Surprise!

Information Surprise or how to discover data

Oleksandr Pryymak /Sasha/
@opryymak
opryymak@gmail.com

What is a ‘surprise’?

surprise

[countable] an event, a piece of news, etc. that is unexpected or that happens suddenly
SYNONYMS: ***shock, … , eye-opener***

[uncountable, countable] a feeling caused by something happening suddenly or unexpectedly
SYNONYMS: ***astonishment, …***

(Oxford Advanced Learner's Dictionary)

Define Surprise!

?

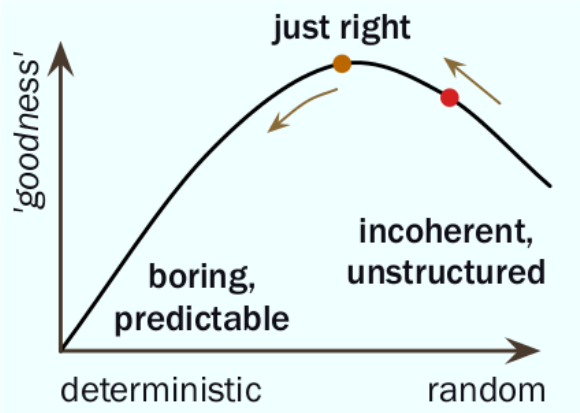
measured in

wows

Quantify Surprise!

Quantify

Need for measuring any type of content: **complexity measure**
Complex is not randomness!



Measures of complexity

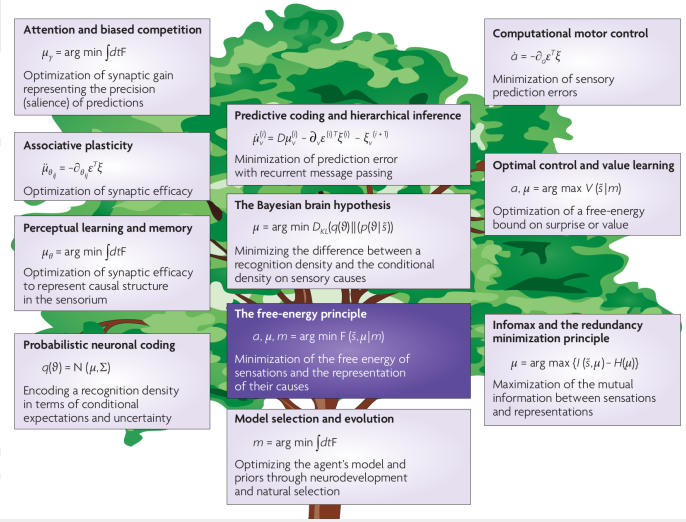
1. Subjective rating
2. #Distinct elements
3. #Dimension
4. #Control parameters
5. Minimal description
6. Information content
7. Minimal generator
8. Minimum energy

Abdallah, S., & Plumbley, M. (2009). ***Information dynamics: patterns of expectation and surprise in the perception of music.*** *Connection Science, 21(2-3), 89-117.*

Surprise Quants in academia

Neuro/Cognitive Science

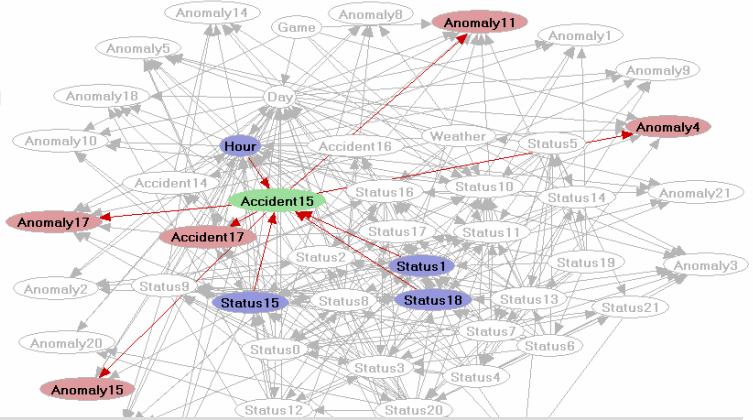
How do we perceive information?



Friston, K. (2010). ***The free-energy principle: a unified brain theory?****. Nature Reviews Neuroscience, 11(2), 127-138.*

Machine Learning

How to measure differences?

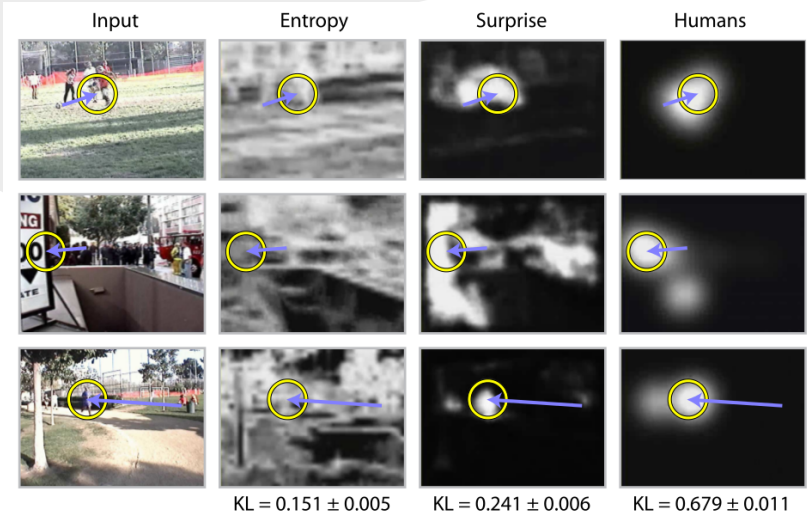


... machine that constantly tells you what you already know is just irritating. So software alerts users only to surprises...

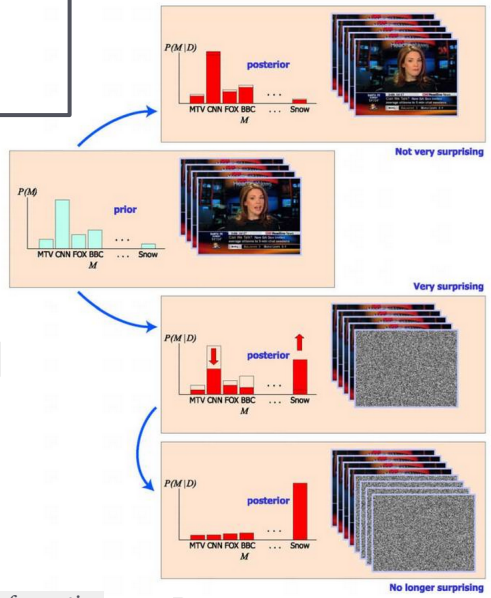
Horvitz, E., Apacible, J., Sarin, R., & Liao, L. ***Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service****.*

Surprise Quants in academia

Neuro/Cognitive Science



Machine Learning



Itti, L., & Baldi, P. F. (2005). ***Bayesian surprise attracts human attention****. In Advances in neural information processing systems (pp. 547-554).*

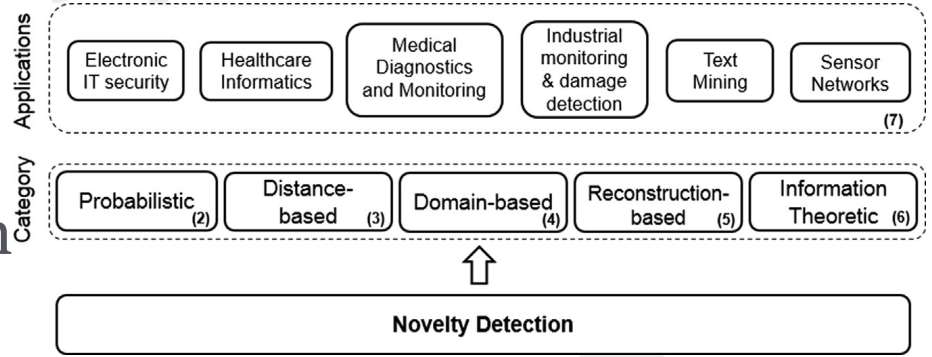
Typical ML applications

Unsupervised Learning

1. Decision trees (inf. gain)
2. MaxEnt principle
3. …

Specifically after ‘surprise’:

4. One-class classification
5. Anomaly detection
6. Novelty measure



Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). ***A review of novelty detection****. Signal Processing, 99, 215-249.*

Quantify surprisal /self-information/

The surprise /*information*/ we get from observing the occurrence of an event having probability p.

Axioms:

$$0 < p \leq 1$$

$$I(p) \geq 0$$

$$I(1) = 0$$

$$I(p_1 * p_2) = I(p_1) + I(p_2) \quad \text{//independent}$$

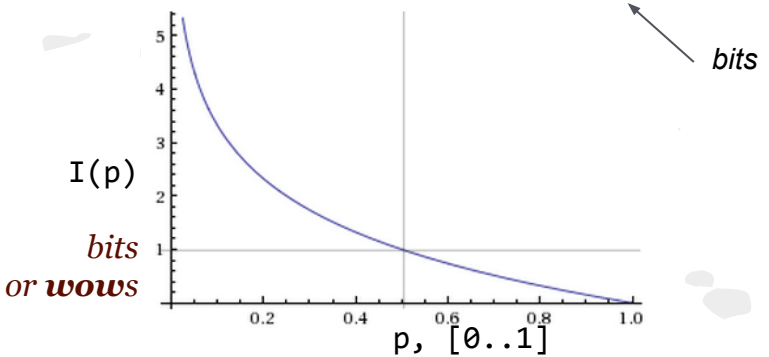
Derive:

$$I(p^2) = I(p*p) = I(p) + I(p) = 2*I(p)$$

$$I(p^a) = a*I(p)$$

Surprisal /self-information/:

$$I(p) = -\log_2(p) = \log_2(1/p)$$



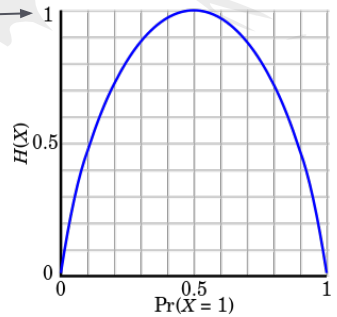
Flipping a fair coin provides 1bit of new information.

Quantify ‘knowledge’ /entropy/

The Shannon *entropy* is the expected value of the *self-information*.

$$H(P) = \langle I(p) \rangle = \sum_{i=1}^{n} p_i * \log_2(1/p_i)$$

max: $log_2(n)$



Entropy of a Bernoulli trial
X ϵ {0,1}

Notes:

1. The maximum entropy distribution is the least informative.

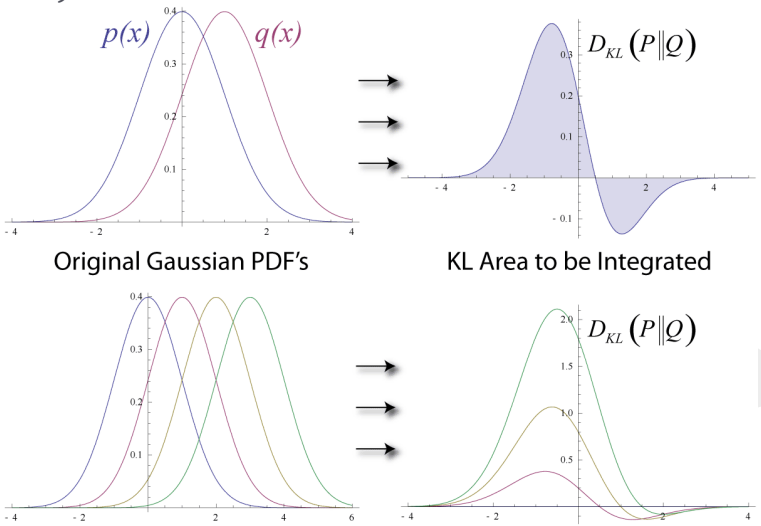
2. The *statistical mechanics* and the *information* entropy are principally the same.

Quantify ‘discovery’ /information gain/

The **Kullback–Leibler divergence** */relative entropy, information gain/*:
is a measure of the information lost when Q is used to approximate P
(measures the expected number of extra bits required to recode)

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i}$$

Assymetric: not a true measure



"KL-Gauss-Example" T. Nathan Mundhenk

Quantify ‘discovery surprise’

Symmetric KL Distances:

$$D_{KL2}(P,Q) = \sum_{i=1}^{n} (p_i - q_i) \log_2 \frac{p_i}{q_i}$$

$$D_{KL1}(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P)$$

$$D_{KL3}(P,Q) = \frac{1}{2}\left[D_{KL}\left(P||\frac{P+Q}{2}\right) + D_{KL}\left(Q||\frac{P+Q}{2}\right)\right]$$

$$D_{KL4}(P,Q) = \max(D_{KL}(P||Q), D_{KL}(Q||P))$$

All result in the same performance:

Pinto, D., Benedí, J. M., & Rosso, P. (2007). **Clustering narrow-domain short texts by using the Kullback-Leibler distance.** In *Computational Linguistics and Intelligent Text Processing*

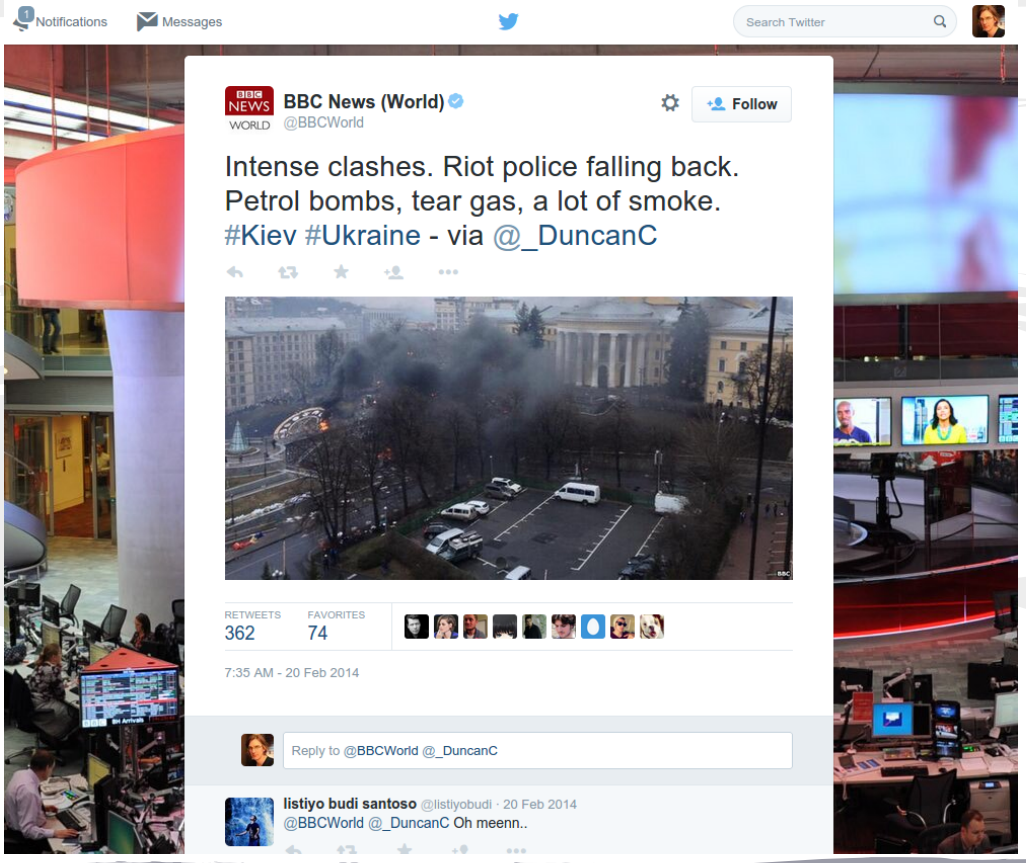
Calculating KLD

Data sparseness problem: often $p_i=0$ -> $KLD(P||Q)=\infty$
Solutions:

- drop components from calculations
- smothing:

$$p_i = \begin{cases} \beta * p_i, & \text{if } p_i > 0 \\ \epsilon = 0.0001 & \text{othewise} \end{cases}$$

$$\beta = 1 - \sum_{\forall p_i = 0} \epsilon$$



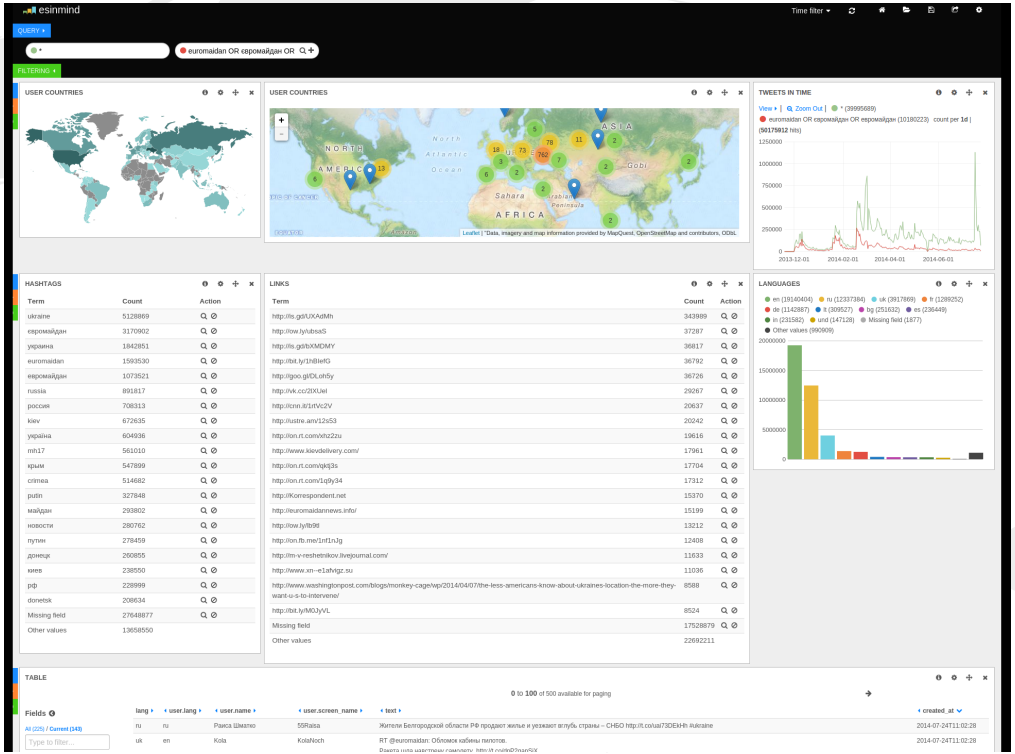
Surprise in Twitter feed

Search engines

Elasticsearch +

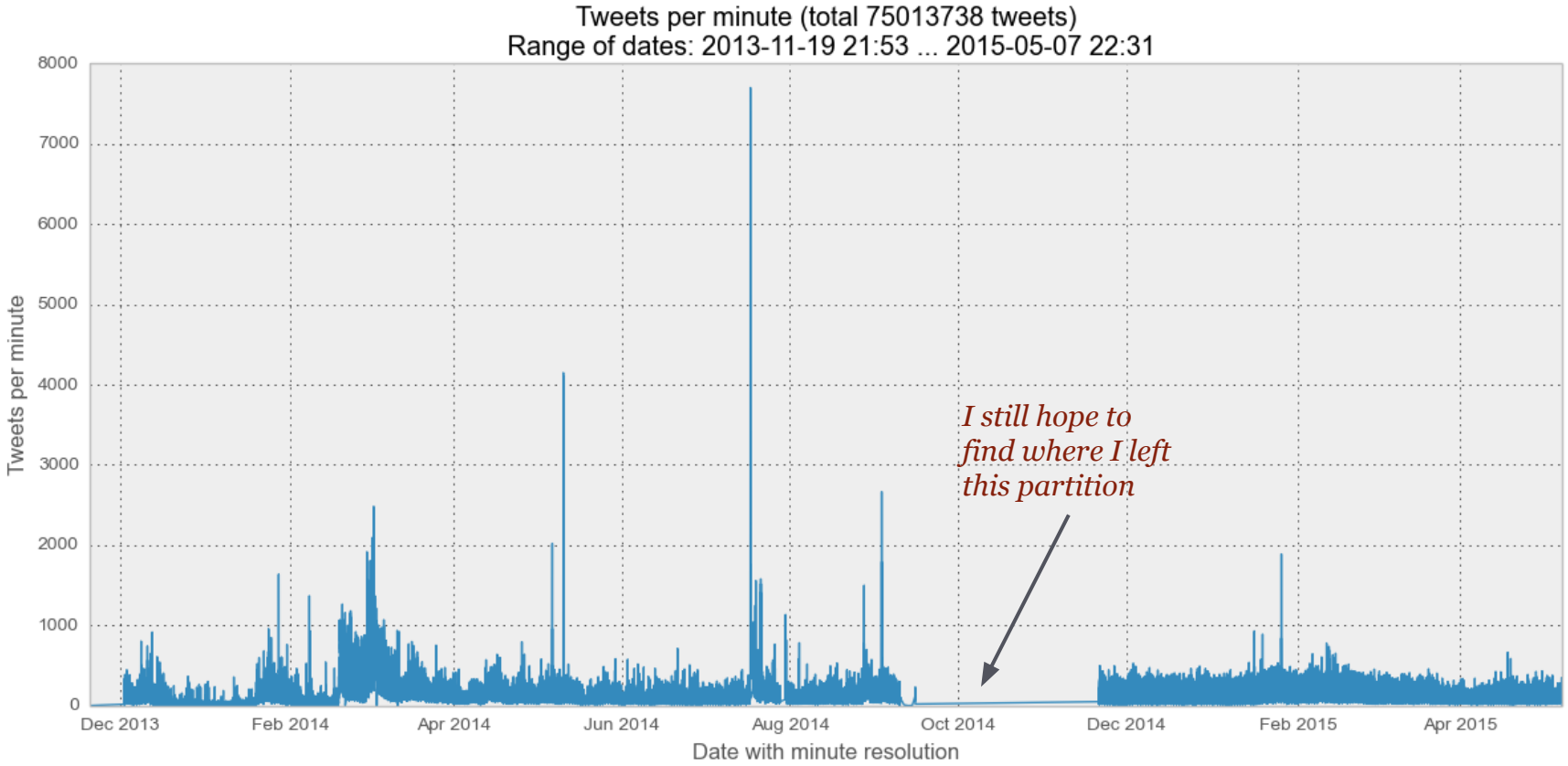
Kibana =

faceted data exploration

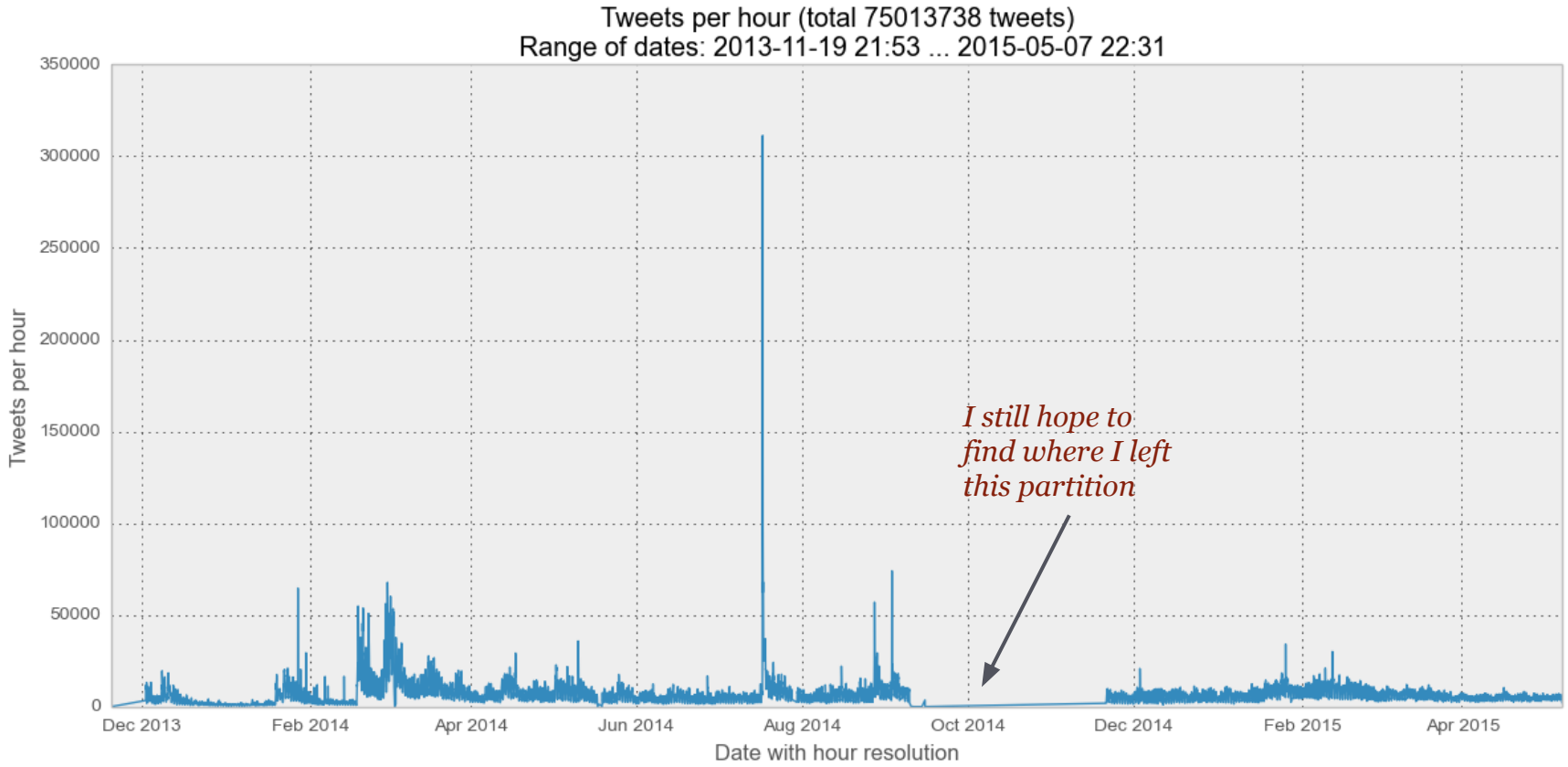


(tweets data)

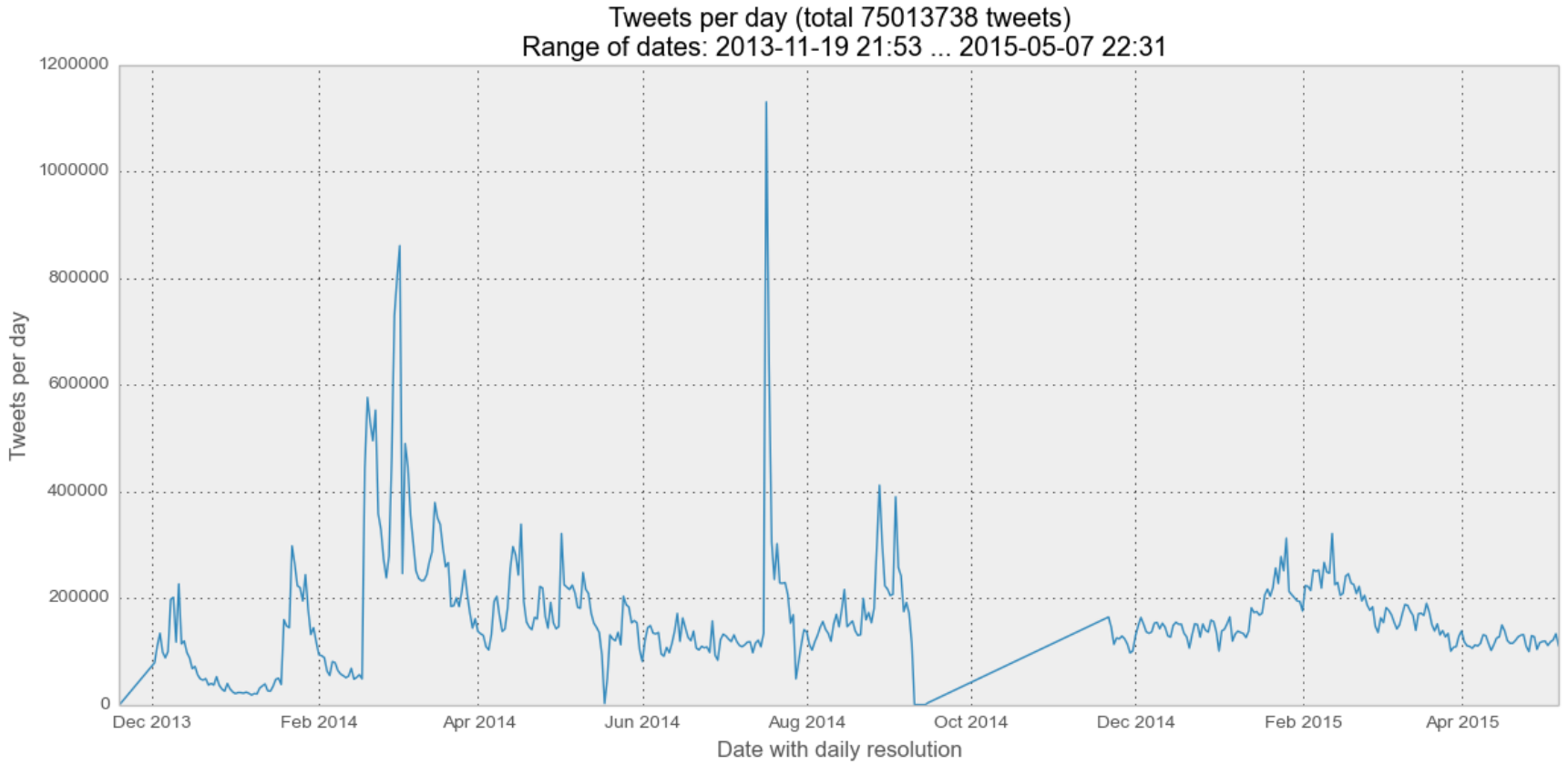
Whole dataset



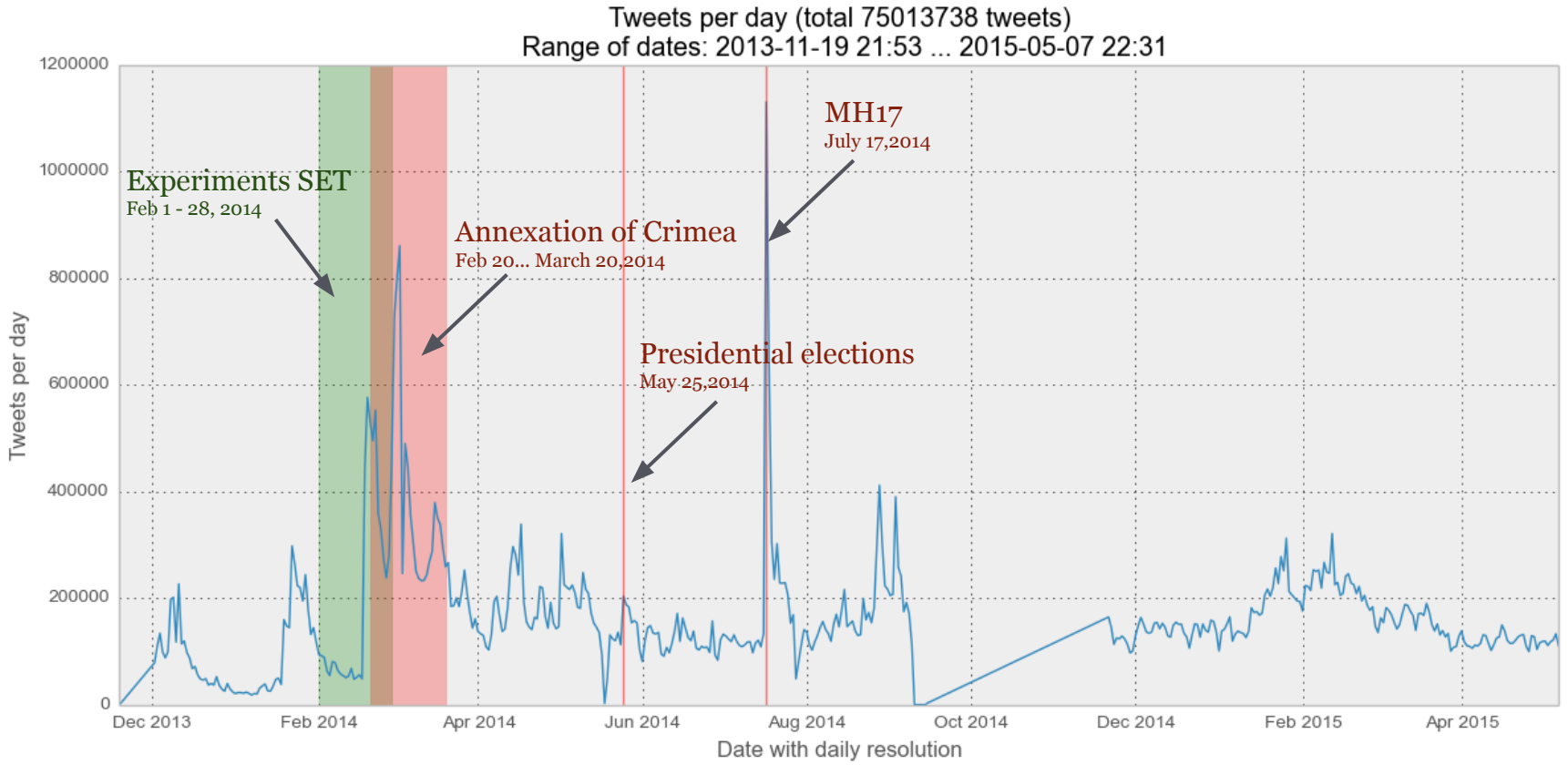
Whole dataset



Whole dataset

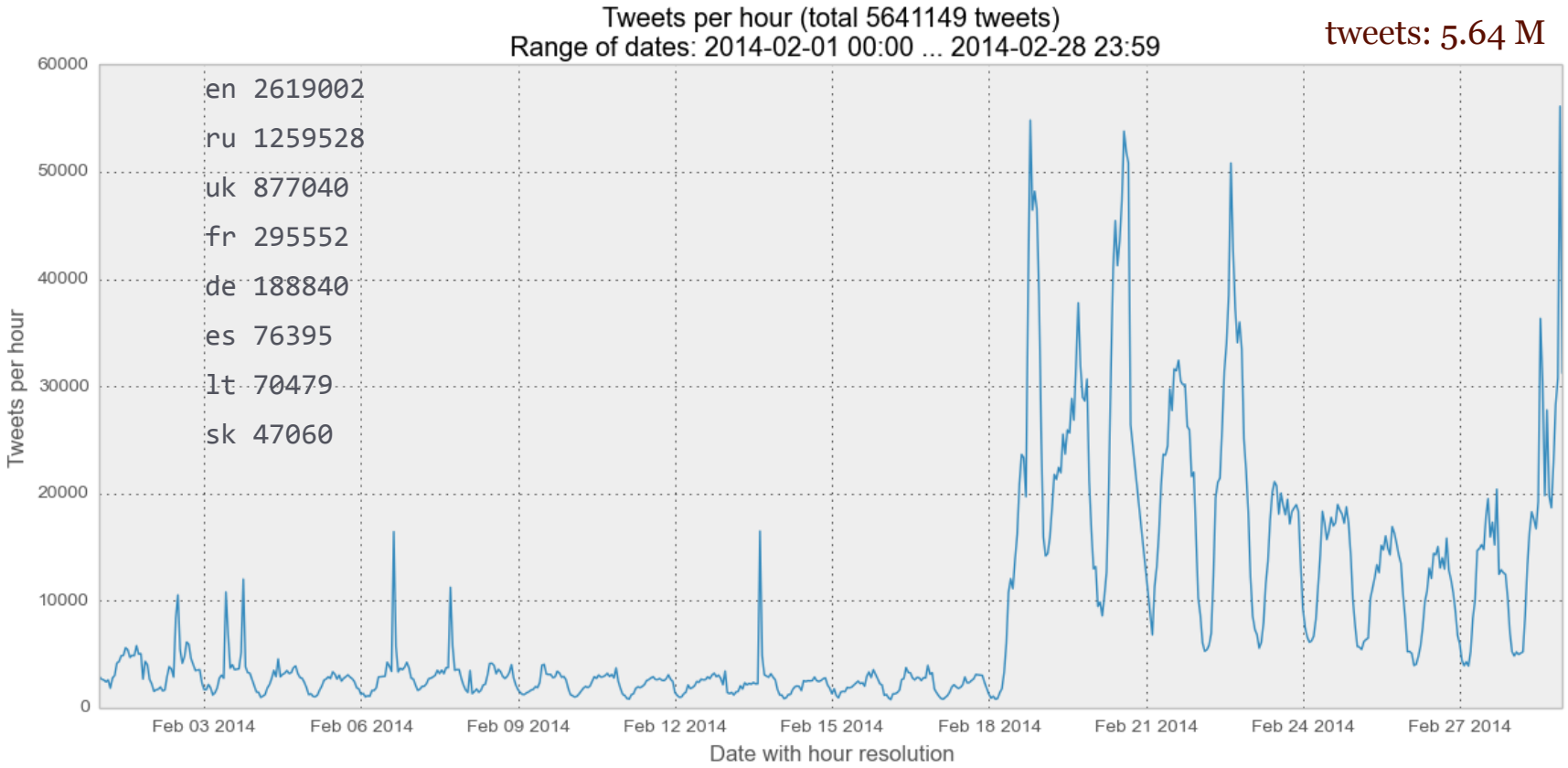


Whole dataset

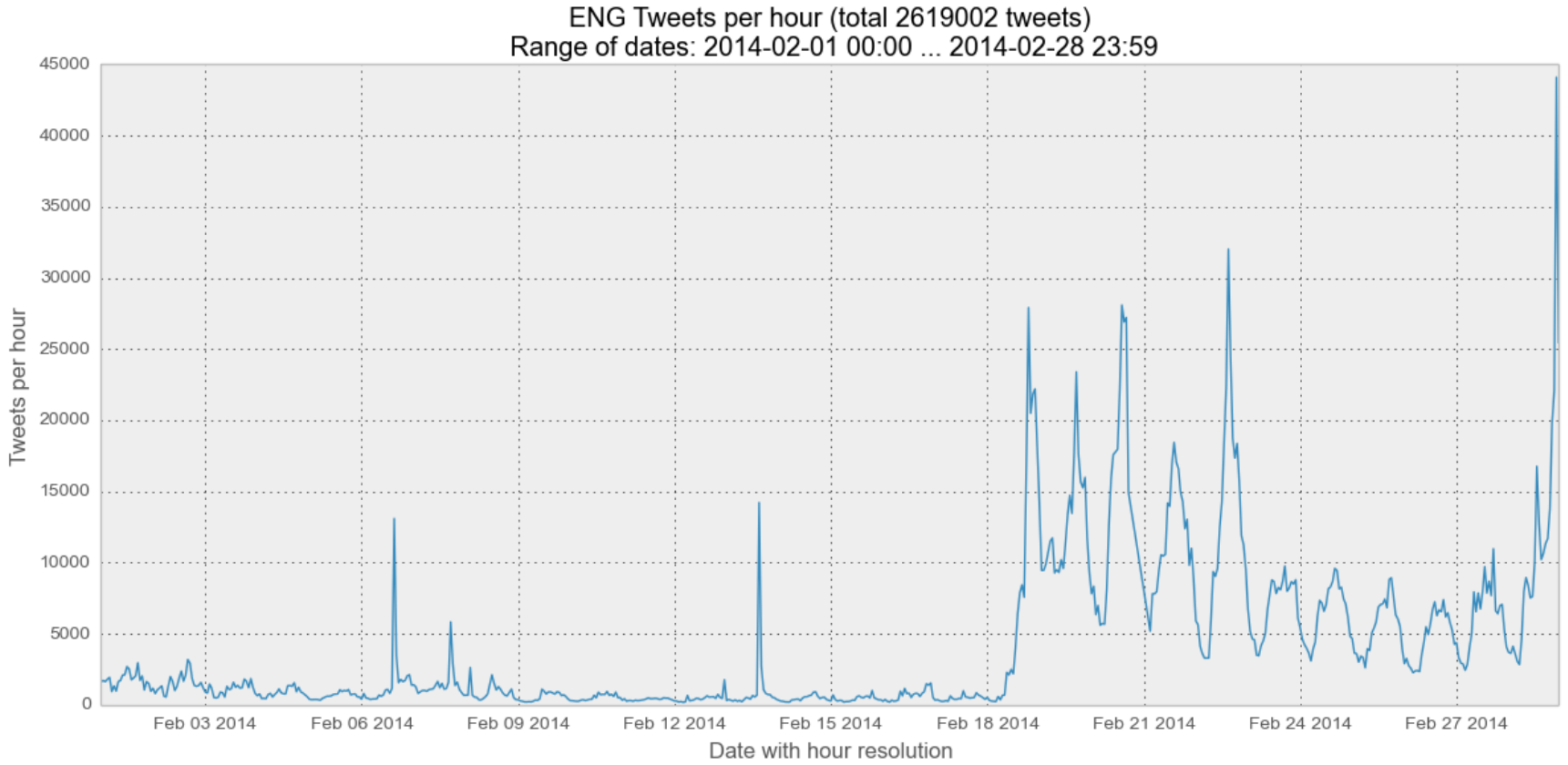


Experiment dataset: Feb2014

tweets: 5.64 M



Experiment dataset: English



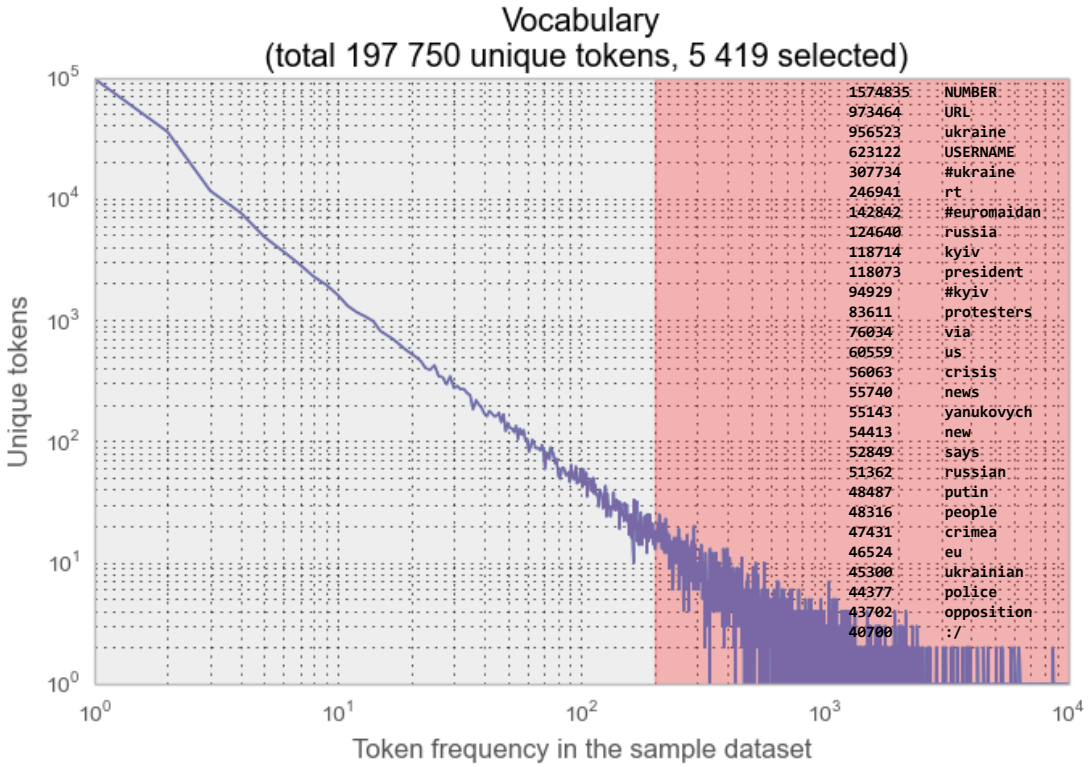
Topic modeling

Tweets are short, but prominent events are widely discussed.

Document is a time slot.

Model:

- bag of words
- freq. threshold > 200 tweets
- term frequency (naive)
- tokenizer: https://github.com/jaredks/tweetokenize
  + a few touches



Event Detection Problem

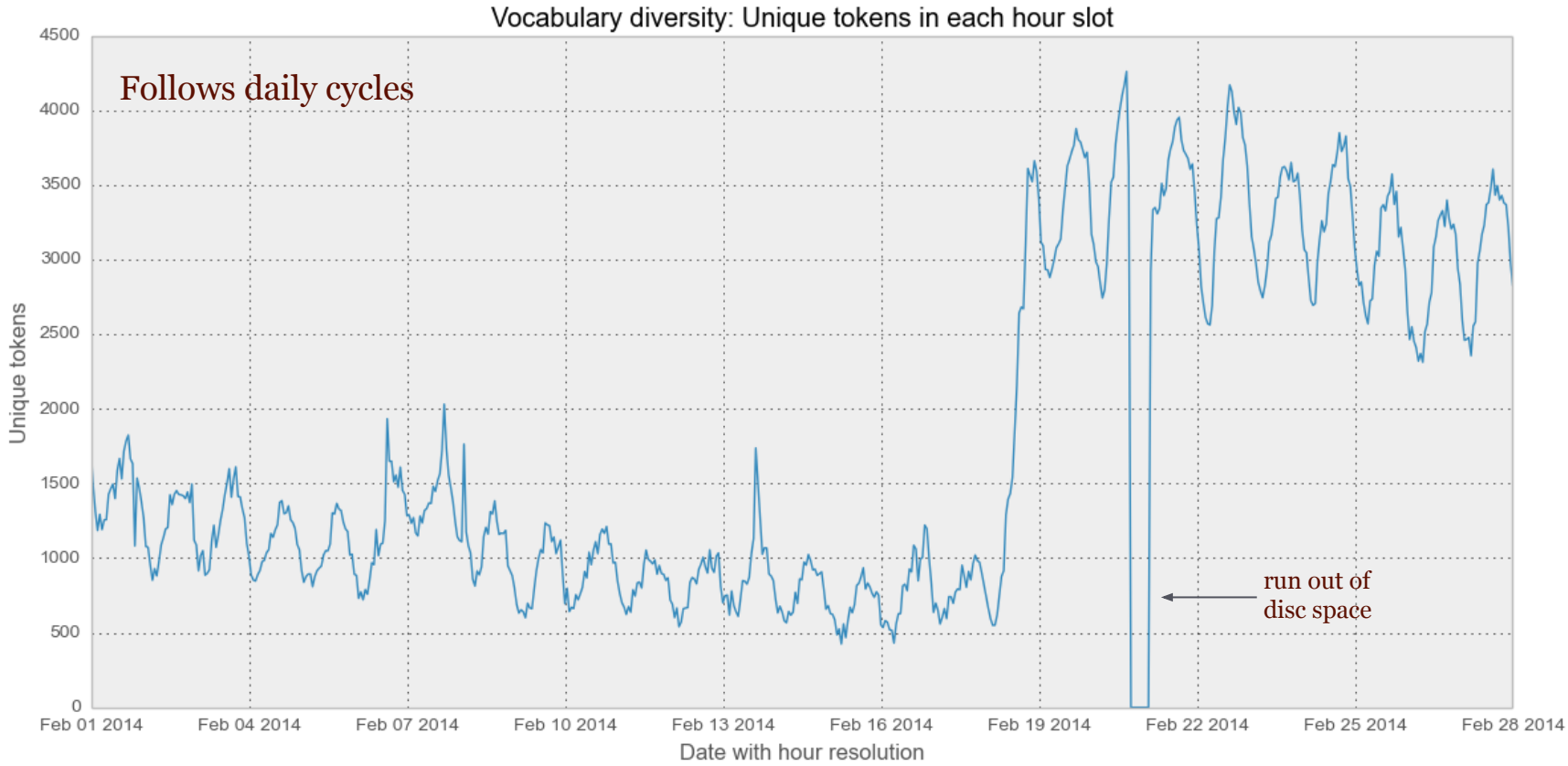
Outliers detection:

- surpriseful
- tweet rate

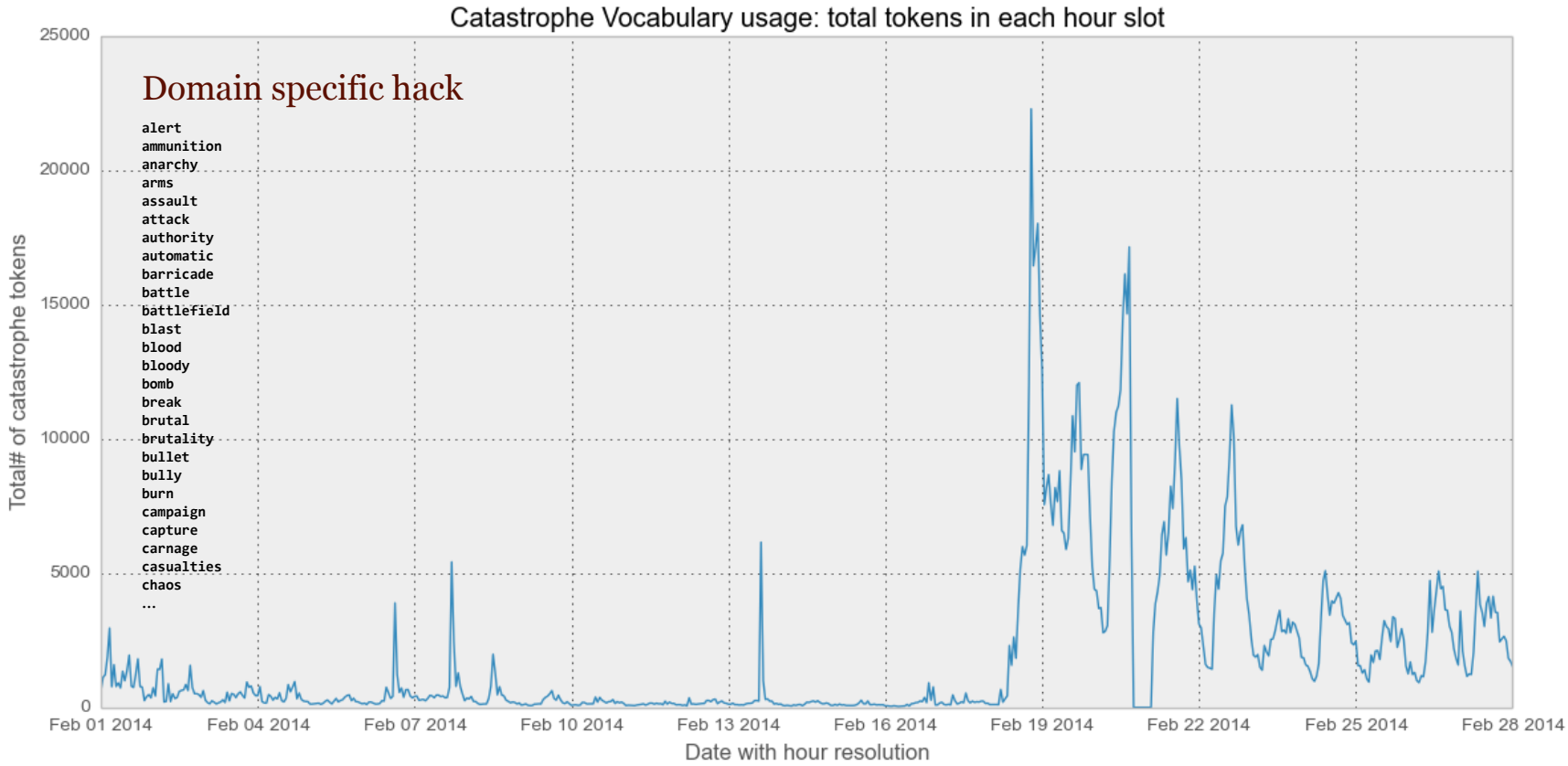
Compare against:

`historic ‘ground truth’ ->`

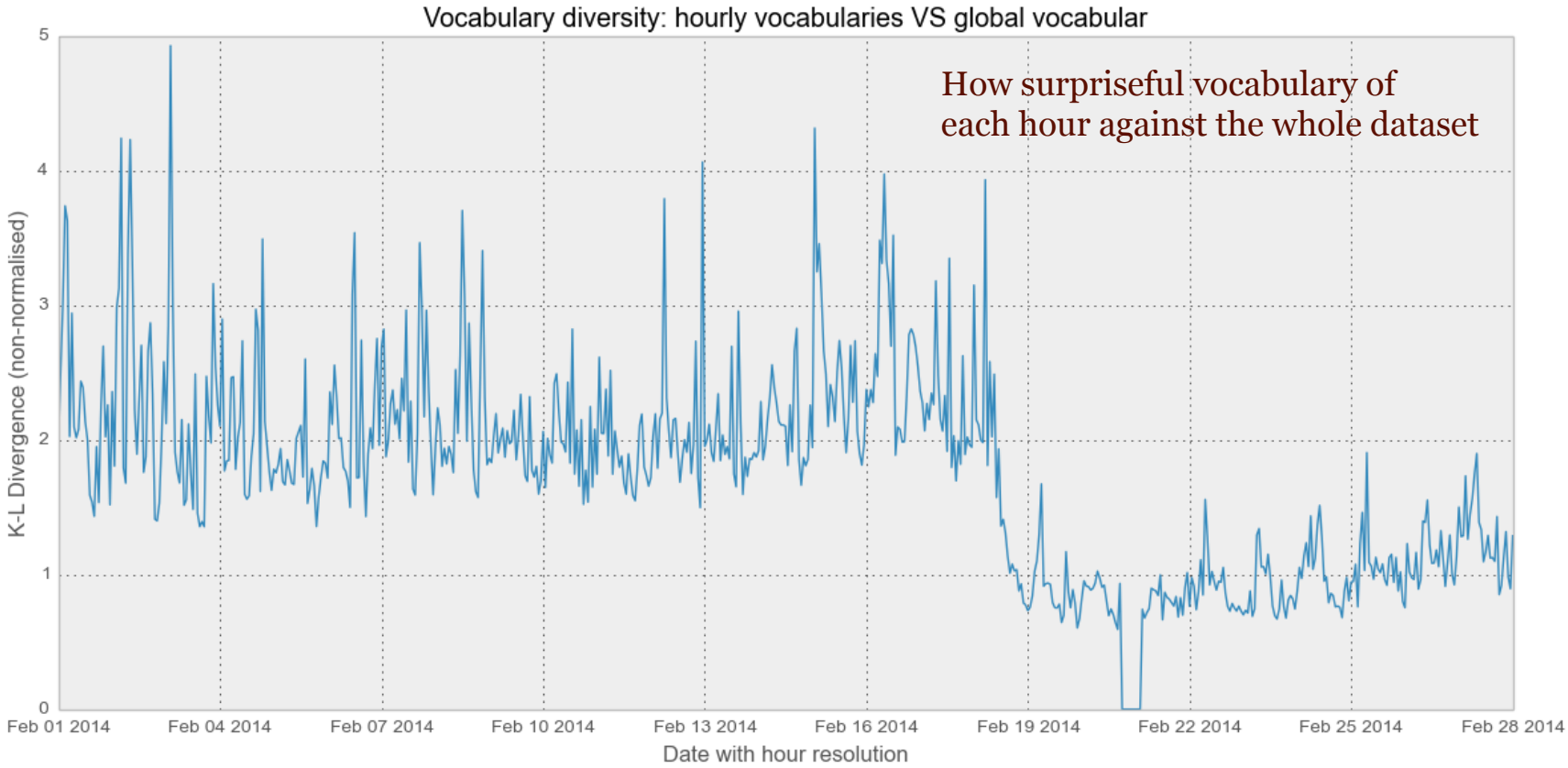
Vocabulary diversity



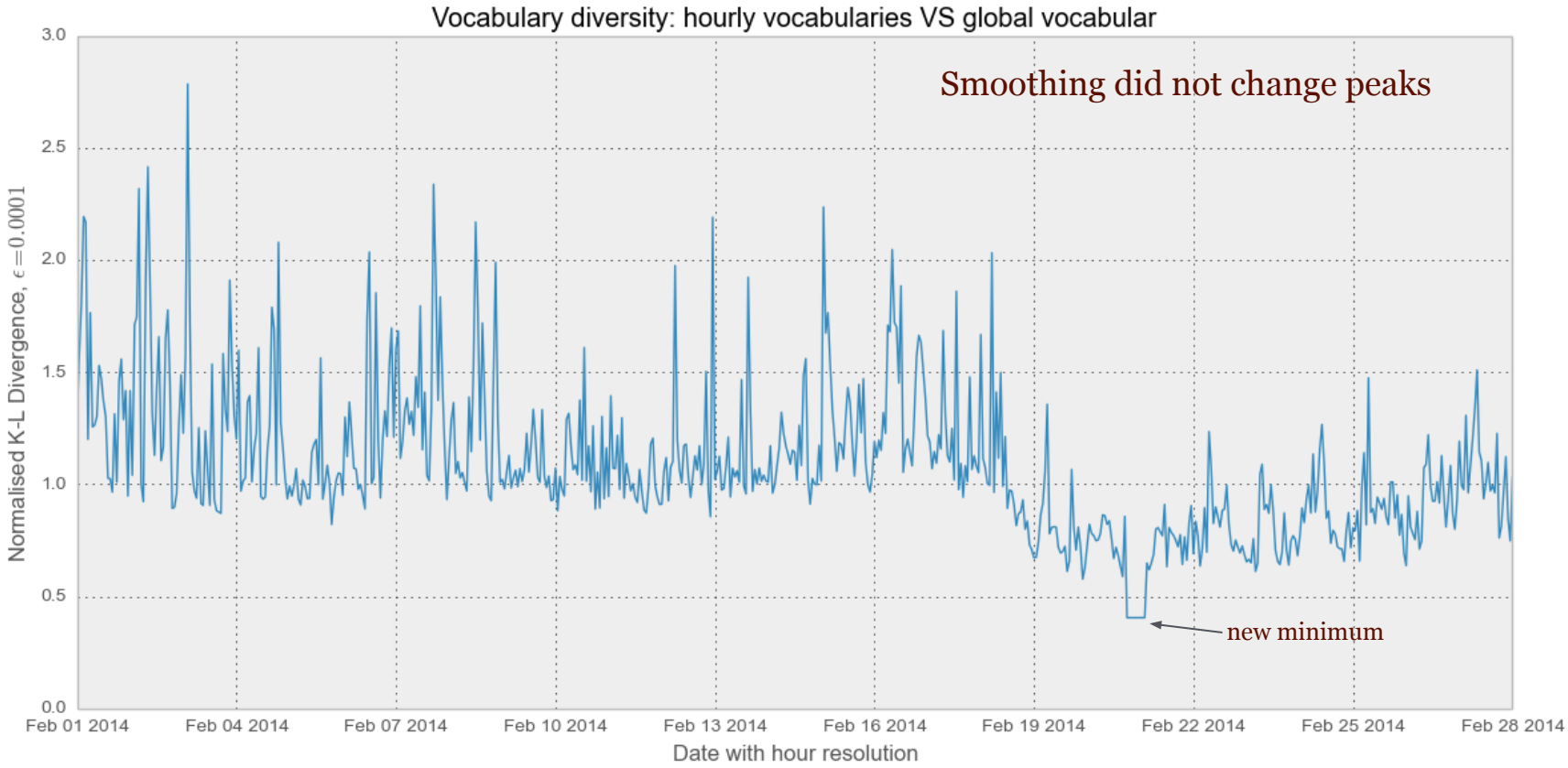
Vocabulary: catastrophe



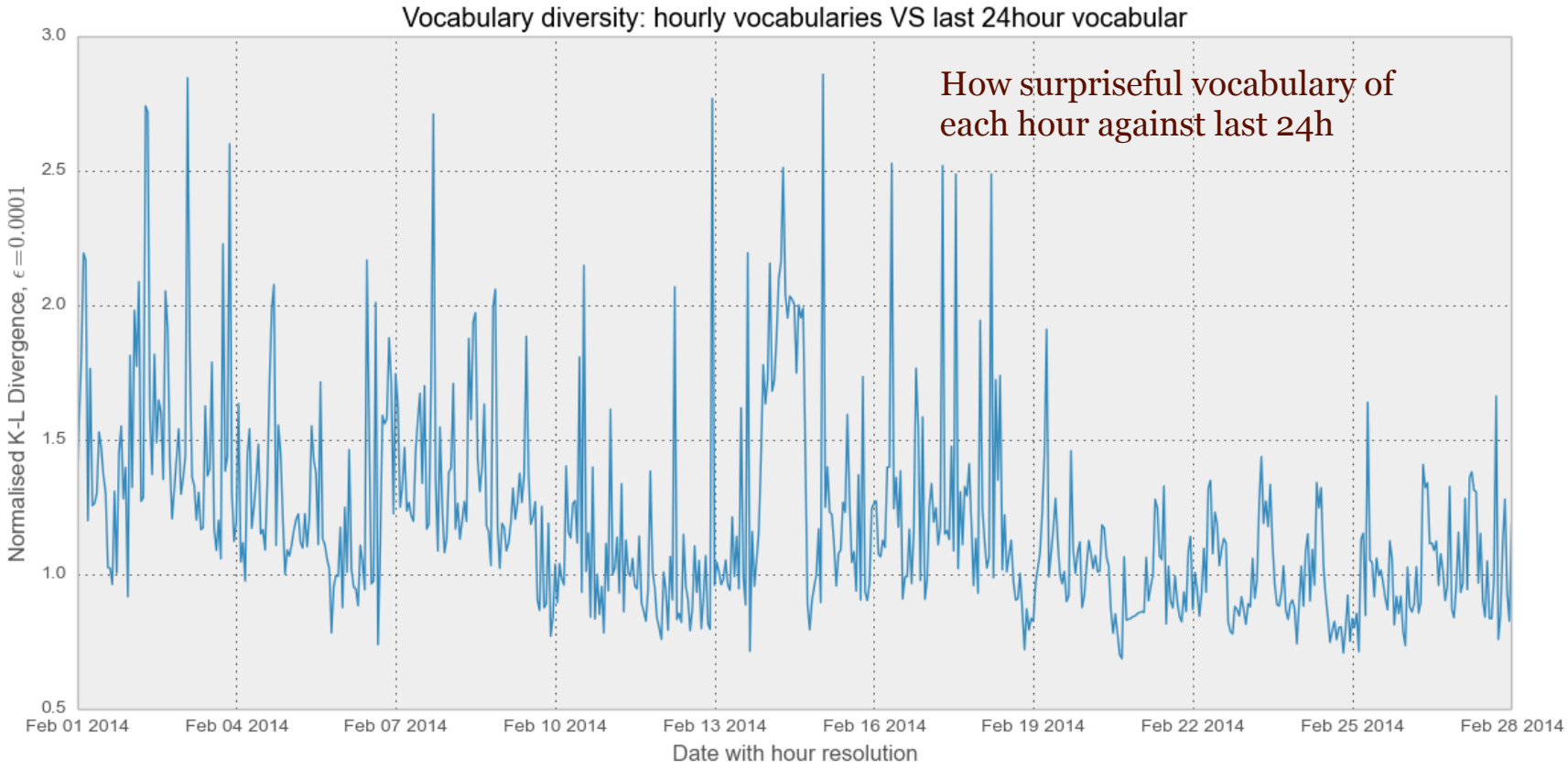
Vocabulary slots: KLD



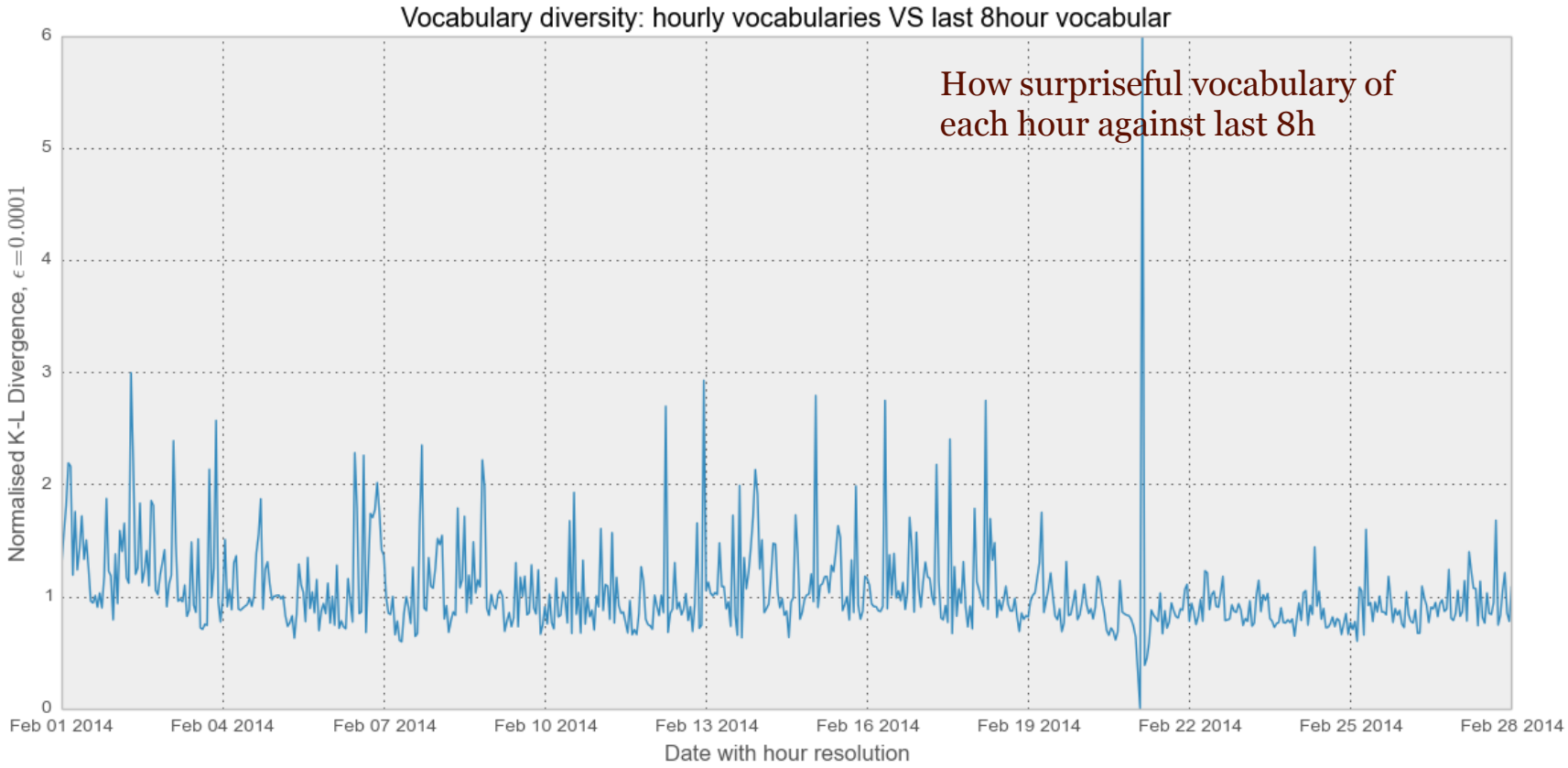
Vocabulary slots: KLD smoothed



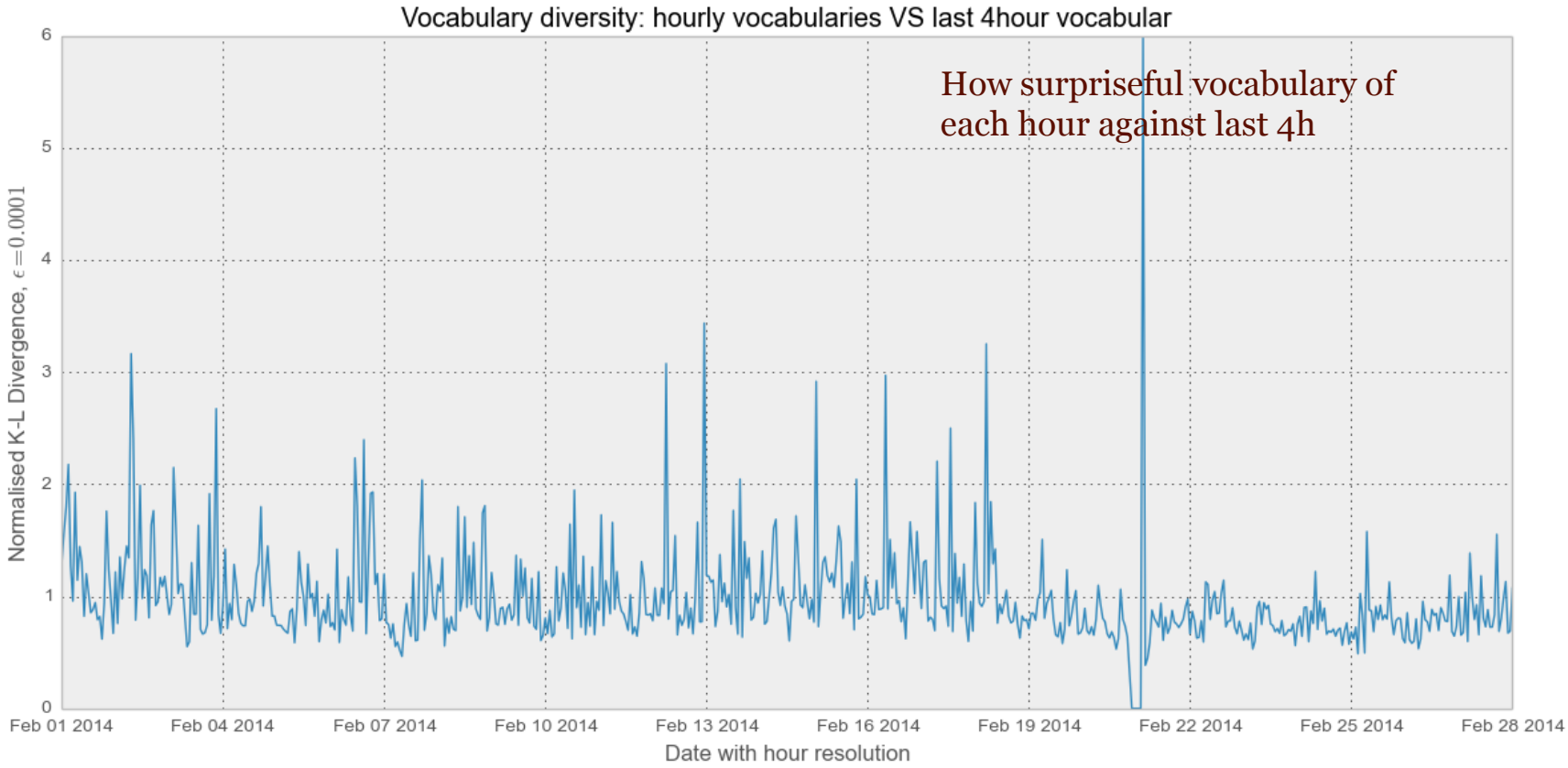
Vocabulary slots: rolling KLD



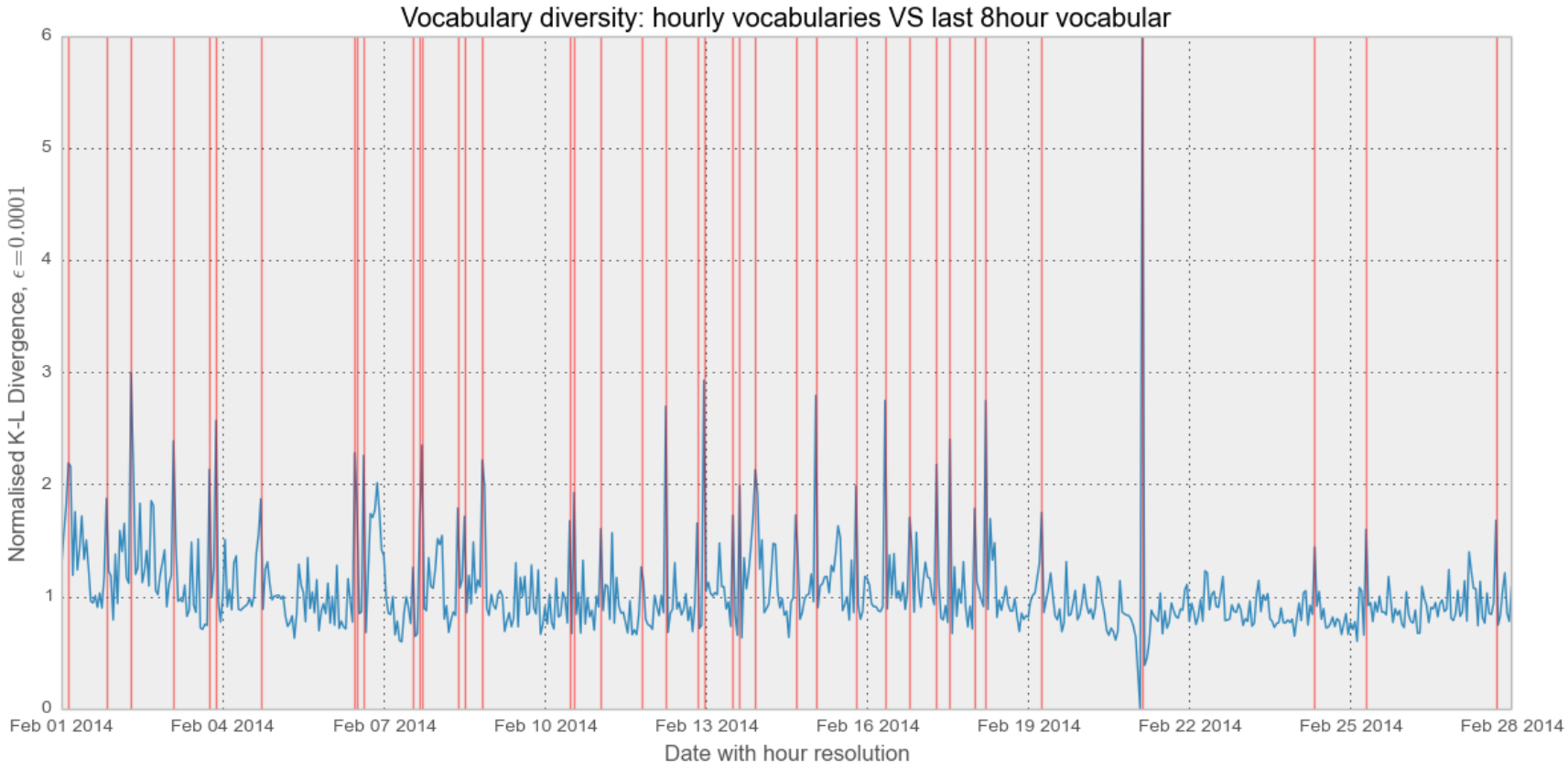
Vocabulary slots: rolling KLD



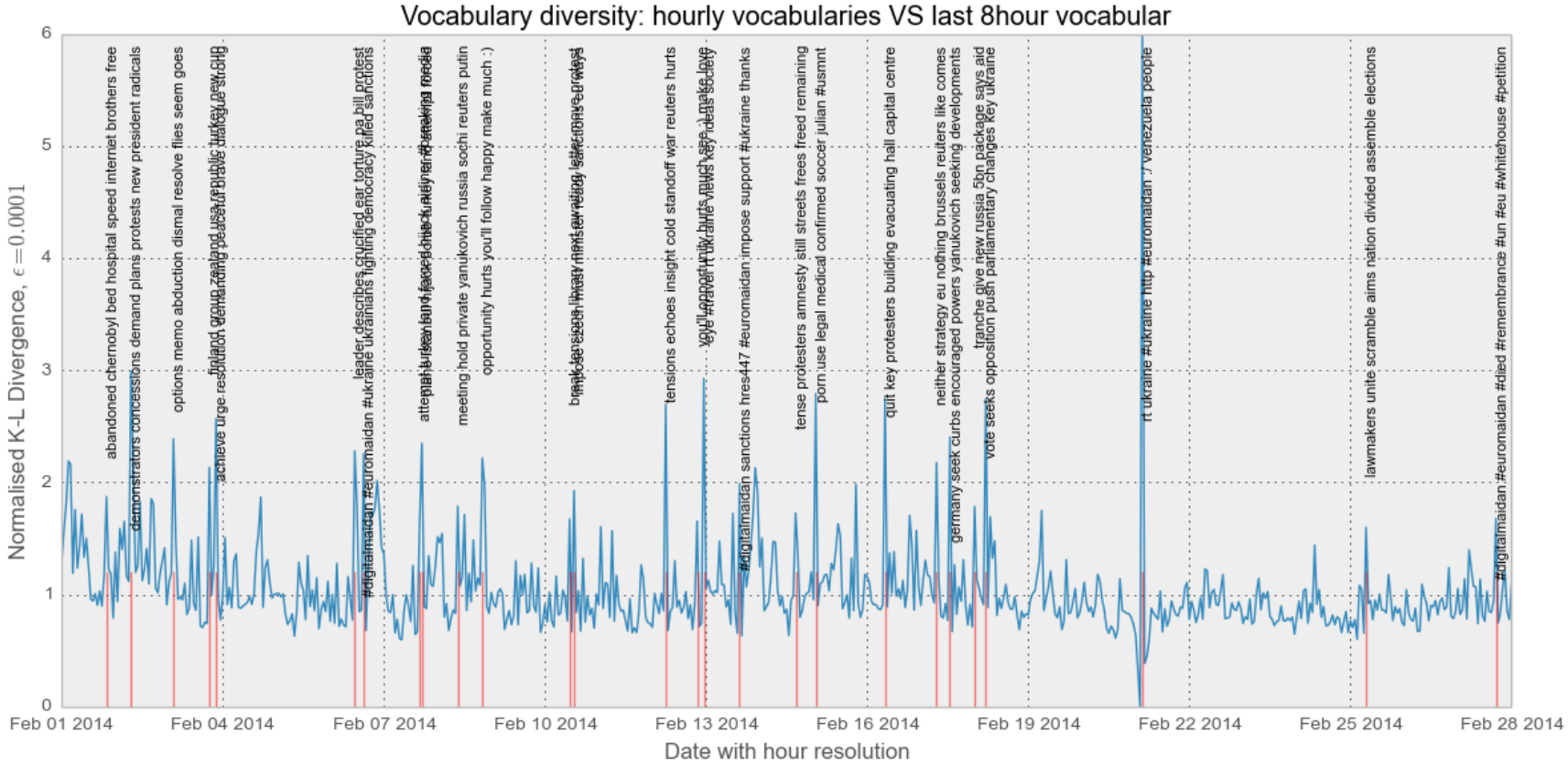
Vocabulary slots: rolling KLD



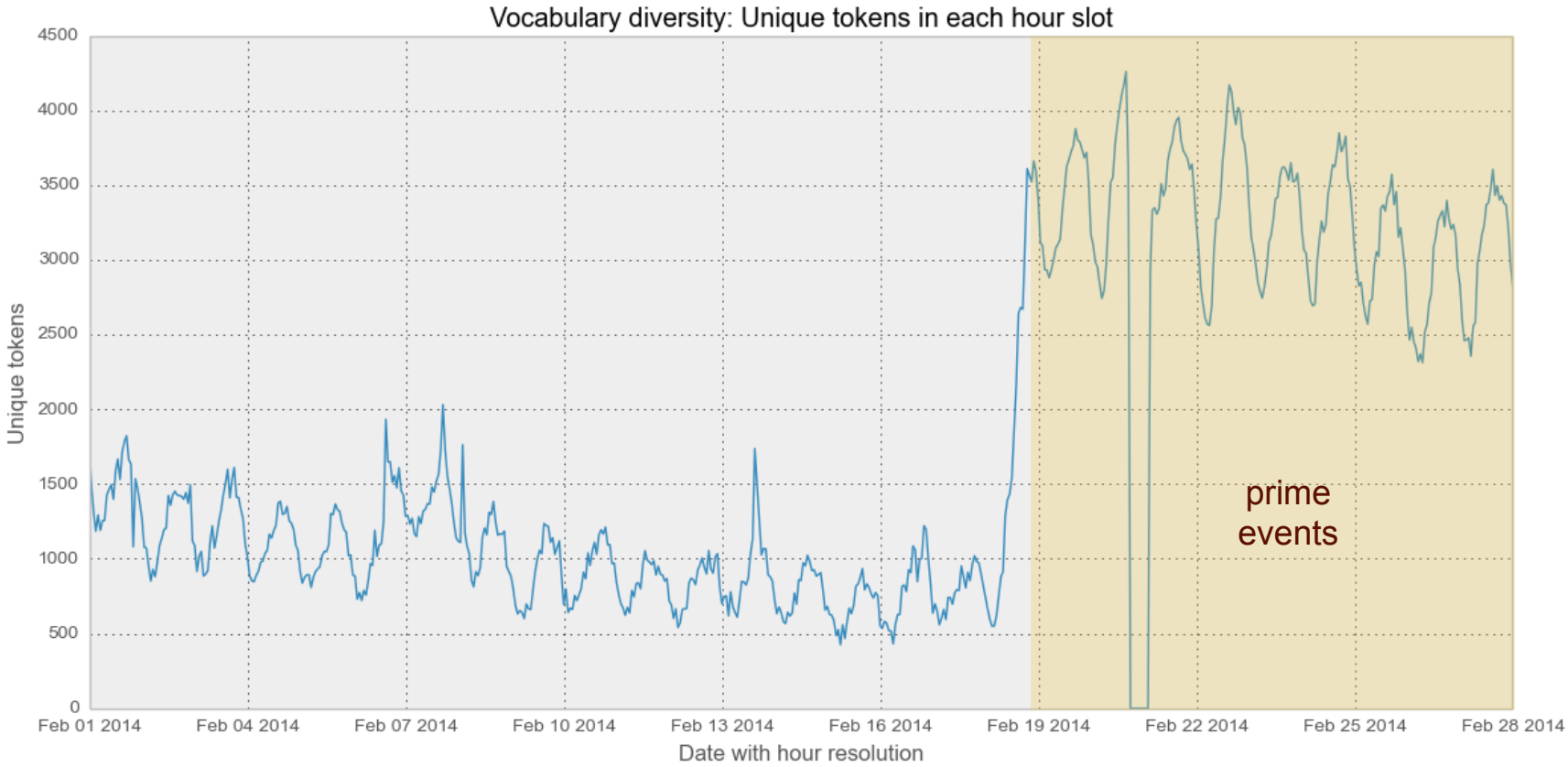
Rolling KLD outliers



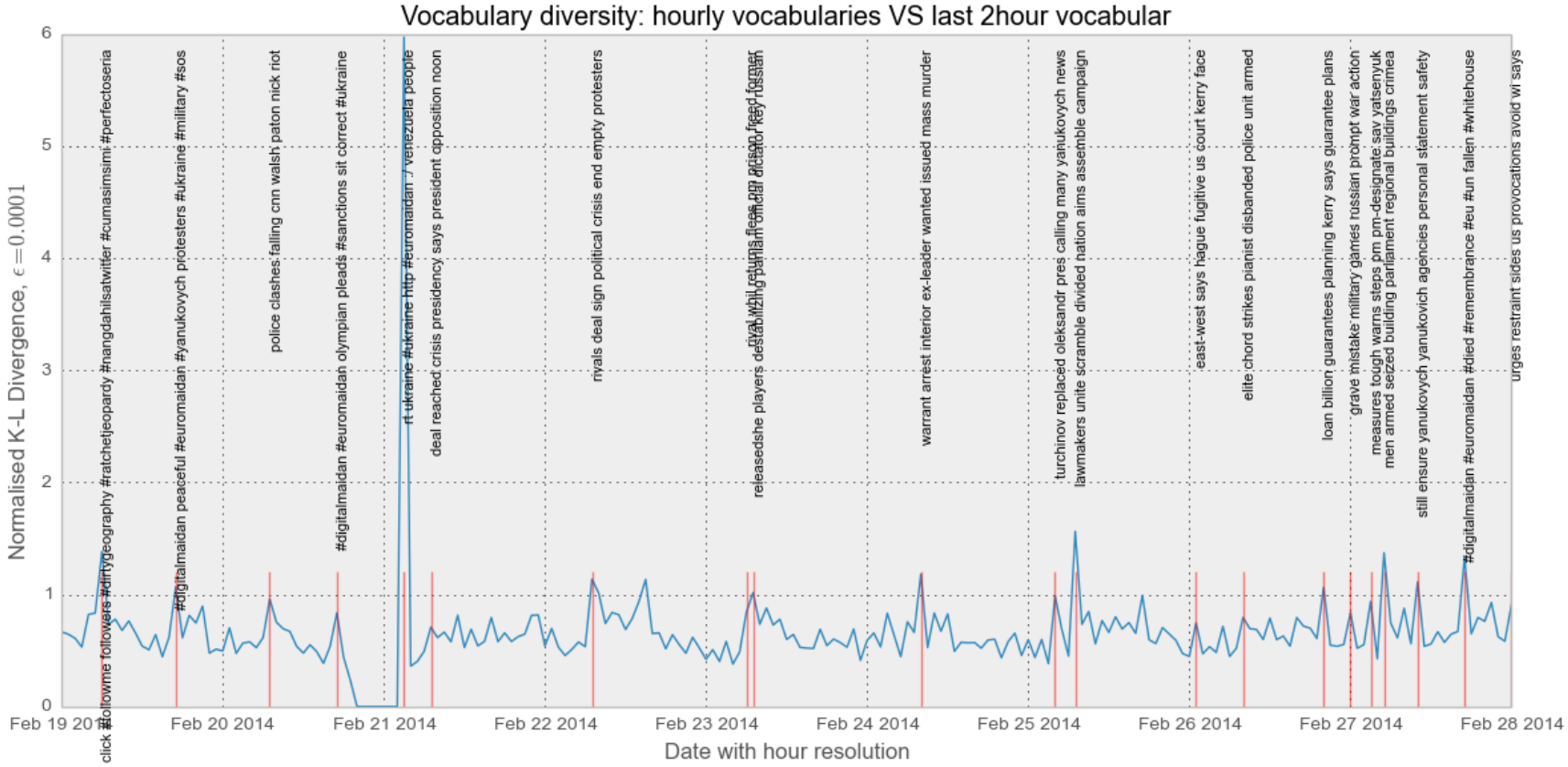
Rolling KLD outliers tokens



Further dataset limitation



Rolling KLD outliers: Feb 19-28



Surpriseful tweets link➥

9p.org.ua/tweets/t1.html

Intense clashes. Riot police falling back. Petrol bombs, tear gas, a lot of smoke. #Kiev #Ukraine - via @_DuncanC http://t.co/WfF3YKR8UD [@biesepnr 24:164]

7:00 AM
February 20, 2014

Intense clashes. Riot police falling back. Petrol bombs, tear gas, a lot of smoke. #Kiev #Ukraine - via @_DuncanC http://t.co/WfF3YKR8UD [@biesepnr 24:164]

Intense clashes. Riot police falling back. Petrol bombs, tear gas, a lot of smoke. #Kiev #Ukraine http://t.co/hLnKfZVfjs - via @?DuncanC [@INAANEWS 1499:257]

RT @BBCWorld: Intense clashes. Riot police falling back. Petrol bombs, tear gas, a lot of smoke. #Kiev #Ukraine - via @_DuncanC http://t.co… [@punkobnoii 79:218] ➥

"@atomiktiger: I stand with the people of the #Ukraine #PJNET #tcot #ORPUW #ccot #teaparty #RedNationRising http://t.co/yNXRMtcxBD" #2A [@Metz_Byron 749:1026]

This is a special #coffee for #Ukraine #bondia #goodmorning #buenosdías #café http://t.co/ZWPIfGBGFG [@valeronline 10:49]

"It's an imp. time for Canada to stand for a Ukraine that's democratic, open & transparent." Cdn. Ukranian Congress http://t.co/aeBpvFlZmY [@Vida_Jay 1621:2002]

5:00 PM
@Marcorubio #SOS #Ukraine #Yanukovych #murders peaceful protesters #Sanctions Now http://t.co/4WqOfJDUzl #euromaidan #digitalmaidan [@RomanMALANKA 141:206]

5:00 PM
Thank you @CBCNews for using correct spelling of KYIV. #EuroMaidan #DigitalMaidan #Ukraine @nytimes @nytimesworld @politico @ABC @abcNews [@Milleeer27 70:60]

@RT_com: MORE: #Ukraine's Yanukovich wanted for mass murder of peaceful citizens - interim interior minister http://t.co/Zr5CGAcVLn true?? [@akatulkhajuria 103:655]

Russian official: West 'destabilizing Ukraine': Ex-PM Tymoshenko

CLICK 500

Intense clashes. Riot police falling back. Petrol bombs, tear gas, a lot of

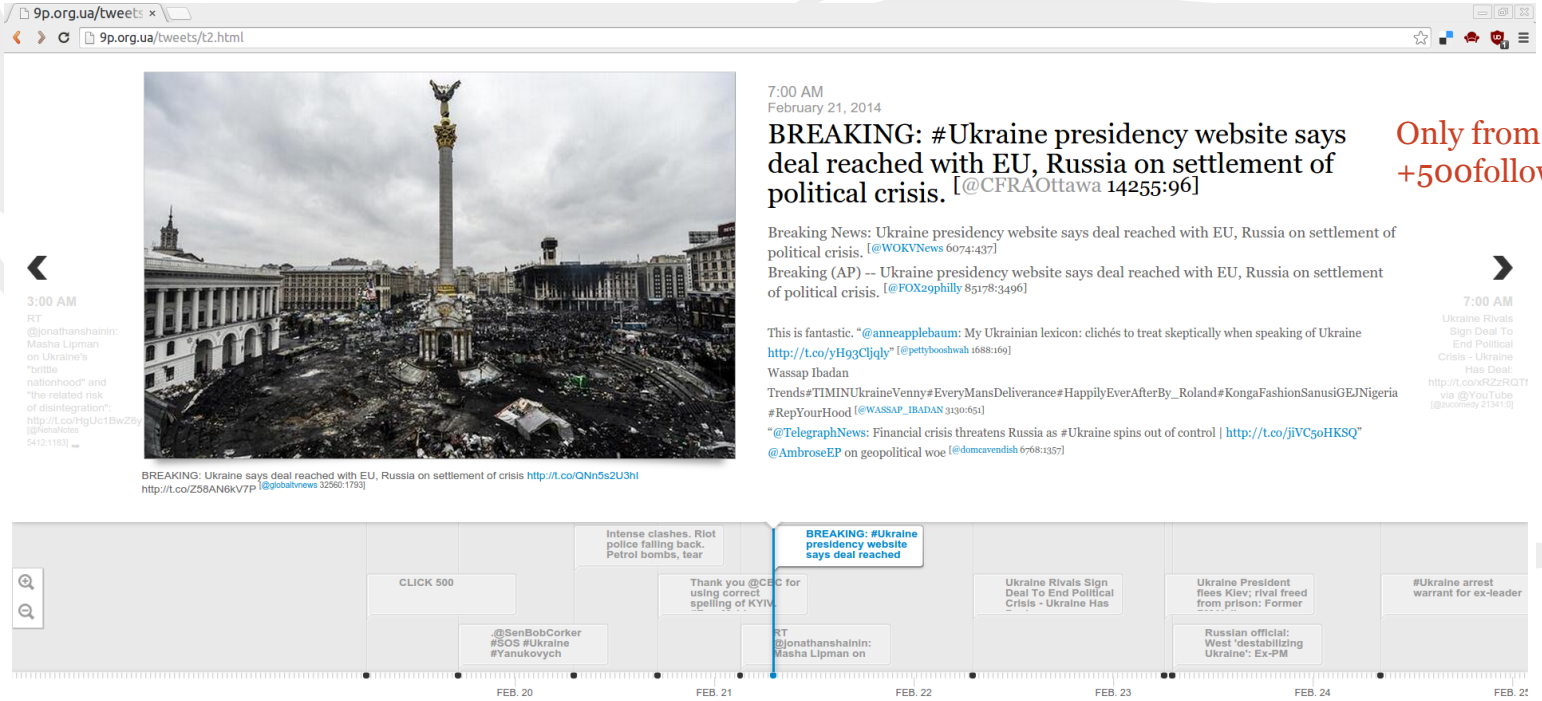
BREAKING: #Ukraine presidency website says deal reached with EU,

Ukraine Rivals Sign Deal To End Political Crisis - Ukraine Has Deal

Ukraine President flees Kiev; rival freed from prison: Former PM Yulia

FEB. 20 FEB. 21 FEB. 22 FEB. 23 FEB. 24

Surpriseful tweets link

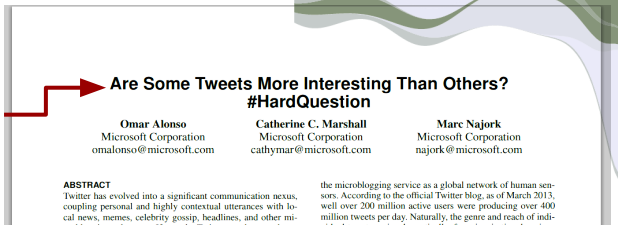


Only from users with +500followers

majdannezalezhnosti.blogspot.com

To improve

1. Benchmark
   a. export ‘hot’ events from online media
2. Fight bots
   a. spam (repetitions, bots)
   b. ‘forced’ opinions
   c. filter low quality
3. Topic model
   a. no Term Frequency
   b. split topics (!)



Are Some Tweets More Interesting Than Others?
#HardQuestion

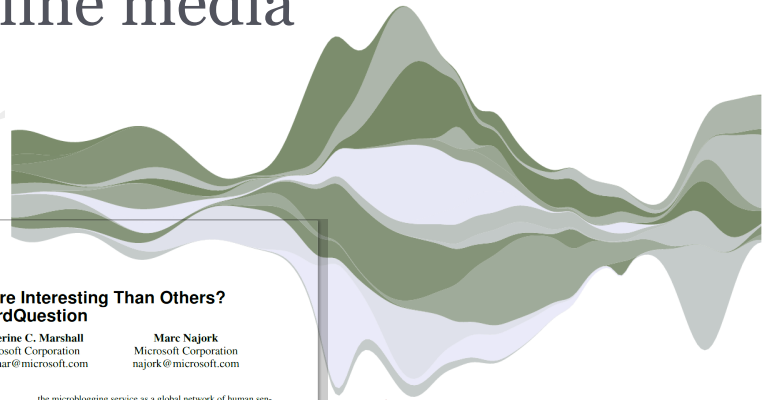
Omar Alonso
Microsoft Corporation
omalonso@microsoft.com

Catherine C. Marshall
Microsoft Corporation
cathymar@microsoft.com

Marc Najork
Microsoft Corporation
najork@microsoft.com

ABSTRACT
Twitter has evolved into a significant communication nexus, coupling personal and highly contextual utterances with local news, memes, celebrity gossip, headlines, and other mi-

the microblogging service as a global network of human sensors. According to the official Twitter blog, as of March 2013, well over 200 million active users were producing over 400 million tweets per day. Naturally, the genre and reach of indi-



Q?