Social Data Science

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

April 24, 2014

Social Data

Where Does Social Data Come From?

GNIP

Community of Users

- Common interests
- Cultural commonalities
- Topic interest/focus
- Common assumption about sharing/privacy interaction
- Contemporaries/peers

Platform

- Interaction modes
- Habits
- Connections
- Recommendations
- Engagement
- Storage/Structure

Social Data Analysts

- Privacy
- T.O.S.
- Enablers of advertising network analysis

Common Social Data Sources

- APIs
- scraping
- firehose

Firehose

Continuous stream
of activities
in near-real time

Firehose volumes

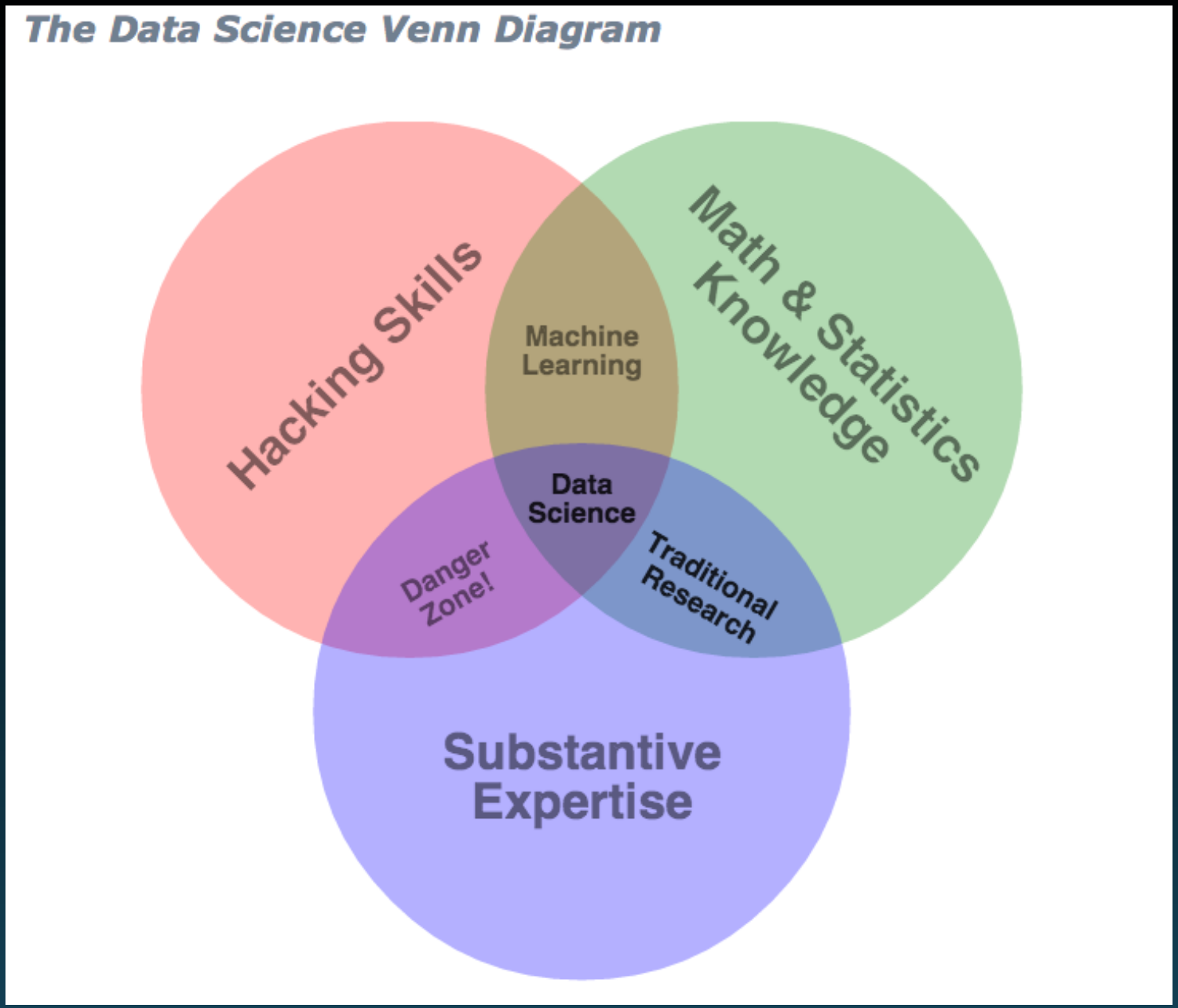
Publisher	Daily Activity
Twitter	520M
Tumblr	110M
Foursquare	4.2M
Wordpress Posts	1M
Wordpress Comments	1.7M
Disqus	1.9M
Engagement (likes, votes)	>60M

Every day @Gnip

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

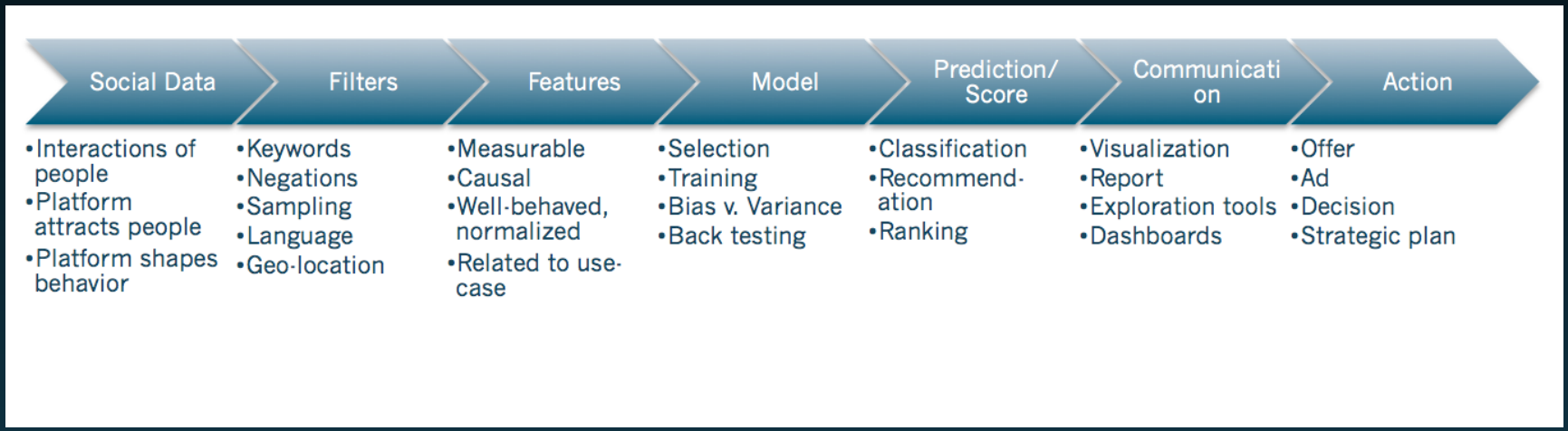
4 Billion + OUT

Data Science



https://s3.amazonaws.com/aws.drewconway.com/latexviz/venn_diagram/data_science.html

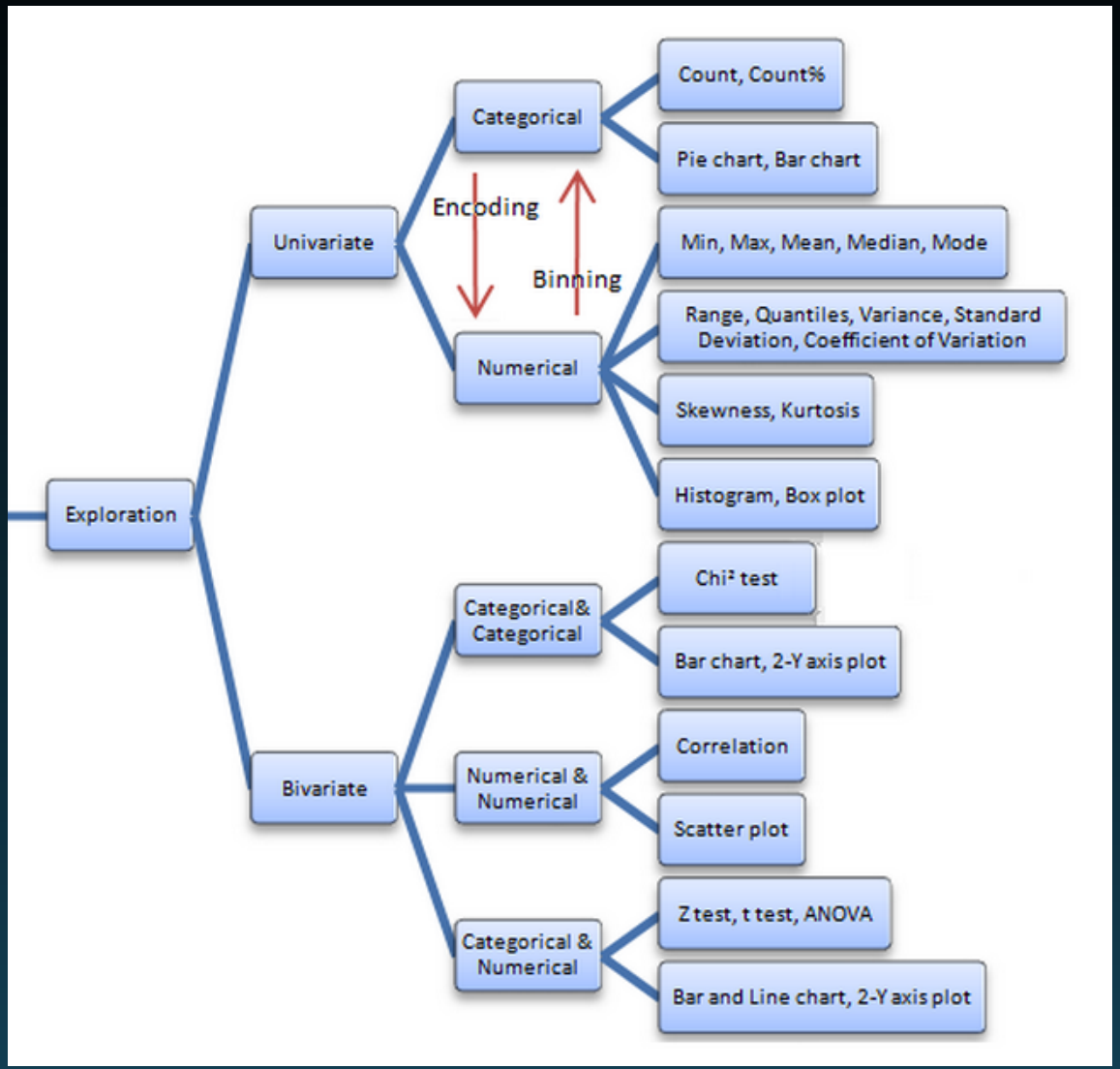
Social Data Processing Pipeline



Three Kinds of Research Projects

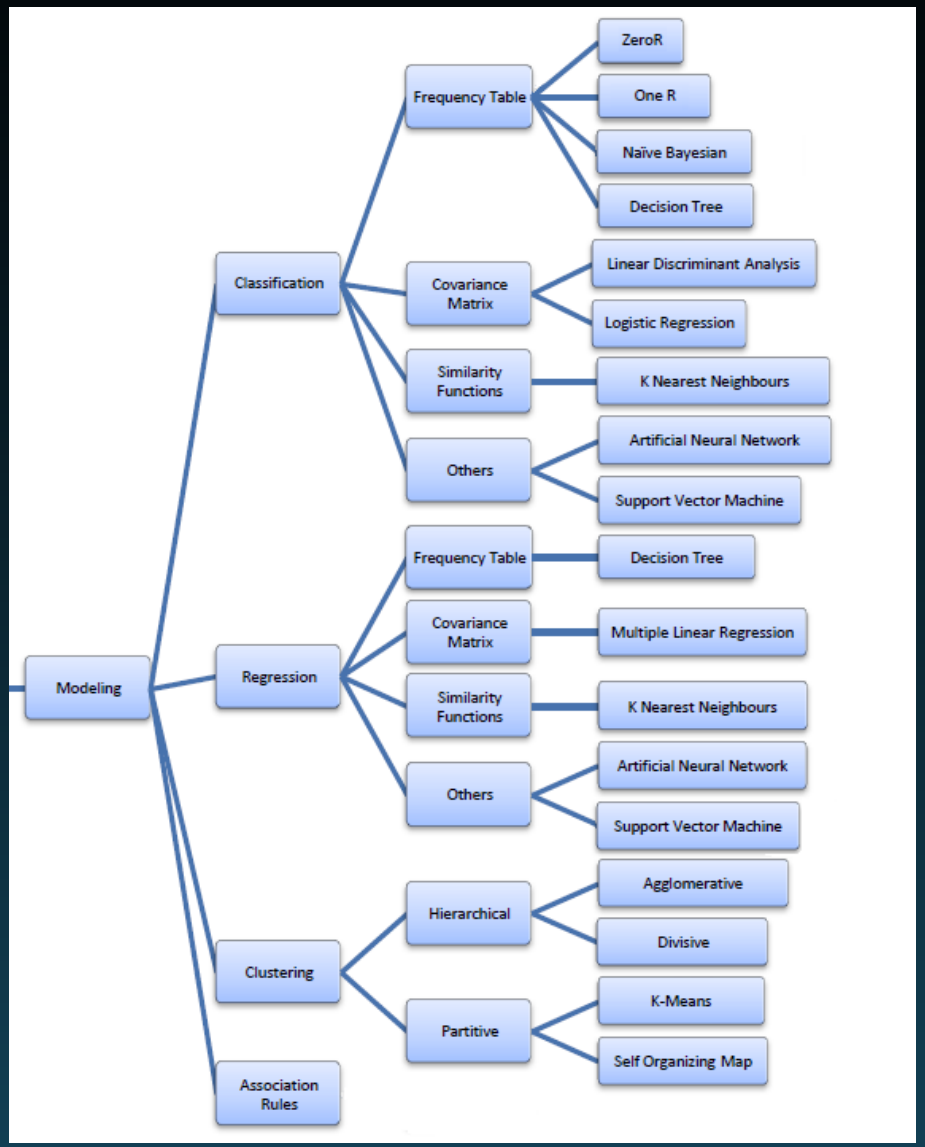
- Descriptive
  - Provides systematic information about a social data set
  - May not begin with hypotheses, but to develop one as you go
  - Detect anomalies
- Exploration
  - Explores a social phenomena
  - Provides background information needed to plan descriptive or explanatory research
  - Trial and error or hypothesis driven
- Explanation
  - 3 Levels: Relationship, Models, Prediction
  - Ideas about the possible causes of a social phenomenon
  - Plan a study that can provide systematic evidence for/against ideas about cause

Data Exploration



http://saedsayad.com/data_mining_map.htm

Data Modeling



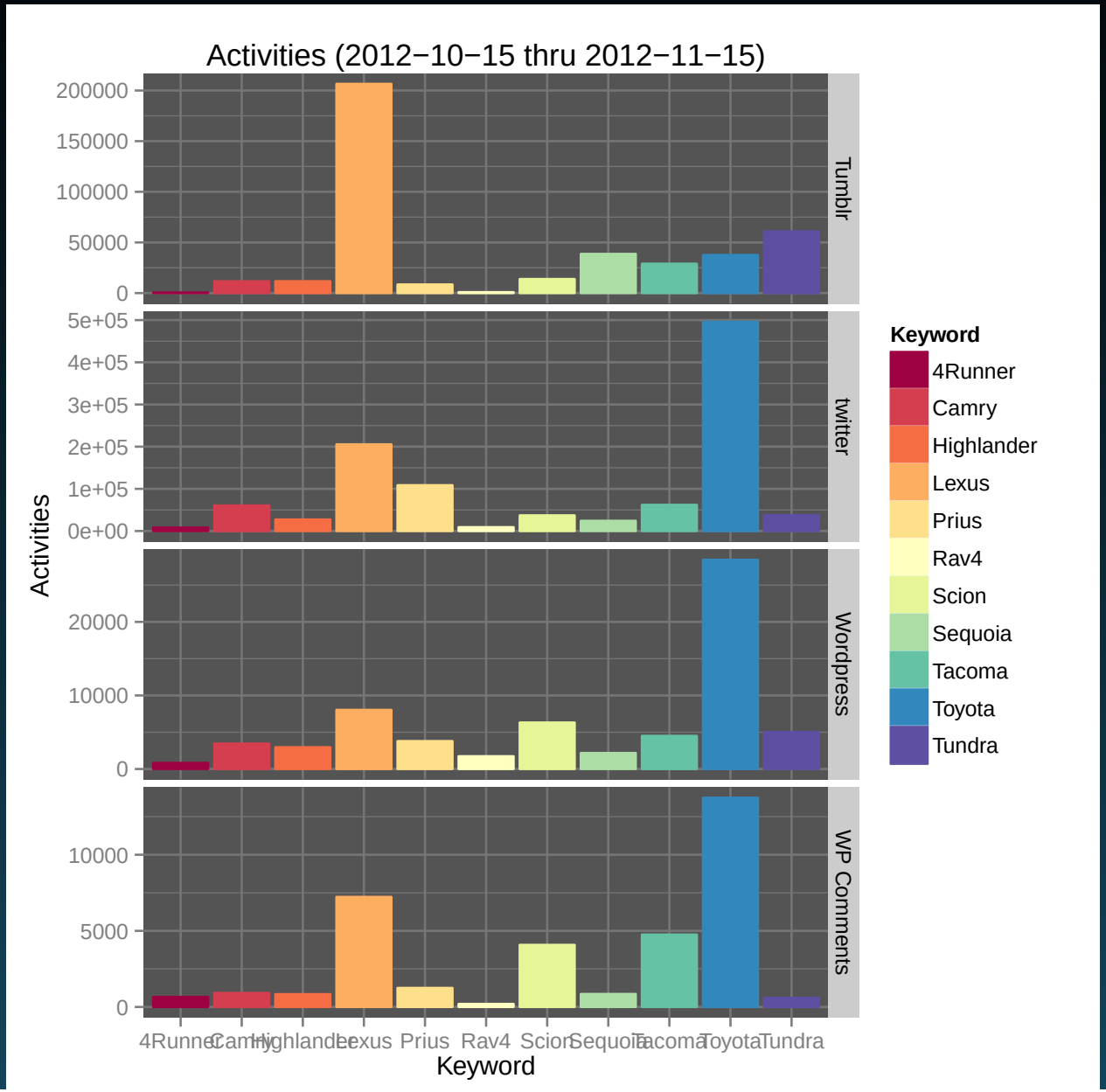
http://saedsayad.com/data_mining_map.htm

Fundamental Processes

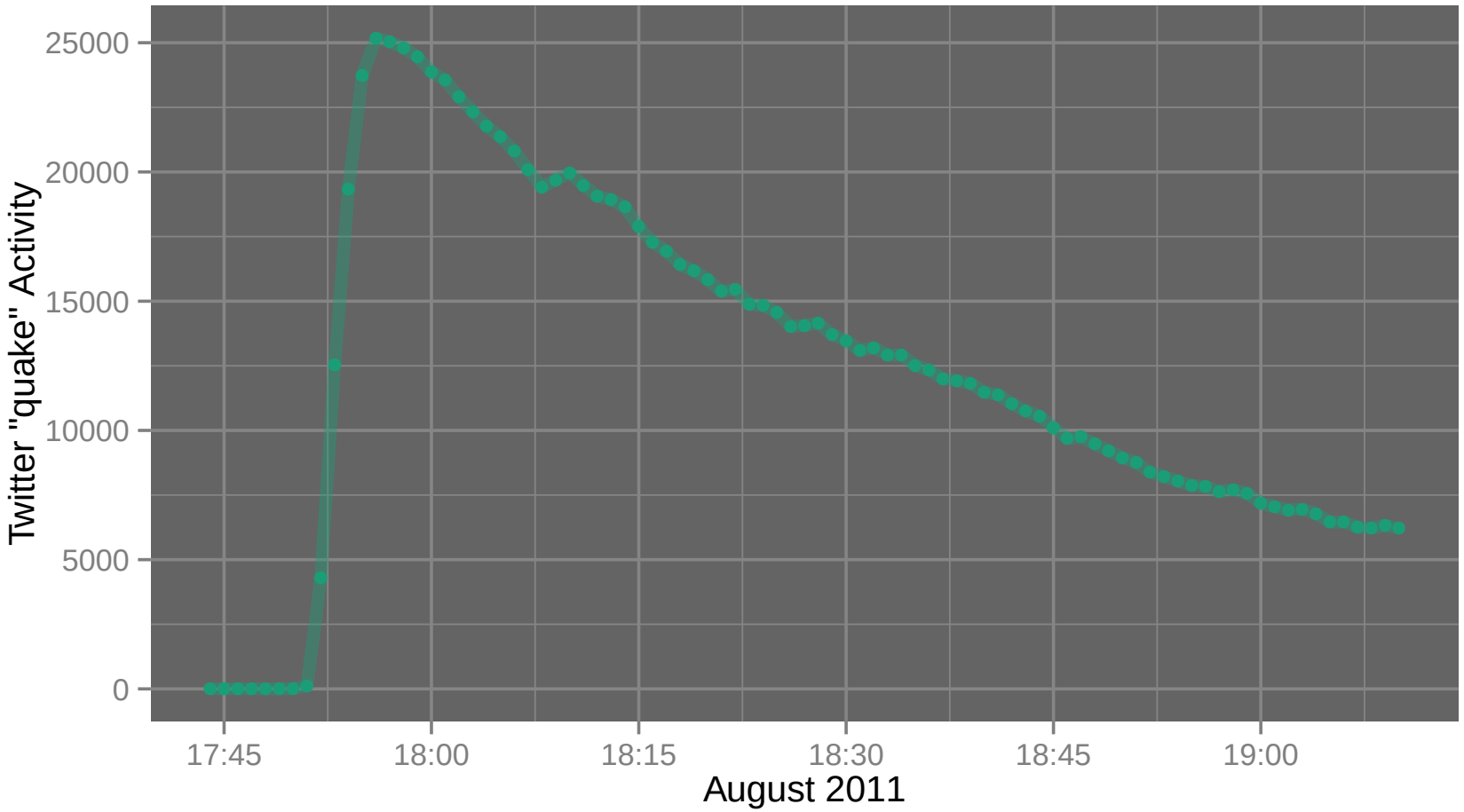
- Networks - e.g. 6-degrees, percolation
- Agent models
- Time series - e.g. “Social Media Pulse”

Time Series: Social Media Pulse

Simple Mention Counts



Unexpected: earthquake



Mentions and time series

We start by bucketing our mention counts by time periods, the activity rate is:

$$\bar{r} = \frac{N}{T},$$

General model for activity rates:

$$p_{activity}(t) = re^{-rt}.$$

give Poisson distribution:

$$P(n) = \frac{e^{-rt}(rt)^n}{n!}.$$

Confidence intervals

Confidence intervals for the Poisson distribution with confidence level equal to $100\%(1-\alpha)$ are given by,

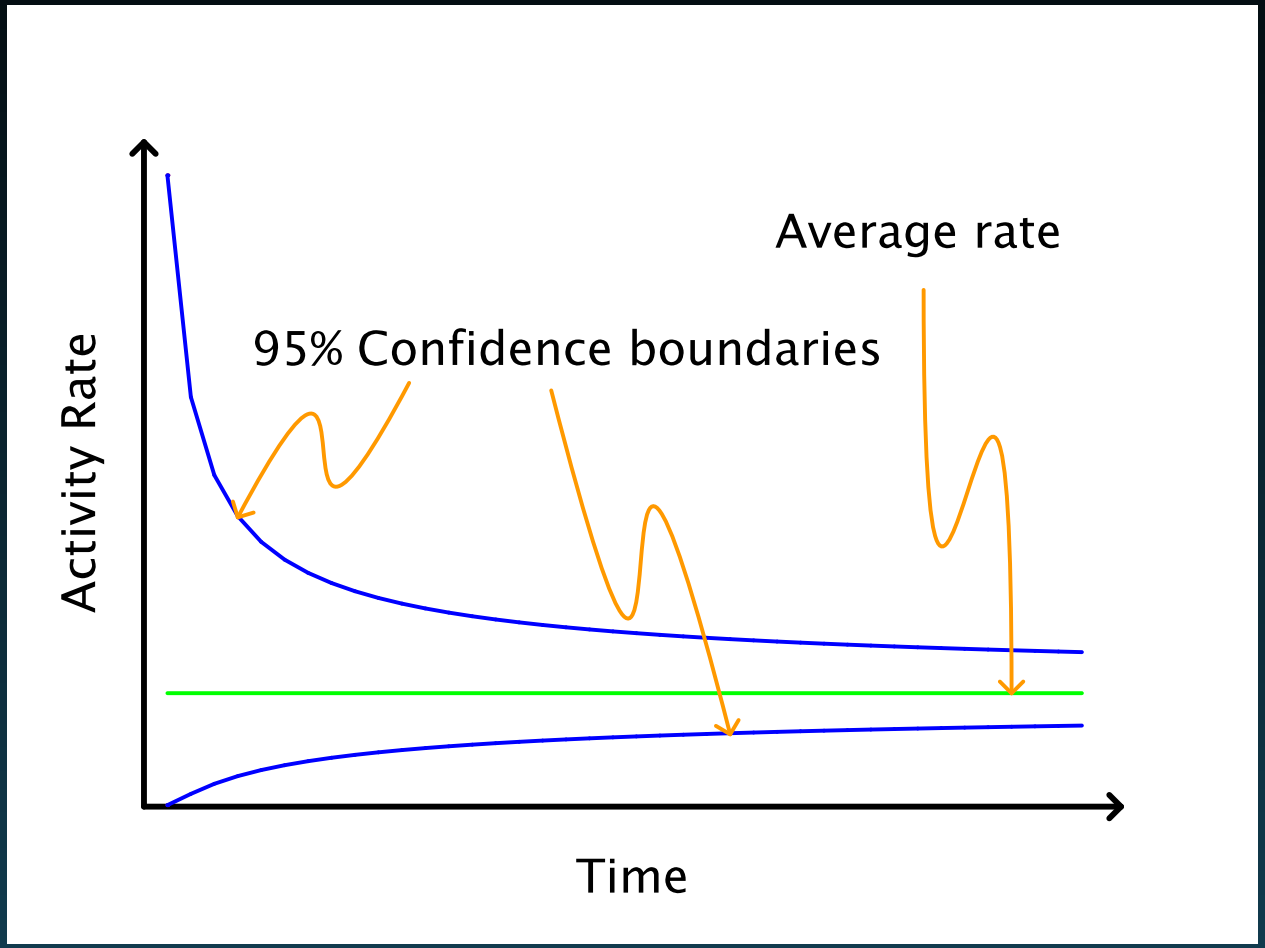
$$\frac{1}{2T}\chi^2(\alpha/2; 2n) \leq r \leq \frac{1}{2T}\chi^2(1-\alpha/2; 2n+2) \qquad (1)$$

where χ^2 is the inverse cumulative distribution function, $CDF^{-1}(p; n)$, of the χ^2 distribution.

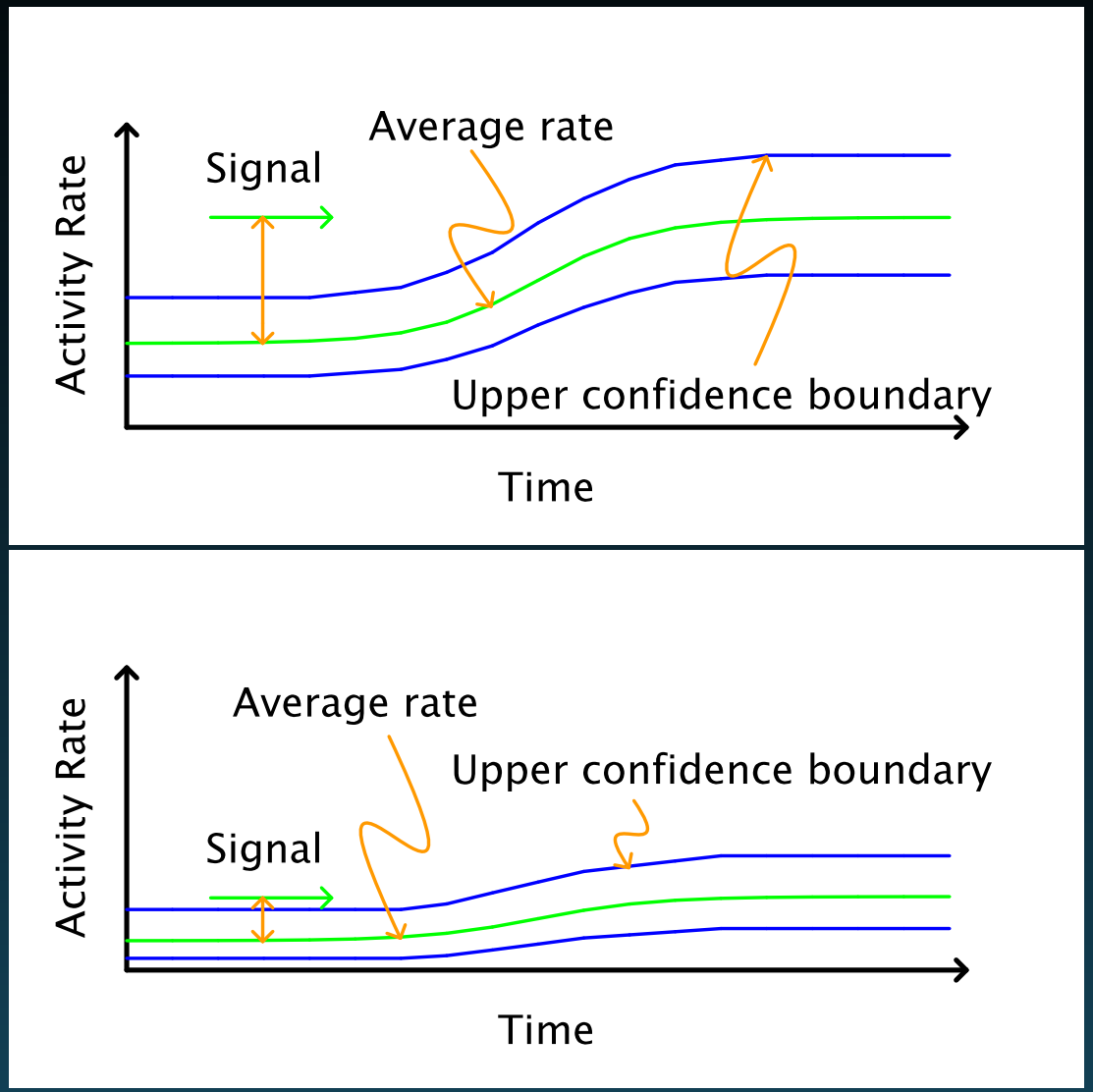
90% confidence intervals?

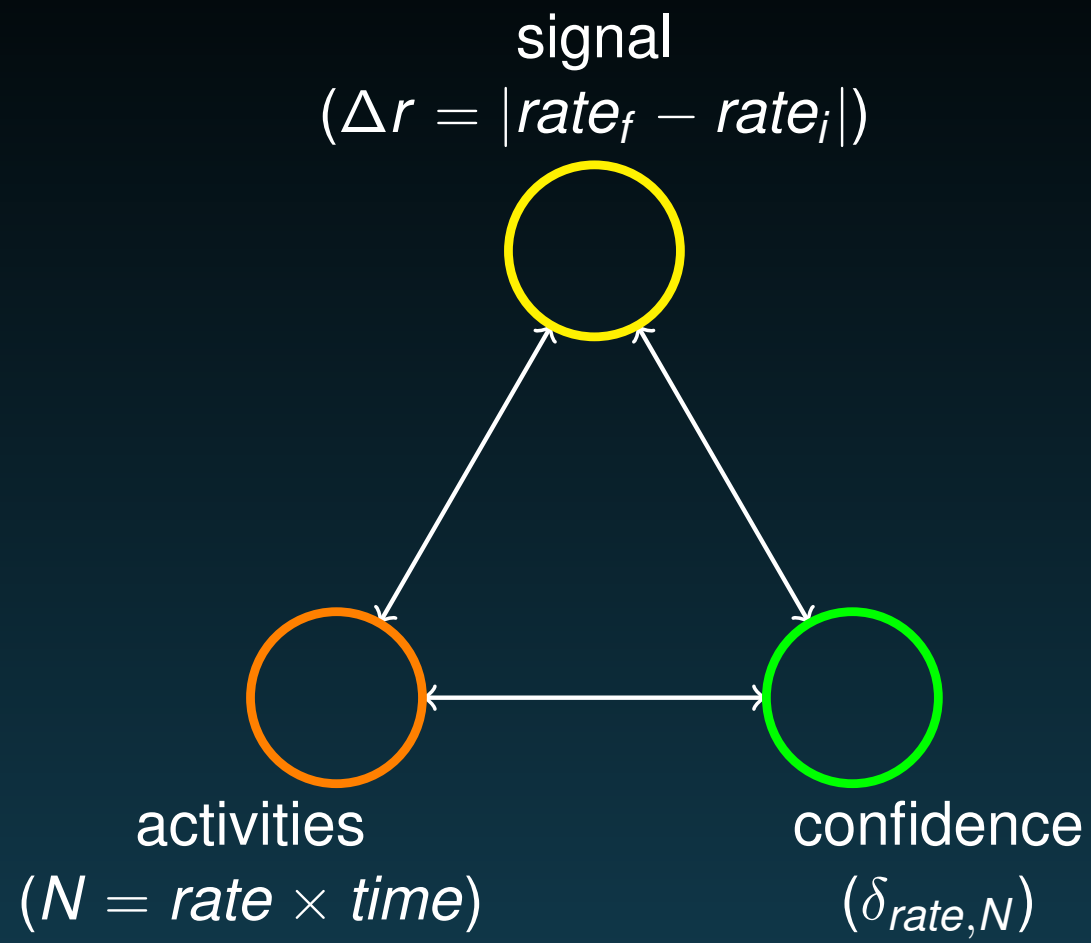
n	Interval Bounds	Interval Size (δn)	Relative Interval
1	0.0513, 4.744	4.693	4.693
2	0.3554, 6.296	5.940	2.970
3	0.8177, 7.754	6.936	2.312
4	1.366, 9.154	7.787	1.947
5	1.970, 10.51	8.543	1.709
10	5.426, 16.96	11.54	1.154
30	21.59, 40.69	19.10	0.6366
40	30.20, 52.07	21.87	0.5468
50	38.96, 63.29	24.32	0.4864
500	463.8, 538.4	74.58	0.1492
750	705.5, 796.6	91.11	0.1215
1000	948.6, 1054.	105.0	0.1050

Make the buckets bigger?

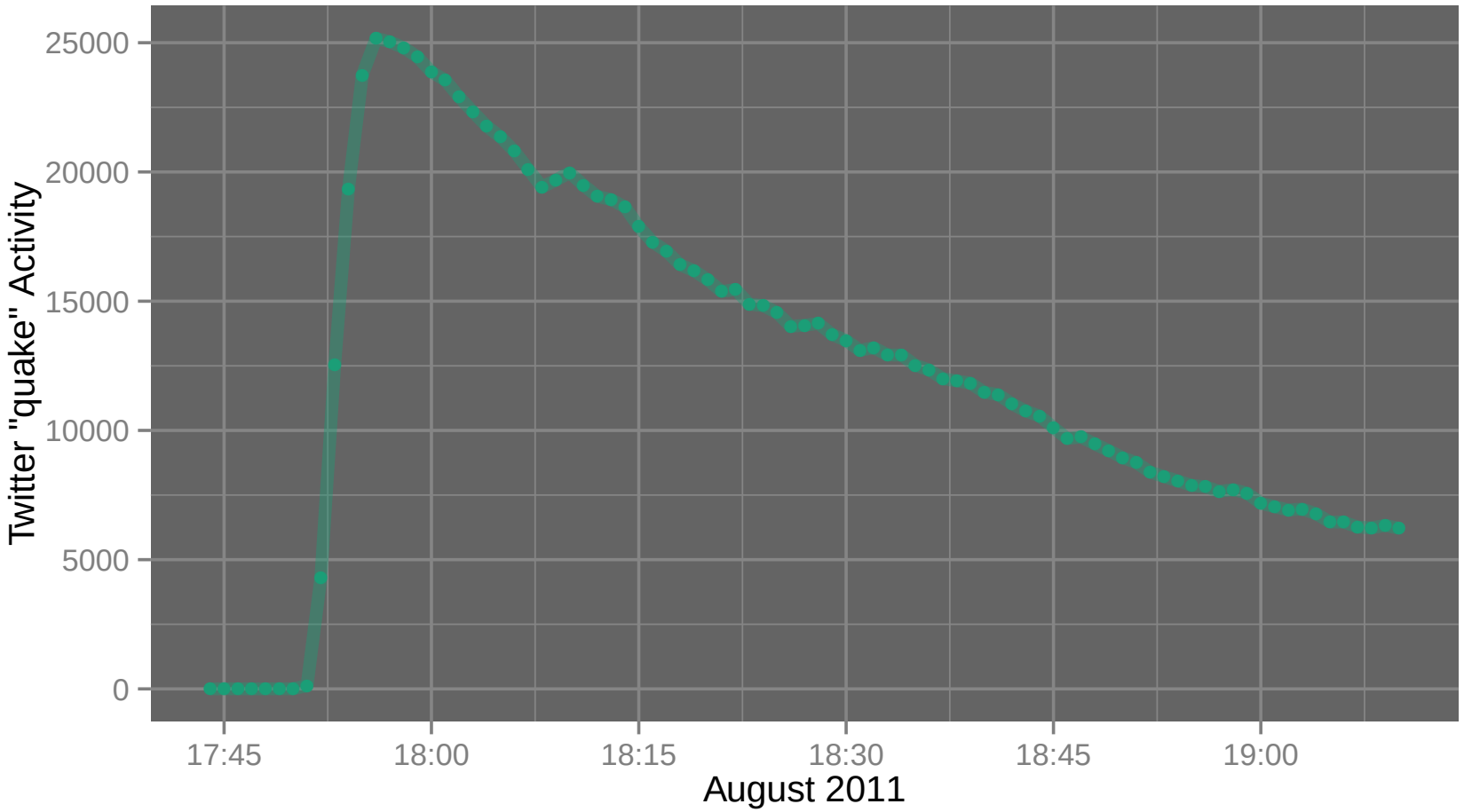


Ignore smaller signals?




signal
$(\Delta r = |rate_f - rate_i|)$
activities
$(N = rate \times time)$
confidence
$(\delta_{rate,N})$

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

Social media pulse

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

$$f(t) = \lambda \exp(-\lambda(t - t_0)), \text{ for } t \geq 0.$$

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

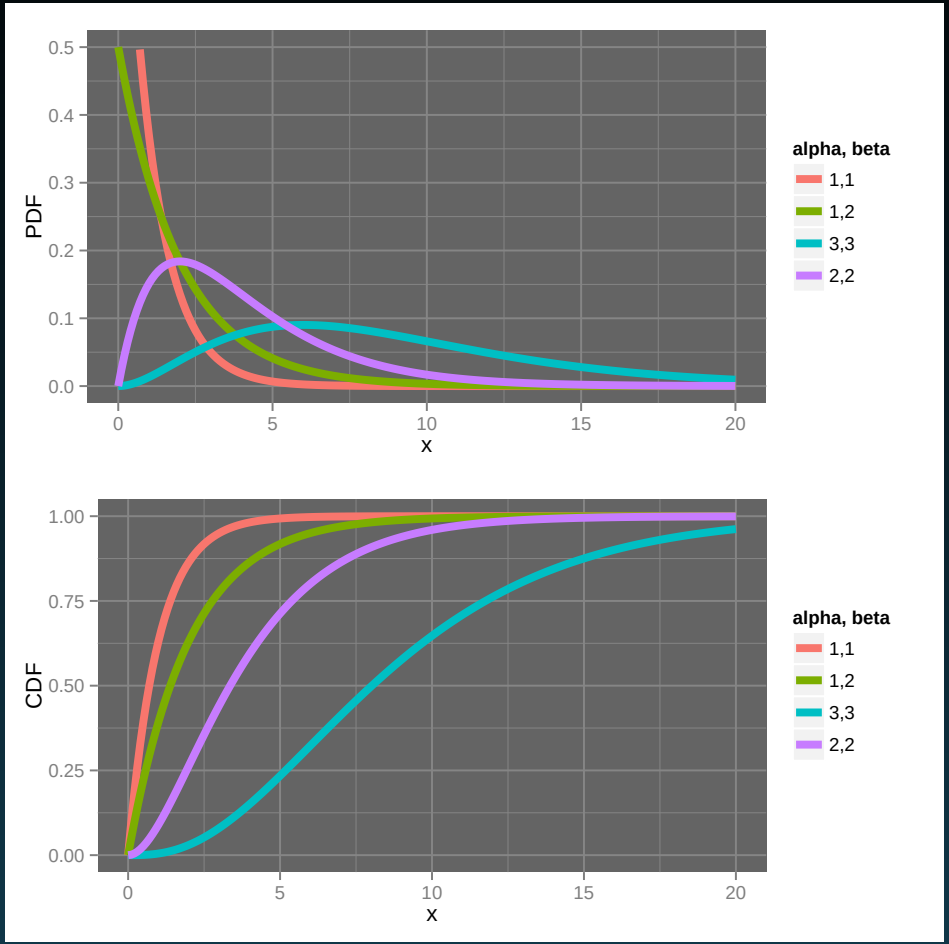
$$S = X_1 + X_2 + \ldots + X_{n \text{ posters}}$$

Gamma probability distribution function,

$$f_S(t) = \frac{\beta^{-\alpha}(t - t_0)^{\alpha-1} \exp(\frac{-(t-t_0)}{\beta})}{\Gamma(\alpha)}$$

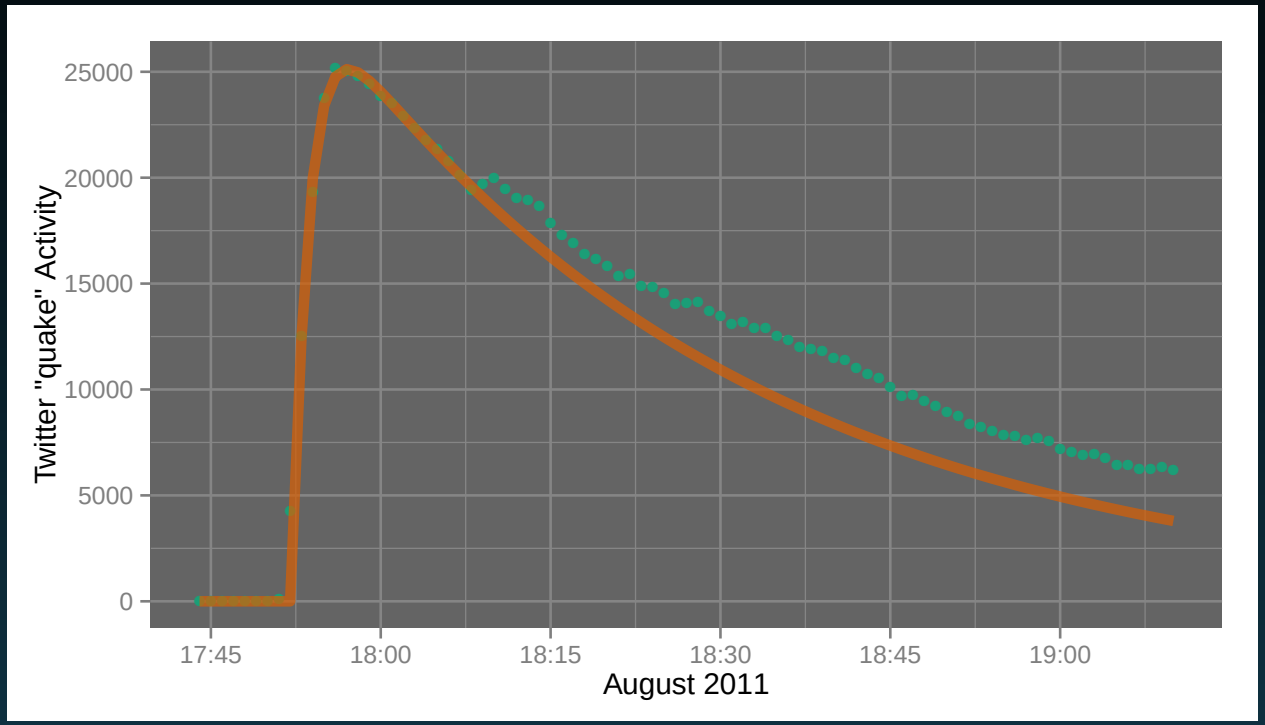
Cumulative distribution is the “generalized regularized incomplete gamma function”,

$$F_S(t) = Q(\alpha, 0, \frac{(t - t_0)}{\beta})$$


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

Why model half-life?

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


25000
20000
15000
10000
5000
0
Twitter "quake" Activity
17:45
18:00
18:15
18:30
18:45
19:00
August 2011

Compare events

We start with a least-squares fit of points up to the $2-3\times$ time-to-peak.

$$t_{time-to-peak} = \beta(\alpha - 1)$$

Average event response time,

$$t_{avg} = \alpha\beta$$

Half-life (time to see half of the activities),

$$t_{\frac{1}{2}life} = F_S^{-1}(\frac{1}{2}).$$

Story size,

$$F_S(t) = Q(\alpha, 0, \frac{t - t_0}{\beta}) \tag{2}$$

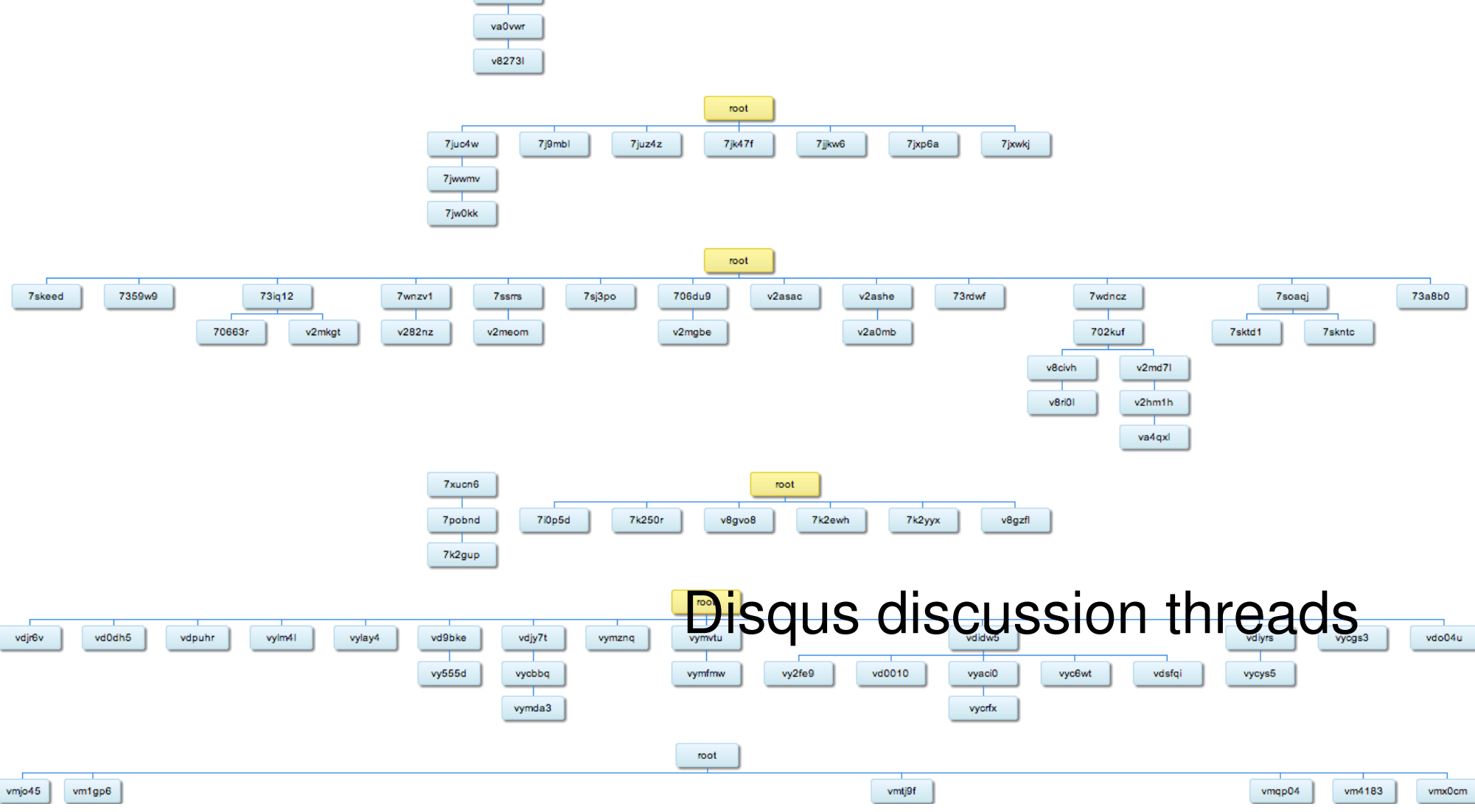
in terms of the incomplete gamma functions,

$$S_{vol} = \int_0^\infty r(t)dt = N_{activities} F_S(t), \tag{3}$$

Topic Modeling – Latent Semantic Indexing

What do we talk about when they talk about X?

Apologies: Raymond Carver



Disqus discussion threads

Disqus Threads

- 7 weeks of Disqus comments data
- Key words: “texting,” “driving” and variants
- Select top threads based on mentions
- 61,406 comments from 365 threads

Disqus Topic Model Approach

- Find comments that mention key words
- Corpus of comments (across many threads)
- tf-idf matrix: terms $\times$ comments
- LSI (rotate space to align with “important” dimensions, reduce dimensions)
- K-means (quick-and-dirty clustering in reduced dimensional space)
- …rinse and repeat (looking for distinction and cohesion)

tf-idf and LSI in one page …

- tf: term frequency
- idf: inverse document frequency

LSI uses singular value decomposition to rotate document matrix from tf-idf to reduce dimensionality in a controlled way. SVD lets us write the document matrix as,

$$D = V\Sigma U^T$$

where Σ is a diagonal matrix and the with values satisfying,

$$\Sigma_{1,1} > \Sigma_{2,2} > \Sigma_{3,3} > \Sigma_{4,4} > \dots$$

To reduce dimensions, truncate the Σ matrix smallest values first.

$$D' \approx V\Sigma' U^T$$

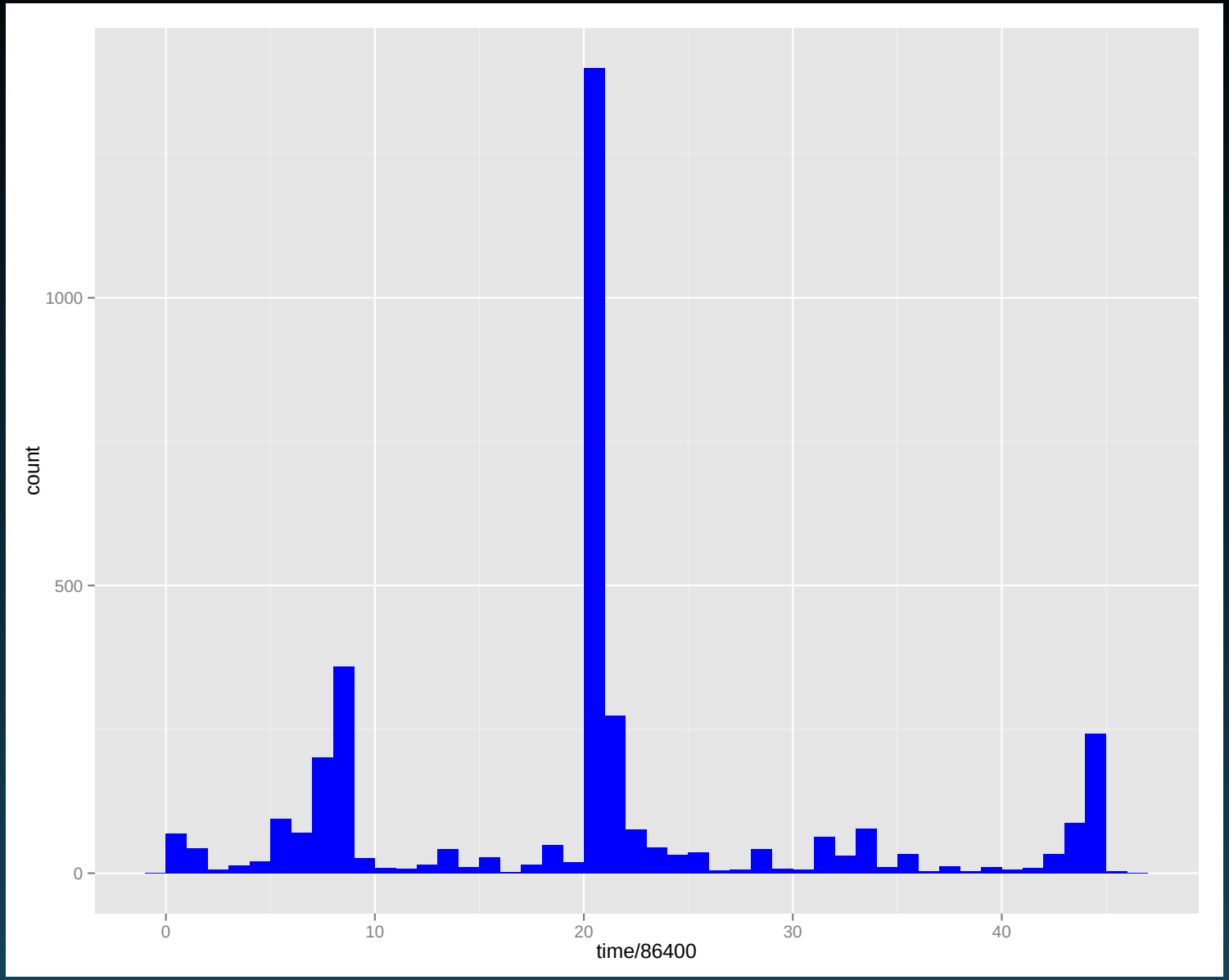
where D' has fewer columns according to how we trimmed Σ.

Disqus Topic Model

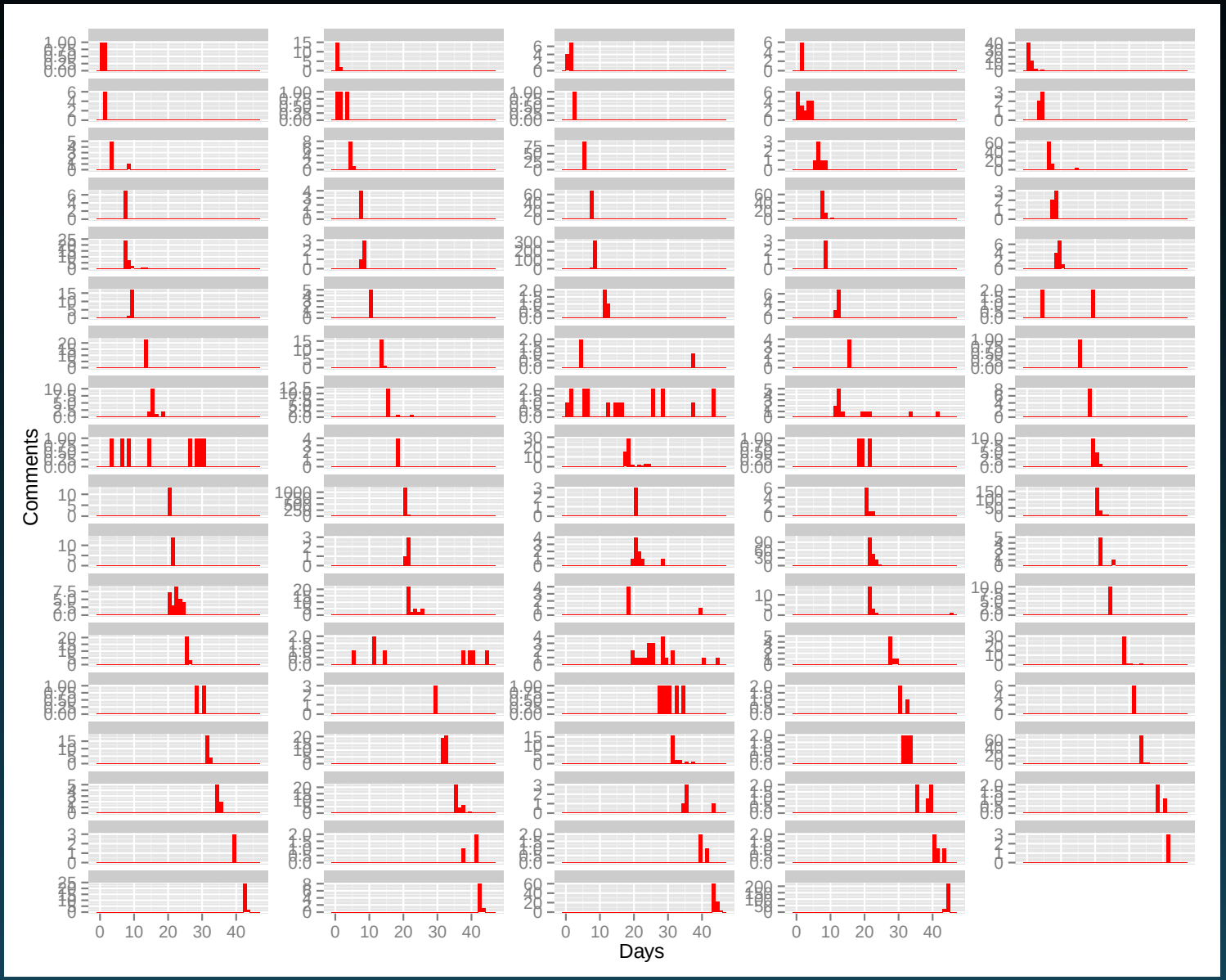
- Same 7 weeks; same keywords
- 32,856 comments from 16,886 threads
- LSI: 500 features $\rightarrow$ 80 features
- K-means: 80 clusters as topics (?!)

Focus on the intersection of Thread and Topic models

Comments with key words over time

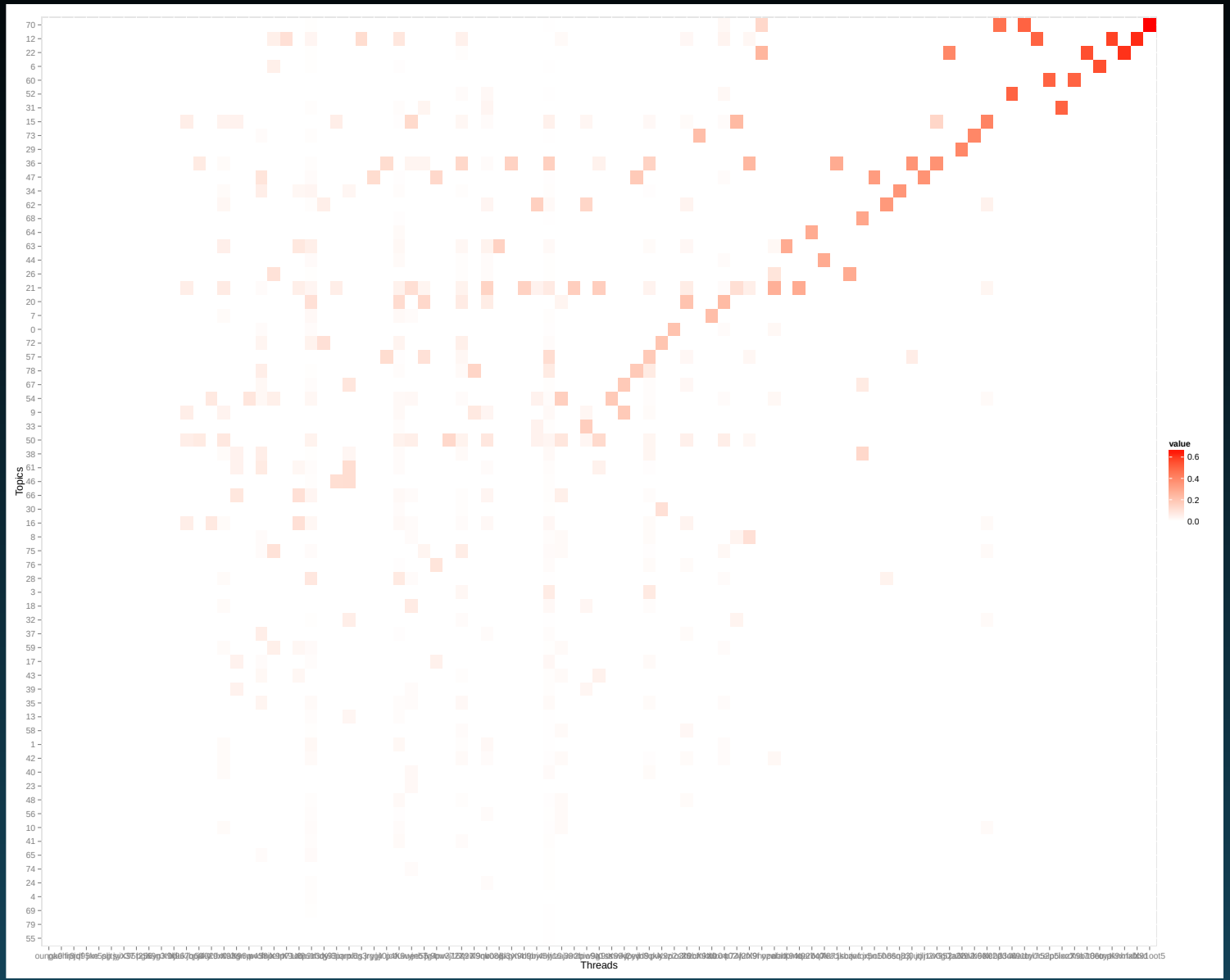


Disqus Thread Activity over Time



Disqus Topics Activity over Time

Dominant 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

Thank you!

- Presentation, data, vis. code at:
  http://github.com/DrSkippy27/CU_2014-04