# 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

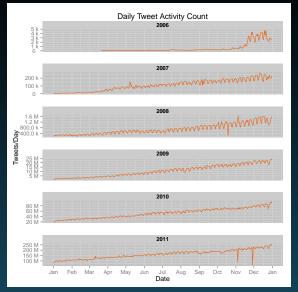
- Firehoses Twitter, Tumblr, Wordpress, Disgus ...
- Data Collectors API access (Gnip !Ping) to Facebook, YouTube Comments,
  Reddit, G+, Flickr ...

#### **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: ~/2012-11-28 EMCToyota/data — ssh — 112×5

atuses/PG34289897188638", "bedy":"Are yes supposed to be a boss because yes' no sitting on a 2004 Toyeth Corolla in your battern of 11" object" ("bole betype" indeed", "bis "big better between 2008 1028289891180808", "bis "many" "Are you supposed to be a boss because you're sitting on a 2004 Toyeth Corolla in your battern 11" "lish ""intropy "bis traces of 2004 2004 10" of 12" "lish "bis traces of 2004 2004 2004 10" of 12" of

{"id":"tag:search.twitter.com.2005:262342895967825920"."objectType":"activity"."actor":{"objectType":"person d":"id:twitter.com:38786420"."link":"http://www.twitter.com/gooner "."displayName":"M Papa"."postedTime":"2009-0 5-08T17:31:45.000Z", "image": "http://a0.twimg.com/profile images/2617052073/5whyv7v8bgksal6p3gp6 normal.jpeg", "su mmary": "Baked goods, Sports, RTing, MTing, General raunchiness, Rhymes, Tweets w/ an accent, Confused, self-effa cing, elitist, prick, "."links"; [{"href"; "http://thoroughlyunimpressed.tumblr.com/"."rel"; "me"}], "friendsCount"; 5 21, "followersCount": 692, "listedCount": 59, "statusesCount": 81173, "twitterTimeZone": "Eastern Time (US & Canada) erified":false,"utcOffset":"-18000","preferredUsername":"gooner ","languages":["en"],"location":{"objectType"; lace", "displayName": "Upper Xanadu"}}, "verb": "share", "postedTime": "2012-10-27723:59:55,0007", "generator": {"displayName": "DisplayName": "Upper Xanadu"}}, "verb": "share", "postedTime": "2012-10-27723:59:55,0007", "generator": {"displayName": "DisplayName": "Upper Xanadu"}}, "verb": "share", "postedTime": "2012-10-27723:59:55,0007", "generator": {"displayName": "DisplayName": "Upper Xanadu"}}, "verb": "share", "postedTime": "2012-10-27723:59:55,0007", "generator": {"displayName": "DisplayName": "DisplayNa wName":"TweetDeck","link":"http://www.tweetdeck.com"},"provider":{"objectType":"service","displayName":"Twitter "link": "http://www.twitter.com"), "link": "http://twitter.com/gooner /statuses/262342895967825920", "body": "RT @Re xHuppke: Was in a Prius and a Dave Matthews song came on and now I'm so white the sun is reflecting off me and b linding airline pi ...", "object": {"id": "tag:search.twitter.com, 2005: 262304847699787776", "objectType": "activity "actor":{"objectType":"person","id":"id:twitter.com:21792170","link":"http://www.twitter.com/RexHuppke","displa Name": "Rex Huppke", "postedTime": "2009-02-24T20:18:55.000Z", "image": "http://a0.twimg.com/profile images/82381602/ catonhead normal.jpeg", "summary": "A guy who writes stuff for money. Huge fan of hypocrisy. Also available on htt p://www.witstream.com.","links":[{"href":null,"rel":"me"}],"friendsCount":702,"followersCount":16605,"listedCour t":1725, "statusesCount":11111, "twitterTimeZone": "Central Time (US & Canada)", "verified":false, "utcOffset": "-2160 0", "preferredUsername": "RexHuppke", "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/RexHuppke/statuses/262384847699787776", "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.", "obj ect":{"objectType":"note","id":"object:search.twitter.com,2005:262304847699787776","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." ink": "http://twitter.com/RexHuppke/statuses/262304847699787776", "postedTime": "2012-10-27721:28:43.000Z"}, "twitte r\_entities":{"user\_mentions":[],"hashtags":[],"urls":[]}},"twitter\_entities":{"user\_mentions":[{"indices":[3,13] "screen\_name":"RexHuppke","id\_str":"21792170","name":"Rex Huppke","id":21792170}],"hashtags":[],"urls":[]},"ret, weetCount":17, "gnip":{"language":{"value":"en"}, "matching\_rules":[{"value":"Prius", "tag":null}], "klout\_score":50

{"id":"tag:search.twitter.com,2005:262342911323172865","objectType":"activity","actor":{"objectType":"person", d":"id:twitter.com:235663769","link":"http://www.twitter.com/tayeechanzzz\_","displayName":"taylorrrrrrr. \*","post edTine":"2011-01-08T19:51:33.0007"."image":"http://a0.twing.com/profile\_images/2545928645/ixeu06vtfvsgozi5iix7\_n ormal.jpeg", "summary": "To be a leader, you have to know how to follow. So why not follow meeeeee?(: i live what i love so cheer is my life. "."links":[{"href":null."rel":"me"}]."friendsCount":228."followersCount":279."listed Count":1."statusesCount":23816."twitterTimeZone":"Pacific Time (US & Canada)"."verified":false."utcOffset":"-288 00", "preferredUsername": "tayeechanzzz\_", "languages": ["en"], "location": {"objectType": "place", "displayName": "Showg irls I, II & III \*"}}, "verb": "post", "postedTime": "2012-10-27723:59:58.0002", "generator": {"displayName": "txt", "l nk":"http://twitter.com/devices"}, "provider":{"objectType":"service", "displayName":"Twitter", "link":"http://www. twitter.com"},"link":"http://twitter.com/tayeechanzzz\_/statuses/262342911323172865","body":"Lolll i never refer to this car as just \"the car\" no. its \"The Prius.\""."object":{"objectType":"note"."id":"object:search.twitte r.com,2005:262342911323172865", "summary": "Lolll i never refer to this car as just \"the car\" no, its \"The Priu s.\""."link":"http://twitter.com/taveechanzzz /statuses/262342911323172865"."postedTime":"2012-10-27723:59:58.00 9Z"}, "twitter entities": {"user mentions": [], "hashtags": [], "urls": []}, "retweetCount": 9, "gnip": {"language": {"value ":"en"}."matching rules":[{"value":"Prius"."tag":null}]."klout score":40}} {"info":{"message":"Replay Request Completed"."sent":"2012-11-30T22:47:18+00:00"."activity count":13372}}

shendrickson@shendrickson4:~/2012-11-28 EMCTovota/data\$

```
"qnip": {
  "klout score": 17.
  "matching rules": [
        "tag": null.
        "value": "Tovota"
  1,
  "language": {
     "value": "en"
},
"object": {
   "postedTime": "2012-10-27T12:07:44.0002".
  "summary": "2 of our 3 cars are racing at silverstone today. Gary first out in the #toyota followed by @c14key in the #renault5 #birkett",
  "link": "http://twitter.com/ChappellRacing/statuses/262163669473972224".
  "id": "object:search.twitter.com,2005:262163669473972224",
  "objectType": "note"
  "actor": {
       *href": "http://www.chappellracing.co.nt", rel": "me" twitter.json
  "twitterTimeZone": null.
  "image": "http://a0.twimg.com/profile images/2654983779/b32323bd8127f48eef997b879b7e4b89 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": 4
  "user mentions": [
        "indices": [
          91.
          98
        "id": 851291652.
         Gnip Inc. @DrSkippy27
```

10

11 12

13 14 15

16

18 19

20

22

24

25

26 27

37 38

39

40

41

42 43

44 45

46 47 48

49

50

52

53

54 55

56 57

58

59 60

61

62

63 64

65 66

67

# 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**

- Rules: 250K rules
- Update individual rules without disconnect, ≤ 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)</li>
- Historical PowerTrack Twitter historical filtering for any time period

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

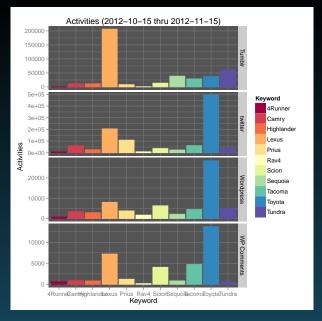
# why social media for breaking stories?

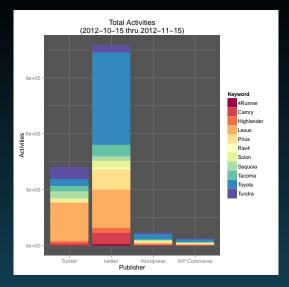
# 1. audience, perspective and coverage

# 2. speed

# 3. richness, diversity

# many publishers: audience, perspective and coverage





#### Darren Aronofsky's "Noah" Delayed Due to Flooding

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



Darren Aronofsky (<u>Black Swan</u>, The Wrestler) has been filmling 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.

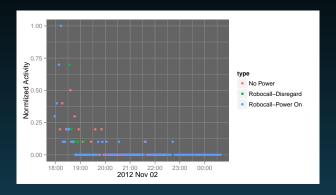


Unfortunately, it was never meant to be sailed.

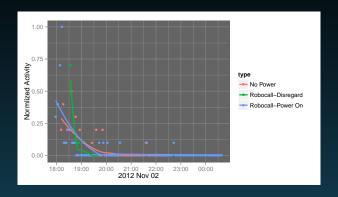
With production wrapping up within the next few weeks, the ark was forced to <u>deal</u> with flooding as Hurricane Sandy passed through Oyster Bay. Emma Watson, one of the actresses on the film, pointed out the irony of flooding being the cause of problems for the production.

# noise or signal?

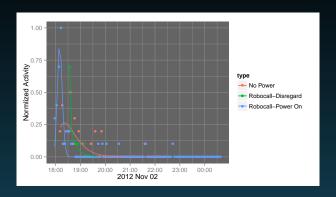
# Sandy - Chelsea Power Outage



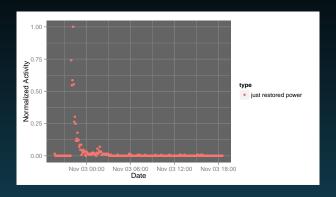
# Chelsea Power Outage: Loess



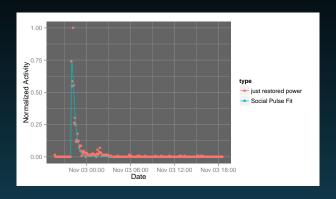
#### Better Statistical Model



#### Real event has much higher volume



#### Real event has much higher volume



# realtime firehose: speed of story

### **Events**

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

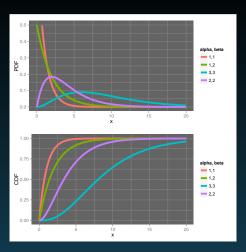
$$f(t) = \lambda \exp(-\lambda t)$$
, for  $t \ge 0$ .

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

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

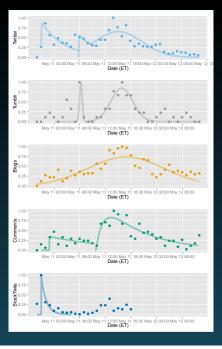
Cumulative distribution is the "generalized regularized incomplete gamma function",

$$F_{\mathcal{S}}(t) = Q(\alpha, 0, \frac{t}{\beta})$$



#### **Publishers**

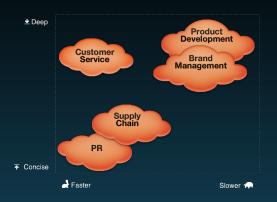
Publisher	Speed
Twitter	Fast
Tumblr	Fast and Slow
Wordpress Posts	Fast and Medium
Wordpress Comments	Fast
Disqus	Fast
Engagement (likes, votes)	Fast



### 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/