Prediction models of Social Media data

The University Of Sheffield.

Daniel Preotiuc-Pietro

daniel@dcs.shef.ac.uk

11.10.2013

Summary

1. Social Media data preprocessing
2. Forecasting political polls
3. Forecasting periodic time series of words

TrendMiner project

- `Large scale, cross-lingual trend mining and summarization of real time media streams’
- 7 organisations; we work with University of Southampton and SORA on machine learning
- application to predicting political polls and financial indicators

www.trendminer-project.eu

1. Text preprocessing

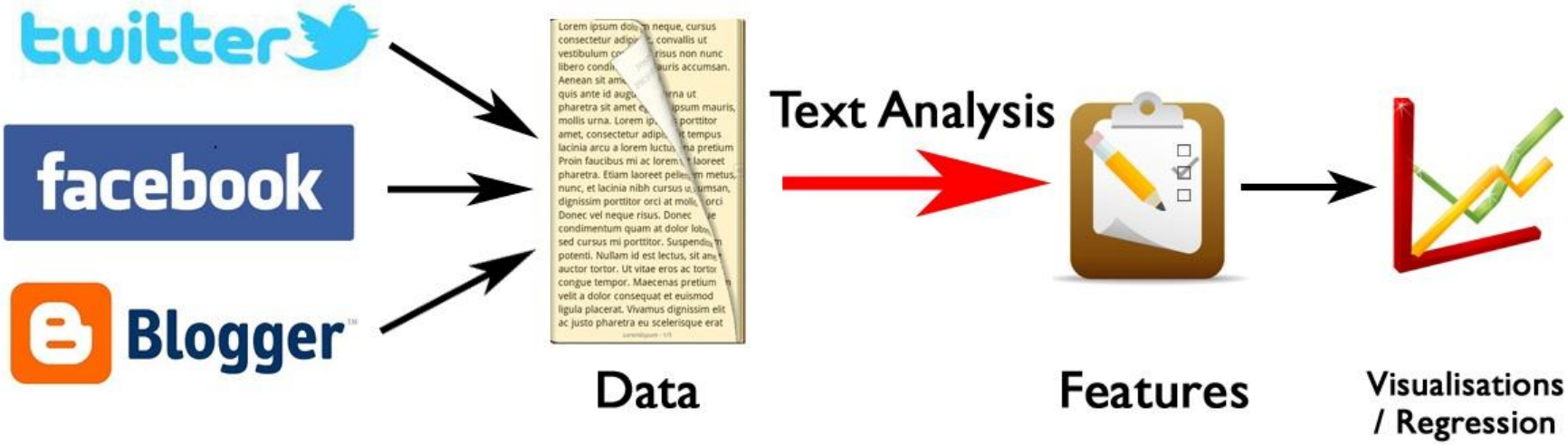
- for Social Media data:
  - Tokenisation
  - Language detection
  - `Sentiment‘ score
  - Geolocation (HT 2013)
  - Deduplication, filters
- pipeline setup, Streaming, MapReduce (ICWSM 2012)

https://github.com/danielpreotiuc

1. Text preprocessing

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

1. Text preprocessing



Texts are short and different in style than from traditional sources

1. Aims

We aim to integrate existing and new tools for OSN data processing in a framework that is:

Fast – real time processing

Modular - easy to add/change modules

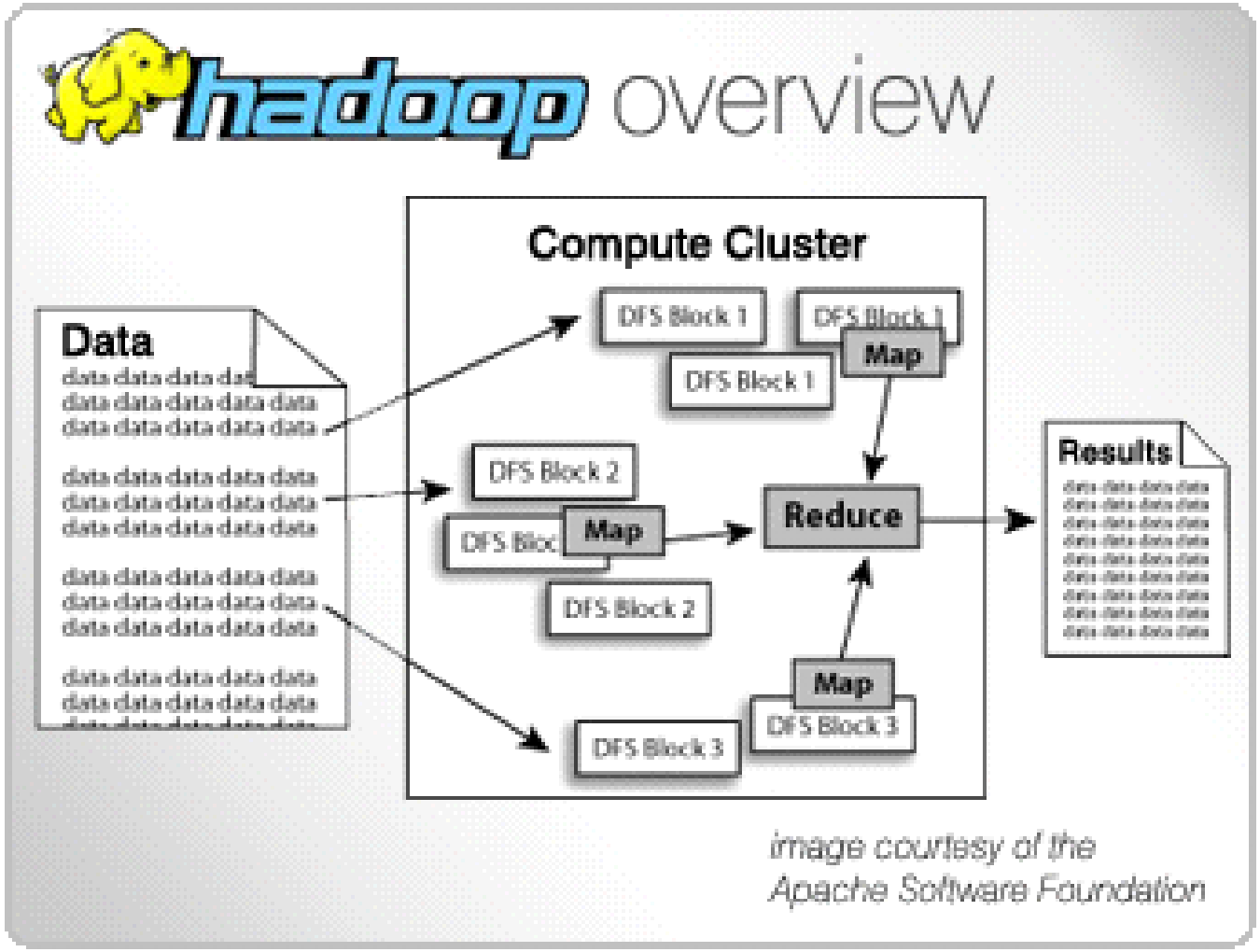
Pipeline architecture - flexible to the user's needs

Extensible - different sources of data (e.g. Facebook)

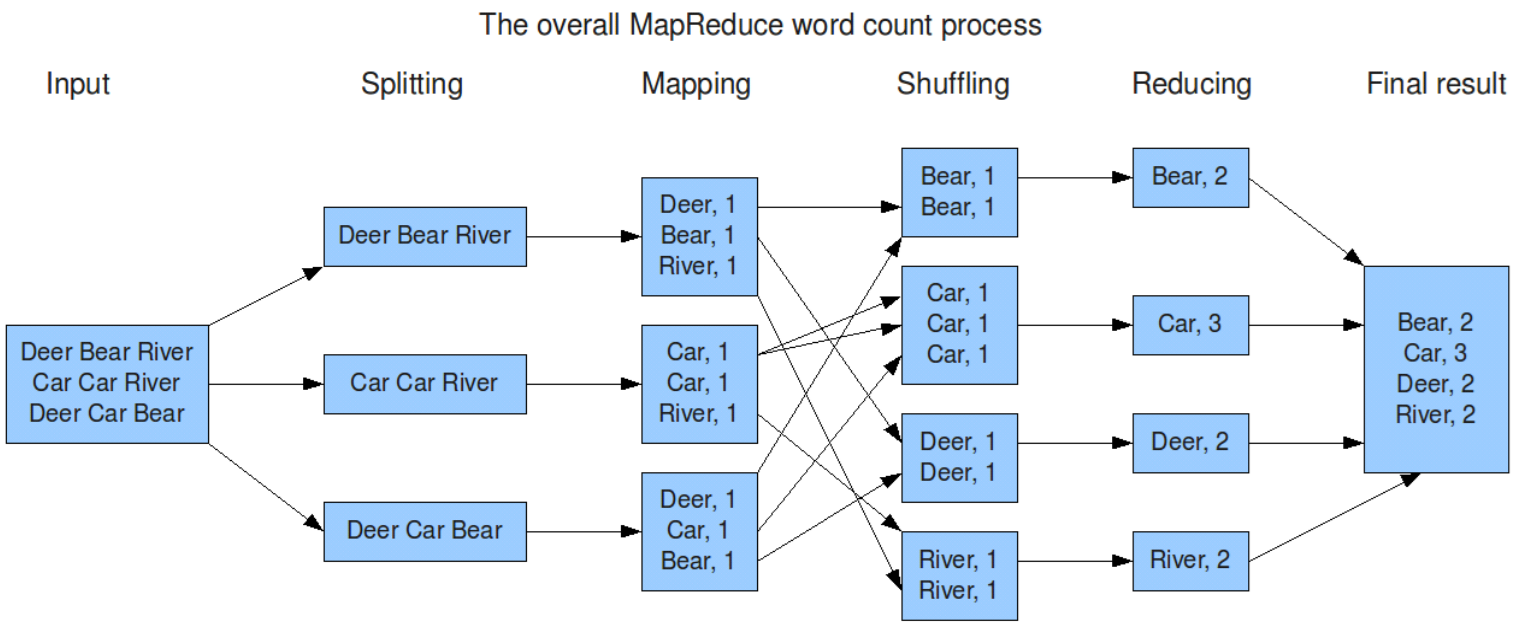
1. Architecture

- I/O bound: analysis takes less than random disk access
- Large data: 20Gb every day – 10% Twitter
  - input files are compressed splittable .lzo
- Many tasks can be done independently to each tweet
- Run in parallel using Apache Hadoop Map-Reduce framework and distributed file-system

1. Architecture



1. Architecture



http://www.searchworkings.org/blog

1. Architecture

Command line tool:

- single node
- distributed

2 types of usage:

- online
- batch analysis

Provided also as a web service

1. Example

Input:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
“user”:{“screen_name”:”abx1”,”location”:”sheffield,uk”, “utc_offset”:0” …},
…}

Output:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
“user”: {“screen_name:”abx1”,[…]},
“analysis”:{
“tokens”:
[“RT”,”@MediaScotland”,”greeeat”,”!!!”,”lvly”,”speech”,”by”,”cameron”,”on”,”scott's”,”indy”,”:)”,”#indyref”
],
“ner”: [“MediaScotland”,”cameron”,”scott's”],
“pos”: [“~”,”@”,”^”,”,”,”,”A”,”N”,”P”,”^”,”P”,”L”,”N”,”E”,”#”],
“spam”: “false”,
“geo”: {“city”: ”Sheffield”, “country”: “England”, “long”:”-1.46”, “lat”:”53.38”, “population”:
“534500”},
“langid”: {“language:” ”en”, “confidence”: 0.51} }

1. Example

Input:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
“user”:{“screen_name”:”abx1”,”location”:”sheffield,uk”, “utc_offset”:0” …},
…}

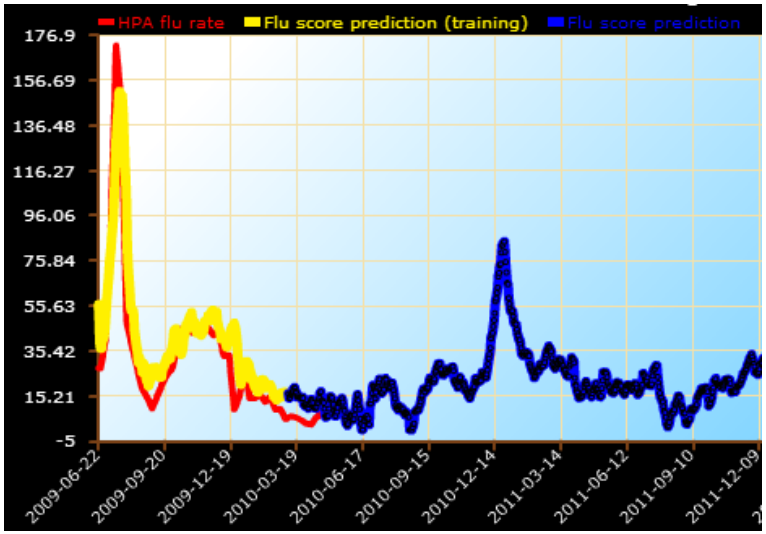
Output:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
“user”: {“screen_name:”abx1”,[…]},
“analysis”:{
“tokens”:
[“RT”,”@MediaScotland”,”greeeat”,”!!!”,”lvly”,”speech”,”by”,”cameron”,”on”,”scott's”,”indy”,”:)”,”#indyref”
],
“ner”: [“MediaScotland”,”cameron”,”scott's”],
“pos”: [“~”,”@”,”^”,”,”,”,”A”,”N”,”P”,”^”,”P”,”L”,”N”,”E”,”#”],
“spam”: “false”,
“geo”: {“city”: ”Sheffield”, “country”: “England”, “long”:”-1.46”, “lat”:”53.38”, “population”:
“534500”},
“langid”: {“language:” ”en”, “confidence”: 0.51} }

2. Text regression

- Task: predict real valued outputs based on textual variables (e.g. word counts)

LASSO on word counts →



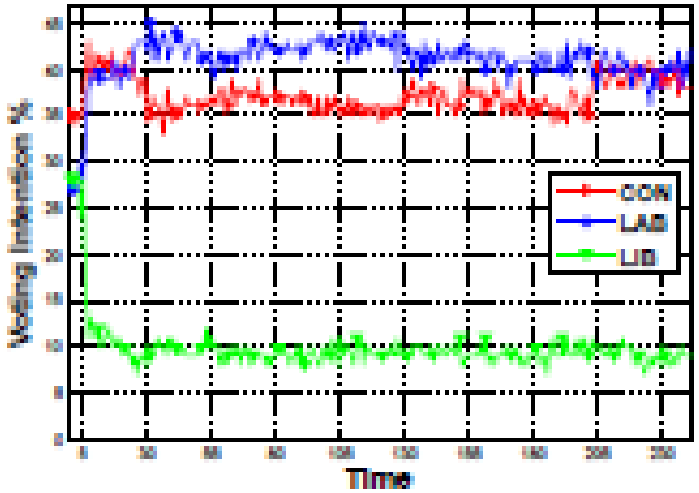
Lampos V., Cristianini N. (2010)

http://geopatterns.enm.bris.ac.uk/epidemics/

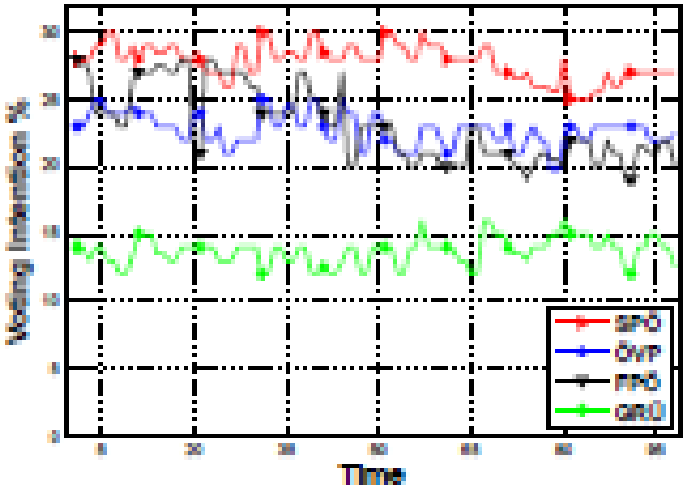
- Other examples: voting intention, financial indicators, weather, etc.

2. Use case

- predicting political polls (not elections!)
- strong baselines, realistic evaluation
- 2 different use cases (U.K. and Austria)



UK polls, 04/2010 – 02/2012



Ö. polls, 01/2012 – 12/2012

2. Motivation

- Twitter and real population demographics are different
- social media has biased opinions, not the most mentioned/positive sentiment party is indicative of real world trends
- more similar setup to traditional polls
- most of the users are not informative for our task and all their tweets represent noise

2. Motivation

- only a few words are informative of the task
- we want to obtain a model of

  sparse users & sparse words
- tune based on existing polls
- regression learns weights for features without using prior knowledge, making models more portable

2. Data

- collection focused on **all** the data from users of Twitter

  40000 U.K. (random)

  60 m. tweets

  1200 Austrian (selected by pol. scientists)

  800k tweets

2. Model

- Standard linear regression framework:

$$f(X) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X} + \beta$$

- The optimization objective is:

$$\{\boldsymbol{w}^*, \beta^*\} = argmin \sum_{i=1}^{n} (\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} + \beta - y_i)^2 + \Psi(\boldsymbol{w}, \rho)$$

where Ψ is the regularisation (e.g. L1, L2, EN)

2. Model

- Bilinear predictive model:

$$f(X) = \boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} + \beta$$

u – user weights, w – word weights, X – word/user counts

- The optimization objective is:

$$\{\boldsymbol{w}^*, \boldsymbol{u}^*, \beta^*\} = argmin \sum_{i=1}^{n} (\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{x_i}\boldsymbol{w} + \beta - y_i)^2$$
$$+\Psi_1\,(\boldsymbol{u}, \boldsymbol{\rho_1}) + \Psi_2\,(\boldsymbol{w}, \boldsymbol{\rho_2})$$

Y – response variable, $\Psi_{1,2}$ - regularisers for users and words

2. Model

BEN (Bilinear Elastic Net)

- Regularizers are both Elastic Nets
- a BEN model for predicting each party's score

<u>Drawback</u>: expect shared information between the tasks (e.g. + LAB is likely to be – CON)

2. Model

- build a bilinear model that learns multiple tasks and shares strength across them
- we use the Group LASSO inside the bilinear framework
- features inside a group have to be all zero/non-zero for all the tasks
- each group is the same word/user for each task

2. Model

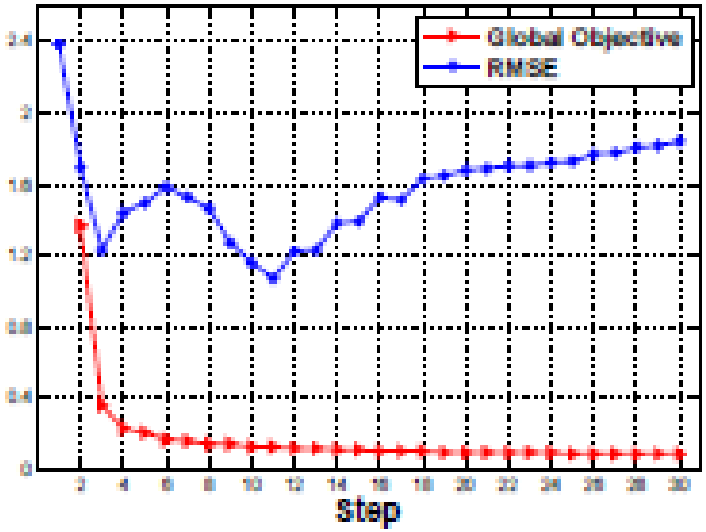
BGL (Bilinear Group Lasso)

- the tasks are predicting each party's score
- optimisation task is:

$$\{w^*, u^*, \beta^*\} = argmin \sum_{t=1}^{T} \sum_{i=1}^{n} (\boldsymbol{u_t} Q_i \boldsymbol{w}_t + \beta_t - y_{ti})^2 + \\ + \lambda_1 \sum_{j=1}^{m} \|W_j\|_2 + \lambda_2 \sum_{k=1}^{p} \|U_k\|_2$$

2. Learning

- Biconvex learning task: solved by a repeated application of 2 convex processes
- Regulariser parameters are fixed and found using grid search on validation
- Empirically choose to stop after 4 steps



2. Learning

- Biconvex learning task: solved by a repeated application of 2 convex processes
- Regulariser parameters

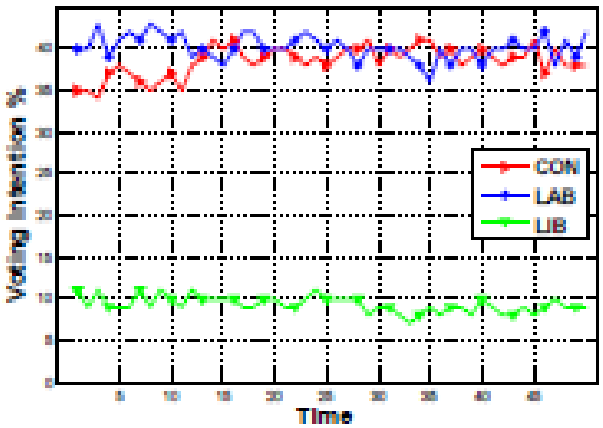
are fixed and found using

grid search on validation

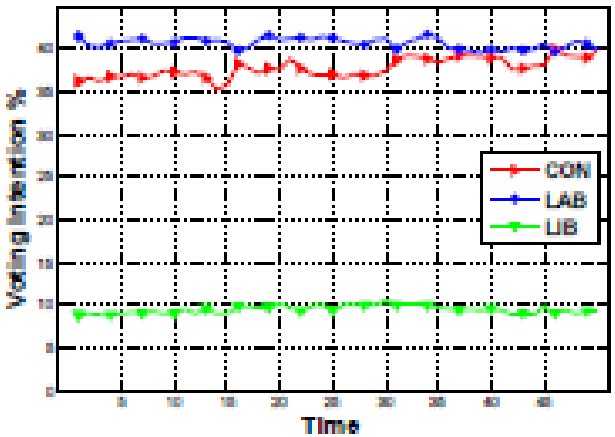
- Empirically choose to

stop after 4 steps

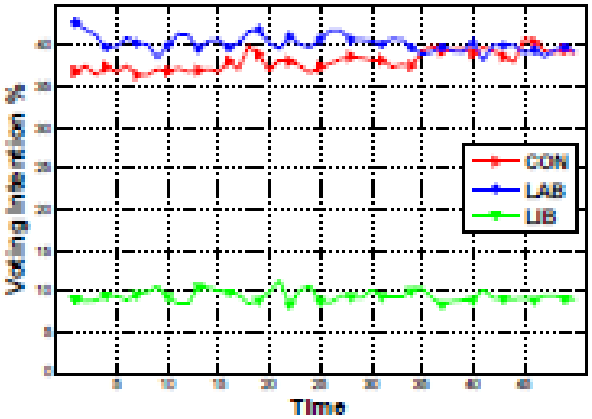
2. Results – U.K.



Ground truth



BEN



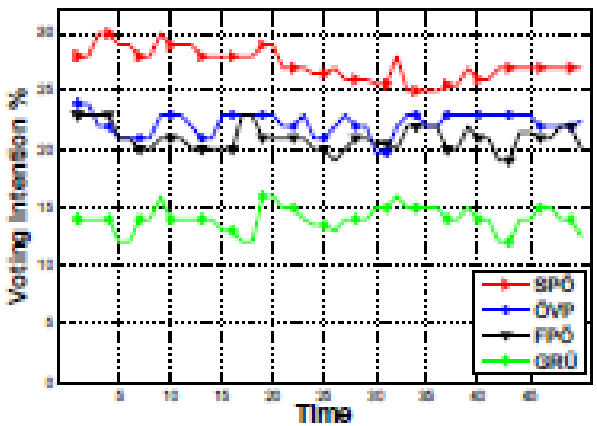
BGL

	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**

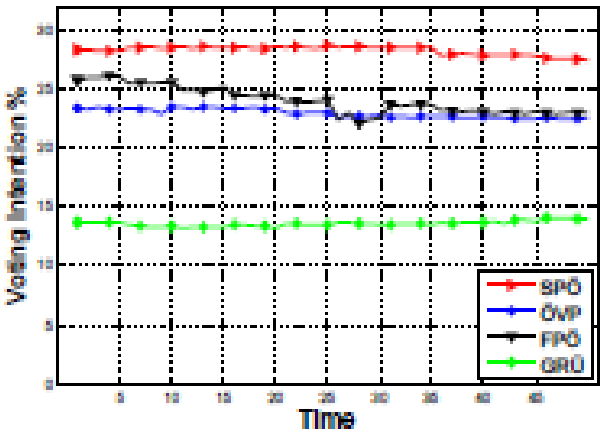
2. Results – U.K.

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

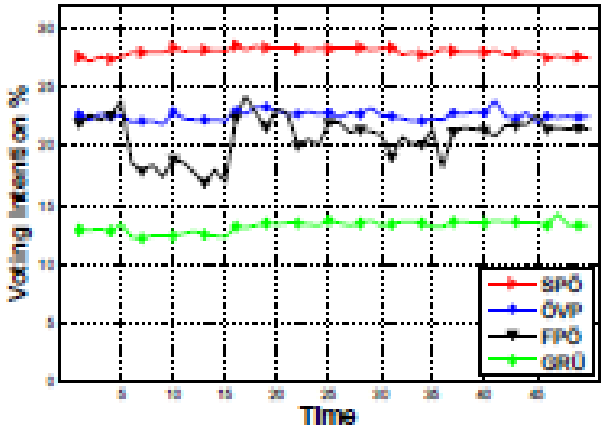
2. Results – Austria



Ground truth



BEN



BGL

	SPÖ	ÖVP	FPÖ	GRÜ	μ
B_μ	1.535	1.373	3.3	1.197	1.851
B_{last}	**1.148**	1.556	**1.639**	1.536	1.47
LEN	1.291	1.286	2.039	**1.152**	1.442
BEN	1.392	1.31	2.89	1.205	1.699
BGL	1.619	**1.005**	1.757	1.374	**1.439**

2. Results – Austria

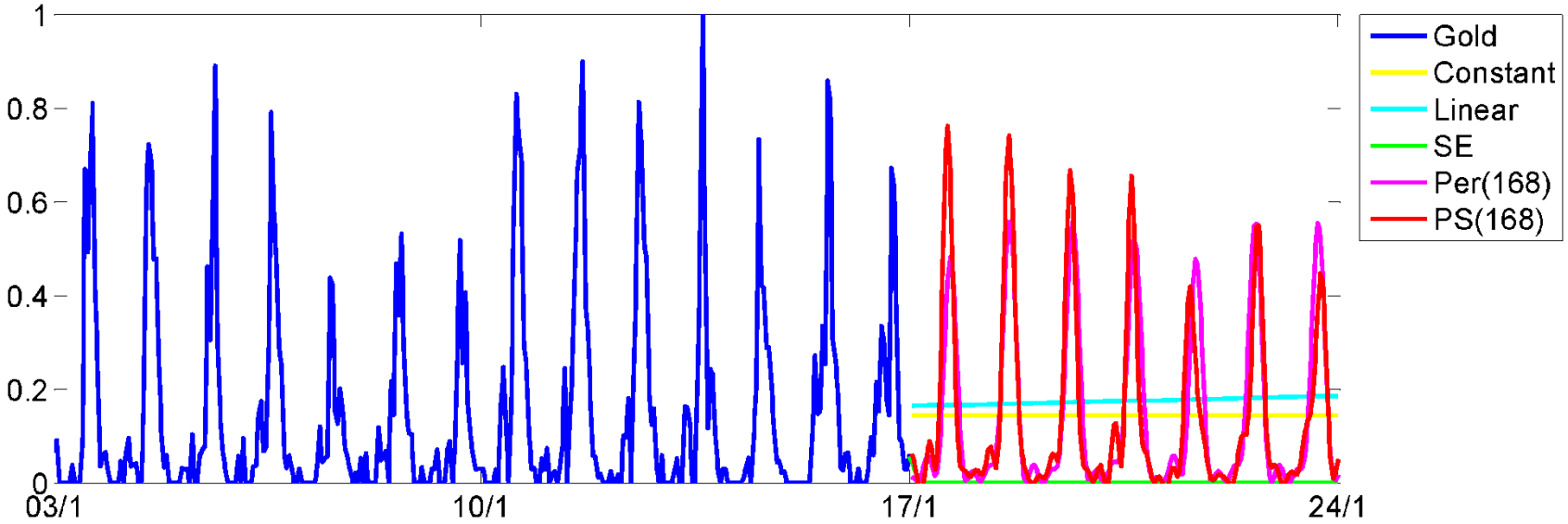
Party	Tweet	Score	Author
SPO	Inflationsrate in O¨ . im Juli leicht gesunken: von 2,2 auf 2,1%. Teurer wurde Wohnen, Wasser, Energie.	0.745	Journalist
	Hans Rauscher zu Felix #Baumgartner “A klaner Hitler” <link>	-1.711	Journalist
OVP	#IchPirat setze mich dafu¨r ein, dass eine große Koalition mathematisch verhindert wird! 1.Geige: #Gruene + #FPOe + #OeVP	4.953	User
	kann das buch “res publica” von johannes #voggenhuber wirklich empfehlen! so zum nachdenken und so... #europa #demokratie	-2.323	User
FPO	Neue Kampagne der #Krone zur #Wehrpflicht: “GIB BELLO EINE STIMME!”	7.44	Political Satire
	Kampagne der Wiener SPO “zum Zusammenleben” spielt Rechtspopulisten in die H¨ande <link>	-3.44	Human Rights
GRU	Protestsong gegen die Abschaffung des Bachelor-Studiums Internationale Entwicklung: <link> #IEbleibt #unibrennt #uniwu	1.45	Student Union
	Pilz “ich will in dieser Republik weder kriminelle Asylwerber, noch kriminelle orange Politiker” - BZO¨ -Abschiebung ok, aber wohin? #amPunkt	-2.172	User

3. Forecasting periodic time series

- Forecasting word time series (i.e. Twitter hashtags) well into the future
- Identify more complex temporal patterns than smoothness i.e. periodicities
- Group time series: periodic vs. non-periodic
- Use in temporally aware text classification

3. Example

Which is the better forecast?



#goodmorning

3. Data

- 1176 hashtags time series from 1 Jan 2011 – 28 Feb 2011
- 6.5 mil deduplicated tweets, 9.55 voc.tokens/tweet
- Hashtags are a proxy for topics on Twitter

#YOLO
Abbr. you only live once
The idiots's excuse for
something stupid that they did.

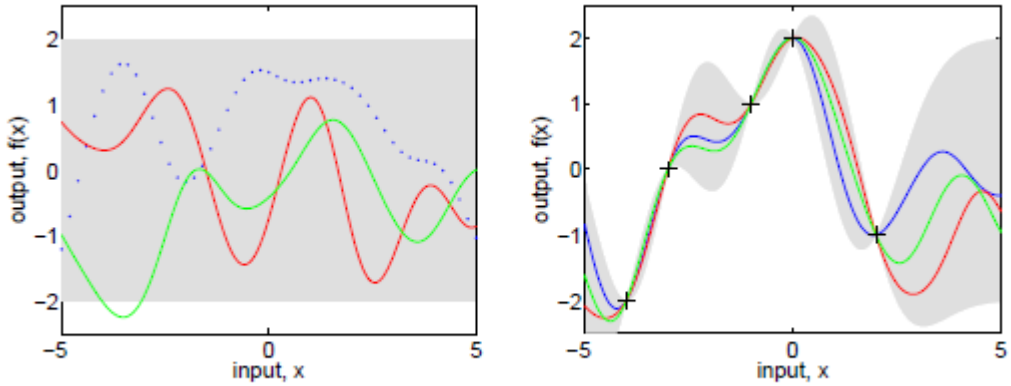
Twitter Power: *The Twitter Hashtag*

"Hey i heard u got that girl pregnant"
"Ya man but hey YOLO"

From www.urbandictionary.com

3. Gaussian processes

- GP - bayesian non-parametric method
- it gives a ‘distribution over functions’
- defined by choice of kernel and its parameters



(a), prior (b), posterior

- Interpolation
  - ‘fill in the gaps’
- Extrapolation
  - forecast future learning from the past

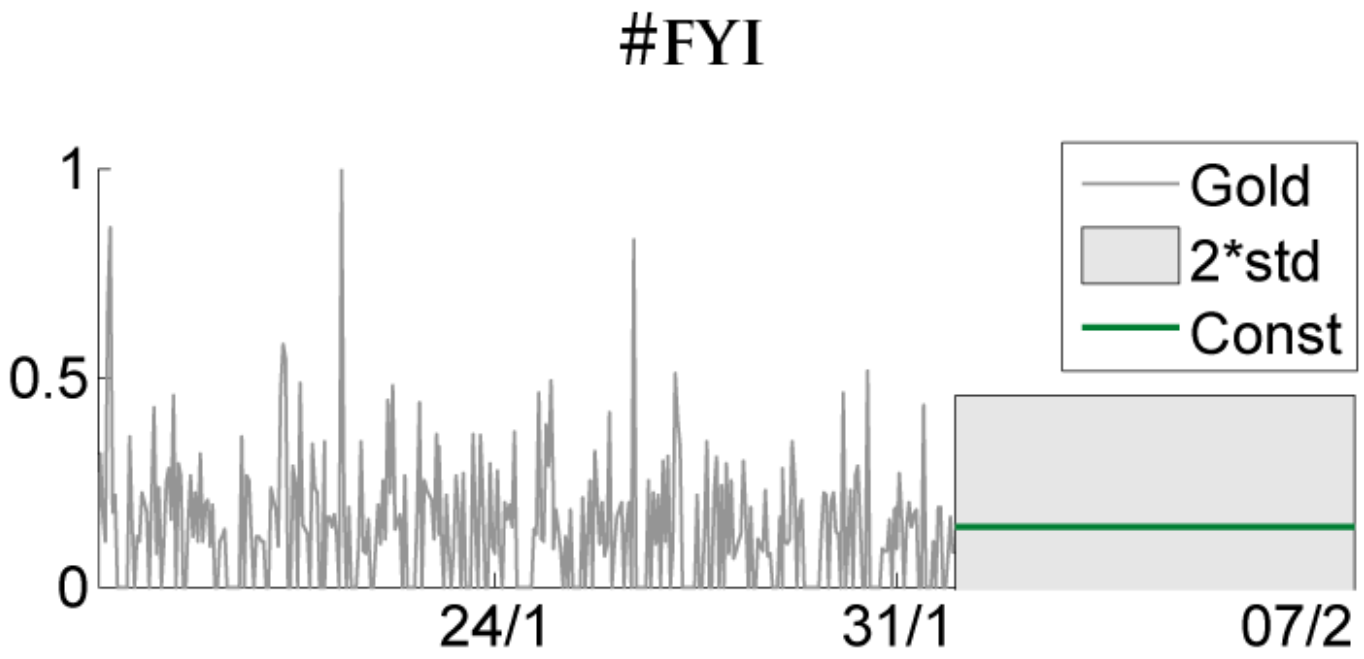
3. Regression task

- <u>Task:</u> Regression
  - has exact inference under the GP
- predict the frequency of a word in the future, given past training data
- for extrapolation, kernel choice is paramount
- intuitively:
  - smooth function -> closer points, high covariance
  - periodic function -> points at period p distance, high covariance

3. Model selection

- given a model, compute probability of the data integrating over the parameter space i.e. Bayesian ‘evidence’; has analytical solution
- balances data fit and model complexity (Occam‘s Razor)
- complex models which can account for many datasets achieve low evidence
- use Negative log Marginal Likelihood (ML-II) for model selection, giving an implicit classification of time series

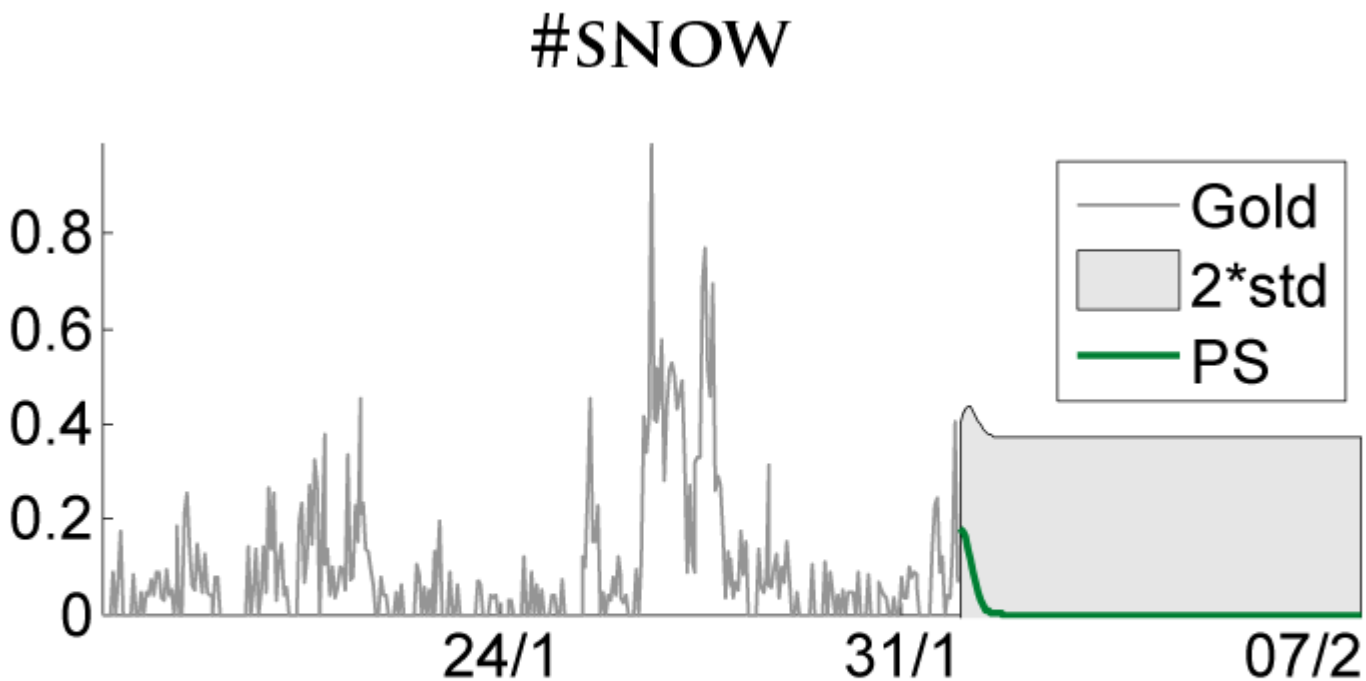
3. Model



$$k_C(t, t') = c$$

#funny
#lego
#likeaboss
#money
#nbd
#nf
#notetoself
#priorities
#social
#true

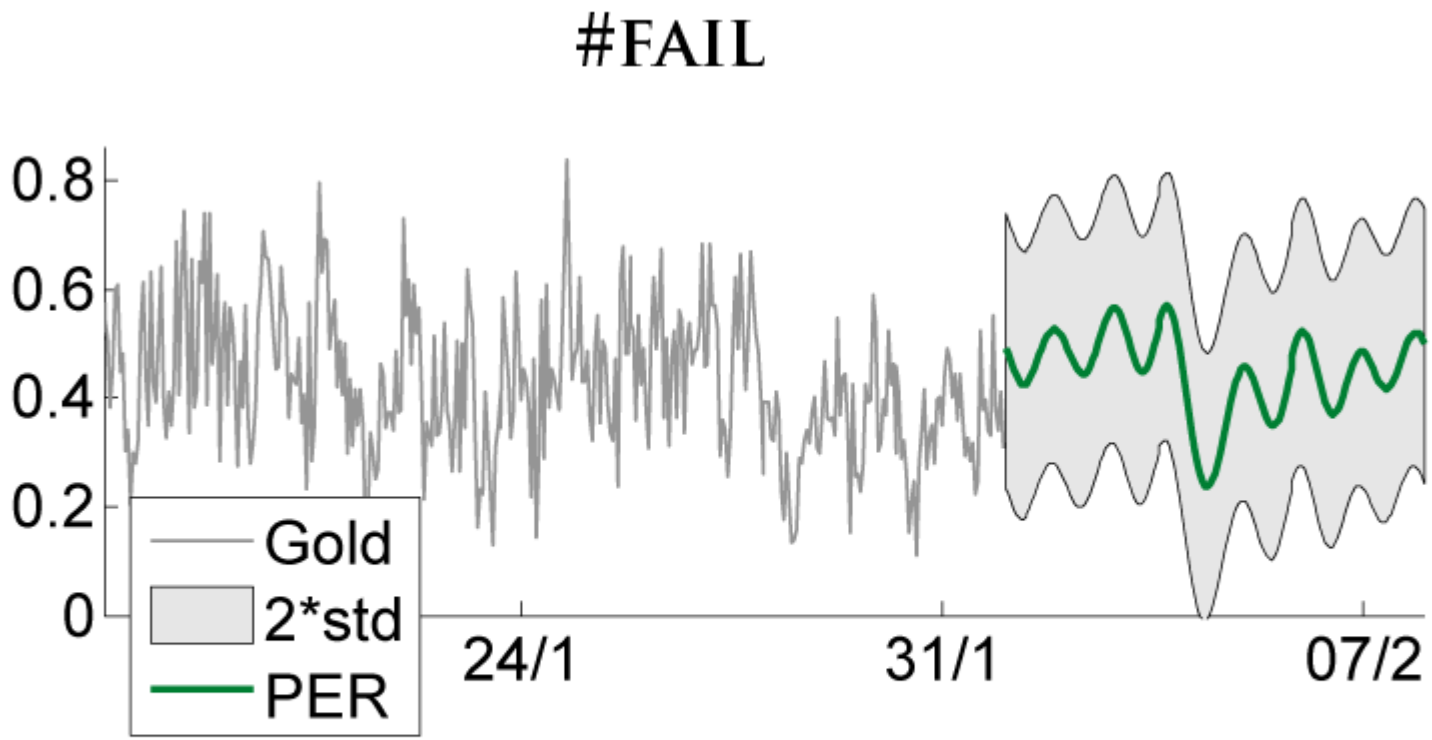
3. Model



#2011
#backintheday
#confessionhour
#februarywish
#haiti
#makeachange
#questionsidontlike
#savelibraries
#snow
#snowday

$$k_{SE}(t,t') = s^2 \cdot \exp - \frac{(t-t')^2}{2l^2}$$

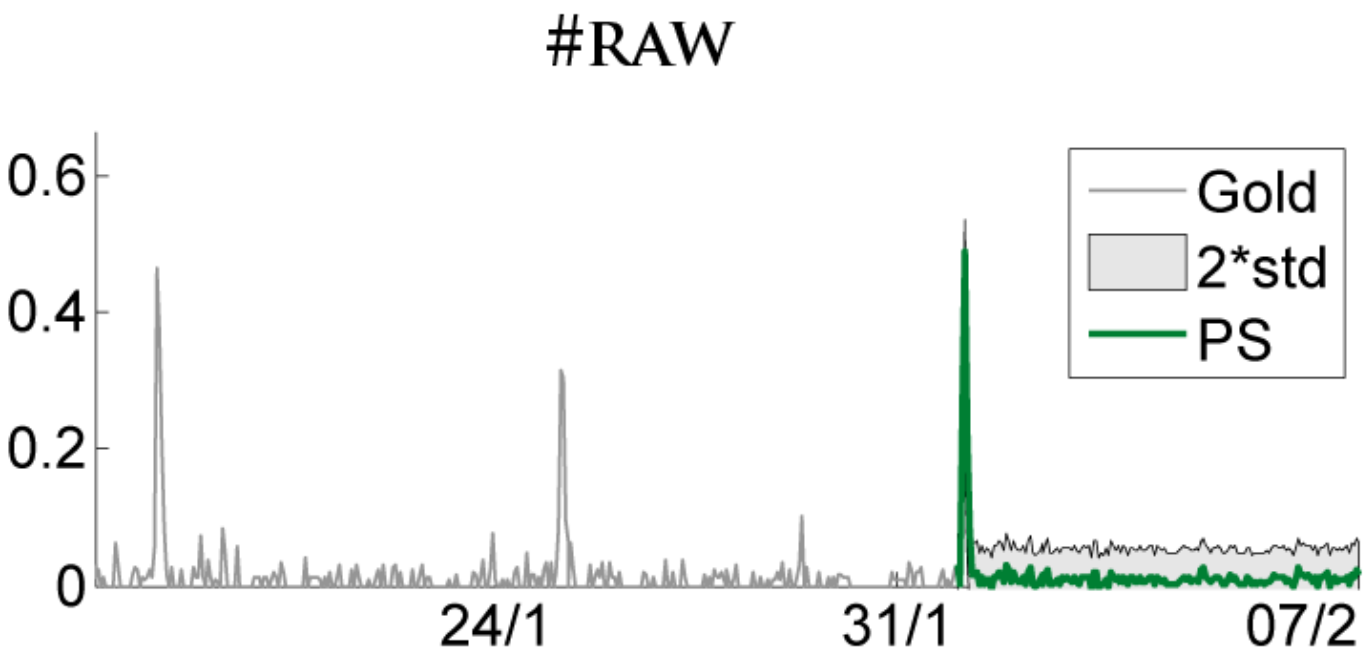
3. Model



$$k_{PER}(t,t') = s^2 \cdot \exp -2 \cdot \left(\frac{\sin^2(2\pi(t-t')^2/p)}{l^2} \right)$$

#brb
#coffee
#facebook
#facepalm
#funny
#love
#rock
#running
#xbox
#youtube

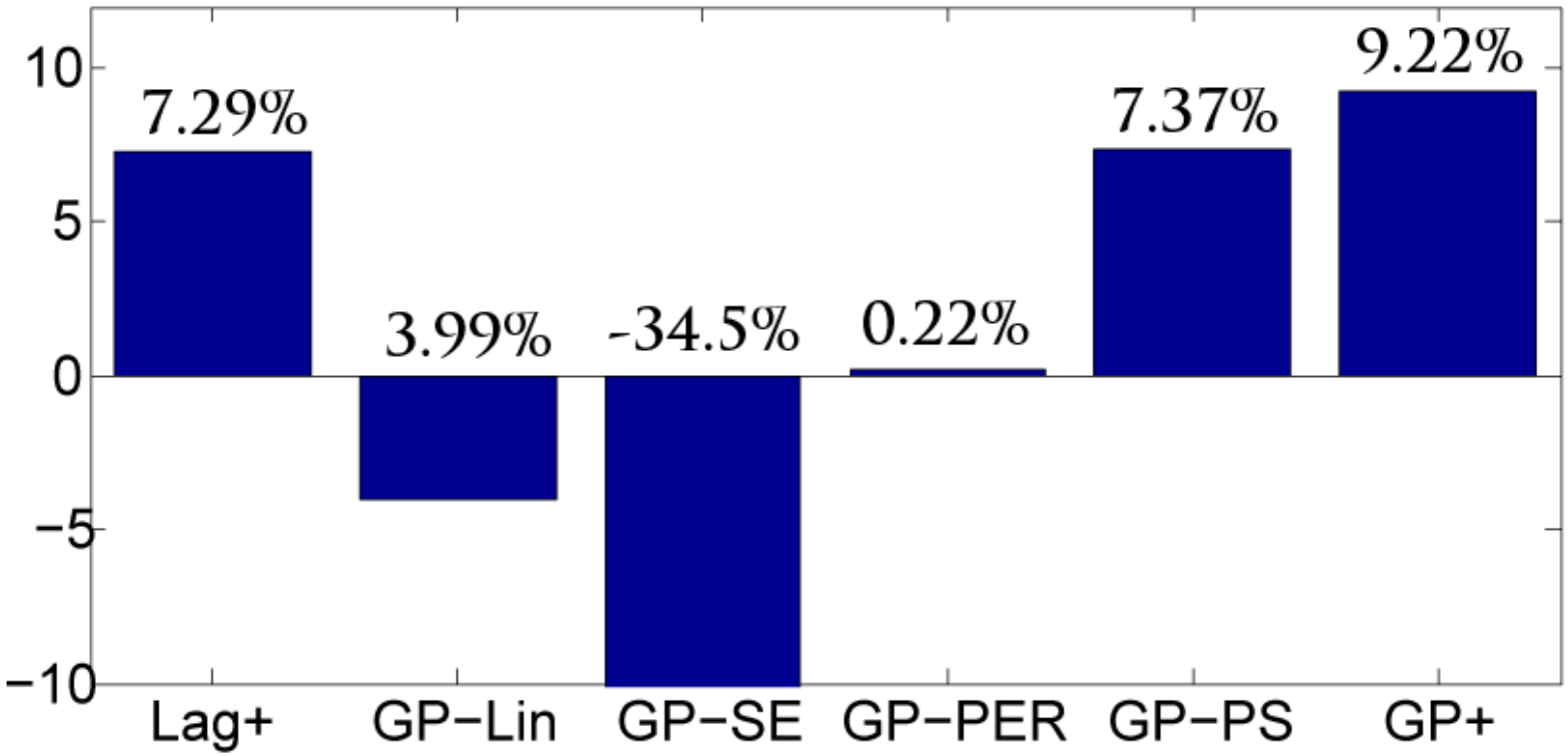
3. Model



#breakfast
#eastenders
#ff
#followfriday
#goodnight
#jobs
#news
#tgif
#thegame
#ww

$$k_{PS}(t,t') = \cos\left(\sin\left(\frac{2\pi \cdot (t-t')^2}{p}\right)\right) \cdot \exp\left(\frac{s\cos(2\pi \cdot (t-t')^2)}{p} - s\right)$$

3. Forecasting



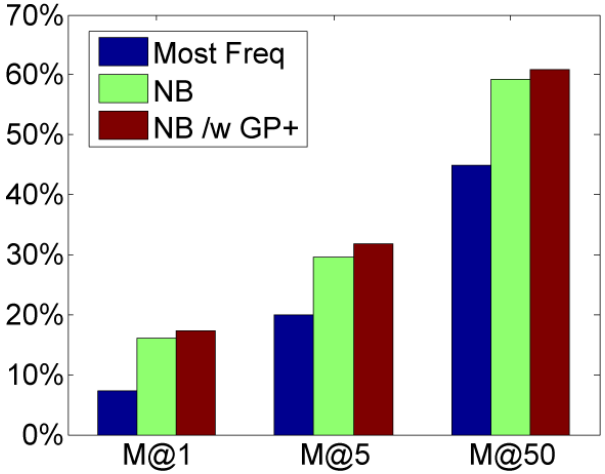
- train on January, forecast February
- performance compared to mean prediction (=GP-Const)
- GP+ performs model selection
- Lag+ AR model that uses the GP determined period

3. Text classification

Task: Predict hashtag based on tweet text

Use GP forecast as prior for Naive Bayes

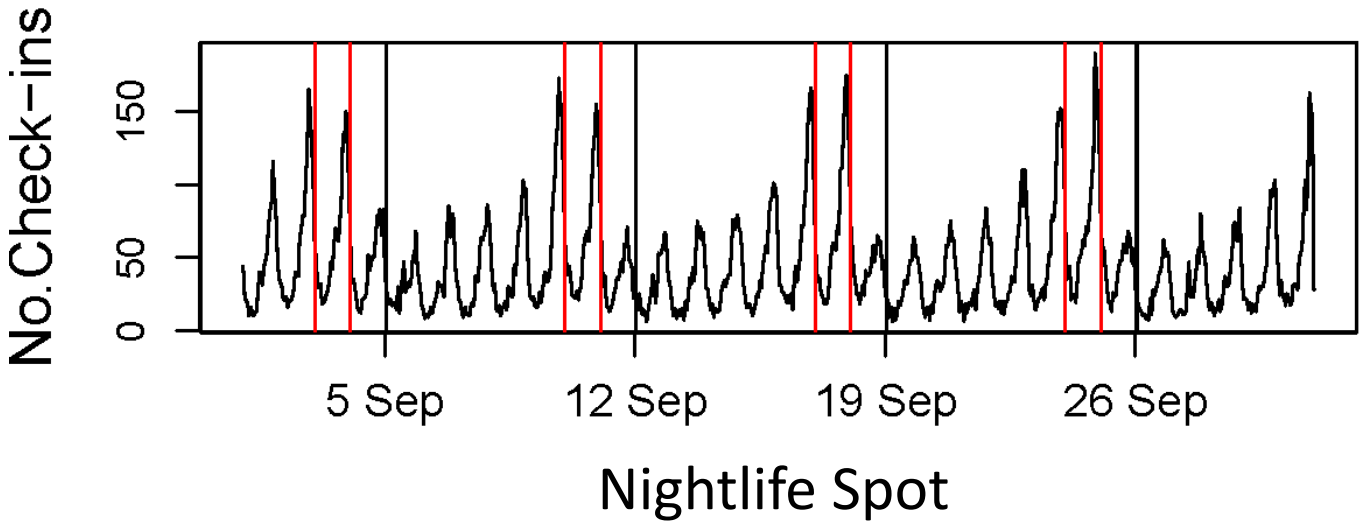
	MF	NB-E	NB-P
Match@1	7.28%	16.04%	17.39%
Match@5	19.90%	29.51%	31.91%
Match@50	44.92%	59.17%	60.85%
MRR	0.144	0.237	0.252



Tweet	Time	Prior	Rank	Prediction
Alfie u doughnut! U didn't confront kay? SMH	7-8pm 3 Feb 2011	E: 0.00027 P: 0.00360	8 1	#nowplaying #eastenders

3. Related work

- Model other periodic time series:
  - User behaviour (WebScience 2013)
  - Download/click-through rates
  - Search queries



Collaborators

Vasileios Lampos
Sheffield
www.lampos.net

Dominic Rout
Sheffield
www.domrout.co.uk

Trevor Cohn
Sheffield
http://dcs.shef.ac.uk/~tcohn/

Sina Samangooei
Southampton
www.sinjax.net

References

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

D. Preotiuc-Pietro, T.Cohn

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

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

(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

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

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

(Public Deliverable) Regression models of trends in streaming data

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

www.preotiuc.ro

Thank you !