From Event Detection to Storytelling on Microblogs

Janani Kalyanam*
UC San Diego

Sumithra Velupillai
KING'S COLLEGE
KTH VETENSKAP OCH KONST

Mike Conway
THE UNIVERSITY OF UTAH

Gert Lanckriet
UC San Diego

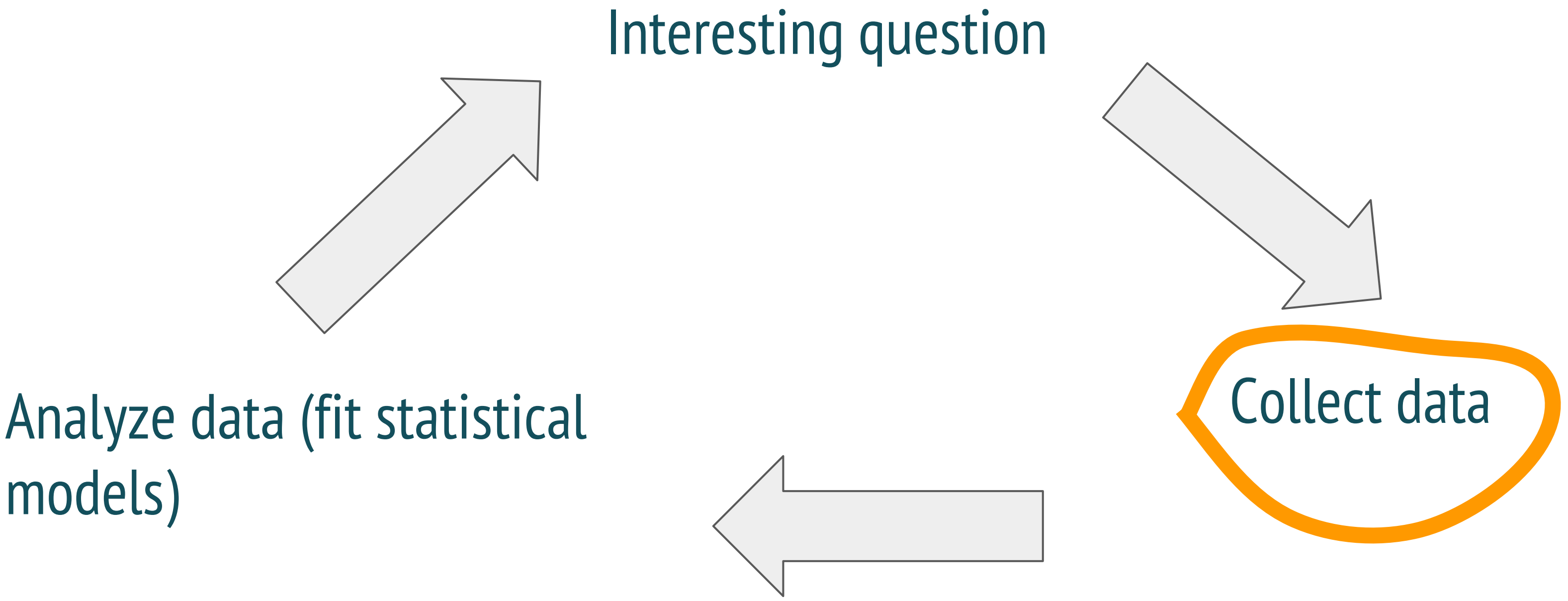
Research Pipeline

Interesting question

Collect data

Analyze data (fit statistical models)

Research Pipeline



How do we make sense of this?

Photo credit: https://marqana.com/wp-content/uploads/embed_toomany.jpg

Some tools

- Topic Modeling

- Event Detection

Some tools - limitations

- Topic Modeling - very short documents

- Event Detection - detect only the onset of events

Our focus in this work...

Our focus

- Topic Modeling - very short documents

- Event Detection - detect only the onset of events

Model

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}$$

doc-term doc-top top-term

Model

$$\mathbf{X} \approx \mathbf{WH}$$

Very sparse

Observation

The size of the vocabulary increases only marginally with increasing number of documents.

(Yan et. al. 2013)

Model

The term-by-term matrix $\mathbf{K}$ is relatively denser.

Model

The term-by-term matrix $\mathbf{K}$ is relatively denser.

Hence, decompose $\mathbf{K}$

$$\underbrace{\mathbf{K}}_{\text{term-term}} \approx \underbrace{\mathbf{Q}^T}_{\text{term-top}} \underbrace{\mathbf{Q}}_{\text{top-term}}$$

Event Progression

Assume a set of documents arrive at every time step

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T} \mathbf{Q}^t$$

Event Progression

Assume a set of documents arrive at every time step

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T} \mathbf{Q}^t$$

Relate the current data to the past history

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T}$$

Event Progression

Assume a set of documents arrive at every time step

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T} \mathbf{Q}^t$$

Relate the current data to the past history

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T} \mathbf{T}^t \mathbf{Q}^{t-1}$$

Event Progression

$$\mathbf{K}^t \approx \underbrace{\mathbf{Q}^{t^T} \mathbf{T}^t}_{\text{top-top}} \underbrace{\mathbf{Q}^{t-1}}_{\text{top-term}}$$

Quick Aside:

$$\mathbf{K}^t \approx \mathbf{Q}^{t^T}\mathbf{T}^t\mathbf{Q}^{t-1}$$

The "tracking matrix" - helps connect the present to the past. This will help build the "timelines" for events.

Model

$$||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{Q}^t||_F^2 + ||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{T}^t\mathbf{Q}^{t-1}||_F^2$$

Model

$$||\mathbf{K}^t - {\mathbf{Q}^t}^T \mathbf{Q}^t||_F^2 + ||\mathbf{K}^t - {\mathbf{Q}^t}^T \mathbf{T}^t \mathbf{Q}^{t-1}||_F^2$$

KNOWN

Optimization

- Collective Matrix Factorization (Singh and Gordon 2008)

- Details in the paper

- Implementation on github

https://github.com/kjanani/matrix_factorization/blob/master/matrix_factorization.py

Interesting Question

Ebola Outbreak 2014. What really happened?

Interesting Question: Goals

Ebola Outbreak 2014. What really happened?

- Want list of all events that occurred.

- How did they progress?

Tasks

- Topic Detection
- Event timelines

Topic Detection

- Estimated topics: Estimate $\mathbf{Q}^t$ at every timestep. Each row is a distribution over words.

- Groundtruth topics: hashtags

Topic Detection Baselines

- Two baselines from the event detection literature

- A few classic topic modeling baselines (NMF, LDA e.t.c.)

Topic Detection Results

$k = 5$

model type	model	NDCG	MAP
[Ours]	MEP	0.2027	0.0953
event detection	trend-detect	0.1823	0.0862
	o-cluster	0.1677	0.0892
topic modeling	O-BTM	0.1745	0.091
	nmf	0.1722	0.0864
	lda	0.1245	0.0589

$k = 7$

model type	model	NDCG	MAP
[Ours]	MEP	0.1626	0.0706
event detection	trend-detect	0.1502	0.0539
	o-cluster	0.1310	0.0534
topic modeling	O-BTM	0.1459	0.0569
	nmf	0.1306	0.0565
	lda	0.0837	0.0366

$k = 10$

model type	model	NDCG	MAP
[Ours]	MEP	0.1430	0.0696
event detection	trend-detect	0.1379	0.0667
	o-cluster	0.1320	0.0606
topic modeling	O-BTM	0.1271	0.0412
	nmf	0.1057	0.0463
	lda	0.0660	0.0164

Event Timelines

How to come up with *timelines* of events?

Event Timelines

How to come up with *timelines* of events?

Look at the tracking matrix $\mathbf{T}^t$

Tracking Matrix

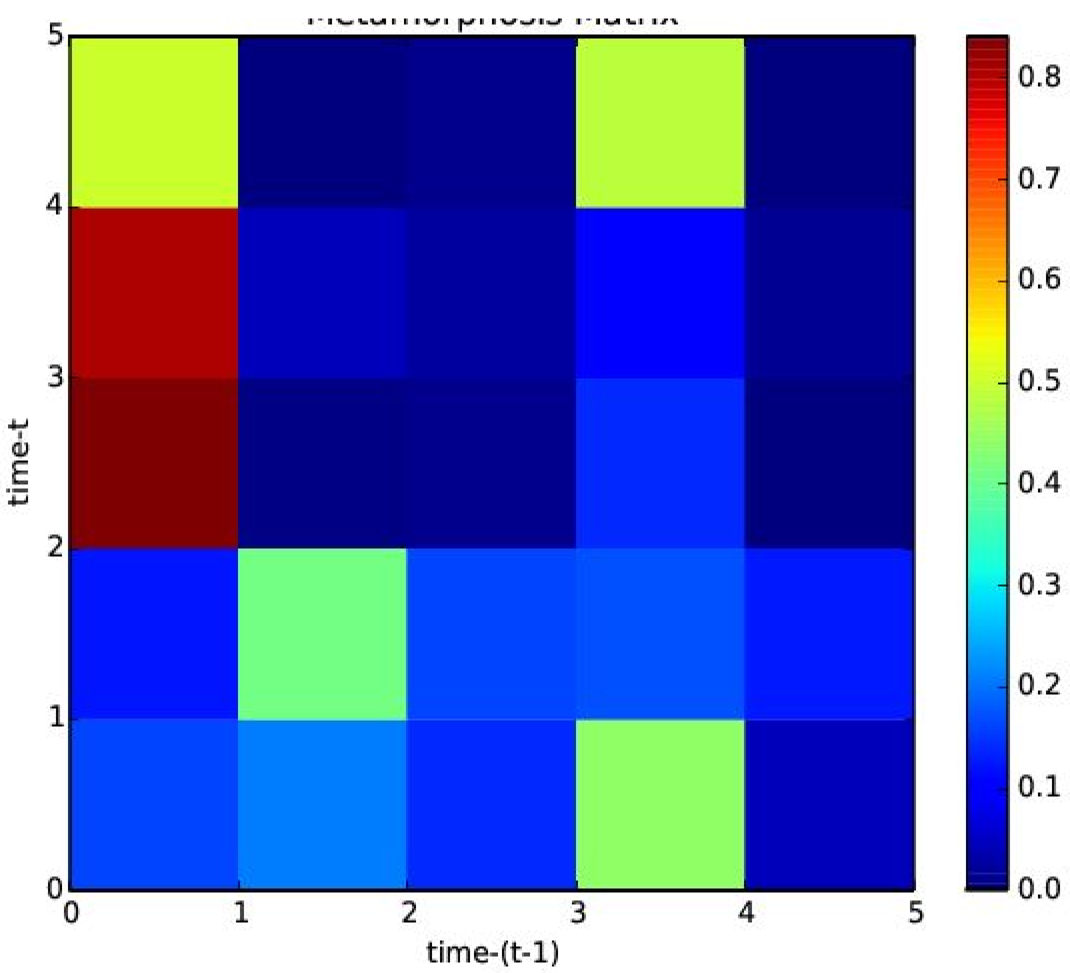
$$||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{Q}^t||_F^2 + ||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{T}^t\mathbf{Q}^{t-1}||_F^2$$

- Tracking matrix is a square matrix of all positive entries

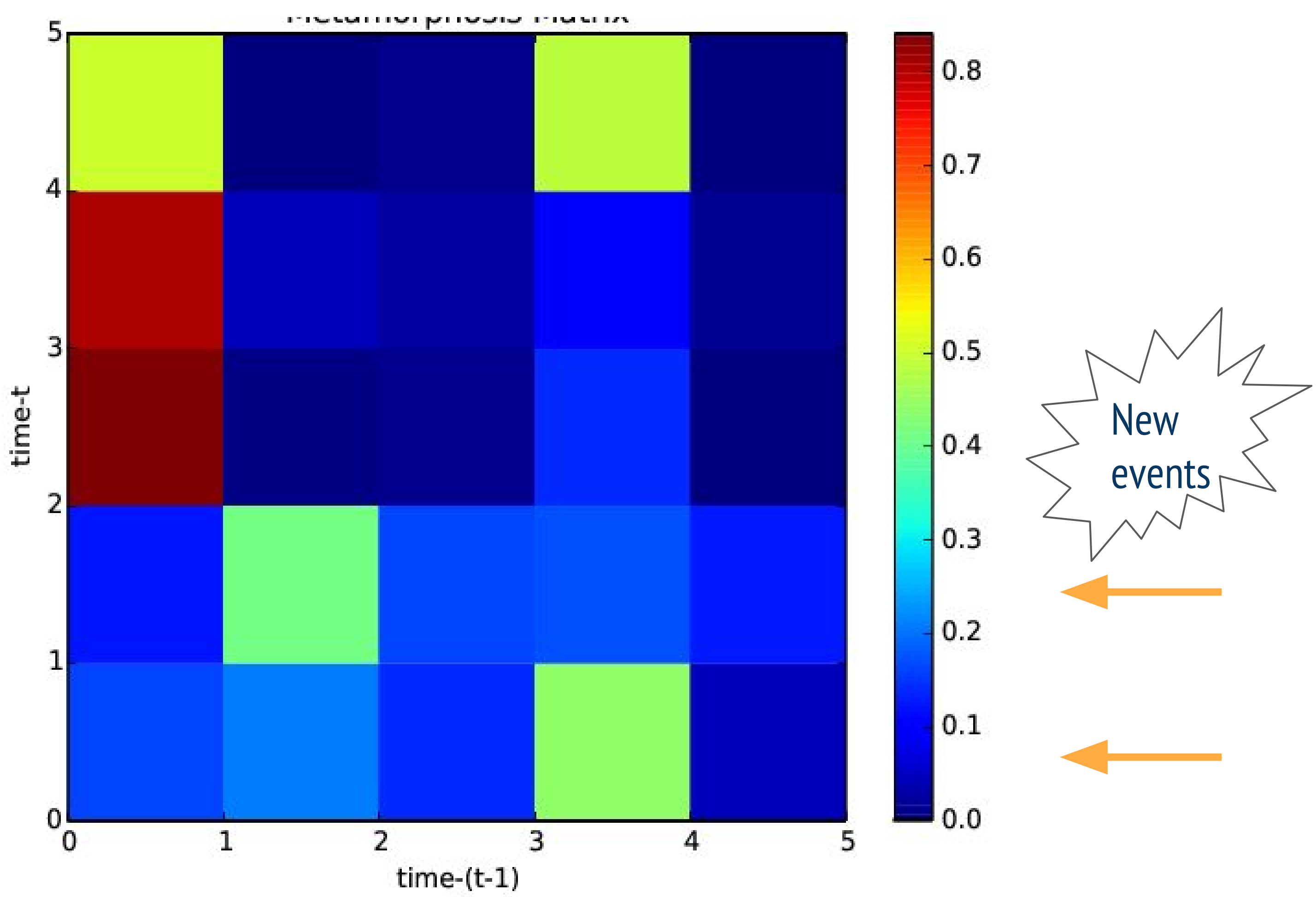
Tracking Matrix

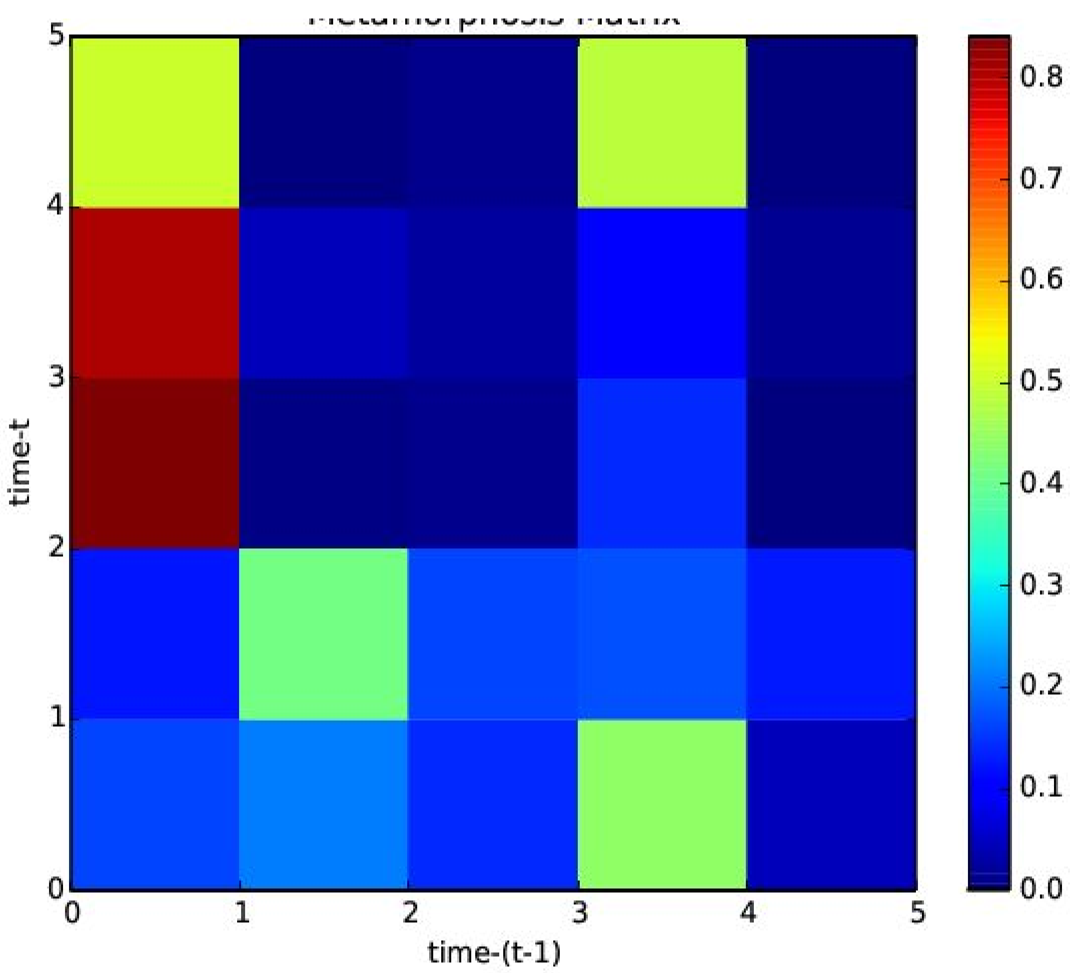
$$||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{Q}^t||_F^2 + ||\mathbf{K}^t - \mathbf{Q}^{t^T}\mathbf{T}^t\mathbf{Q}^{t-1}||_F^2$$

- Tracking matrix is a square matrix of all positive entries
- It show how the topics have changed from $t-1$ to t
- Row-i tells how topic-i at time t is related to all the topics at time $t-1$

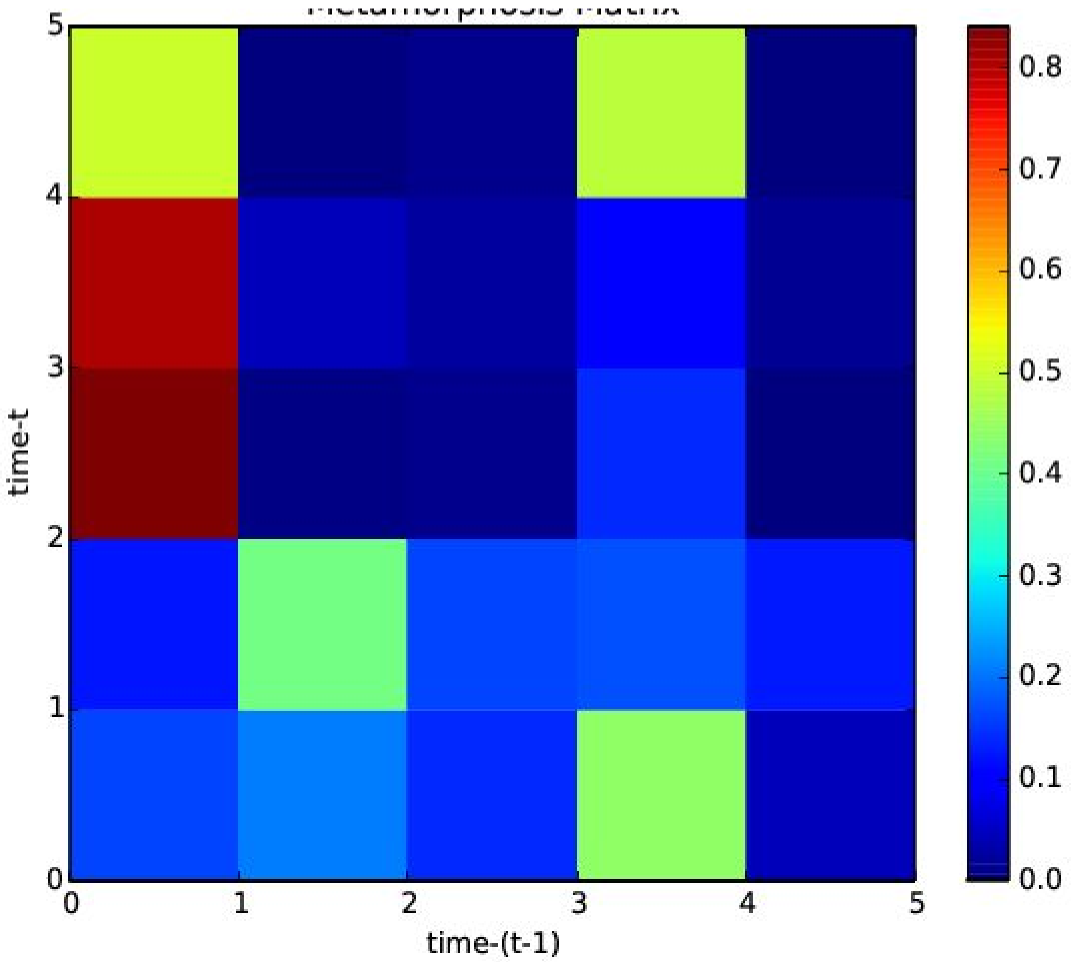


time-t
time-(t-1)
0
1
2
3
4
5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


time-t
time-(t-1)
0
1
2
3
4
5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
New
events


Metamorphosis Matrix
time-t
time-(t-1)
0 1 2 3 4 5
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8


Continuing events



Evolving events

Entropy H(X)

$$H(X) := -P(x_i) \sum_{i=1}^{n} \log(P(x_i))$$

- Quantifies the amount of “randomness”
- Range [0, 2.32]

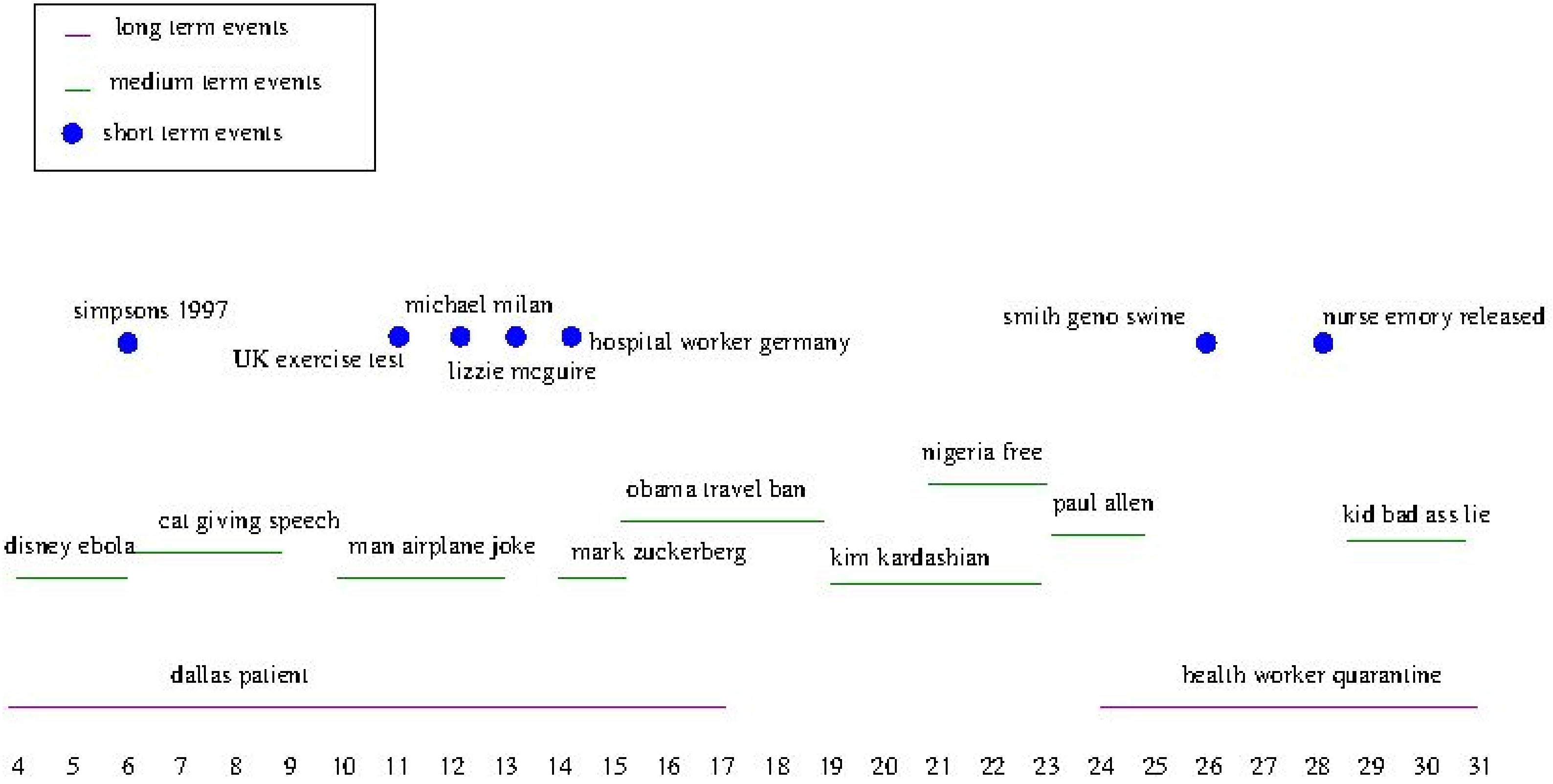
Based on Entropy

- $< 1 \longrightarrow$ continuing events
- $1 <= H(X) <= 2 \longrightarrow$ evolving events
- $> 2 \longrightarrow$ new events
- Also ending events

Timeline Generation

- Look at heatmap
- Connect dots to the previous timesteps (if possible)
- Else, new events, or noise

Timeline



long term events
medium term events
short term events
simpsons 1997
UK exercise test
michael milan
lizzie mcguire
hospital worker germany
smith geno swine
nurse emory released
nigeria free
obama travel ban
paul allen
kid bad ass lie
cat giving speech
disney ebola
man airplane joke
mark zuckerberg
kim kardashian
dallas patient
health worker quarantine
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Example continuing events - memes

2014/10/19	kim, kardashian, married, american, died
2014/10/20	kim, kardashian, married, american, died
2014/10/21	kim, kardashian, married, american, died
2014/10/22	kim, kardashian, married, american, died
2014/10/23	kim, kardashian, married, american, died

Example of evolving events

2014/10/07	kidney, dialysis
2014/10/08	thomas, eric, duncan, died, first, patient
2014/10/09	died, patient
2014/10/10	duncan, fever, nurse
2014/10/11	nurse, symptoms
2014/10/12	health, care, worker, positive
2014/10/13	health, care, worker, protocol
2014/10/14	nurse, dallas, nina, pham
2014/10/15	health, care, worker, 2nd, positive
2014/10/16	nurse, flight, ohio
2014/10/17	virus, flight, nina, pham

Thank you!

Questions?

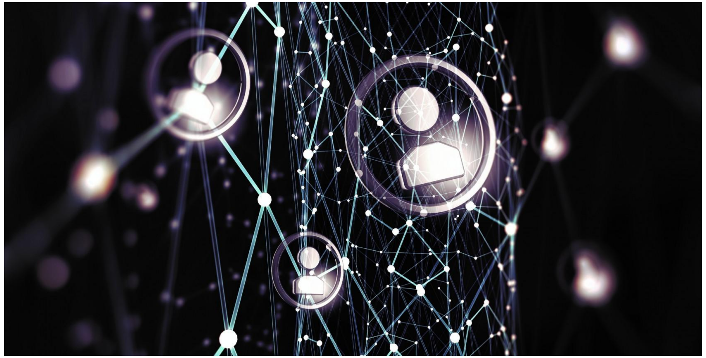


Photo credit: http://www.realityblurred.com/realitytv/images/2015/03/mtv-the-real-world.jpg AND http://globalpse.org/en/wp-content/uploads/2016/04/Social-media-e1459493030525.jpg

SUPER FANTASTIC

Social Media Era