Temporal models of streaming Social Media data

Daniel Preotiuc-Pietro
Supervisor: Trevor Cohn

10.03.2014

Context

- vast increase in user generated content
- Online Social Networks

  most time-consuming activity
- multiple modalities: text, time, location, user info, images, etc.
- social network structure
- Challenges:
  - Engeneering: data volume
  - Algorithmic: restricted information, grounded in context, streaming, noise

Motivation

- SM data allows to study fine grained time
- Effect of time usually ignored in NLP, with few exceptions and on historical corpora

  sequence models for word sequences

  smoothly varying parameters in topic models & text regression
- Supervised forecasting applications

  internal, external
- Unsupervised methods based on underlying temporal effects

Aims

i. Social Media text is time dependent.
ii. Modelling the temporal dimension is beneficial for a better understanding of real world effects.
iii. Modelling time is useful in downstream applications.
iv. Replicable & Portable methods

   independent of language and external resources.

Online Social Networks

Social Networks are based on sharing a piece of generated content with your social network

Microblogs
Short text (140 char.)

Location Based Social Networks
Check-in (venue oriented)

Data collection:

- using public APIs
- datasets:
    - general (Gardenhose - 10% Twitter – 15 Tb total)
    - focused on a set of users (e.g. 20k freq. Foursquare users)
    - focused on locations (e.g. UK, Austria)

Text Processing

new conventions

lack of context

creative spellings

RT @MediaScotland greeeat!!!lvly speech by cameron on scott's indy :) #indyref

shortenings

unorthodox capitalisation

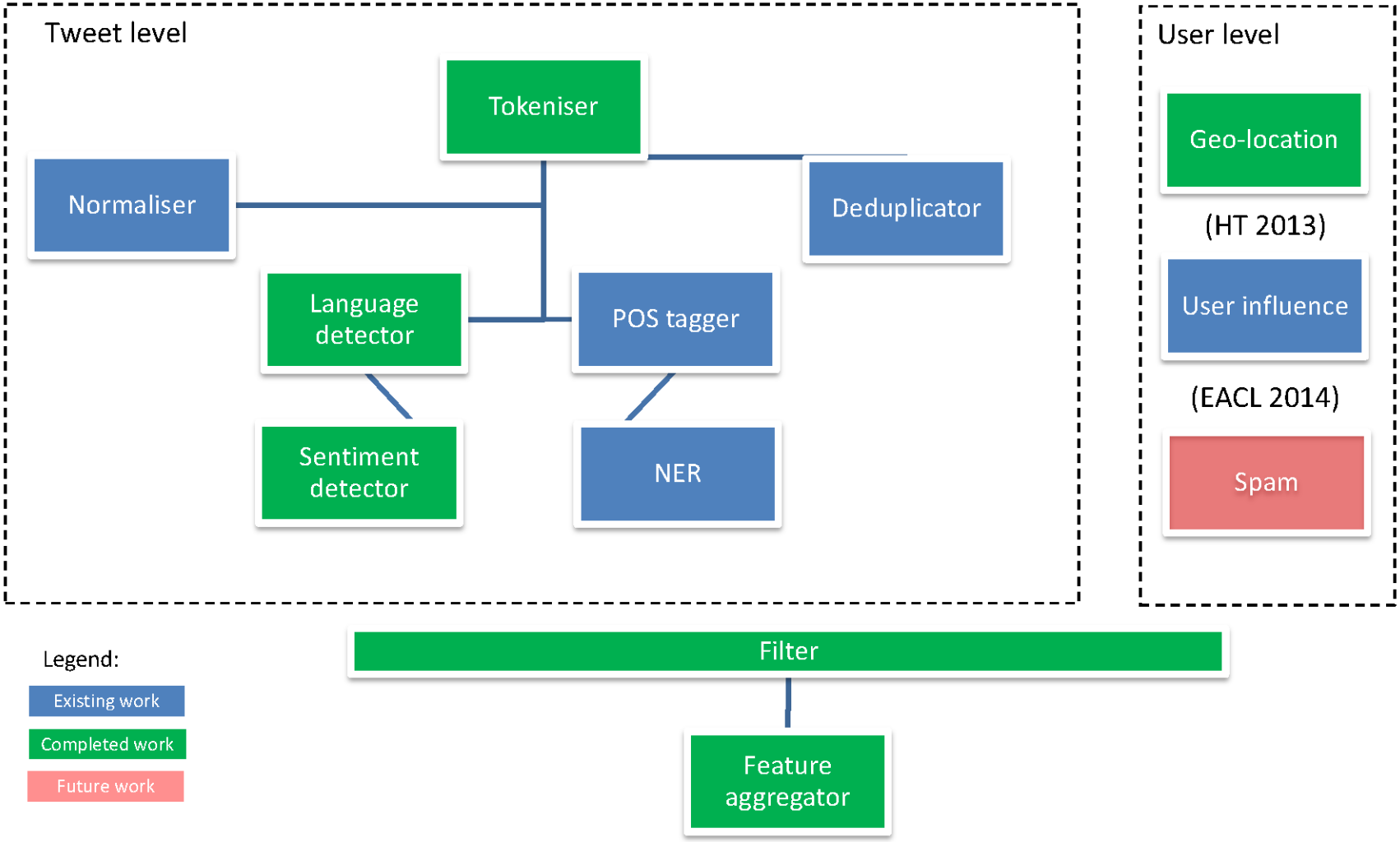
OOV words

Processing Architecture

- Fast: real time processing, Hadoop MapReduce (I/O bound), online and batch processing
- Scalable: adding more machines
- Modular: easy to add new modules
- Pipeline: the user specifies his needs
- Extensible: different sources of data (USMF format)
- Data consistency: JSON format, append to ‘analysis’
- Reusable: open-source

(ICWSM 2012)

Components


Tweet level
Tokeniser
Normaliser
Deduplicator
Language detector
POS tagger
Sentiment detector
NER
User level
Geo-location
(HT 2013)
User influence
(EACL 2014)
Spam
Filter
Feature aggregator
Legend:
Existing work
Completed work
Future work

Text based forecasting

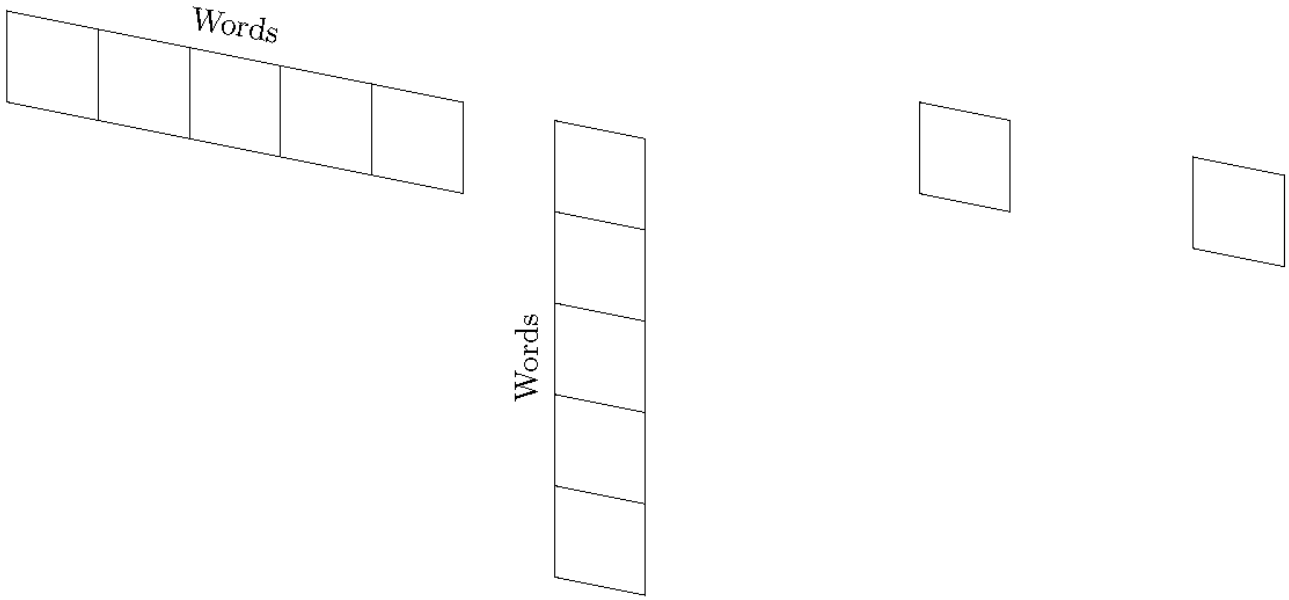
Task: predicting real world outcomes

Aim: replace expensive polls with social media

- predict political voting intention (not elections!)
- based on social media (Twitter) text
- strong baselines (last day, mean)
- 2 different use cases (U.K. and Austria)
- U.K. 42k users, 60m tweets, 3 parties, 2 years

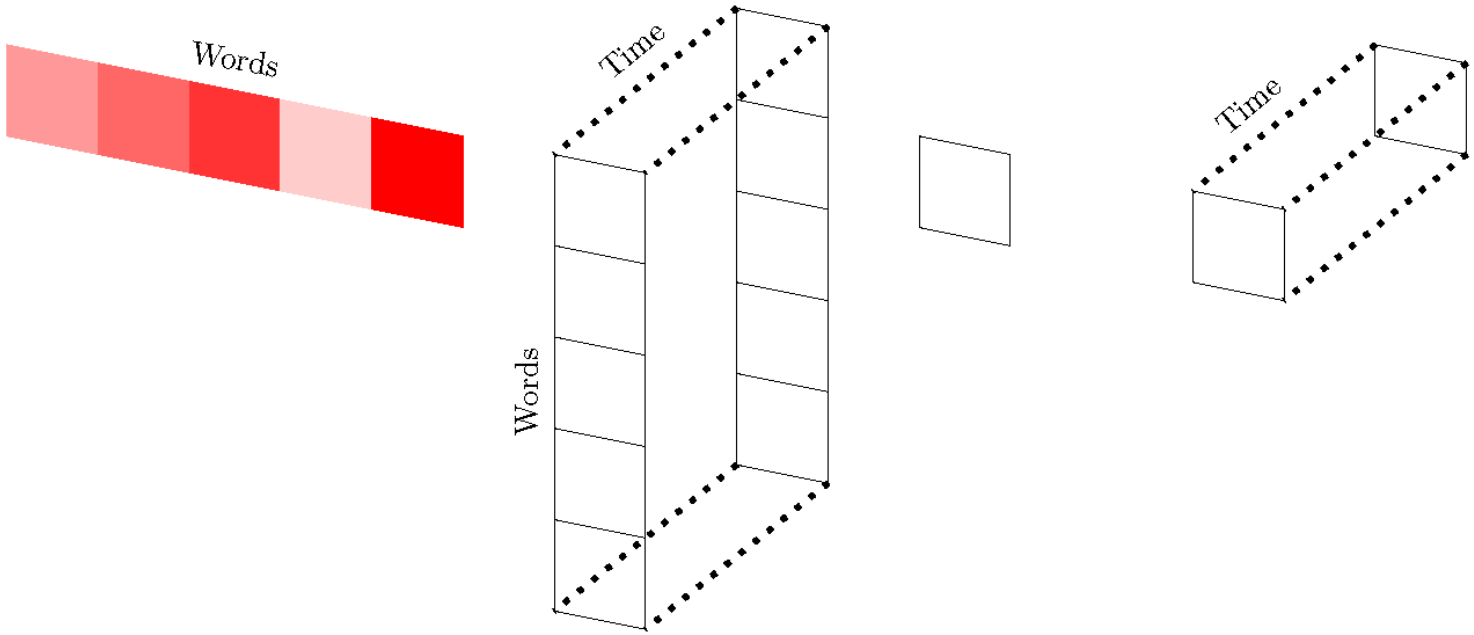
(ACL 2013)

Linear regression



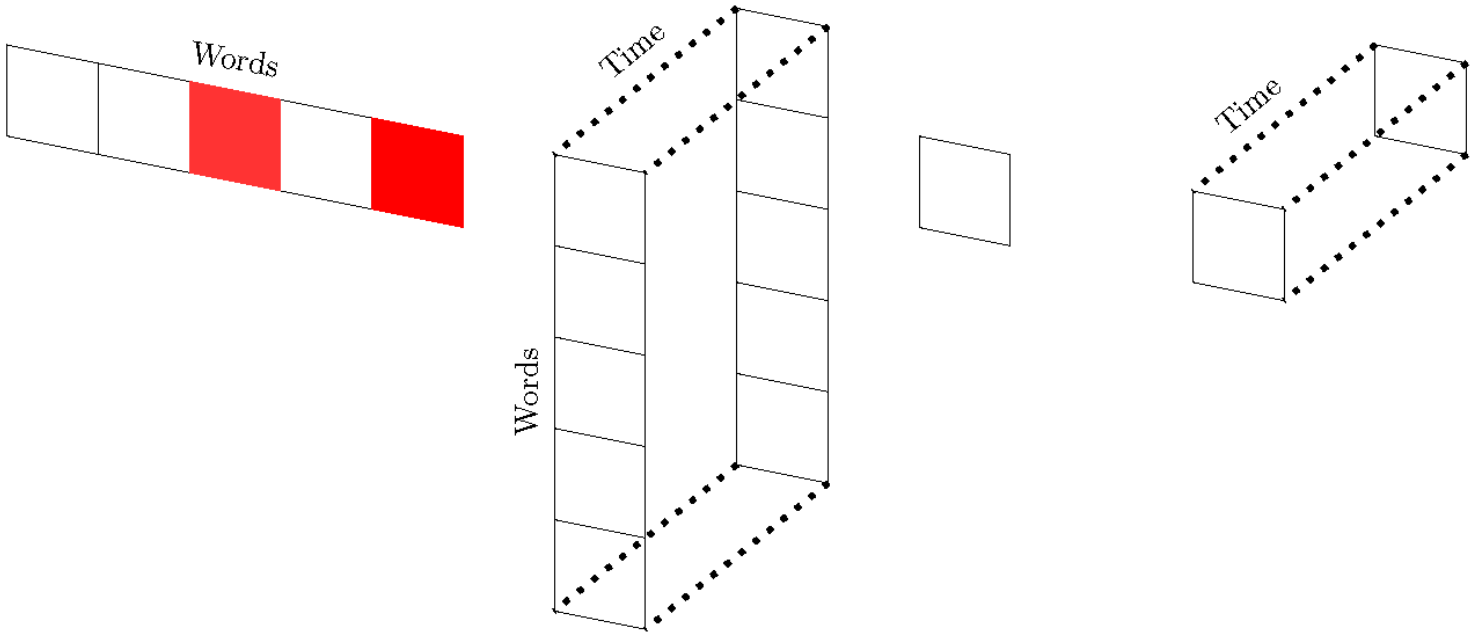
$$w \quad x_t \quad + \quad \beta \quad = \quad y_t$$

Linear regression



$$\{\mathrm{w}, \beta\} = \mathrm{argmin} \sum_{i=1}^{n} (wx_i + \beta - y_i)^2$$

Linear regression



$$\{\mathrm{w}, \beta\} = \mathrm{argmin} \sum_{i=1}^{n} (wx_i + \beta - y_i)^2 + \psi_{el}(w, \rho)$$

LEN – Elastic Net

Bilinear regression

- main issue is noise:

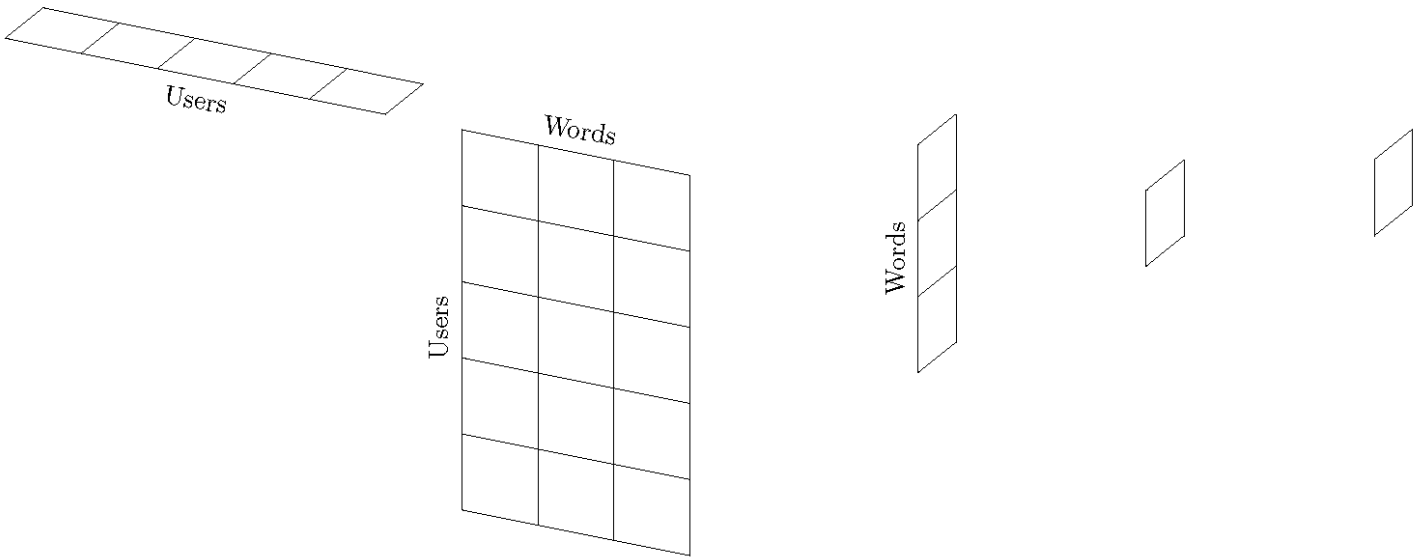
  many non-informative users

- we look for a model of

  sparse words & sparse users

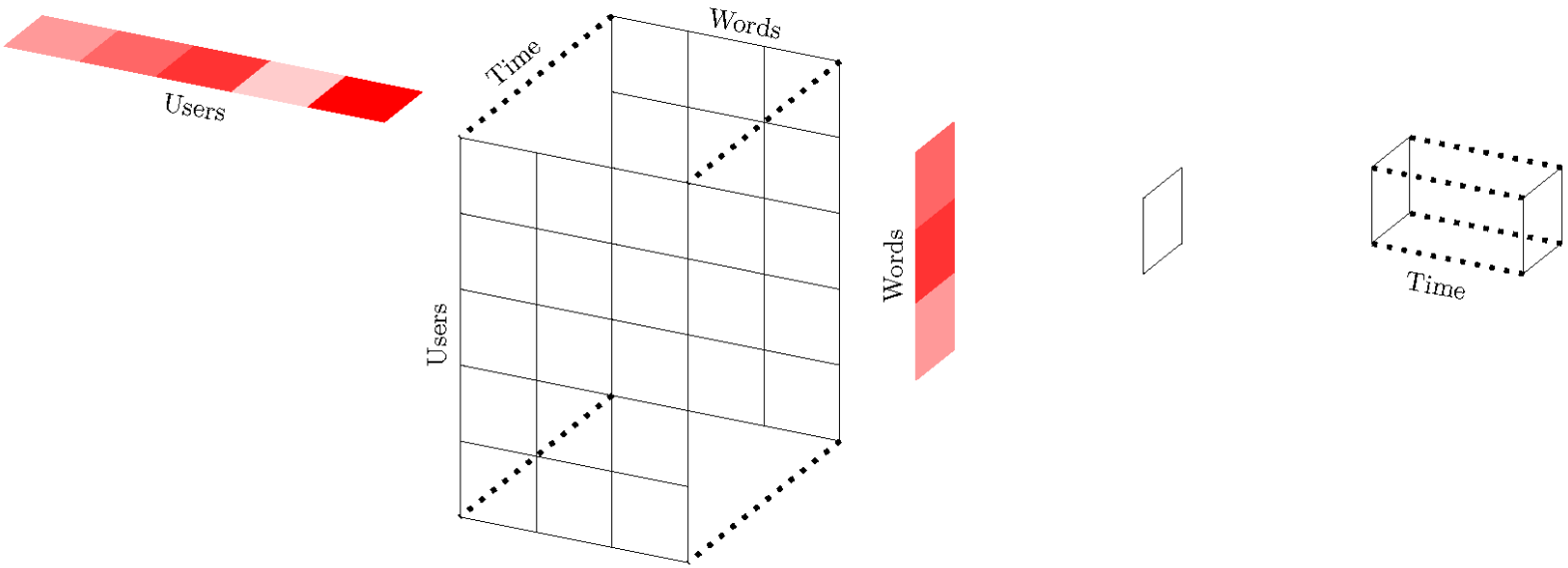
- bi-convex optimisation problem
- solved by alternatively fixing each set of weights and iterating until convergence

Bilinear regression



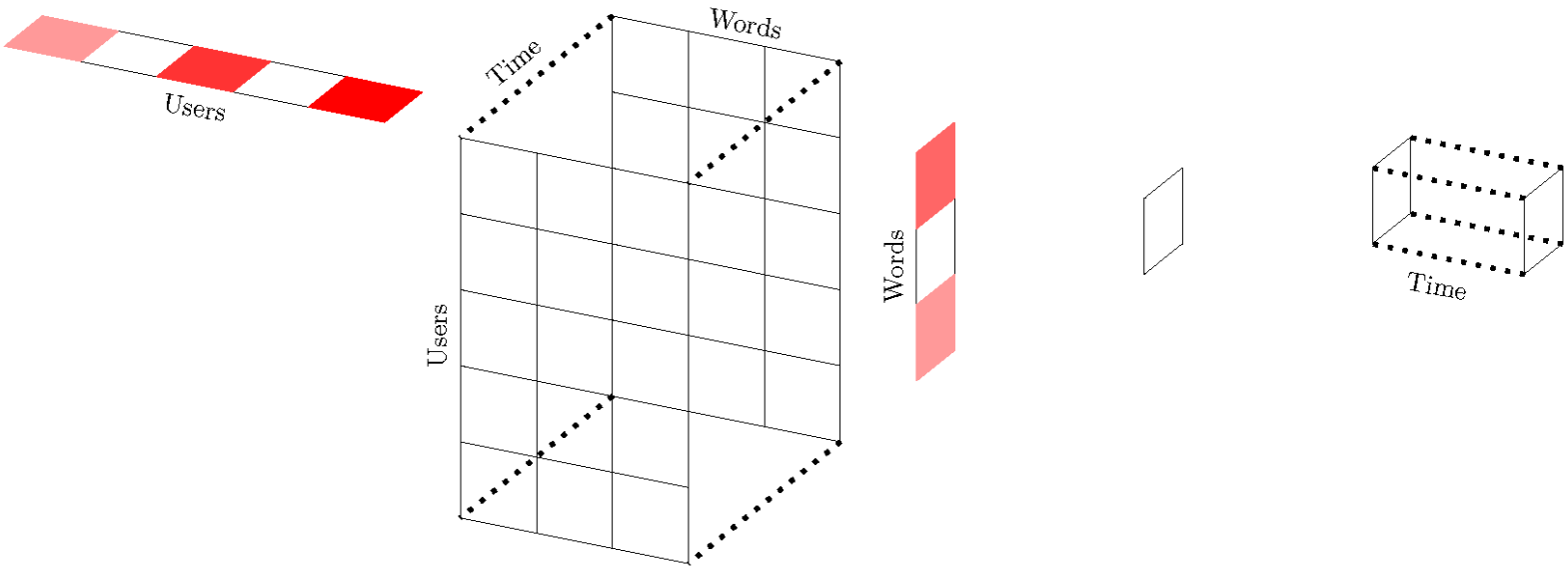
$$u \qquad X_t \qquad w^T + \beta = y_t$$

Bilinear regression



$$\{\mathrm{w}, \mathrm{u}, \beta\} = \mathrm{argmin} \sum_{i=1}^{n} (uX_i w^T + \beta - y_i)^2$$

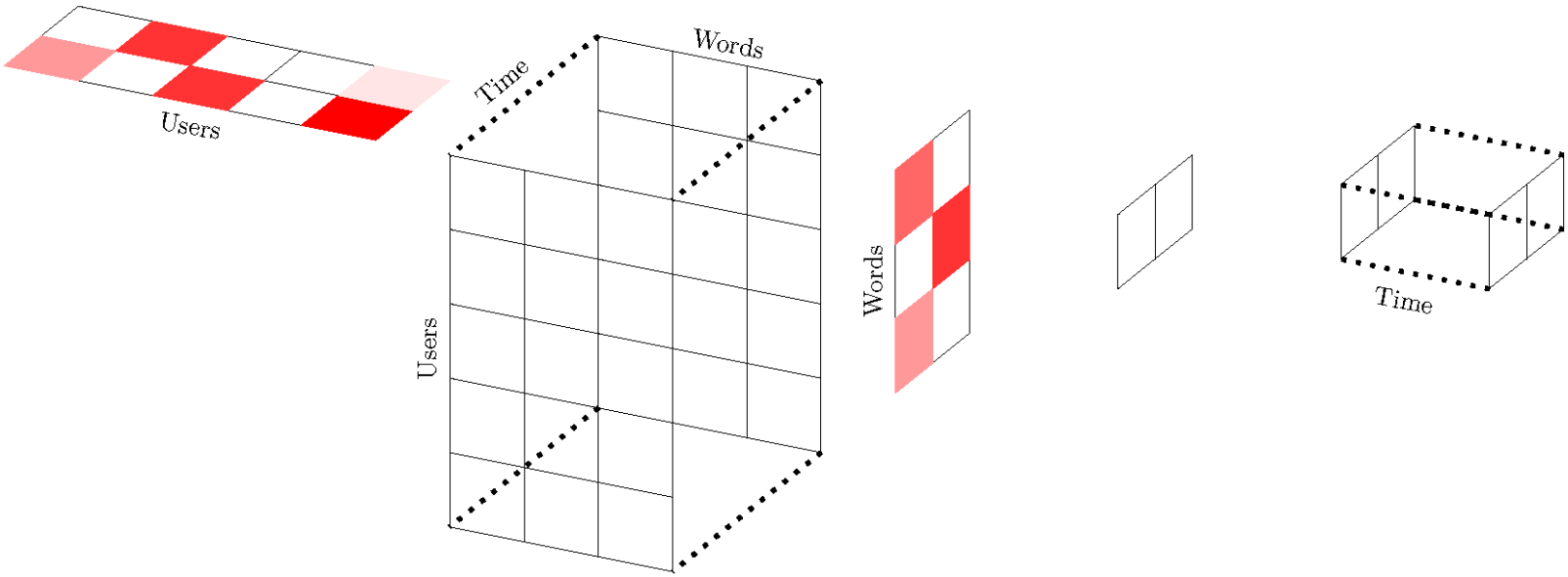
Bilinear regression



$$\{\mathrm{w}, \mathrm{u}, \beta\} = \mathrm{argmin} \sum_{i=1}^{n} (uX_i w^T + \beta - y_i)^2 + \psi_{el}(w, \rho_1) + \psi_{el}(u, \rho_2)$$

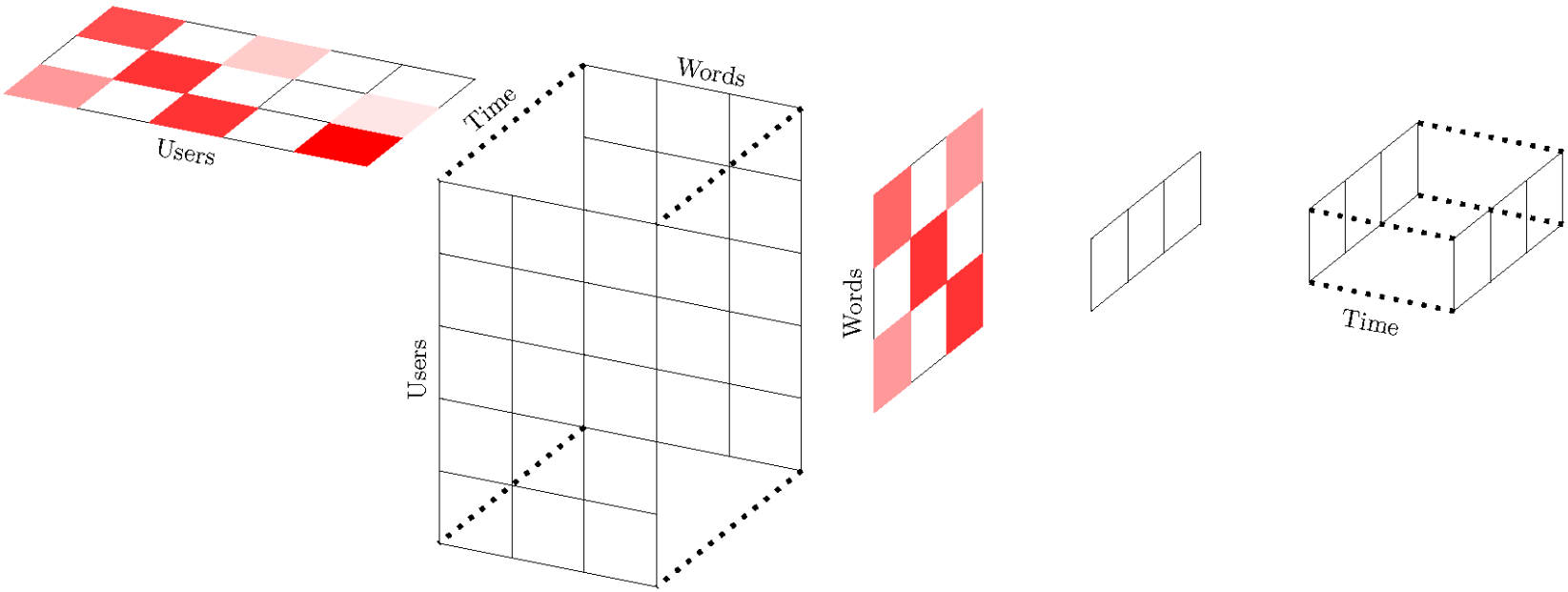
BEN – Bilinear Elastic Net

Bilinear regression



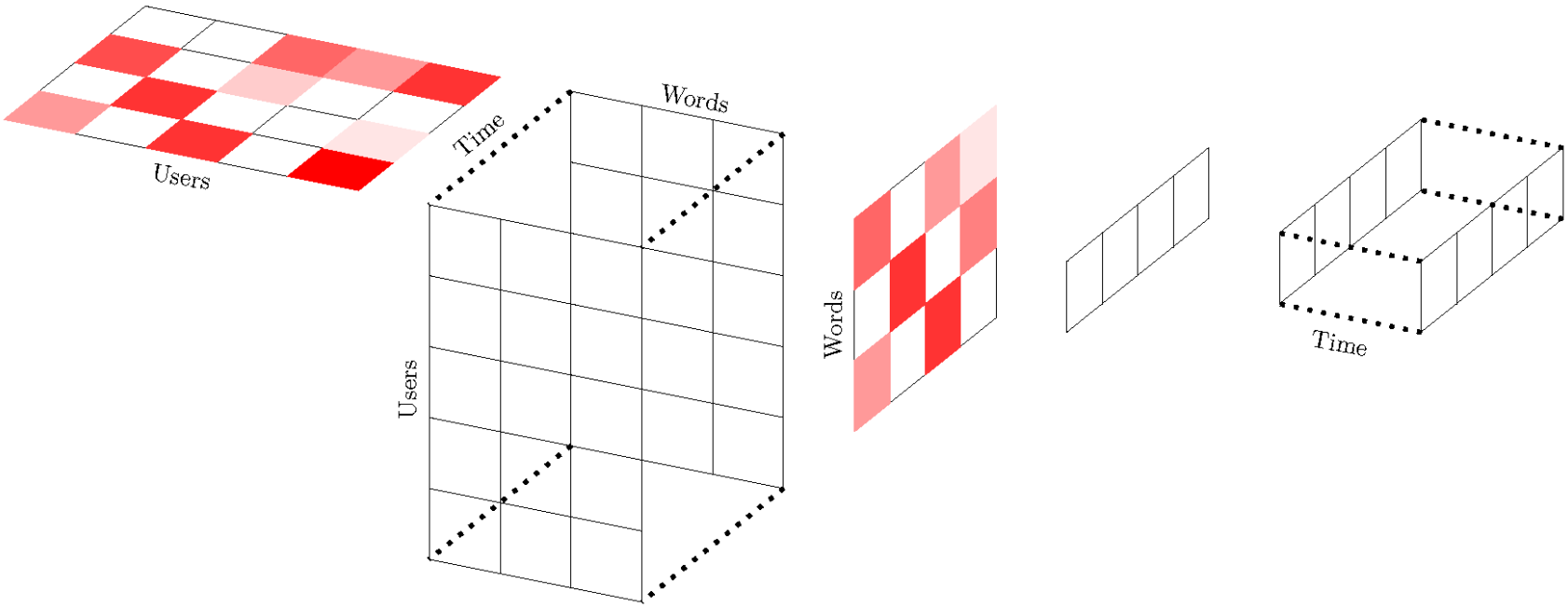
$$\{w_t, u_t, \beta\} = \text{argmin} \sum_{i=1}^{n} (u_t X_i w_t + \beta - y_{ti})^2 + \psi_{el}(w_t, \rho_1) + \psi_{el}(u_t, \rho_2)$$

Bilinear regression



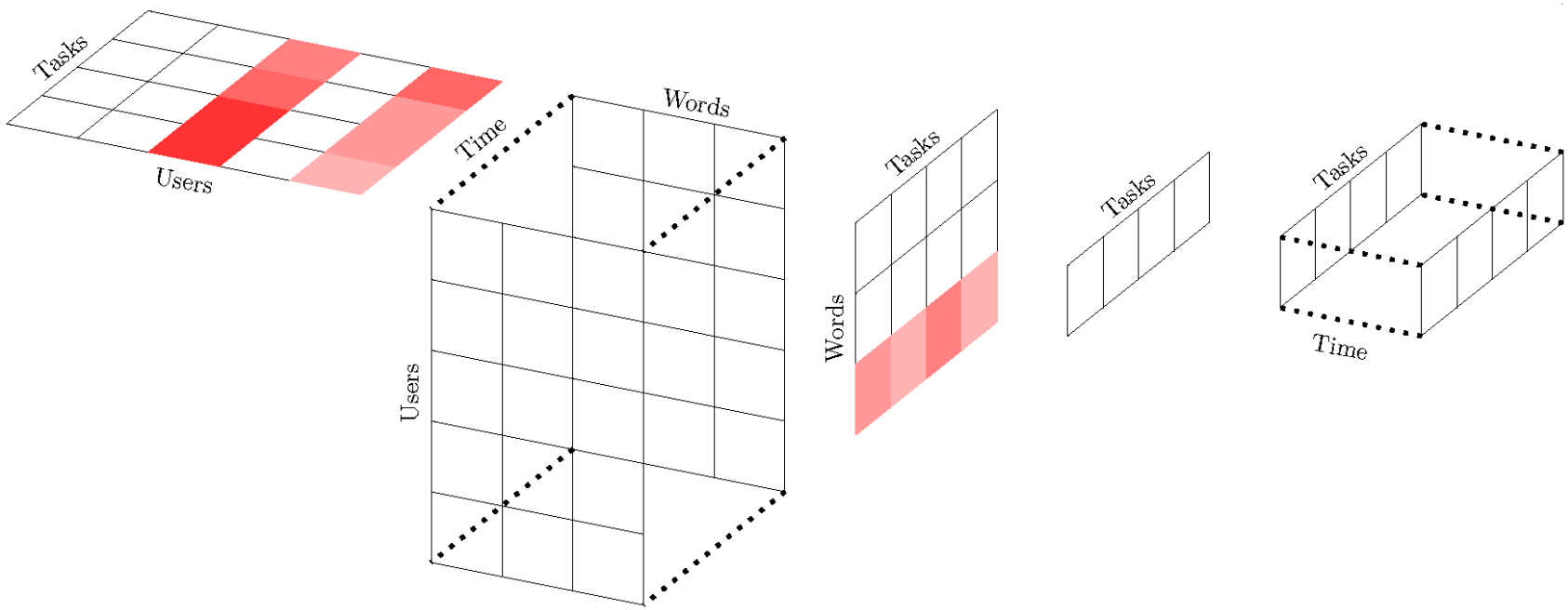
$$\{w_t, u_t, \beta\} = \text{argmin} \sum_{i=1}^{n} (u_t X_i w_t + \beta - y_{ti})^2 + \psi_{el}(w_t, \rho_1) + \psi_{el}(u_t, \rho_2)$$

Bilinear regression



$$\{w_t, u_t, \beta\} = \text{argmin} \sum_{i=1}^{n} (u_t X_i w_t + \beta - y_{ti})^2 + \psi_{el}(w_t, \rho_1) + \psi_{el}(u_t, \rho_2)$$

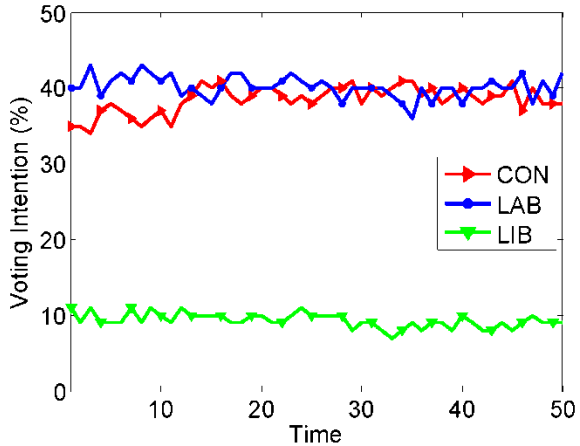
Bilinear regression



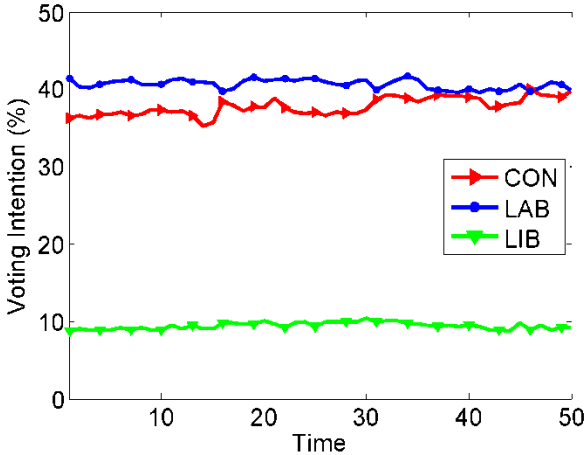
$$\{\mathrm{w}, \mathrm{u}, \beta\} = \operatorname{argmin} \sum_{t=1}^{\tau} \sum_{i=1}^{n} (u_t X_i w_t + \beta - y_{ti})^2 + \psi_{l_1 l_2}(w, \rho_1) + \psi_{l_1 l_2}(u, \rho_2)$$

BGL – Bilinear Group LASSO

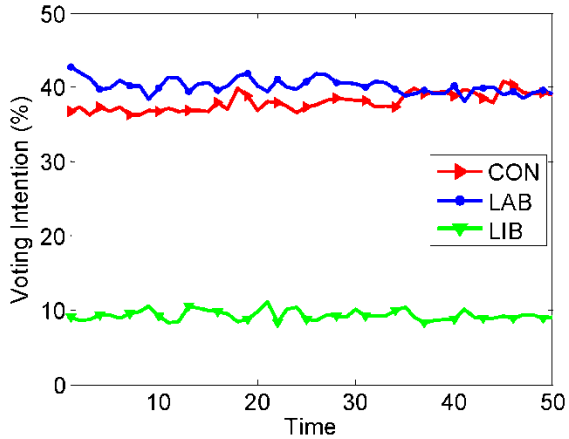
Results



Ground truth



BEN



BGL

	CON	LAB	LBD	μ
B_μ	2.272	1.663	1.136	1.69
B_{last}	2	2.074	1.095	1.723
LEN	3.845	2.912	2.445	3.067
BEN	1.939	1.644	1.136	1.573
BGL	**1.785**	**1.595**	**1.054**	**1.478**

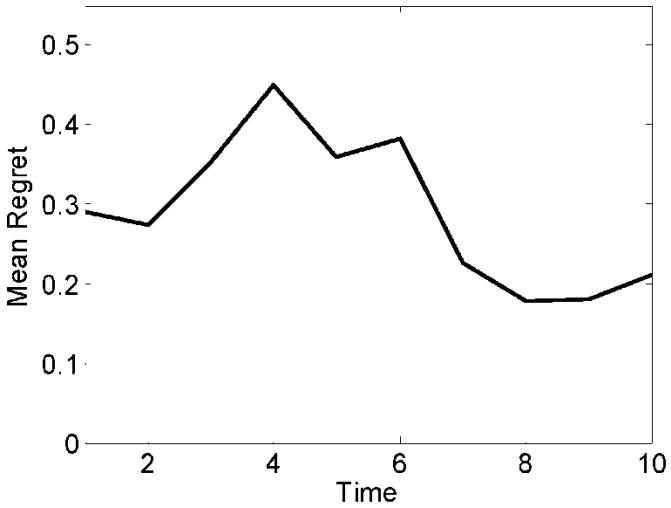
Qualitative analysis

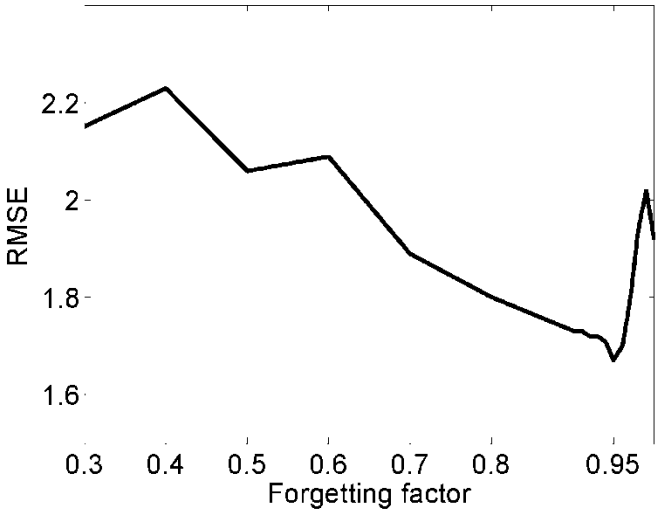
Party	Tweet	Score	Author
CON	PM in friendly chat with top EU mate, Sweden's Fredrik Reinfeldt, before family photo	1.334	Journalist
	Have Liberal Democrats broken electoral rules? Blog on Labour complaint to cabinet secretary	-0.991	Journalist
LAB	Blog Post Liverpool: City of Radicals Website now Live <link> #liverpool #art	1.954	Art Fanzine
	I am so pleased to head Paul Savage who worked for the Labour group has been Appointed the Marketing manager for the baths hall GREAT NEWS	-0.552	Politicial (Labour)
LBD	RT @user: Must be awful for TV bosses to keep getting knocked back by all the women they ask to host election night (via @user)	0.874	LibDem MP
	Blog Post Liverpool: City of Radicals 2011 – More Details Announced #liverpool #art	-0.521	Art Fanzine

Online learning

One-pass online learning algorithm:

- more realistic setup
- Stochastic Gradient Descent with proximal steps
- results are worse, but comparable
- `forgetting factor' incorporates temporal smoothing: new data is more relevant than old data





Gaussian Processes

Task: Forecast hashtag frequency in Social Media

- identify and categorise complex temporal patterns

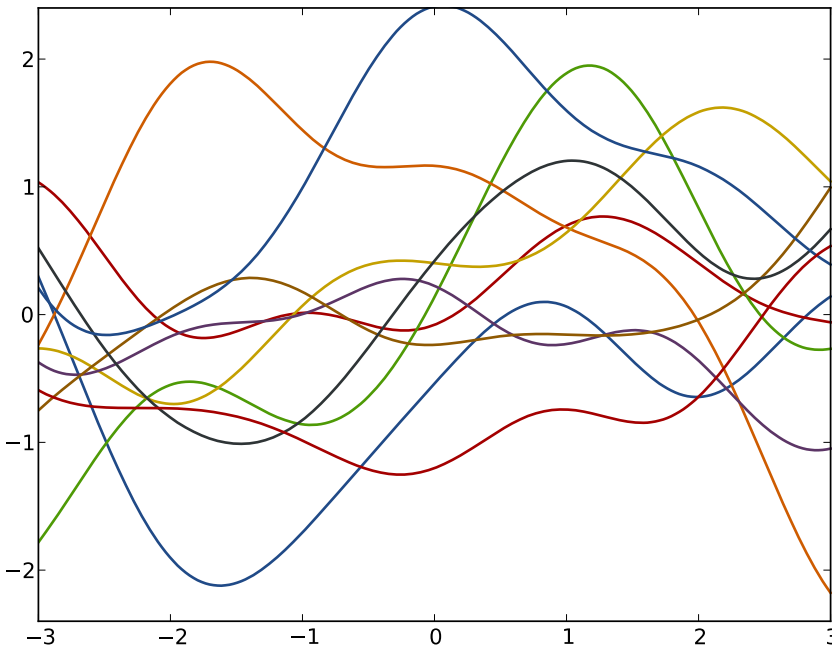
(EMNLP 2013)

Non-parametric Bayesian framework

- kernelised
- probabilistic formulation
- propagation of uncertainty
- exact posterior inference for regression
- Non-parametric extension of Bayesian regression
- very good results, but hardly used in NLP

Gaussian Processes

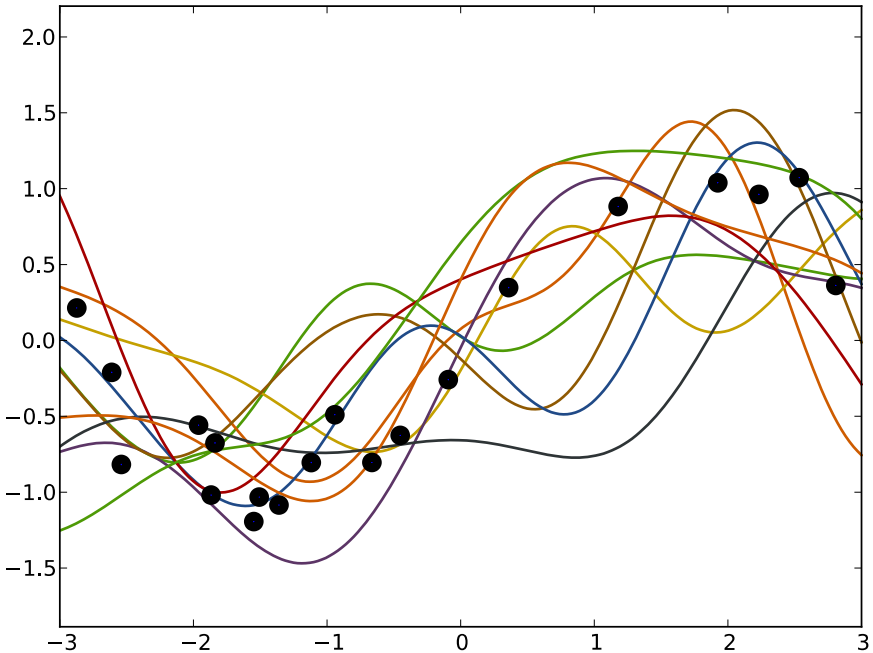
Define prior over functions



$$p(f)$$

Compute posterior

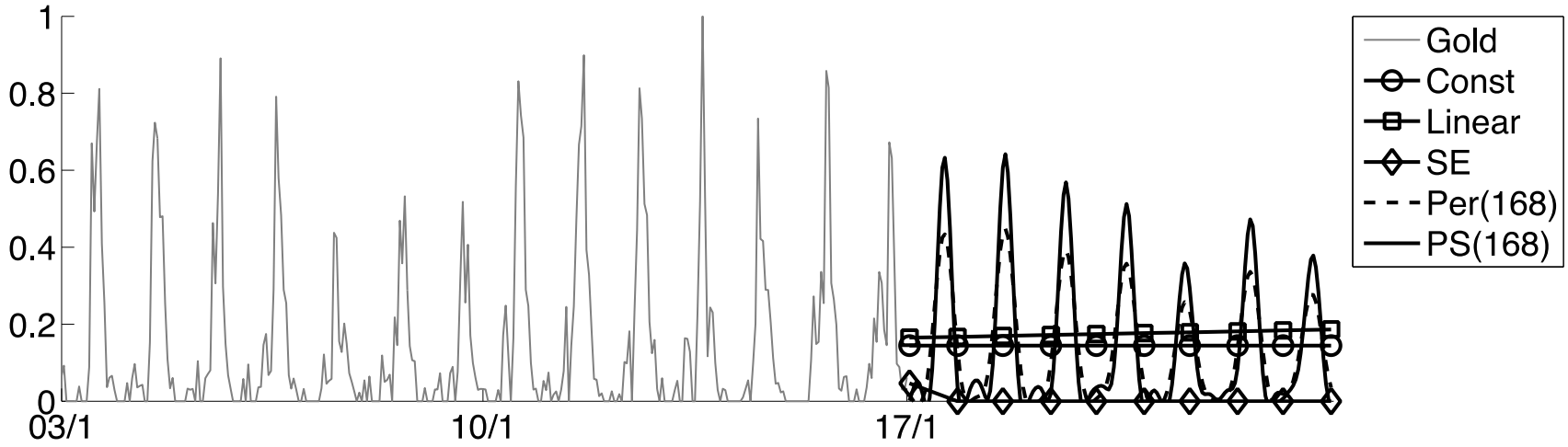
$$p(f|X, \mathbf{t})$$



GP Kernel

- Defines the covariance between two points
  i. constant
  ii. SE (aka RBF): smoothly varying outputs
  iii. PER: smooth periodic
  iv. PS: spiking periodic
- Select the model (kernel) with highest marginal likelihood
  - Bayesian model selection
  - balances data fit with model capacity
- automatically identifies the period (if exists)
- allows learning of different flavours of temporal phenomena

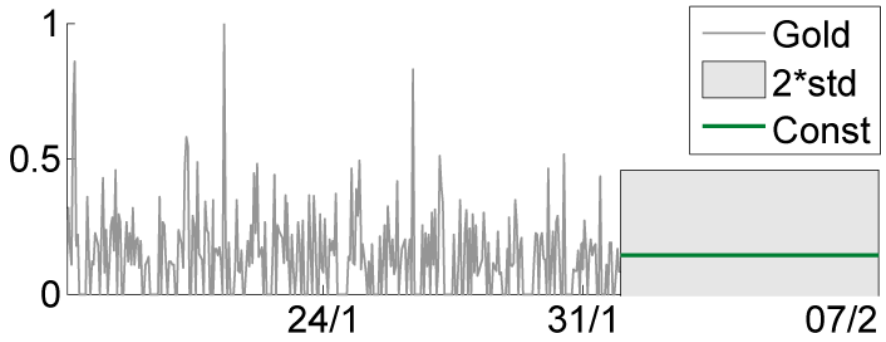
Extrapolation



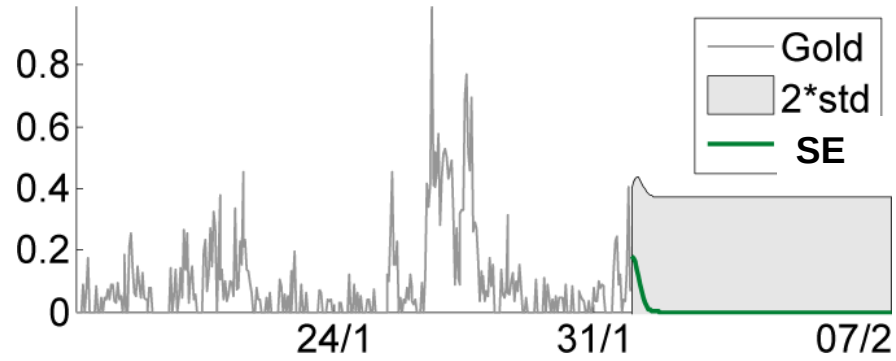
	Const	Lin	SE	PER	PS
NLML	-41	-34	-176	-180	-192
NRMSE	0.213	0.214	0.262	0.119	0.107

Examples of time series

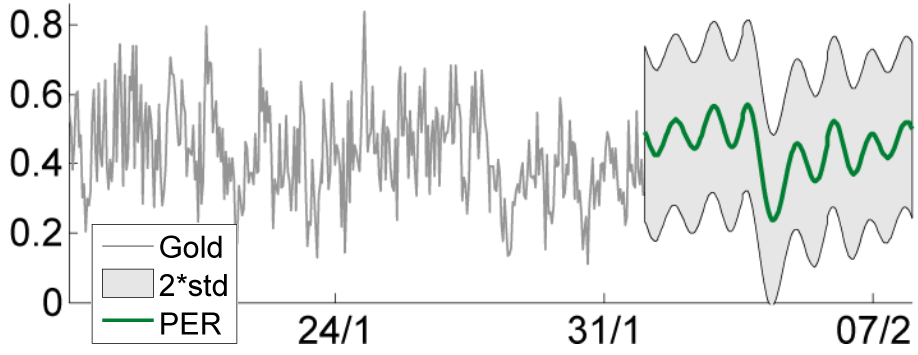
#FYI



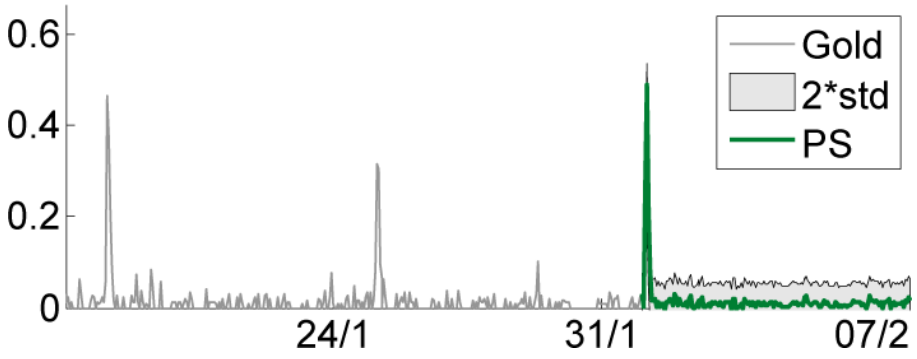
#SNOW



#FAIL



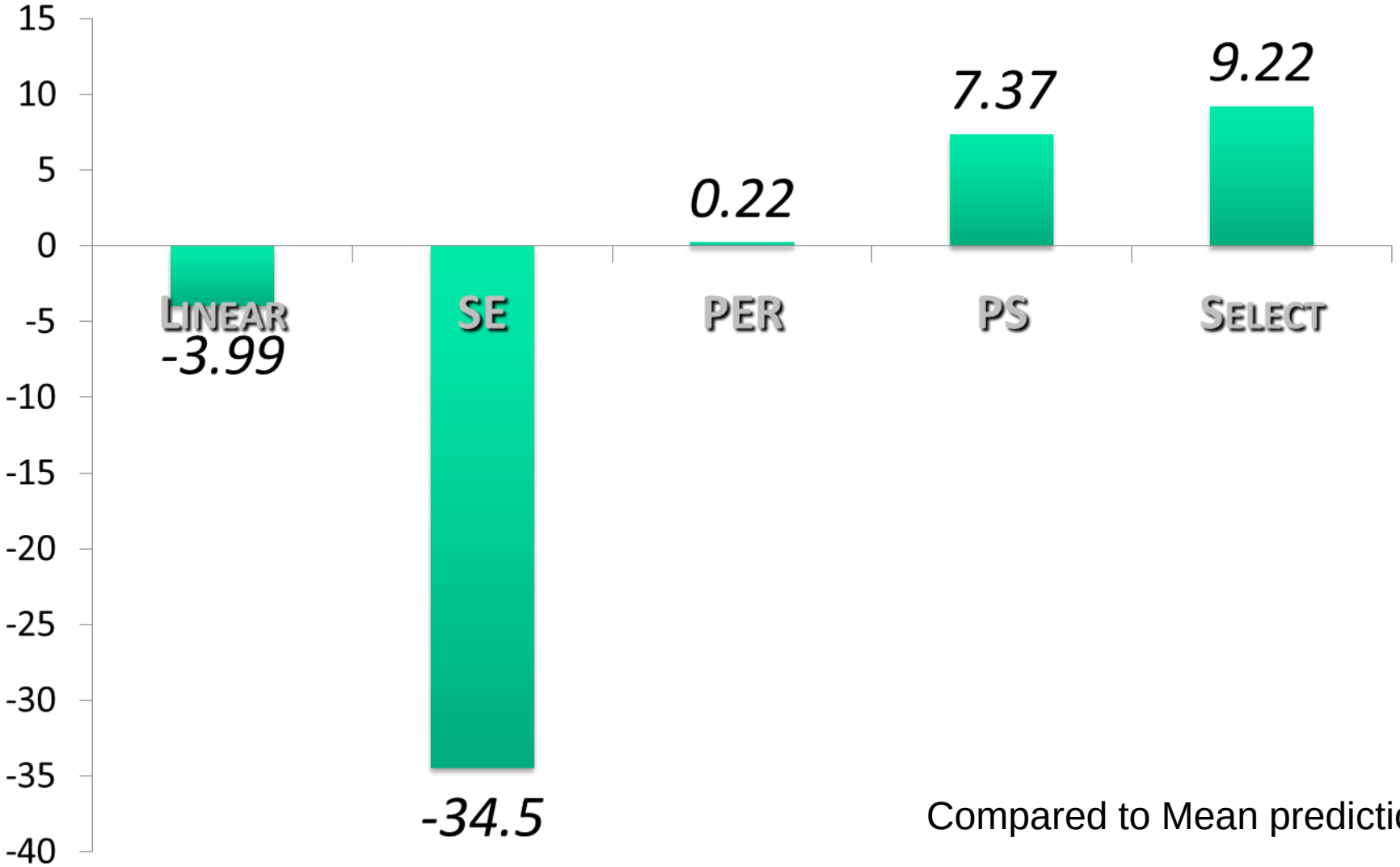
#RAW



Experimental results

Const	SE	PER	PS
#funny	#2011	#brb	#ff
#lego	#backintheday	#coffee	#followfriday
#likeaboss	#confessionhour	#facebook	#goodnight
#money	#februarywish	#facepalm	#jobs
#nbd	#haiti	#funny	#news
#nf	#makeachange	#love	#nowplaying
#notetoself	#questionsidontlike	#rock	#tgif
#priorities	#savelibraries	#running	#twitterafterdark
#social	#snow	#xbox	#twitteroff
#true	#snowday	#youtube	#ww
49	**268**	**493**	**366**

Experimental results



Compared to Mean prediction

Text classification

Task: Assign the hashtag to a given tweet

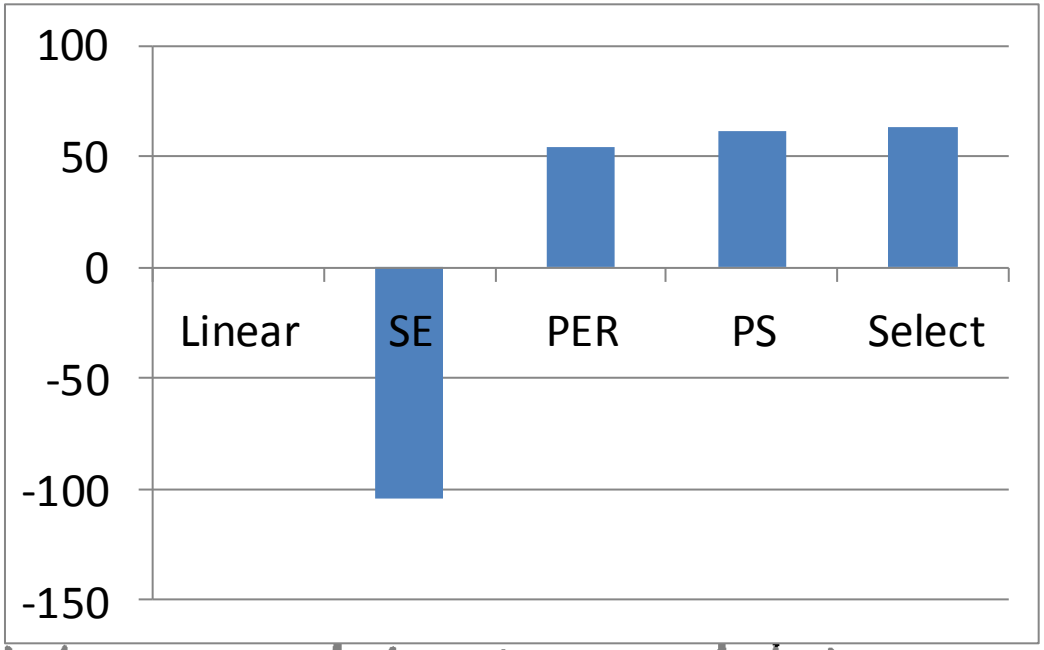
- Most frequent (MF)
- Naive Bayes model (NB-E)
- Naive Bayes with GP forecast as prior (NB-P)

	MF	**NB-E**	**NB-P**
Match@1	7.28%	16.04%	17.39%
Match@5	19.90%	29.51%	31.91%
Match@50	44.92%	59.17%	60.85%
MRR	0.144	0.237	0.252

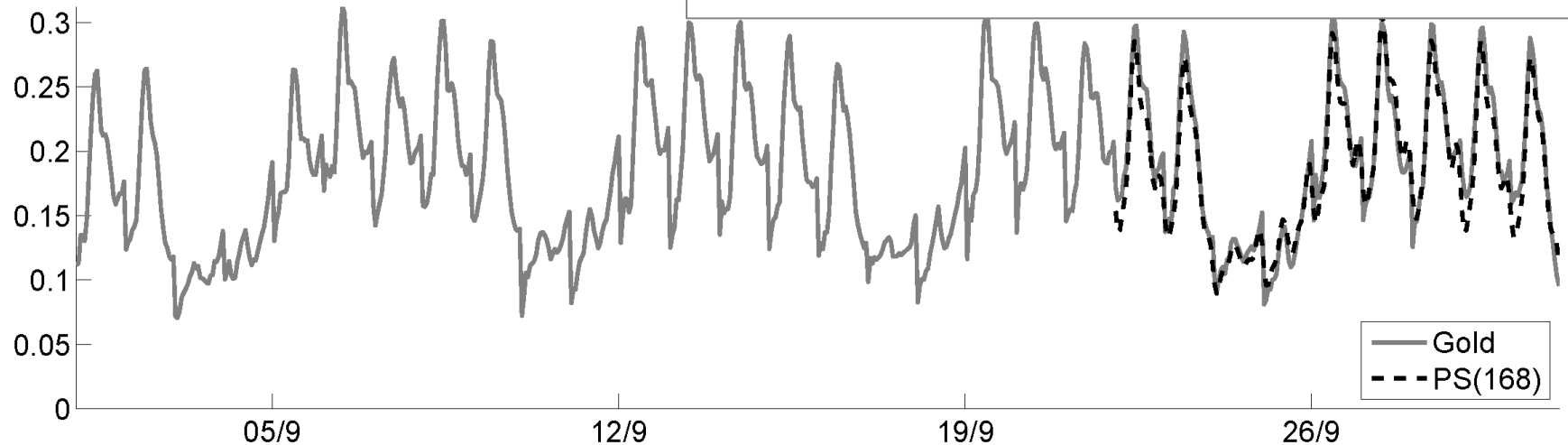
User behaviour

Task: Predict venue check-in frequencies

- Modelled using GPs
- Compared to Mean



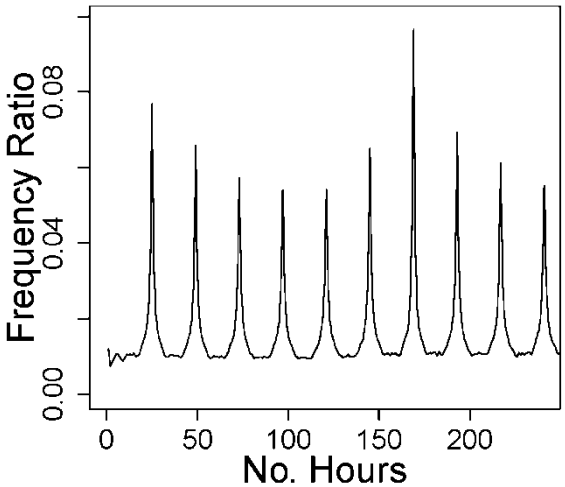
Professional Venues



Individual user behaviour

Task: Predict venue type of user check-in

- highly periodic
- compared to standard Markov predictors



Method	Accuracy
Random	11.11%
M.Freq Categ.	35.21%
Markov-1	36.13%
Markov-2	34.21%
Daily period	38.92%
Weekly period	40.65%

(WebScience 2013)

Word co-occurences

Discover events based on temporal text variation

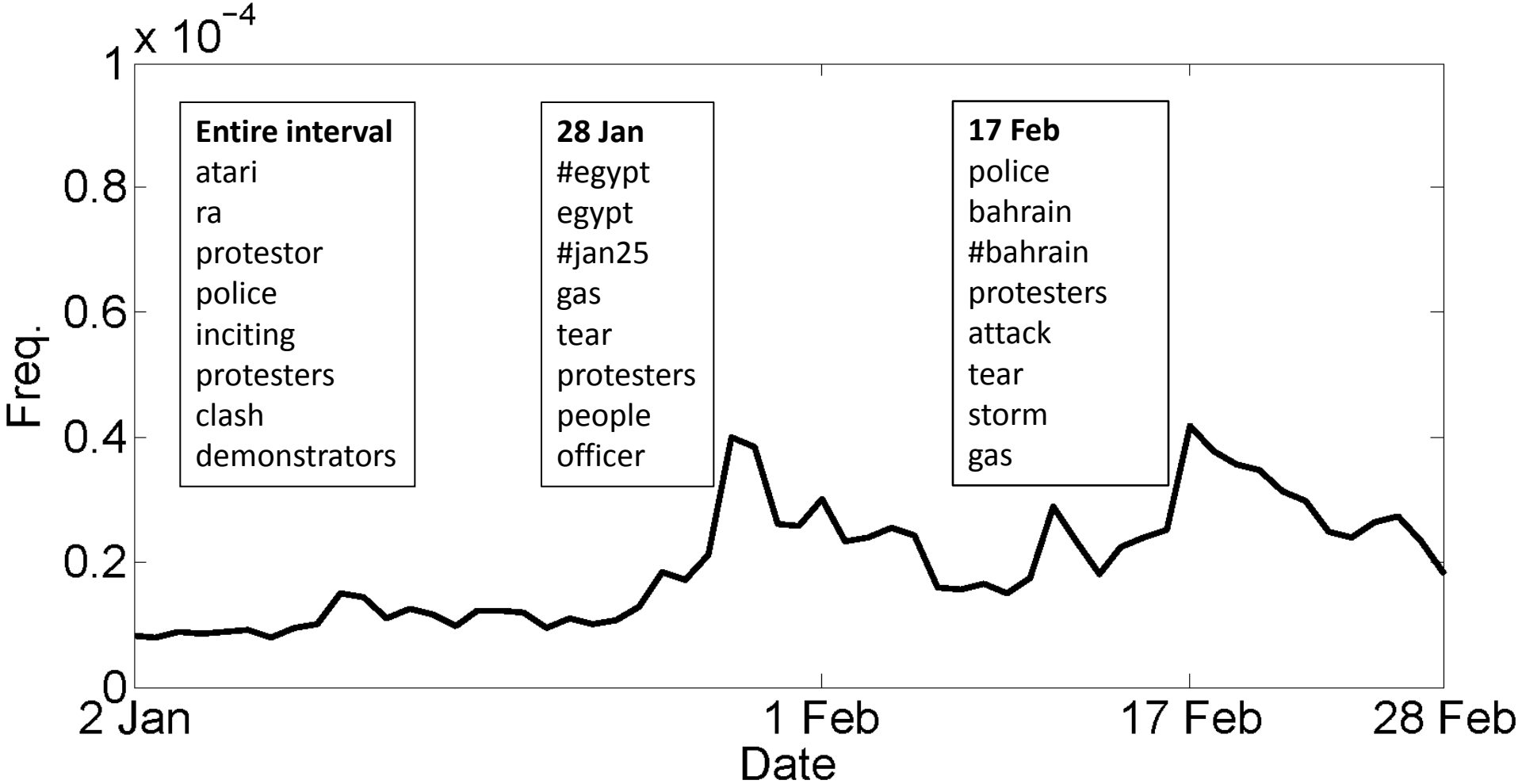
- word co-occurrence (e.g. NPMI) computed over large, static corpora:

  similar concepts or collocations

- computed over data from social media that reflects timely events (e.g. Twitter)

  current events & news

Co-occurences over time



`riot'

Method

- cluster words (cf. messages) in a time interval
- spectral clustering using NPMI as similarity measure
- coherent clusters corresponds to an event
- central words are important concepts

  used to extract relevant tweets

Sample event

Query: Kubica crash

Label: Formula 1 driver Robert Kubica injured in rally crash http://ow.ly/3R71Q

Coherence: 0.47, **Magnitude:** 140

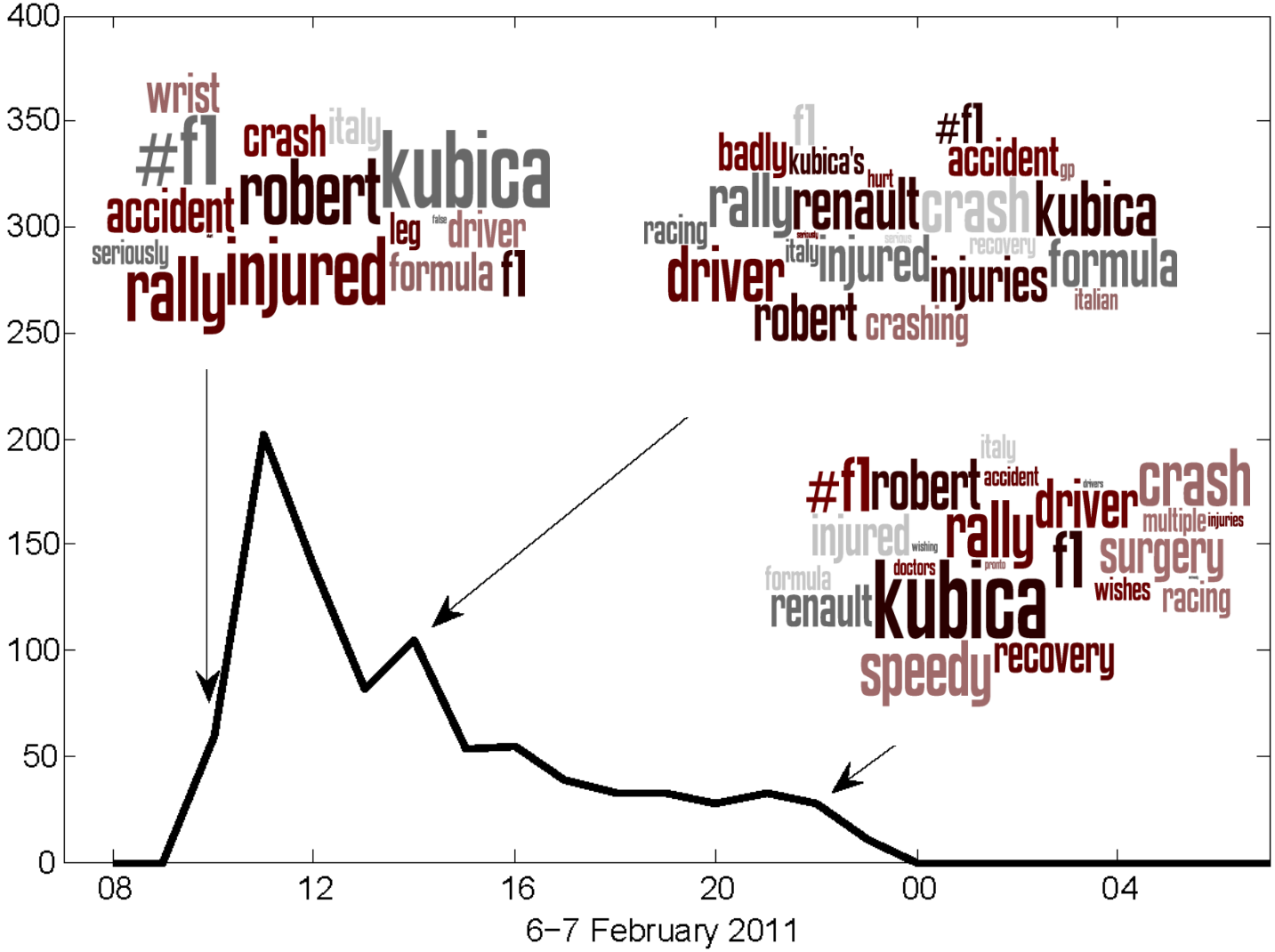
Date: 06 Feb 2011, 12-1pm

Longitudinal analysis

- discovers event evolution and persistence
- shows content drift over time
- evolutionary spectral clustering

  create consistent clusters across consecutive time windows

Longitudinal analysis



Conclusions

- Social Media data is highly time dependent
  - text has different proprieties conditioned on time
- By modelling time we gain a better understanding of real world effects
  - SM can be used to uncover real world events
  - SM can be used for ‘nowcasting’ indicators
  - complex temporal patterns play an important role in SM

Future directions

- Models incorporating regional and demographic variation
- Different domains of application: economics
- Introduce complex patterns to topic models
- Integration in downstream applications: IR
- Text + User behaviour

References

(ICWSM 2012) Trendminer: An Architecture for Real Time Analysis of Social Media Text.

D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, M. Niranjan

(HT 2013) Where's @wally: A classification approach to Geolocating users based on their social ties.

D. Rout, D. Preotiuc-Pietro, K.Bontcheva, T. Cohn ('Ted Nelson' award)

(WebScience 2013) Mining User Behaviours: A study of check-in patterns in Location Based Social Networks.

D. Preotiuc-Pietro, T. Cohn

(ACL 2013) A user-centric model of voting intention from Social Media.

V. Lampos, D. Preotiuc-Pietro, T. Cohn

(EMNLP 2013) A temporal model of text periodicities using Gaussian Processes.

D. Preotiuc-Pietro, T. Cohn

(EACL 2014) Predicting and Characterising User Impact on Twitter.

V. Lampos, N. Aletras, D. Preotiuc-Pietro, T.Cohn

Thank you !