

Gnip, with examples...

Data Science/Business Analytics Meetup - Denver

Scott Hendrickson
Data Scientist
@DrSkippy27

December 5, 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

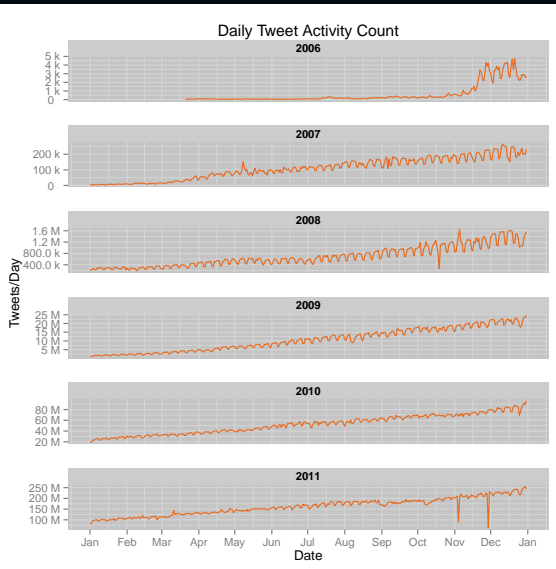
- Firehoses -
Twitter, Tumblr, Wordpress, Disqus ...
- Data Collectors -
API access (Gnip !Ping) to Facebook, YouTube Comments, Reddit, G+, Flickr ...

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

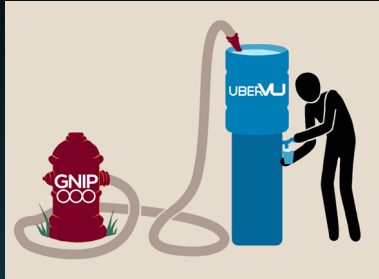
Twitter Growth (2006 - 2010)...



Without Gnip...



With Gnip...



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

Streaming firehose with cURL ...

```

@scotthendrickson shendrickson@shendricksOn4:~$ curl -s https://api.twitter.com/1/statuses/user_timeline.json?screen_name=shendricksOn4&count=100
atuser/26234289451106304", "body": "Are you supposed to be a boss because you're sitting on a 2004 Toyota Corolla in your state? lol", "id": "26234289451106304", "text": "Are you supposed to be a boss because you're sitting on a 2004 Toyota Corolla in your state? lol", "summary": "Are you supposed to be a boss because you're sitting on a 2004 Toyota Corolla in your state? lol", "link": "http://twitter.com/Shubbstatus/statuses/26234289451106304", "postedTime": "2012-10-27T23:59:54.000Z", "twitter_entities": {"user_mentions": [], "hashtags": [], "urls": [], "retweetCount": 0, "gnip": {"language": {"value": "en"}, "matching_rules": [{"value": "Toyota", "tag": null}], "klout_score": 130}}}, {"id": "tag:search.twitter.com,2005:262342895967825920", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:38786428", "link": "http://www.twitter.com/gooner", "displayName": "M Papa", "postedTime": "2009-08-27T18:00:00Z", "summary": "Baked goods. Sports. RTing, MTing. General randomness. Rhymes. Tweets w/ an accent. Confused, self-effacing, elitist, prick.", "links": [{"href": "http://thoroughlyimpressed.tumblr.com/", "rel": "me"}], "friendsCount": 521, "followersCount": 692, "listedCount": 59, "statusCount": 81173, "twitterTimezone": "Eastern Time (US & Canada)", "verified": false, "utcOffset": "-18000", "preferredUsername": "gooner", "languages": "en", "location": {"objectType": "place", "displayName": "Upper Xanadu", "verb": "share", "postedTime": "2012-10-27T23:59:55.000Z", "generator": {"displayName": "TwitterDeck", "link": "http://www.twitterdeck.com/"}}, "provider": {"objectType": "service", "displayName": "Twitter"}, "url": "http://twitter.com/38786428/statuses/262342895967825920", "body": "RT @RexHupke: Was in a Prius and a Dave Matthews song came on and now I'm so white the sun is reflecting off me and blinding airline pi..."}, {"object": {"id": "tag:search.twitter.com,2005:262304847699787776", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:21927918", "link": "http://www.twitter.com/RexHupke", "displayName": "Rex Hupke", "postedTime": "2009-02-24T0:18:55.000Z", "image": "http://ab.twimg.com/profile_images/82381602/catonhead_noraml.jpeg", "summary": "A guy who writes stuff for money. Huge fan of hypocrisy. Also available on http://www.wtstream.com.", "links": [{"href": null, "rel": "me"}], "friendsCount": 782, "followersCount": 16665, "listedCount": 10, "statusCount": 23816, "twitterTimezone": "US (Canada)", "verified": false, "utcOffset": "-21600", "preferredUsername": "RexHupke", "languages": "en", "location": {"objectType": "place", "displayName": "Chicago", "verb": "post", "postedTime": "2012-10-27T21:28:43.000Z", "generator": {"displayName": "Twitter for Android", "link": "http://twitter.com/download/android"}, "provider": {"objectType": "service", "displayName": "Twitter", "link": "http://www.twitter.com/"}}, "link": "http://twitter.com/RexHupke/statuses/262304847699787776", "body": "Was in a Prius and a Dave Matthews song came on and now I'm so white the sun is reflecting off me and blinding airline pilots."}, {"object": {"id": "tag:search.twitter.com,2005:26234289487699787776", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:21927918", "link": "http://www.twitter.com/RexHupke", "displayName": "Rex Hupke", "postedTime": "2012-10-27T21:28:43.000Z", "summary": "Was in a Prius and a Dave Matthews song came on and now I'm so white the sun is reflecting off me and blinding airline pilots.", "link": "http://twitter.com/RexHupke/statuses/26234289487699787776", "postedTime": "2012-10-27T21:28:43.000Z", "twitter_entities": {"user_mentions": [], "hashtags": [], "urls": [], "retweetCount": 0, "gnip": {"language": {"value": "en"}, "matching_rules": [{"value": "Prius", "tag": null}], "klout_score": 140}}}, {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2011-01-08T19:51:33.000Z", "image": "http://ab.twimg.com/profile_images/2545928645/xeoe06tvfysp0zjx7n_noraml.jpeg", "summary": "To be a leader, you have to know how to follow. So why not follow someone? I live what I love so there's my life.", "links": [{"href": null, "rel": "me"}], "friendsCount": 1220, "followersCount": 279, "listedCount": 1, "statusCount": 23816, "twitterTimezone": "Pacific Time (US & Canada)", "verified": false, "utcOffset": "-28800", "preferredUsername": "tyeechazz", "languages": "en", "location": {"objectType": "place", "displayName": "Show", "verb": "share", "postedTime": "2012-10-27T23:59:58.000Z", "generator": {"displayName": "txt", "link": "http://twitter.com/tyeechazz"}, "provider": {"objectType": "service", "displayName": "Twitter", "link": "http://www.twitter.com/"}}, "link": "http://twitter.com/tyeechazz/statuses/262342911323172865", "body": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:262342911323172865", "objectType": "activity", "actor": {"objectType": "person", "id": "id:twitter.com:236663769", "link": "http://twitter.com/tyeechazz", "displayName": "Tyler", "postedTime": "2012-10-27T23:59:58.000Z", "summary": "Lolll I never refer to this car as just \"the car\" no, its \"The Prius.\""}, {"object": {"id": "tag:search.twitter.com,2005:2623
```

```

9  "gnip": {
10    "klout_score": 17,
11    "matching_rules": [
12      {
13        "tag": null,
14        "value": "Toyota"
15      }
16    ],
17    "language": {
18      "value": "en"
19    }
20  },
21  "object": {
22    "postedTime": "2012-10-27T12:07:44.000Z",
23    "summary": "2 of our 3 cars are racing at silverstone today. Gary first out in the #toyota followed by @cl4key in the #renault5 #birkett",
24    "link": "http://twitter.com/ChappellRacing/statuses/262163669473972224",
25    "id": "object:search.twitter.com,2005:262163669473972224",
26    "objectType": "note"
27  },
28  "actor": {
29    "preferredUsername": "ChappellRacing",
30    "displayName": "ChappellRacing",
31    "links": [
32      {
33        "href": "http://www.chappellracing.co.uk",
34        "rel": "me"
35      }
36    ],
37    "twitterTimeZone": null,
38    "image": "http://a0.twimg.com/profile_images/2654983779/b32323bd8127f48eef967b879b7e4b89_normal.png",
39    "verified": false,
40    "location": {
41      "displayName": "Kent",
42      "objectType": "place"
43    },
44    "statusesCount": 41,
45    "summary": "Home to champions of the BARC Cannons Tin Top Challenge. Follow us as we race around the country!",
46    "languages": [
47      "en"
48    ],
49    "utcOffset": null,
50    "link": "http://www.twitter.com/ChappellRacing",
51    "followersCount": 18,
52    "friendsCount": 22,
53    "listedCount": 0,
54    "postedTime": "2011-10-06T15:53:01.000Z",
55    "id": "id:twitter.com:386052286",
56    "objectType": "person"
57  },
58  "twitter_entities": {
59    "user_mentions": [
60      {
61        "indices": [
62          91,
63          98
64        ],
65        "id": "851291652",
66        "screen_name": "DrSkippy27"
67      }
68    ]
69  }
70}

```

Link: twitter.json

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:

newline = OR

space = AND

"AND" = AND

"OR" = OR

"_" = NOT

"(...)" = grouping

"..." = literal

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

- 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
- Rules: 250K rules
- Update individual rules without disconnect, \lesssim 1s update time (100s of rules)
- Rule tagging

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

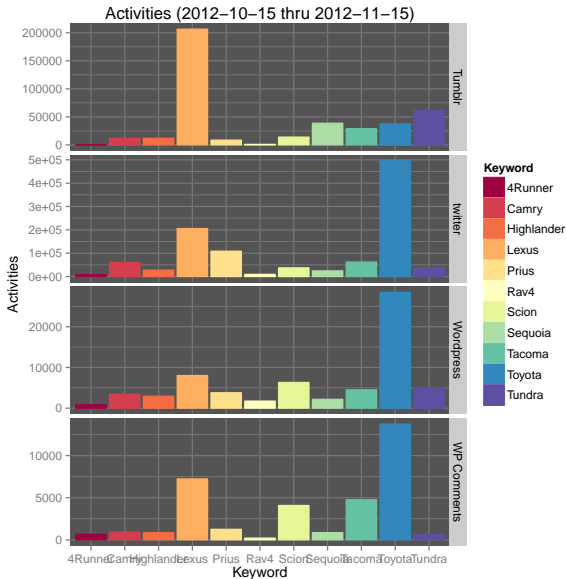
why social media for
breaking stories?

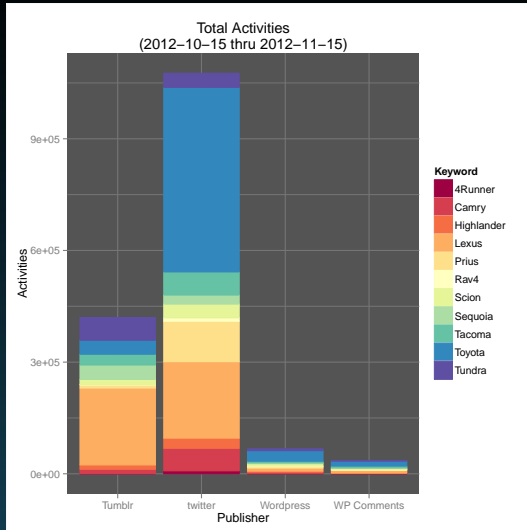
1. audience, perspective and coverage

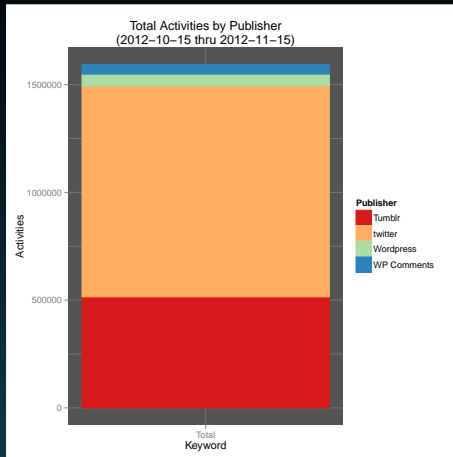
2. speed

3. richness, diversity

many publishers: audience,
perspective and coverage







Darren Aronofsky's "Noah" Delayed Due to Flooding

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



[SUBMIT COMMENT](#)

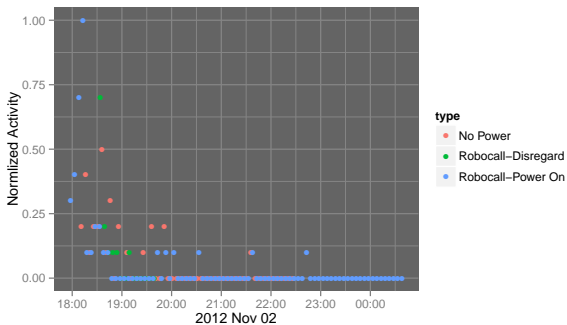
Darren Aronofsky ([Black Swan](#), [The Wrestler](#)) has been filming his "Noah" film, based on the Biblical tale of Noah's Ark, at Oyster Bay, NY. To make it as realistic as possible, the director built a massive ark, which measures 450 feet long, 75 feet tall and 45 feet wide.

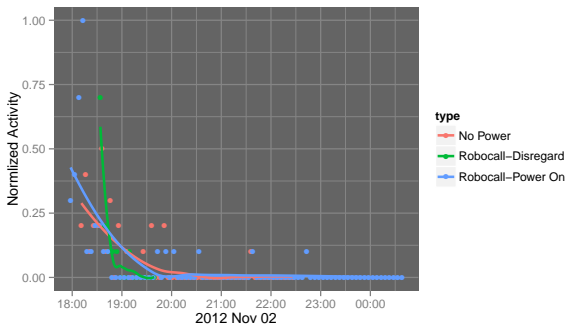


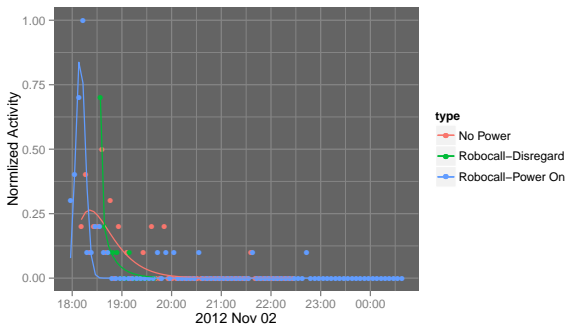
Unfortunately, it was never meant to be sailed.

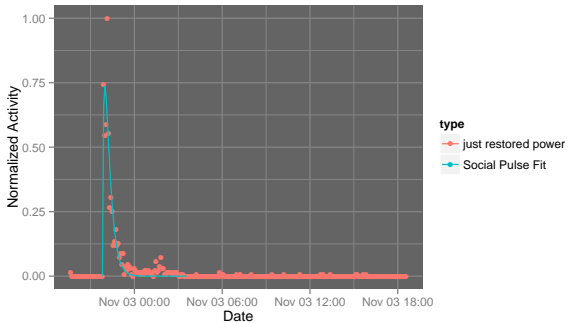
With production wrapping up within the next few weeks, the ark was forced to [deal](#) with flooding as Hurricane Sandy passed through Oyster Bay. Emma Watson, one of the actresses on the film, pointed out the irony of flooding being the cause of problems for the production.

noise or signal?









realtime firehose: speed

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

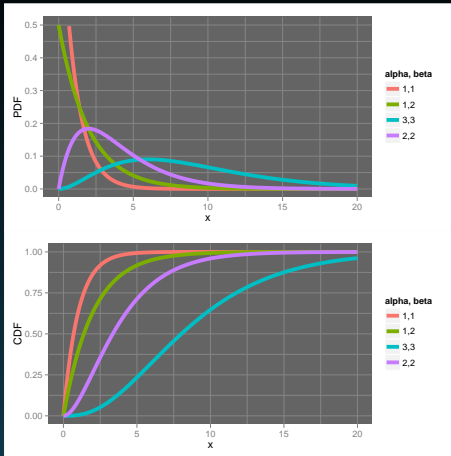
$$S = X_1 + X_2 + \dots + X_{n \text{ posters}}.$$

Probability distribution function,

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

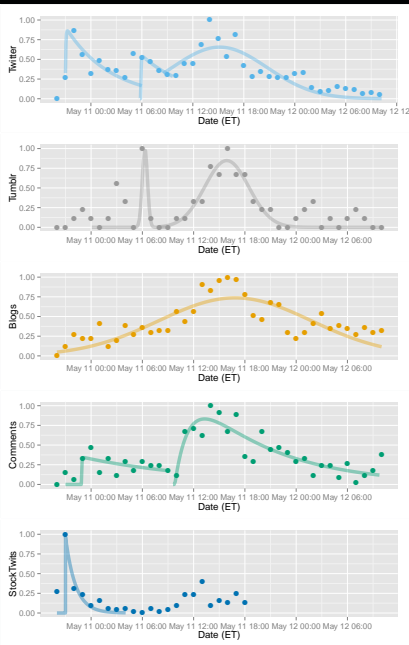
Cumulative distribution is the “generalized regularized incomplete gamma function”,

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



Publishers

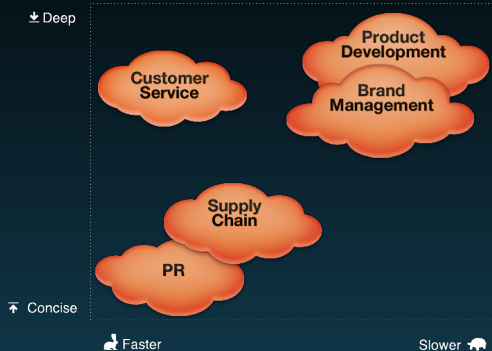
Publisher	Speed
Twitter	Fast
Tumblr	Fast and Slow
Wordpress Posts	Fast and Medium
Wordpress Comments	Fast
Disqus	Fast
Engagement (likes, votes)	Fast



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/>