

# About 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 ...

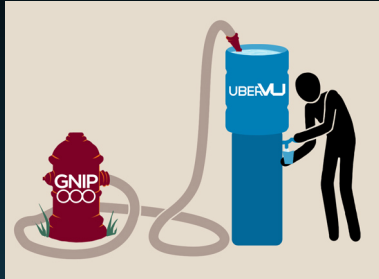
Firehose: Continuous stream of flexibly structured social media activities in near-real time.

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

# Without Gnip...



# With Gnip...



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

Filter: match terms (AND, OR, NOT), metadata



Shape: sample,  
redundancy, replay

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

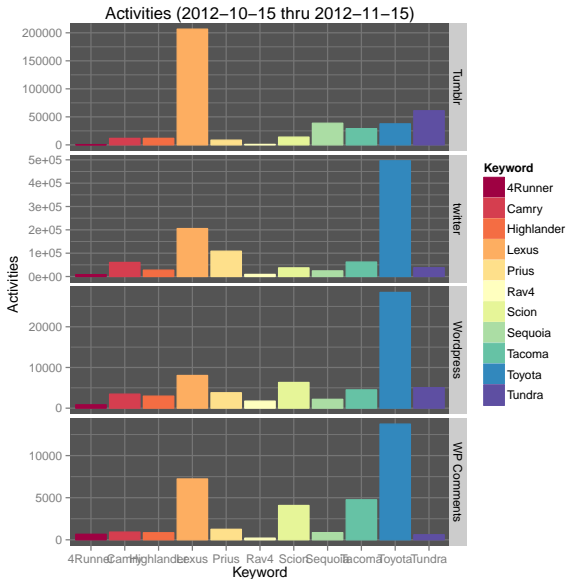
why social media for  
breaking stories?

# 1. audience, perspective and coverage

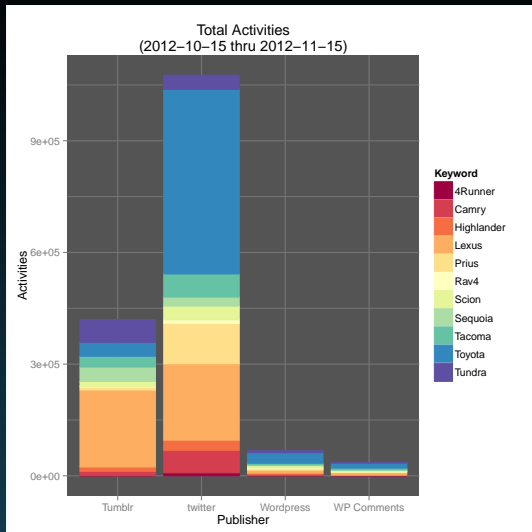
## 2. speed

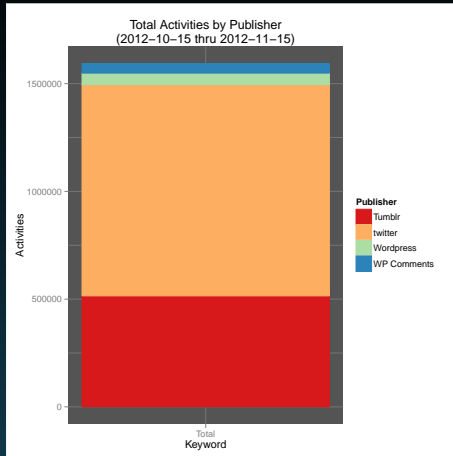
### 3. richness, diversity

many publishers: audience,  
perspective and coverage









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

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



[SUBMIT COMMENT](#)

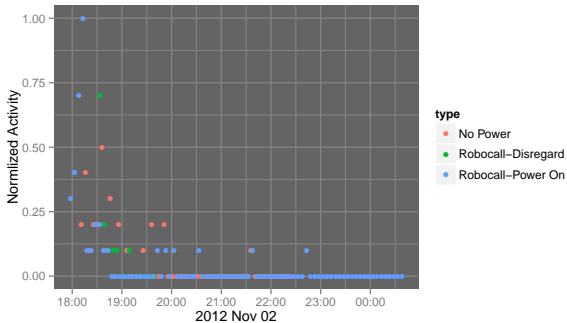
Darren Aronofsky ([Black Swan](#), [The Wrestler](#)) has been filming his "Noah" film, based on the Biblical tale of Noah's Ark, at Oyster Bay, NY. To make it as realistic as possible, the director built a massive ark, which measures 450 feet long, 75 feet tall and 45 feet wide.

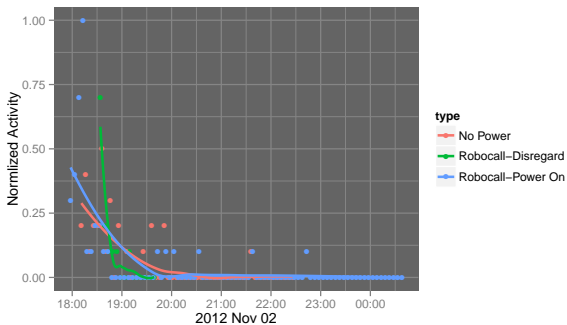


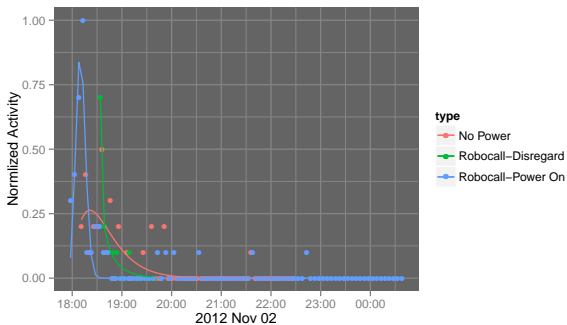
Unfortunately, it was never meant to be sailed.

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?

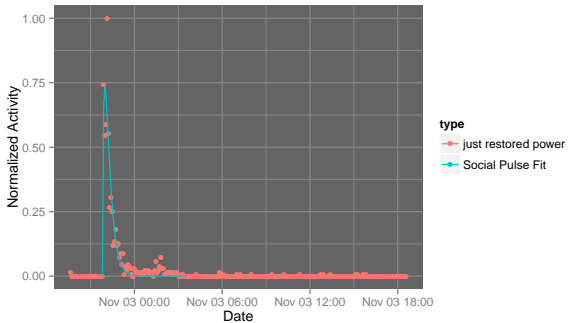












# realtime firehose: speed

# 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

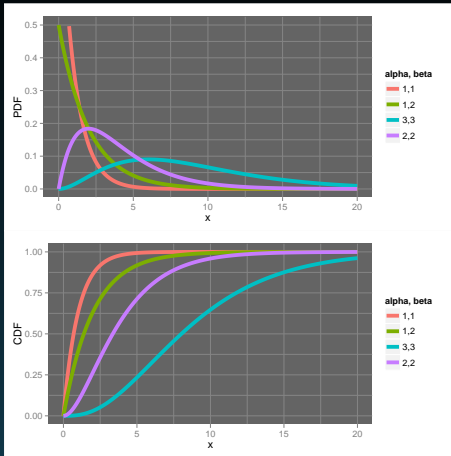
$$S = X_1 + X_2 + \dots + X_{n \text{ posters}}.$$

Probability distribution function,

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

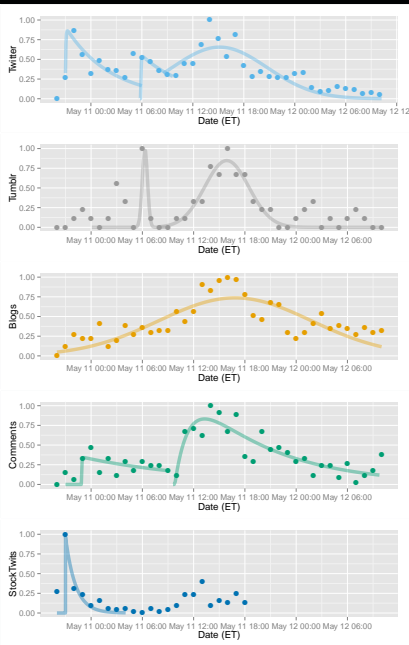
Cumulative distribution is the “generalized regularized incomplete gamma function”,

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



# Publishers

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

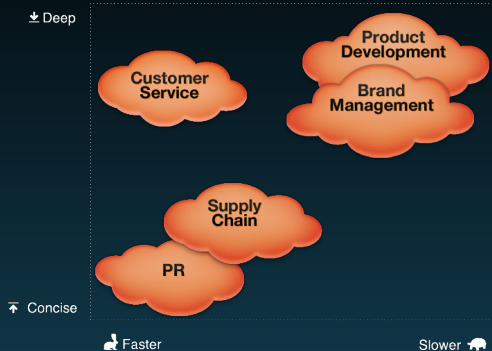




# Speed and Richness

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/>