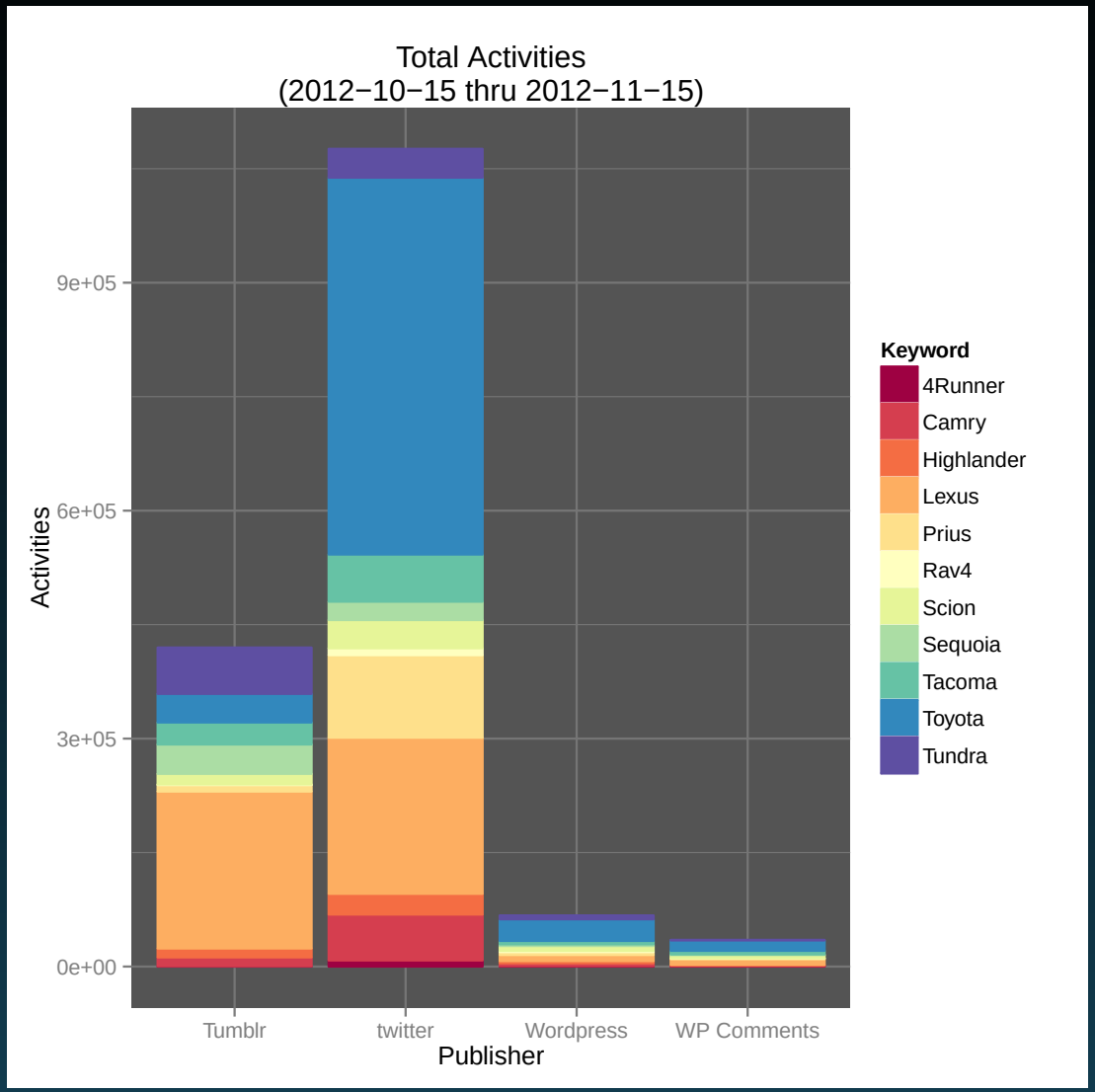
1. audience, perspective and coverage

2. speed

3. richness, diversity

many publishers: audience, perspective and coverage


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


Total Activities
(2012−10−15 thru 2012−11−15)
Activities
9e+05
6e+05
3e+05
0e+00
Tumblr
twitter
Wordpress
WP Comments
Publisher
Keyword
4Runner
Camry
Highlander
Lexus
Prius
Rav4
Scion
Sequoia
Tacoma
Toyota
Tundra

Darren Aronofsky's "Noah" Delayed Due to Flooding

Posted: October 31st, 2012 by WorstPreviews.com Staff

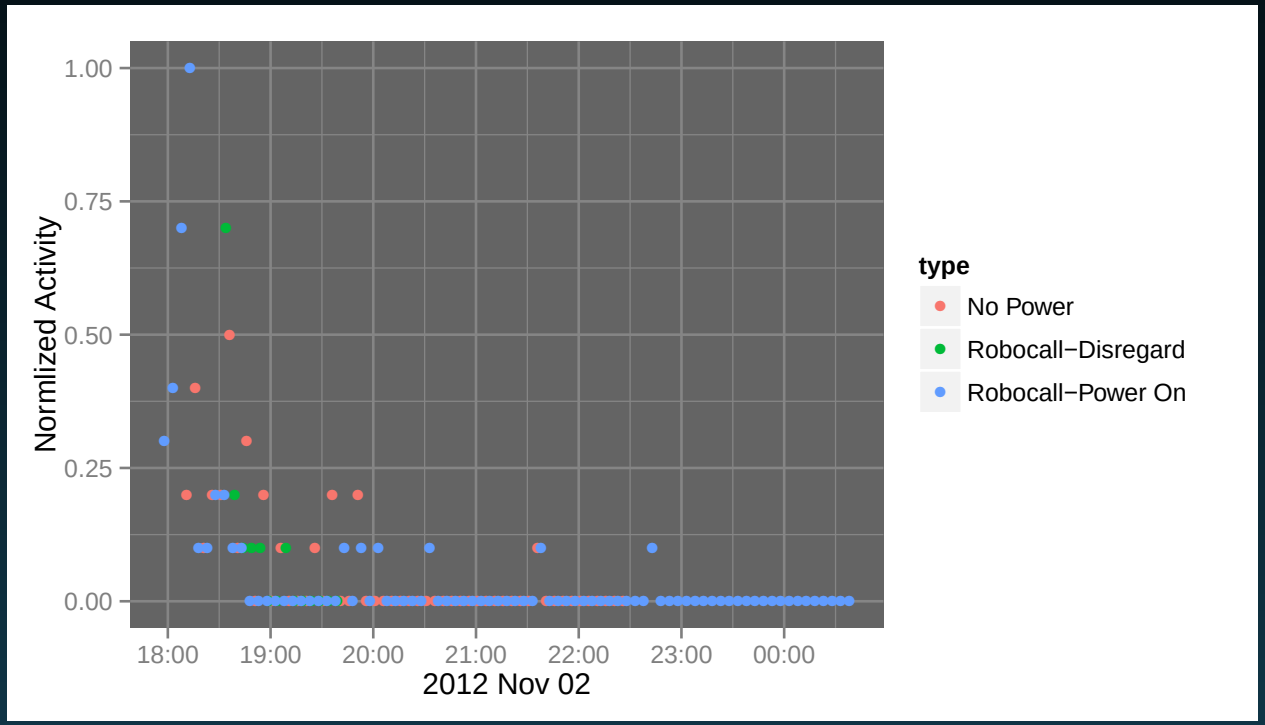
SUBMIT COMMENT

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.

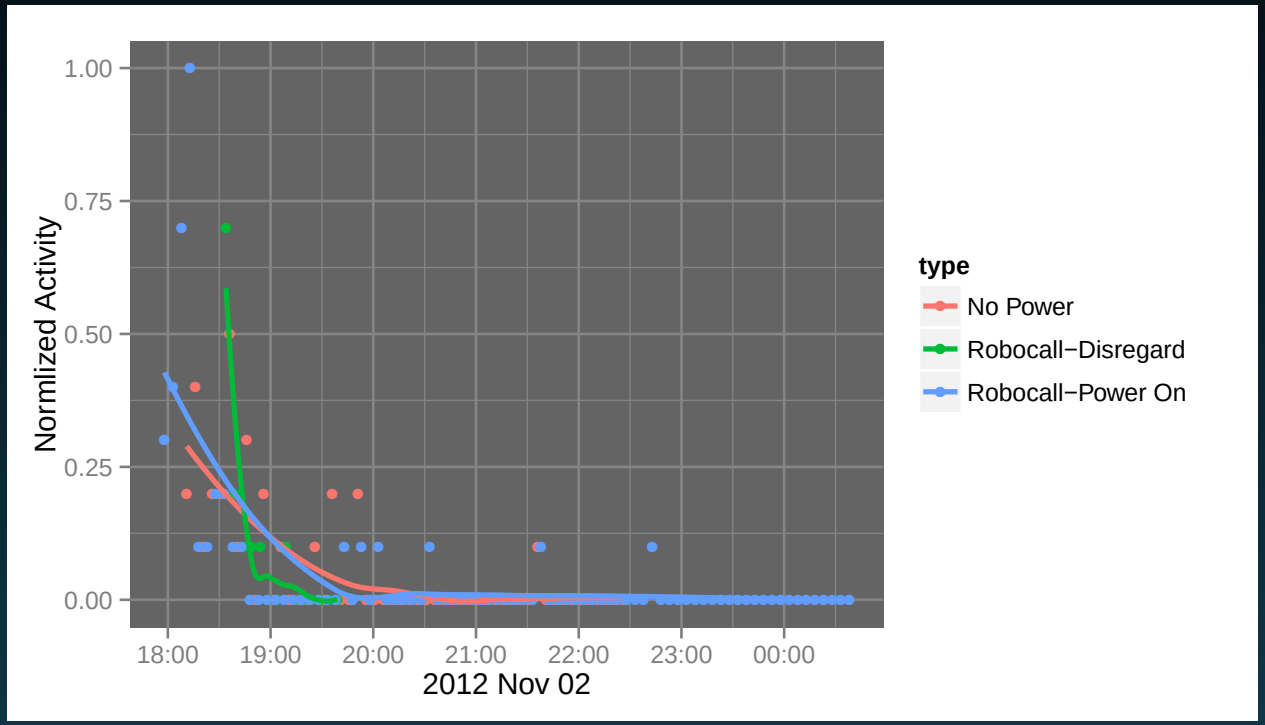
With production wrapping up within the next few weeks, the ark was forced to deal with flooding as Hurricane Sandy passed through Oyster Bay. Emma Watson, one of the actresses on the film, pointed out the irony of flooding being the cause of problems for the production.

noise or signal?

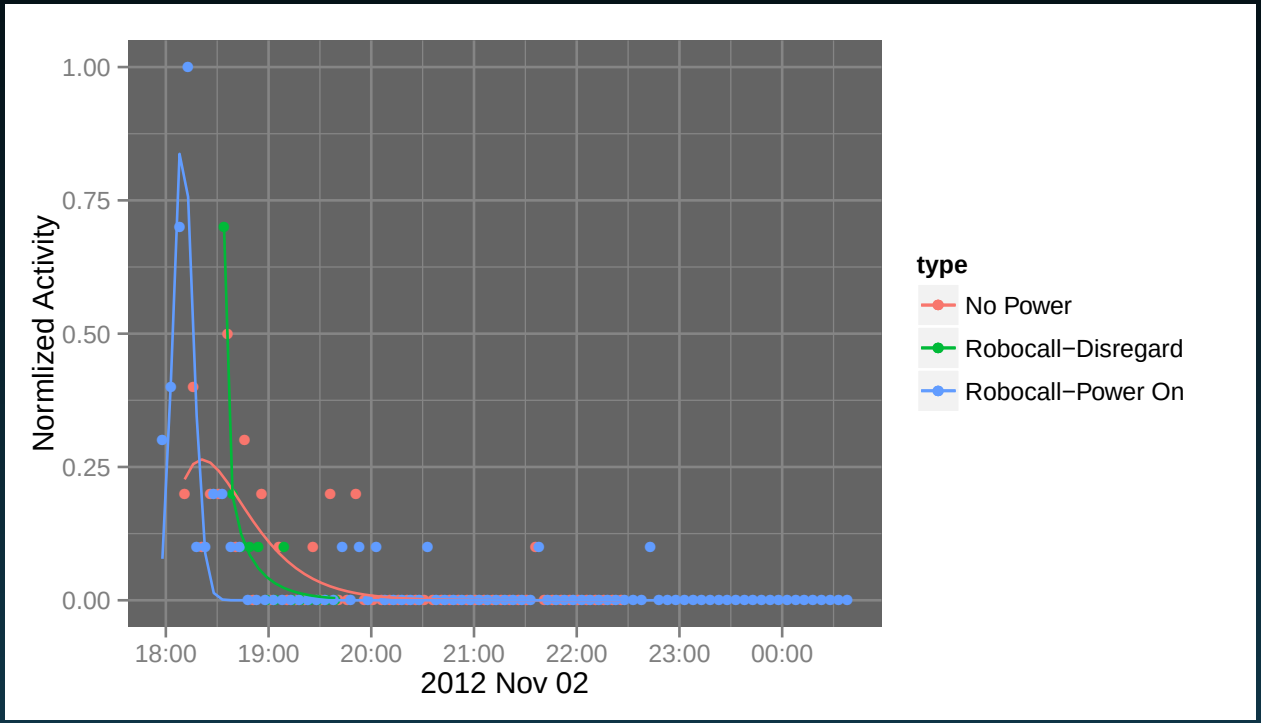
Sandy – Chelsea Power Outage



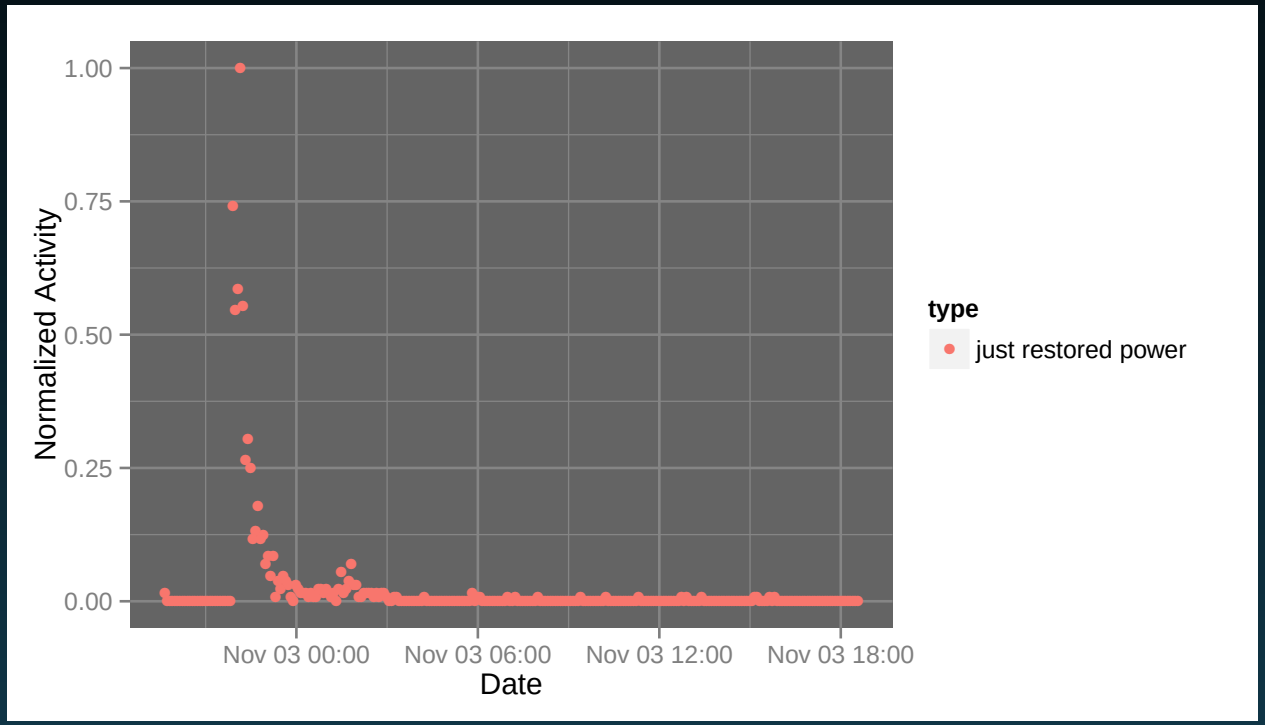
Chelsea Power Outage: Loess



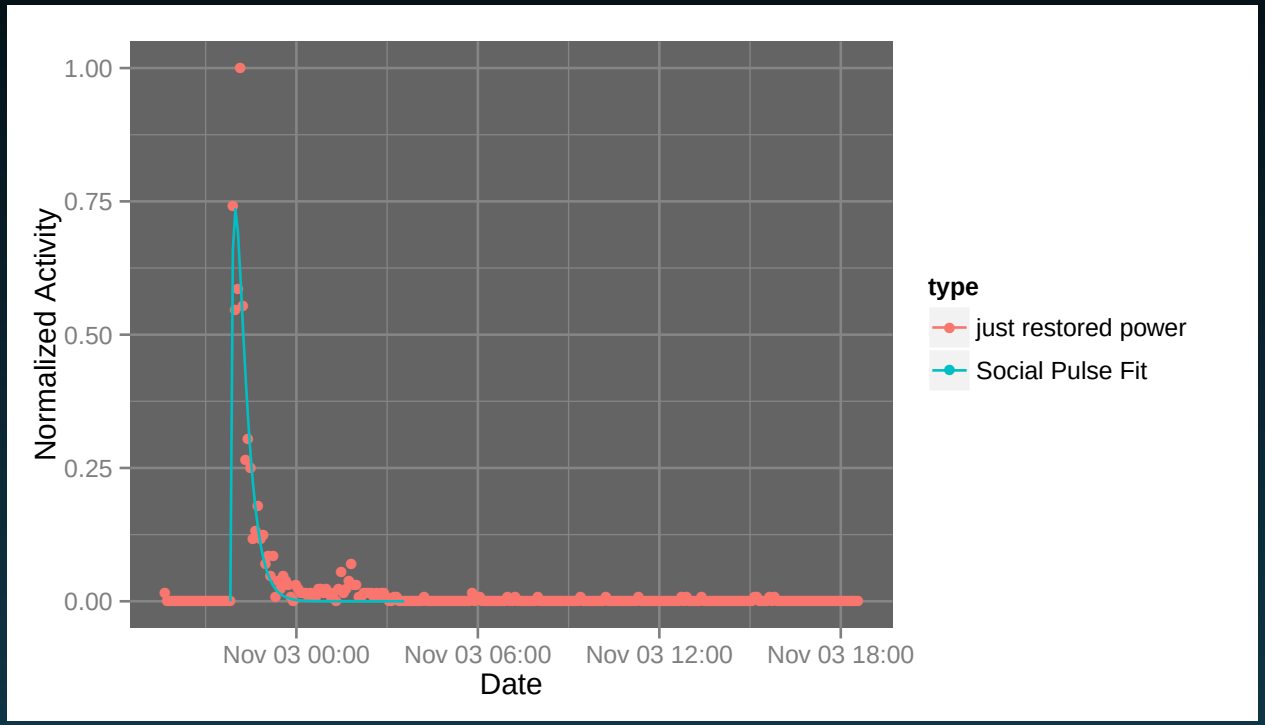
Better Statistical Model



Real event has much higher volume



Real event has much higher volume



realtime firehose: speed of story

Events

Type	Response	Examples
Expected	Approx. Symmetric	Hurricane Sandy Olympics
Unexpected (many obs.)	Social Media Pulse	Beyonce' VMAs Mexico earthquake Steve Jobs
Unexpected (spread)	Sigmoid Pulse	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,

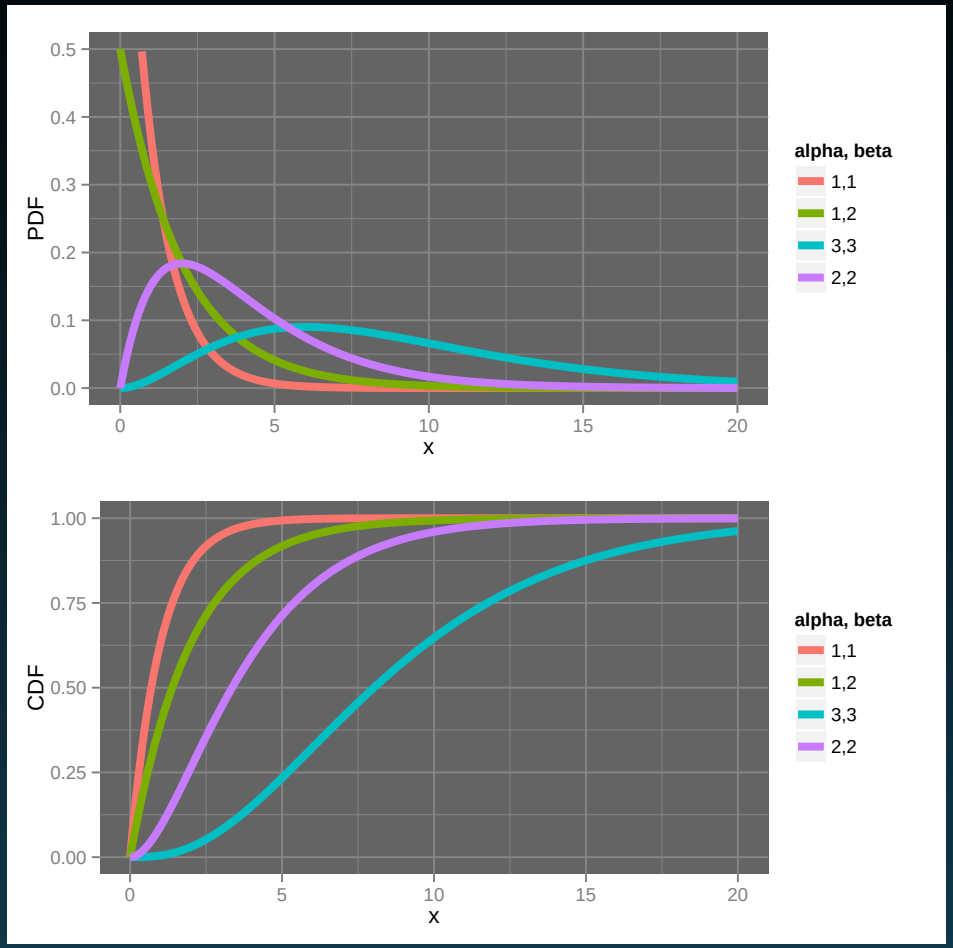
$$f(t) = \lambda \exp(-\lambda t), \text{ for } t \geq 0.$$

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

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

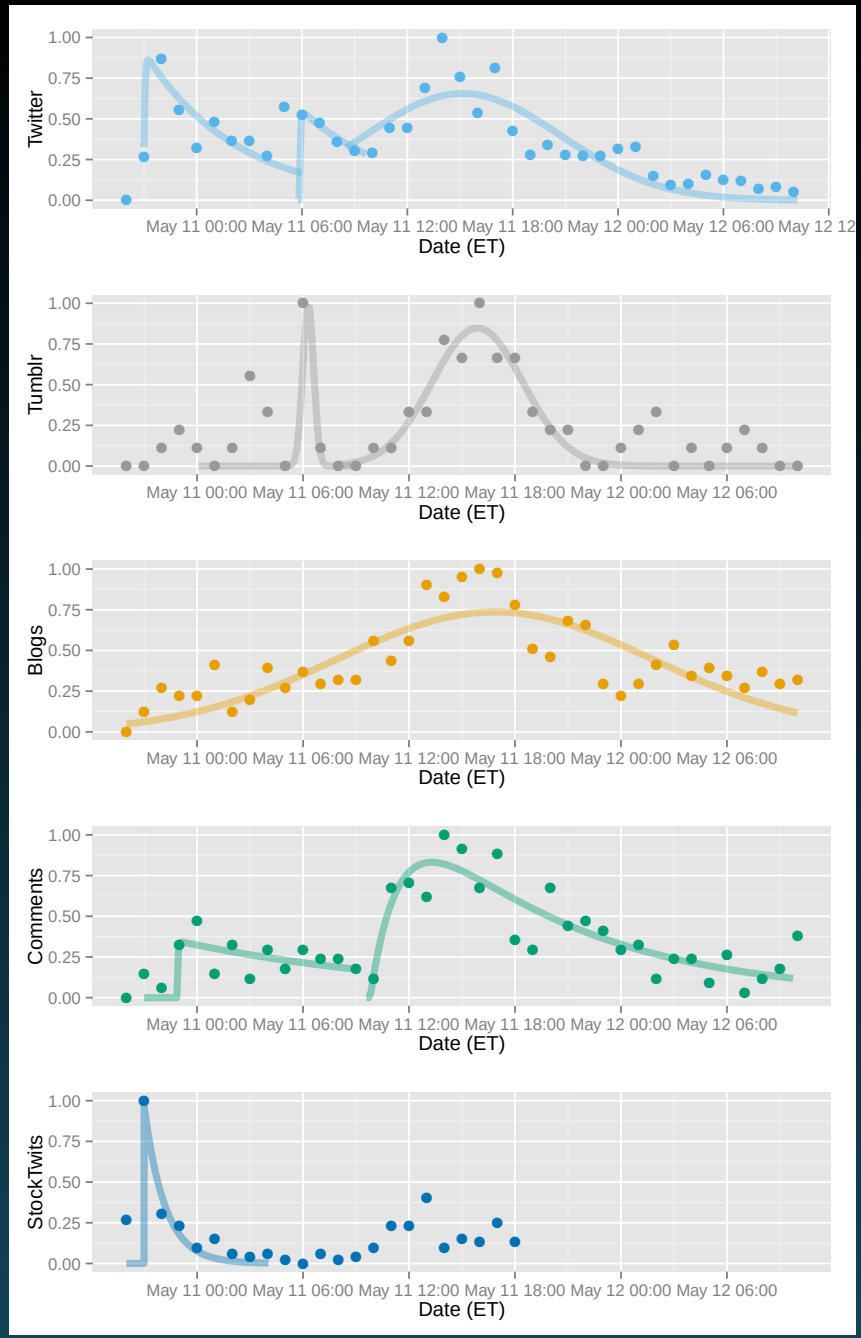
Cumulative distribution is the “generalized regularized incomplete gamma function”,

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


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

Publishers

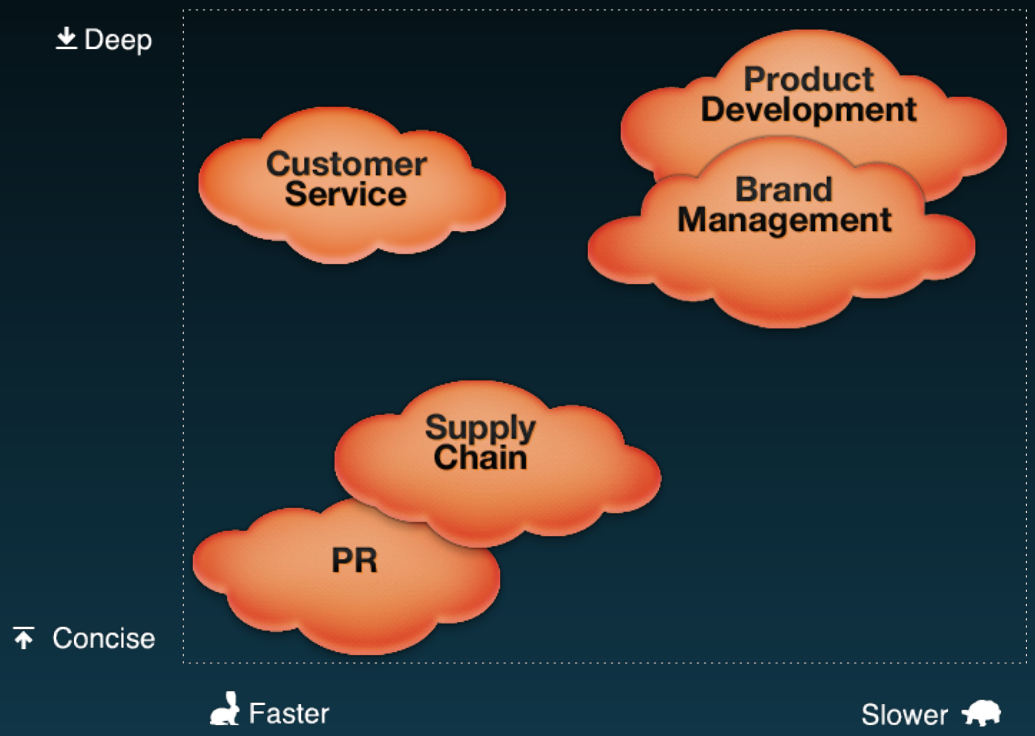
Publisher	Speed
Twitter	Fast
Tumblr	Fast and Slow
Wordpress Posts	Fast and Medium
Wordpress Comments	Fast
Disqus	Fast
Engagement (likes, votes)	Fast


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)
StockTwits
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)

Speed and Richness

Publisher	Speed	Richness
Twitter	Fast	Concise
Tumblr	Fast, Slow	Rich, multimedia
Wordpress Posts	Fast, Medium	Rich, text
Wordpress Comments	Fast	Reactive, small-to-medium
Disqus	Fast	Reactive, small-to-medium
Engagement	Fast	Terse

Social Cocktail



Thank you!

- Presentation, data, code at:
  http://github.com/DrSkippy27/DSBAMeetup2012
- Gnip is hiring: http://gnip.com/careers/