User-generated content: *collective* and *personalised* inference tasks

Vasileios Lampos

Department of Computer Science
University College London

(March, 2016; @ DIKU)

Structure of the talk

1. **Introductory remarks**

2. **Collective inference tasks** from user-generated content
   — Nowcasting flu rates from Twitter / Google
   — Modelling voting intention (*bilinear text regression*)

3. **Personalised inference tasks** using social media
   — Occupation, income, socioeconomic status & impact

4. **Concluding remarks**

Context and motivation

+ the Internet, the World Wide Web and connectivity

+ numerous successful web products feeding from user activity

+ lots of user-generated content & activity logs, e.g. ***social media*** and ***search engine query logs***

+ large volumes of digitised data (‘***Big Data***’), birth of Data Science (*nothing new in principal*)

How can we use online data to improve our society, interpret human behaviour, and enhance our understanding about our world?

Context and motivation

+ the Internet, the World Wide Web and connectivity
+ numerous successful web products feeding from user activity
+ lots of user-generated content & activity logs, e.g. ***social media*** and ***search engine query logs***
+ large volumes of digitised data (‘***Big Data***’), birth of Data Science (*nothing new in principal*)

How can we use online data to improve our society, interpret human behaviour, and enhance our understanding about our world?

User-generated content: *Ongoing* applications

+ **Health**
  > disease surveillance, intervention impact
+ **Finance & Commerce**
  > financial indices
  > consumer satisfaction, market share
+ **Politics**
  > estimation of voting intentions
  > public opinion barometers
+ **Social and behavioural sciences**
  > complement questionnaire based studies
  > approach answers to unresolved questions

Added value of user-generated content for health

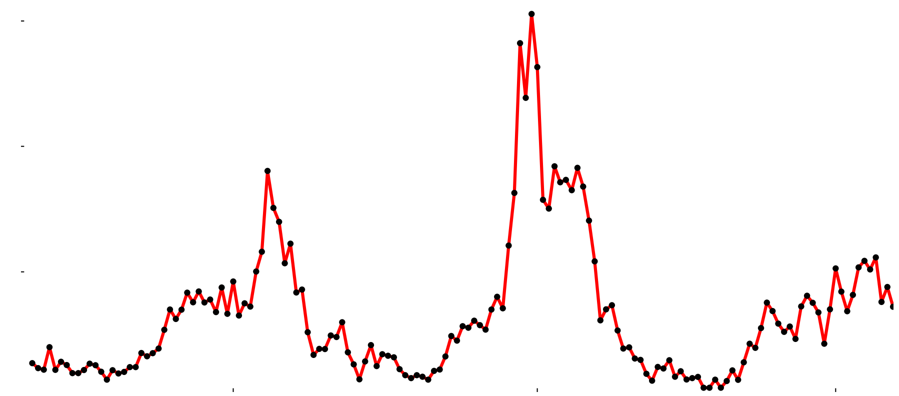
- \+ Online content can potentially access a larger and **more representative** part of the population
  Note: *Traditional health surveillance schemes are based on the subset of people that actively seek medical attention*
- \+ More **timely** information (*almost instant*) about a disease outbreak in a population
- \+ Geographical regions with **less established health monitoring systems** can greatly benefit
- \+ Small **cost** when data access and expertise are in place

Collective inference tasks from user-generated content

Lampos & Cristianini, 2012;
Lampos, Preotiuc-Pietro & Cohn, 2013;
Lampos, Miller, Crossan & Stefansen, 2015

Flu rates from Twitter: The task

n-gram frequency
time series



$f:$ $\longrightarrow$ *Flu surveillance disease rates from a health agency*

$\mathbf{X} \in \mathbb{R}^{M \times N}$ $\mathbf{y} \in \mathbb{R}^{M}$

Flu rates from Twitter: Lasso for feature selection

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

$$\underset{\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta - \sum_{j=1}^{m} x_{ij} w_j \right)^2 + \boxed{\lambda \sum_{j=1}^{m} |w_j|} \right\}$$

or $$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \boxed{\lambda \|\boldsymbol{w}\|_{\ell_1}} \right\}$$

also known as **lasso** or **L1-norm regularisation**

(*Tibshirani, 1996*)

Flu rates from Twitter: Bootstrap lasso

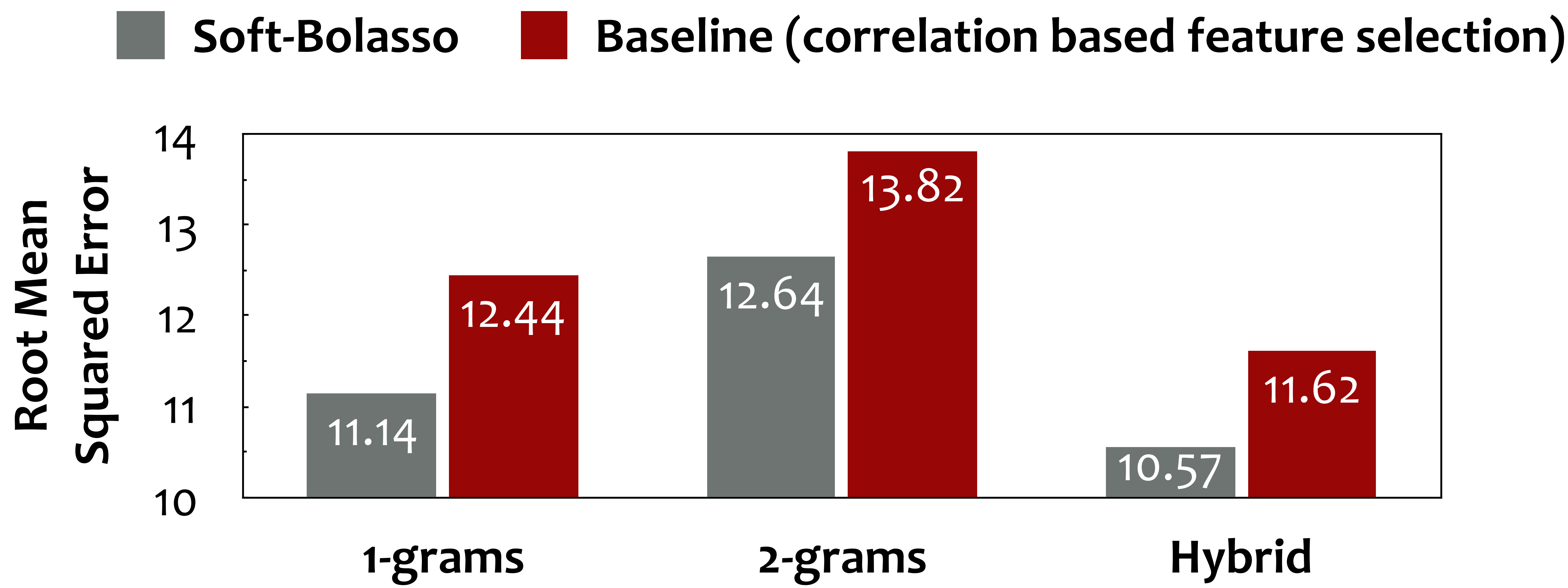
Lasso may not always select the *true model* due to collinearities in the feature space (*Zhao & Yu, 2006*)

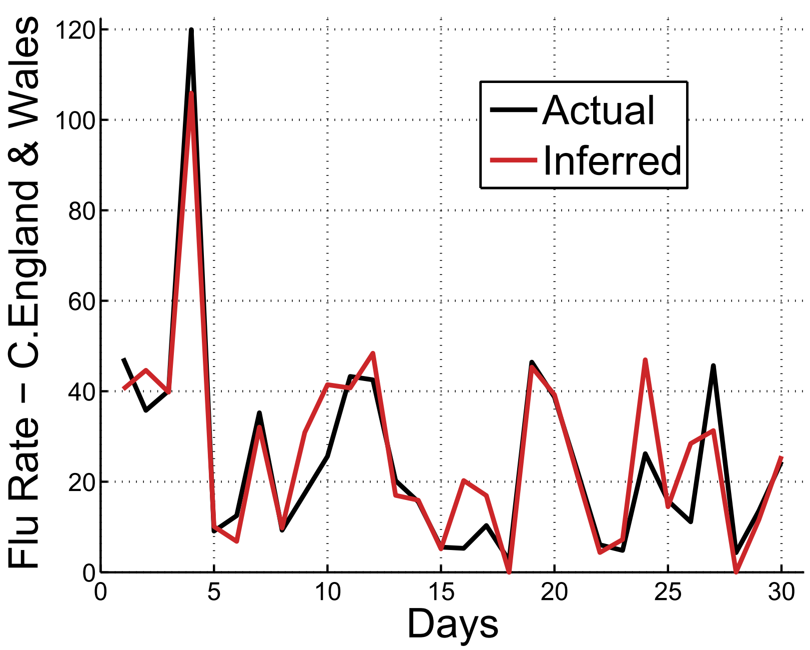
(*Bach, 2008*)

Bootstrapping lasso (‘*bolasso*’) for feature selection

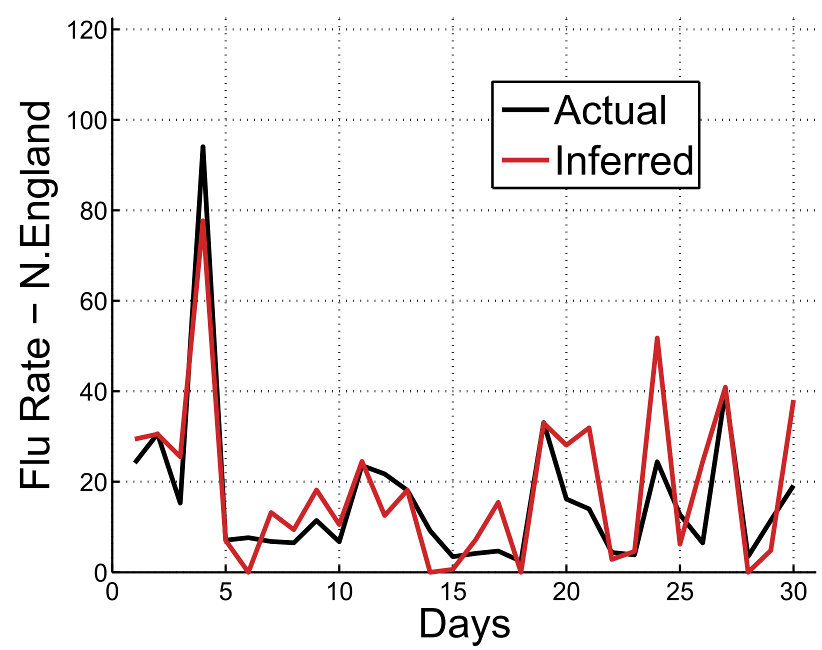
+ For a number (N) of bootstraps, i.e. iterations
  > Sample the feature space with replacement ($\boldsymbol{X}_i$)
  > Learn a new model ($\boldsymbol{w}_i$) by applying lasso on $\boldsymbol{X}_i$ and $\boldsymbol{y}$
  > Remember the *n*-grams with nonzero weights
+ Select the *n*-grams with nonzero weights in p% of the N bootstraps
+ p can be optimised; if p<100%, then ‘*soft bolasso*’

Flu rates from Twitter: Performance

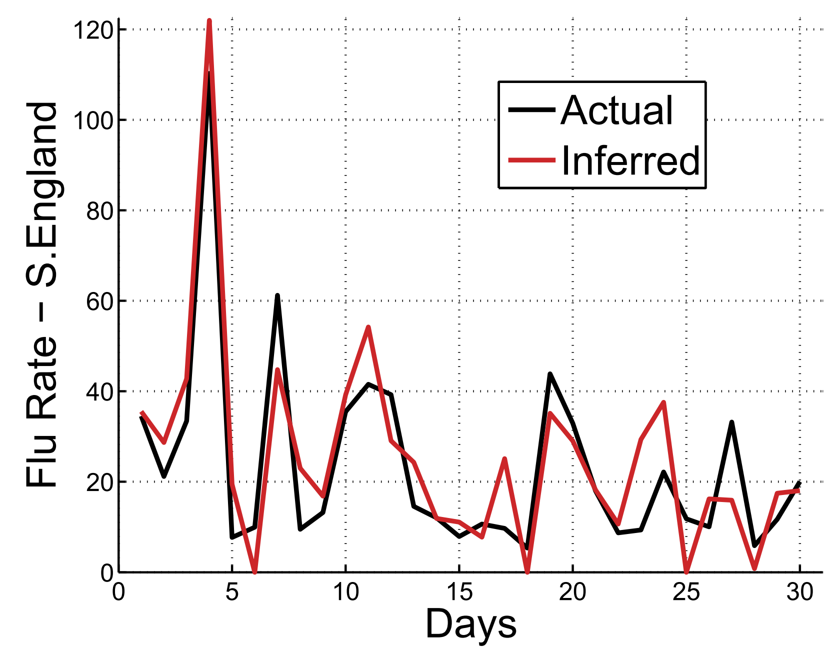




(a) *C. England & Wales* – RMSE: 8.36



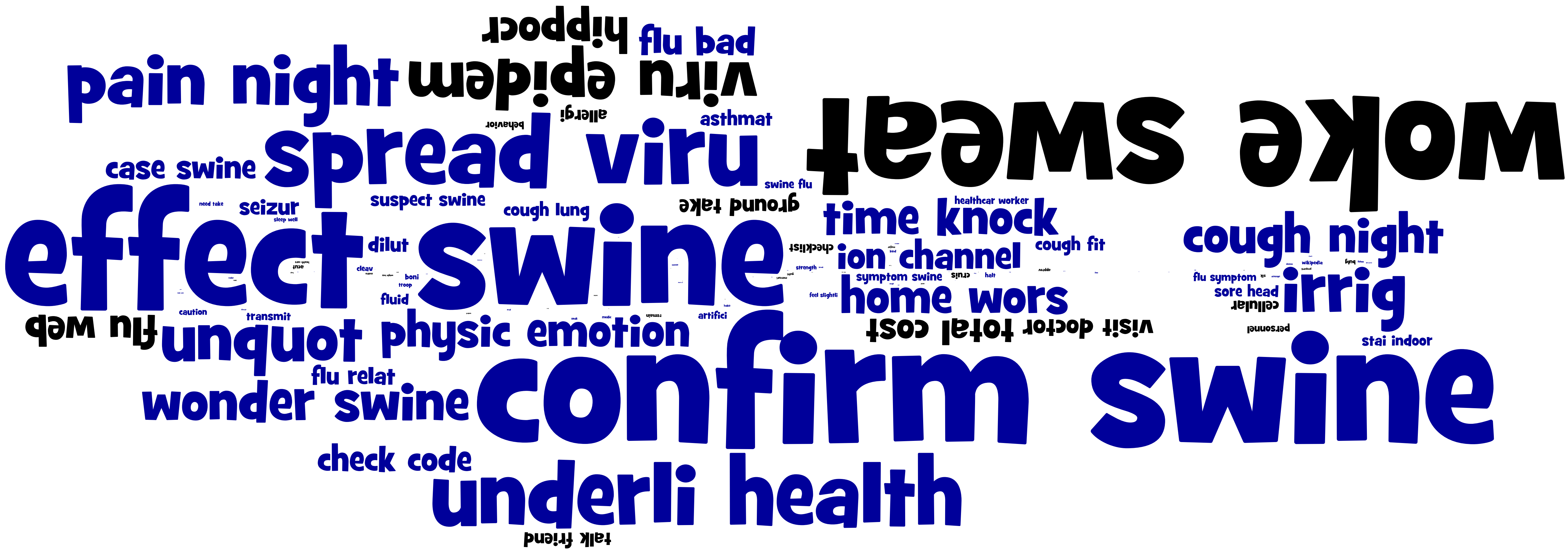
(b) *N. England* – RMSE: 9.782



(c) *S. England* – RMSE: 9.86

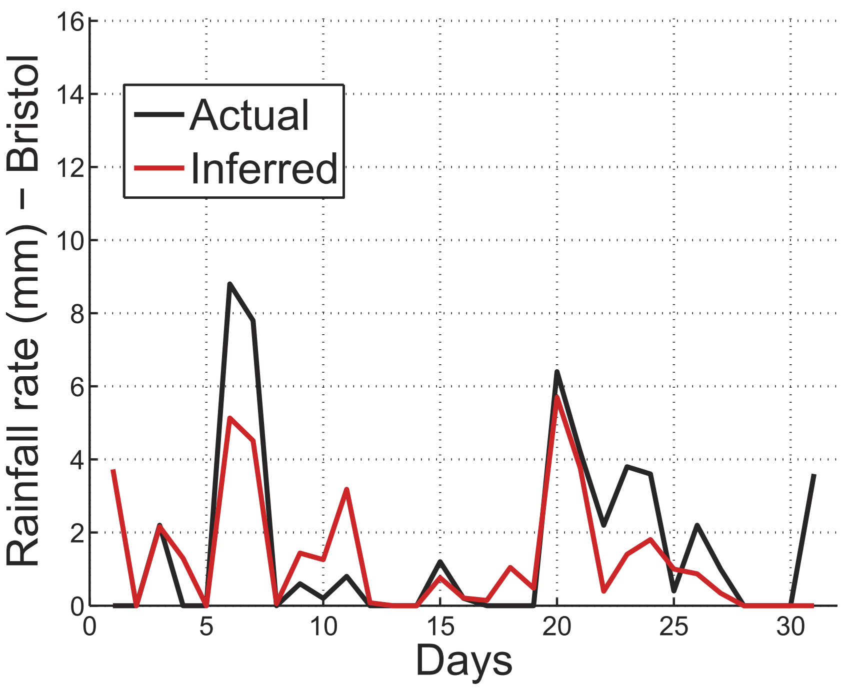
(*Lampos & Cristianini, 2012*)

Flu rates from Twitter: Selected features

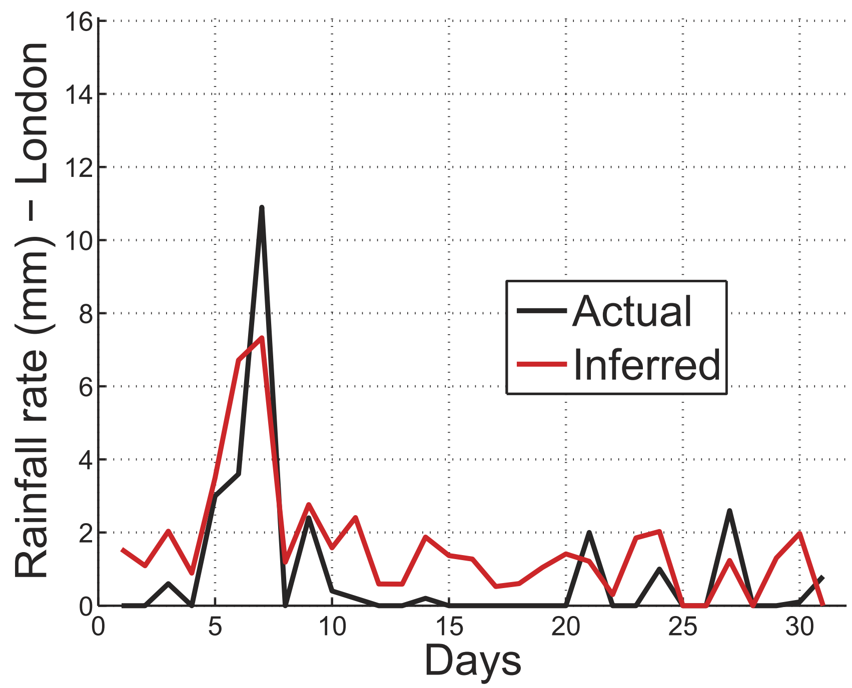


Word cloud with selected n-grams. Font size is proportional to the regression's weight; n-grams that are upside-down have a negative weight.

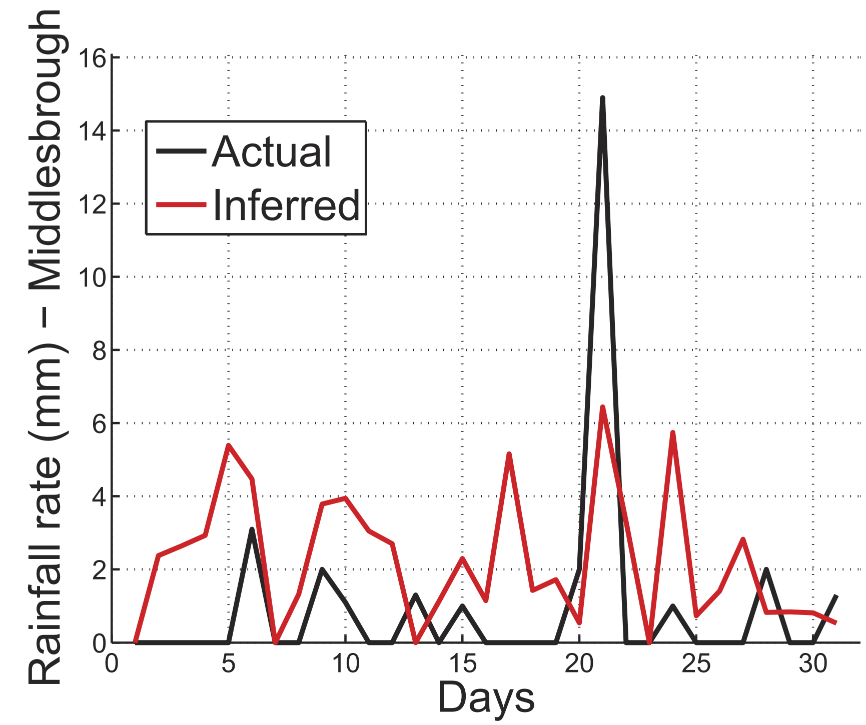
Rainfall rates from Twitter: *Generalisation*



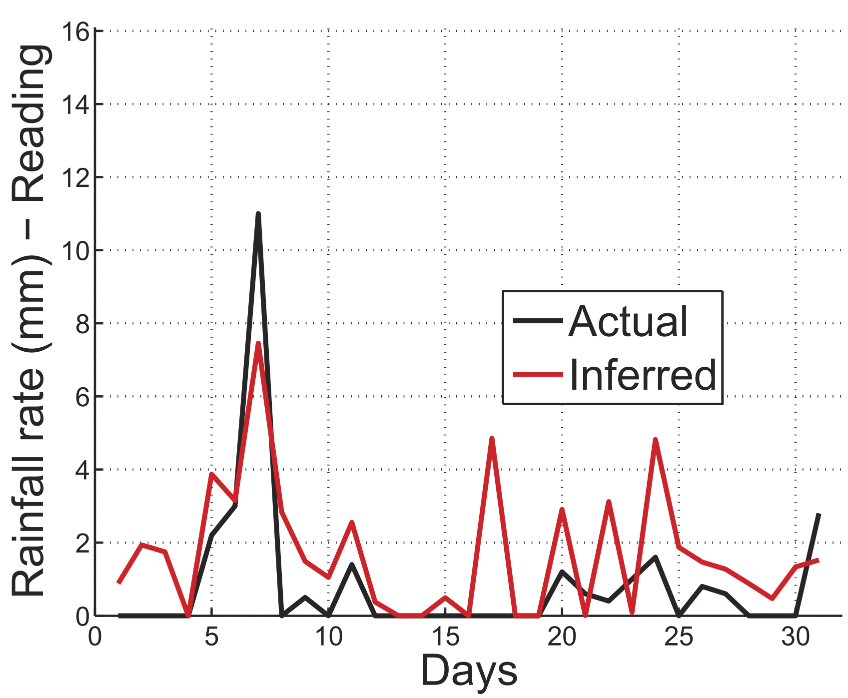
(a) *Bristol* – RMSE: 1.579



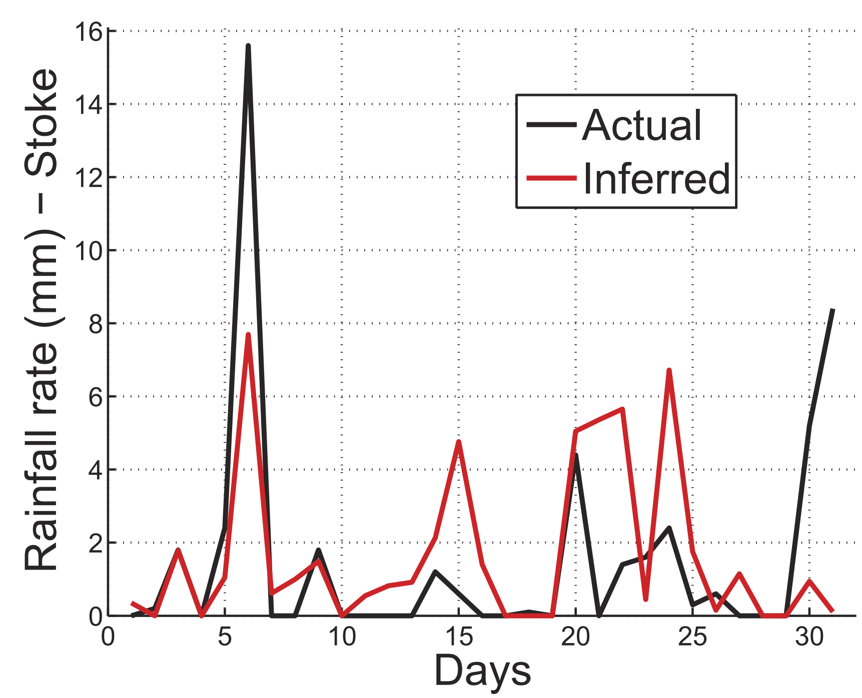
(b) *London* – RMSE: 1.399



(c) *Middlesbrough* – RMSE: 2.785

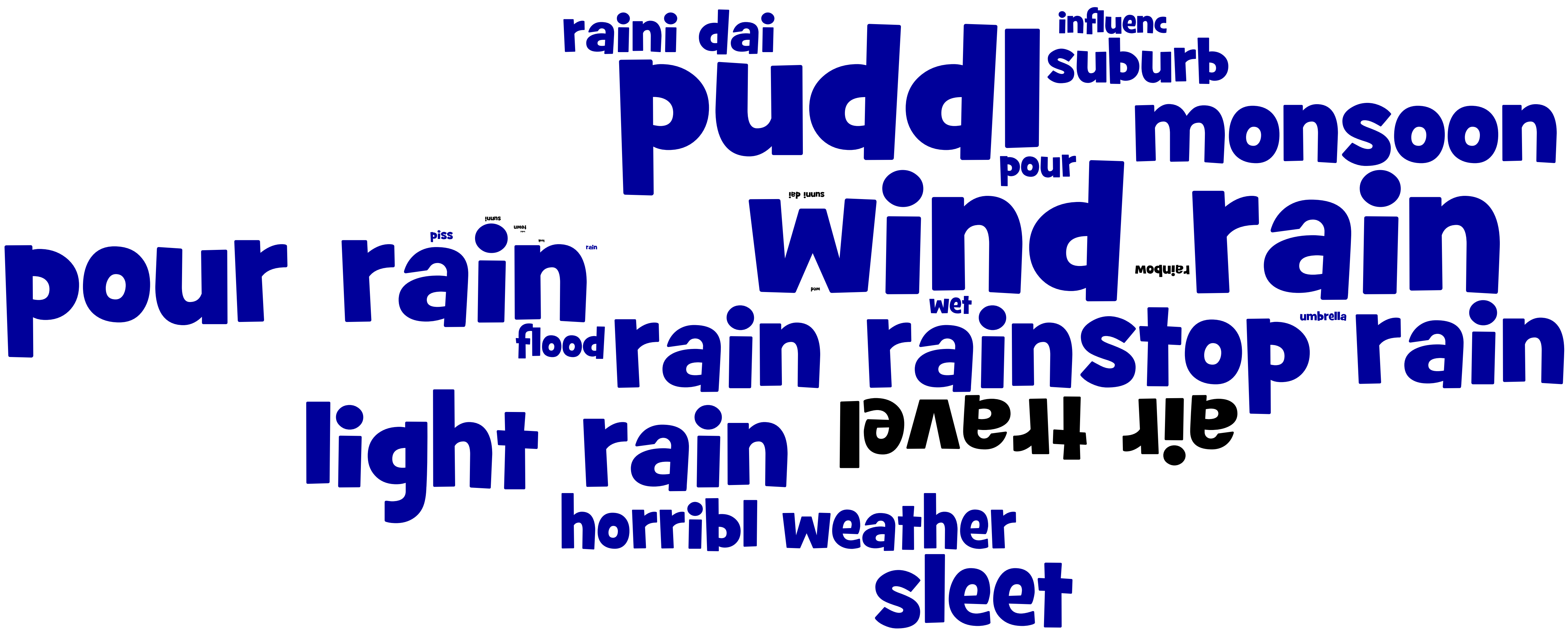


(d) *Reading* – RMSE: 1.695



(e) *Stoke-on-Trent* – RMSE: 2.815

Rainfall rates from Twitter: Selected features

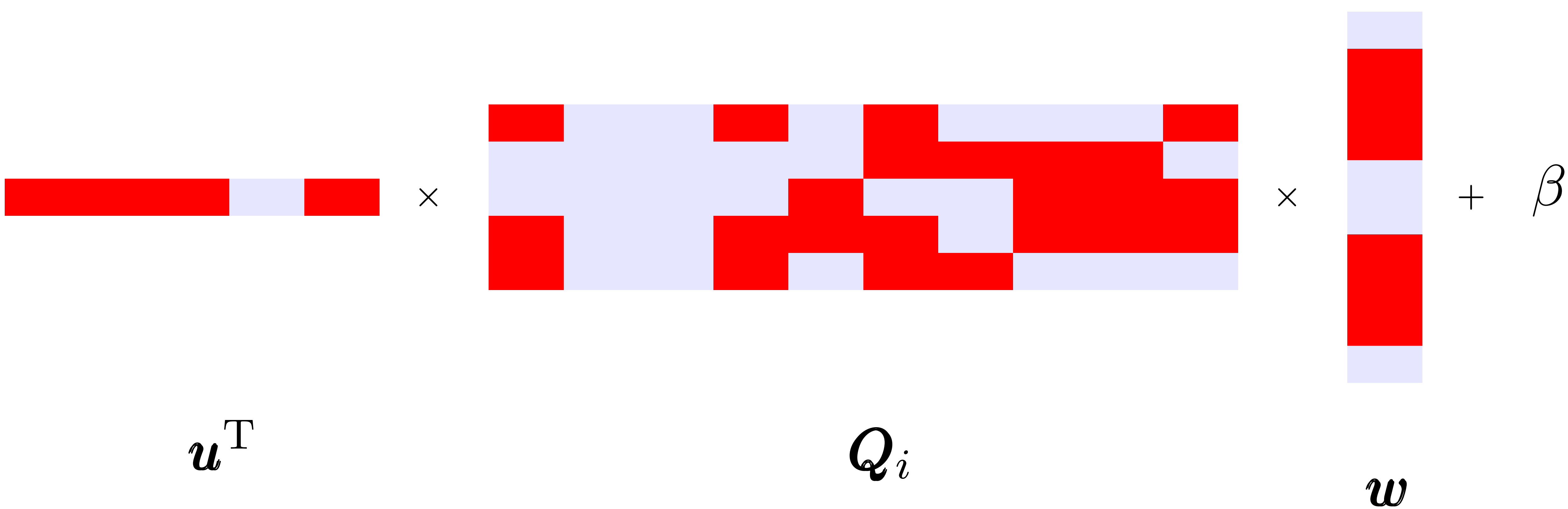


Word cloud with selected n-grams. Font size is proportional to the regression's weight; n-grams that are upside-down have a negative weight.

Bilinear regression

- users $p \in \mathbb{Z}^{+}$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, \ldots, n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, \ldots, n\}$ — $\boldsymbol{y}$
- weights, bias $u_k, w_j, \beta \in \mathbb{R}$, $k \in \{1, \ldots, p\}$, $j \in \{1, \ldots, m\}$ — $\boldsymbol{u}$, $\boldsymbol{w}$, β

$$f\left(\boldsymbol{Q}_i\right) = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta$$



Bilinear regularised regression

• users	$p \in \mathbb{Z}^+$			
• observations	$\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{\mathcal{X}}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$u_k, w_j, \beta \in \mathbb{R}$,	$k \in \{1, ..., p\}$ $j \in \{1, ..., m\}$	—	$\boldsymbol{u}$, $\boldsymbol{w}$, β

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \psi(\boldsymbol{u}, \theta_u) + \psi(\boldsymbol{w}, \theta_w) \right\}$$

$\psi(\cdot)$: **regularisation function** with a set of hyper-parameters (θ)

- if $\psi(\boldsymbol{v}, \lambda) = \lambda \|\boldsymbol{v}\|_{\ell_1}$ Bilinear Lasso
- if $\psi(\boldsymbol{v}, \lambda_1, \lambda_2) = \lambda_1 \|\boldsymbol{v}\|_{\ell_2}^2 + \lambda_2 \|\boldsymbol{v}\|_{\ell_1}$ Bilinear Elastic Net (**BEN**)

(*Lampos, Preotiuc-Pietro & Cohn, 2013*)

Bilinear elastic net (BEN): *training a model*

BEN's objective function

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}_i\boldsymbol{w}+\beta-y_i\right)^2 + \lambda_{u_1}\|\boldsymbol{u}\|_{\ell_2}^2+\lambda_{u_2}\|\boldsymbol{u}\|_{\ell_1} + \lambda_{w_1}\|\boldsymbol{w}\|_{\ell_2}^2+\lambda_{w_2}\|\boldsymbol{w}\|_{\ell_1}\right\}$$

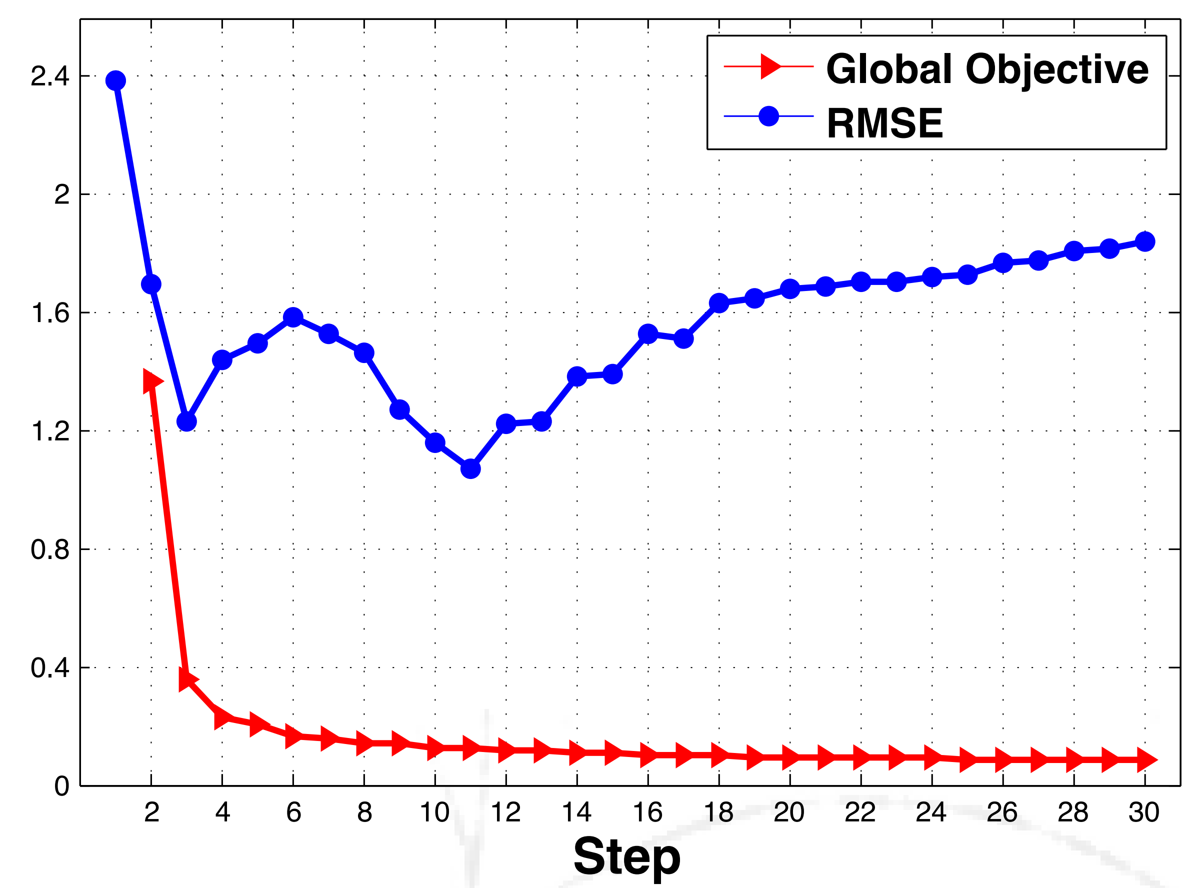
Biconvex problem

+ fix ***u***, learn ***w*** and vice versa
+ iterate through convex optimisation tasks

Large-scale solvers in SPAMS (*Mairal et al., 2010*)

Global objective function during training (***red***)

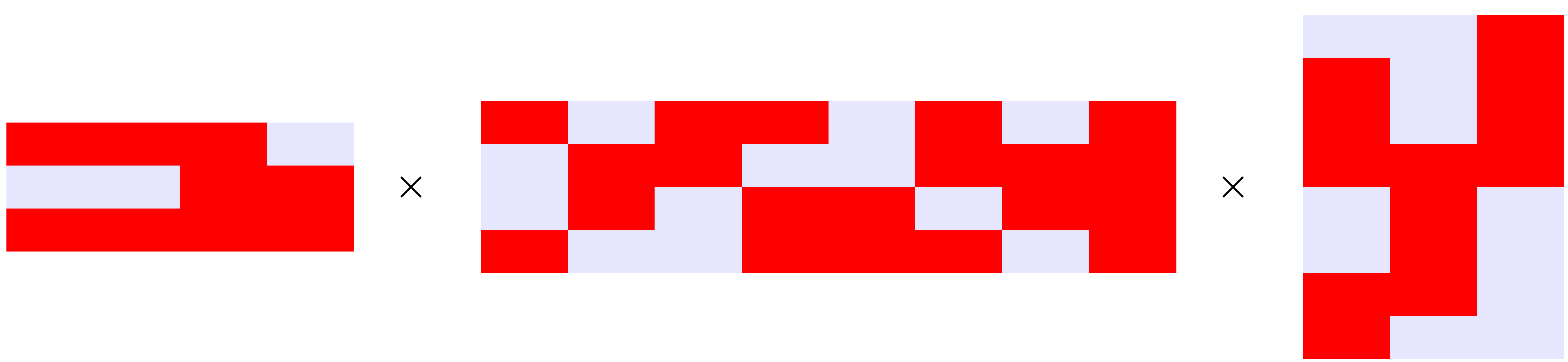
Corresponding prediction error on held out data (***blue***)



Bilinear multi-task learning

- tasks $\tau \in \mathbb{Z}^{+}$
- users $p \in \mathbb{Z}^{+}$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, \ldots, n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $\boldsymbol{y}_i \in \mathbb{R}^{\tau}$, $i \in \{1, \ldots, n\}$ — $\boldsymbol{Y}$
- weights, bias $\boldsymbol{u}_k, \boldsymbol{w}_j, \boldsymbol{\beta} \in \mathbb{R}^{\tau}$, $k \in \{1, \ldots, p\}$, $j \in \{1, \ldots, m\}$ — $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{\beta}$

$$f\left(\boldsymbol{Q}_i\right) = \operatorname{tr}\left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{W}\right) + \boldsymbol{\beta}$$

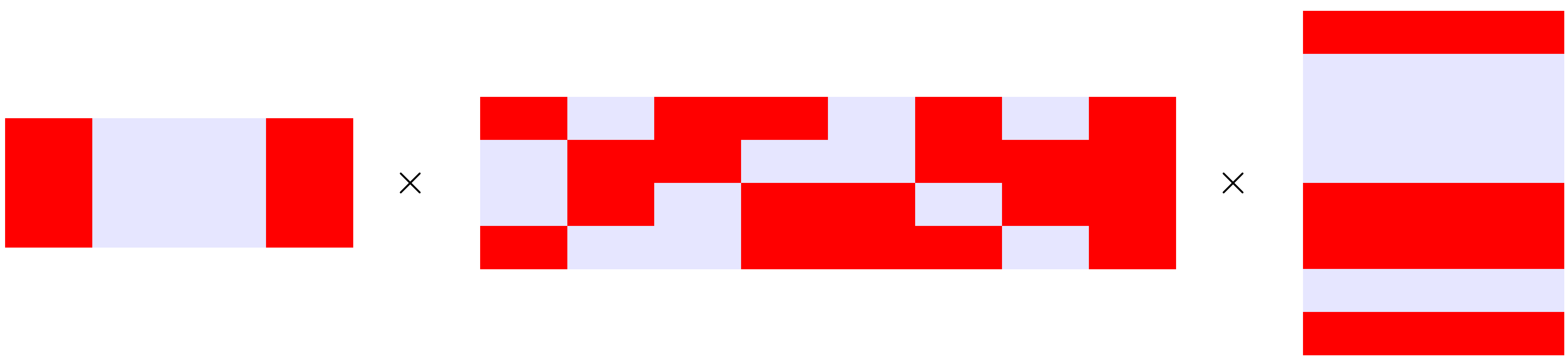


$\boldsymbol{U}^{\mathrm{T}}$ $\quad$ $\boldsymbol{Q}_i$ $\quad$ $\boldsymbol{W}$

Bilinear Group $\ell_{2,1}$ (BGL)

$$\underset{\boldsymbol{U},\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\boldsymbol{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_2 \right\}$$

(*Argyriou et al., 2008*)



$\boldsymbol{U}^{\mathrm{T}}$ $\boldsymbol{Q}_i$ $\boldsymbol{W}$

+ a feature (user or word) is usually **selected** (activated) **for all tasks,** but with different weights
+ useful in the domain of ***political preference inference***

Inferring voting intention from Twitter: Data

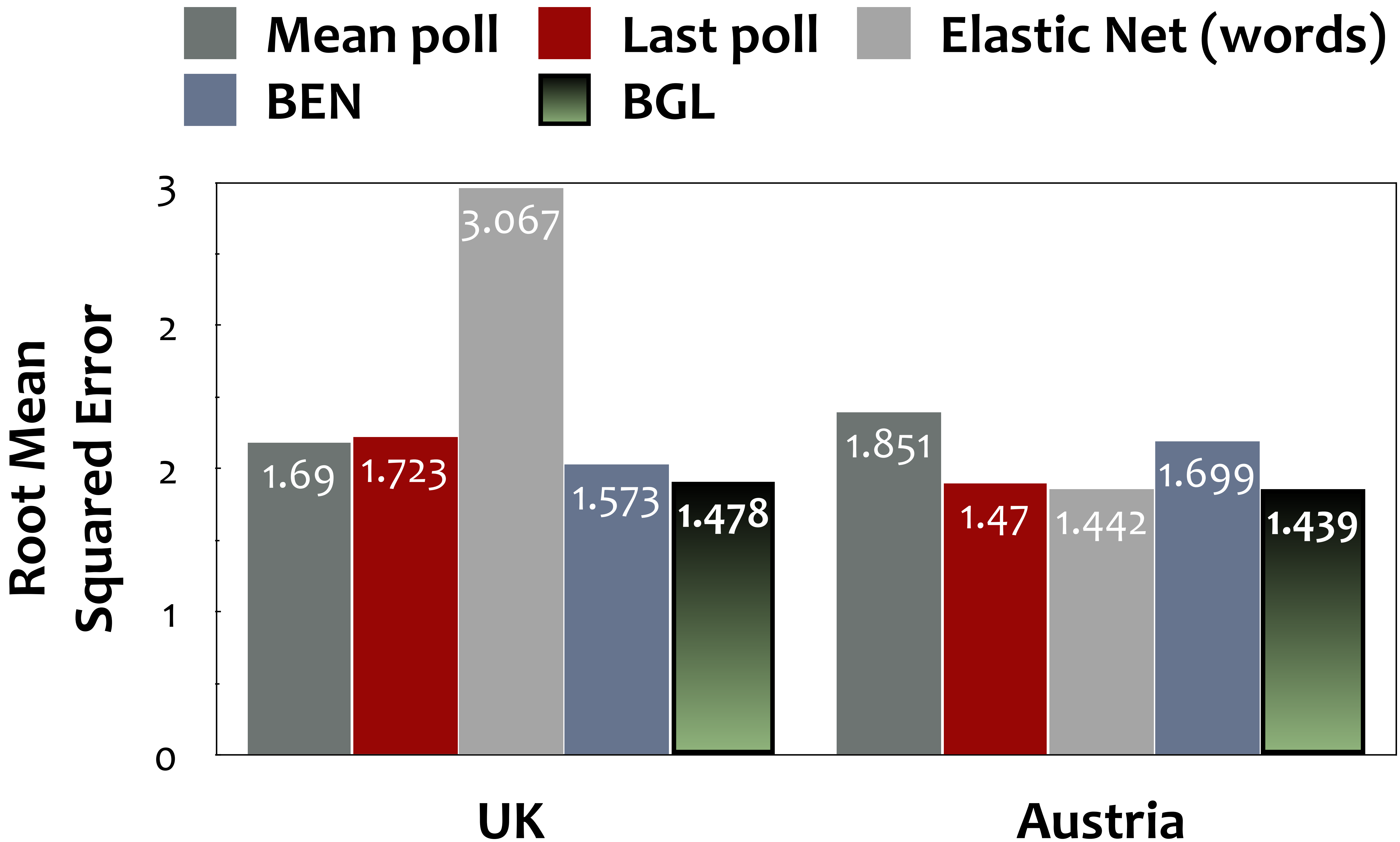
United Kingdom

+ 3 parties (Conservatives, Labour, Lib Dem)
+ **42,000** Twitter **users** distributed proportionally to UK's regional population figures
+ **60 million** tweets & **80,976** 1-grams extracted
+ 240 polls from 30 Apr. 2010 to 13 Feb. 2012

Austria

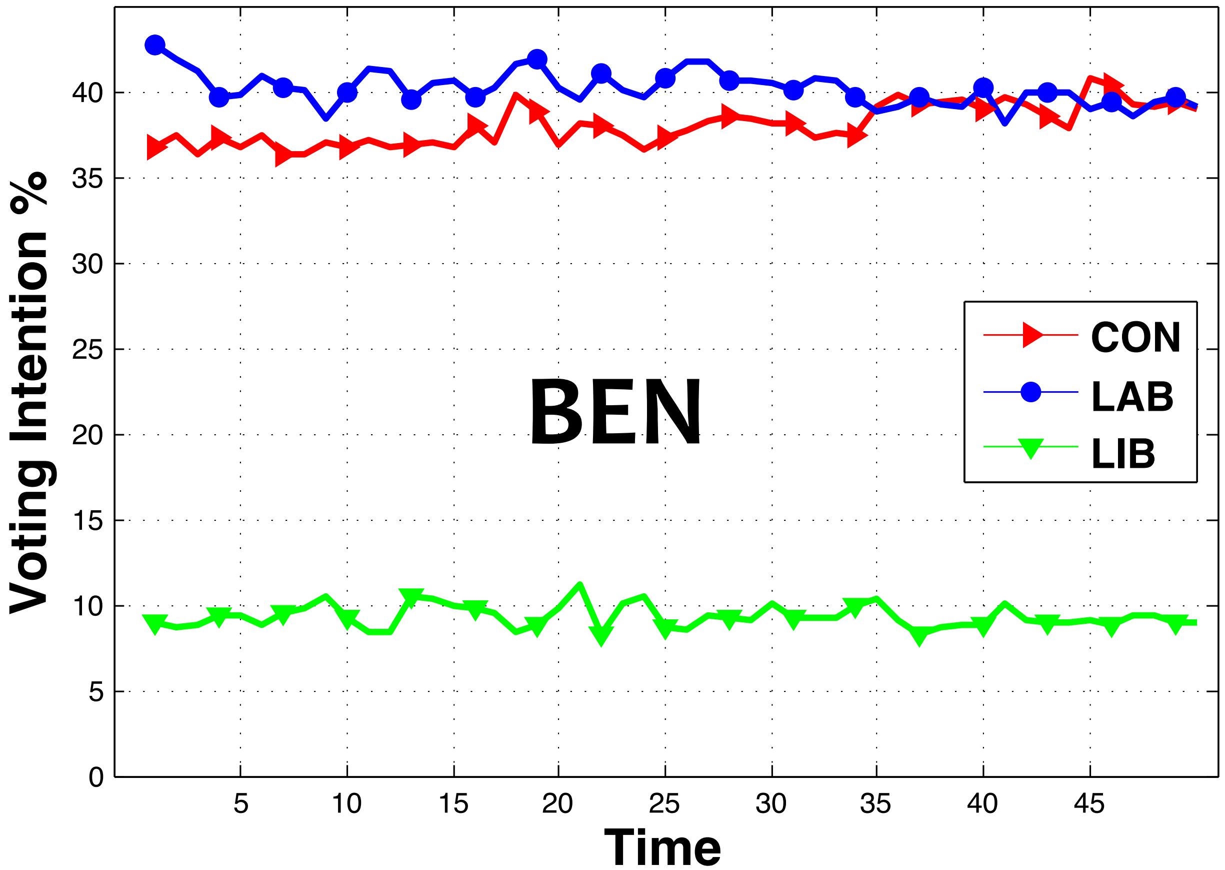
+ 4 parties (SPO, OVP, FPO, GRU)
+ **1,100** politically active Twitter **users** selected by political scientists
+ **800,000** tweets & **22,917** 1-grams extracted
+ 98 polls from 25 Jan. to 25 Dec. 2012

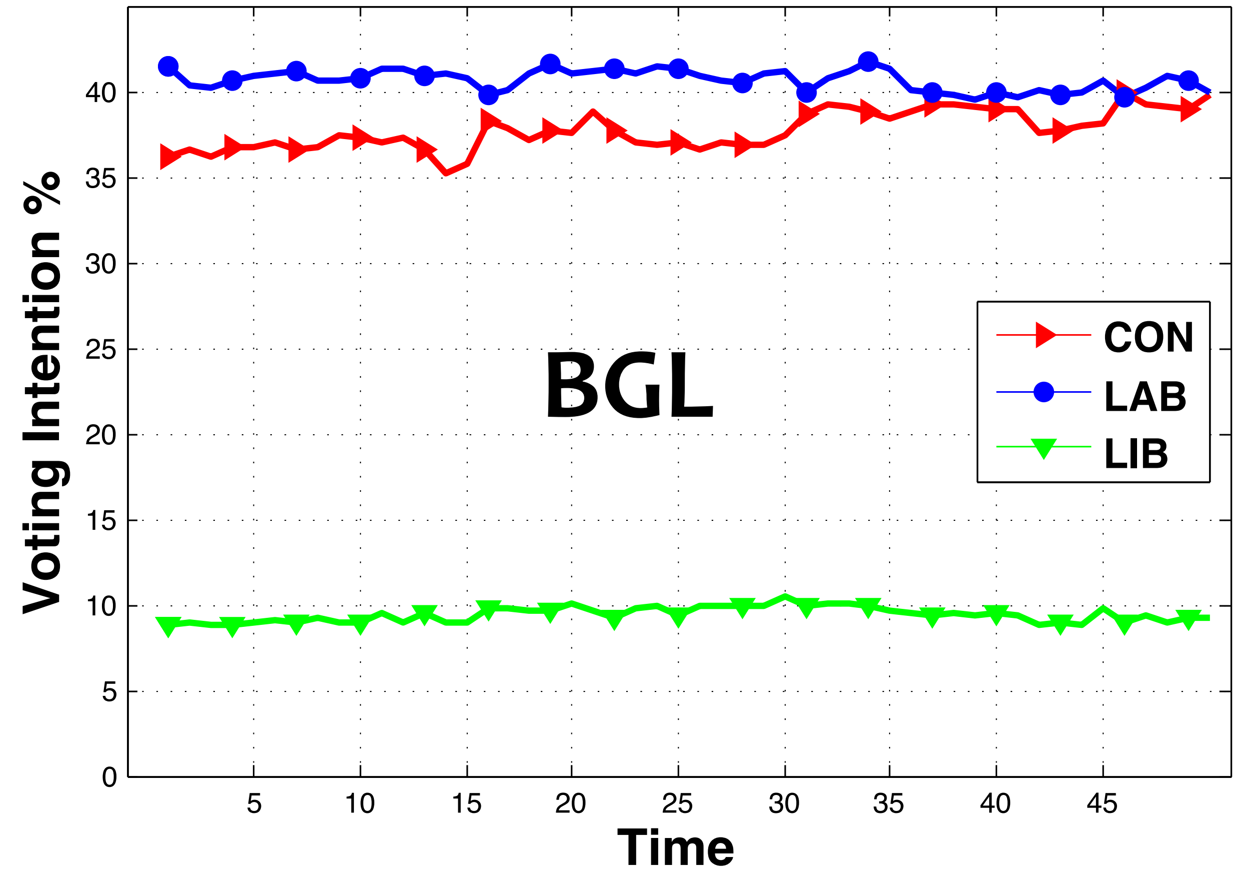
Inferring voting intention from Twitter: Performance

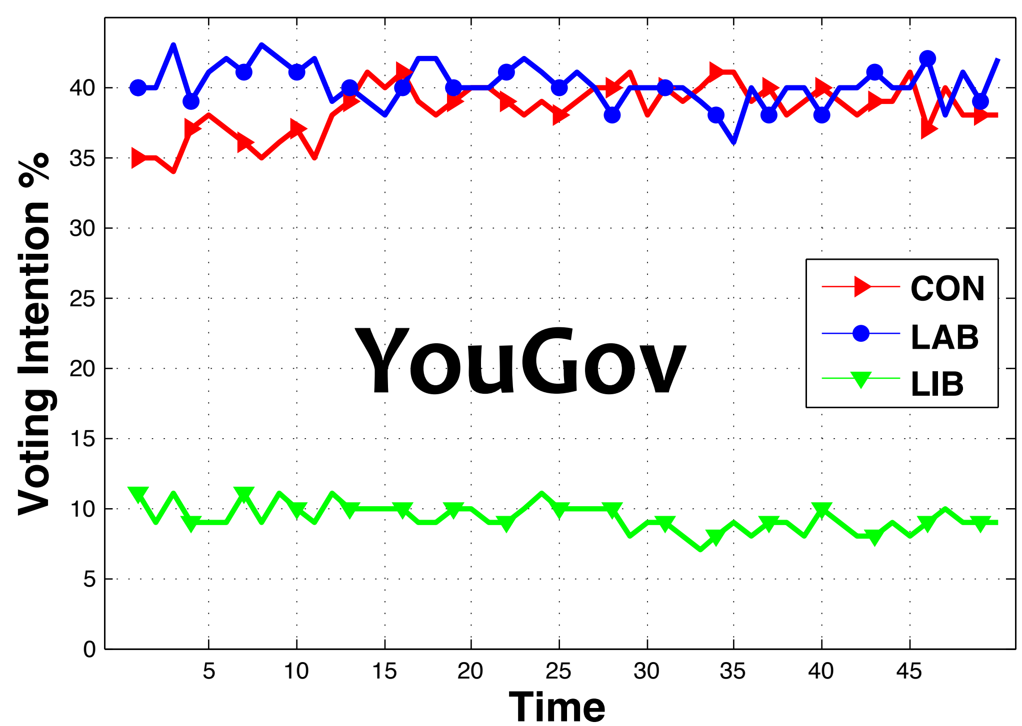


(*Lampos, Preotiuc-Pietro & Cohn, 2013*)

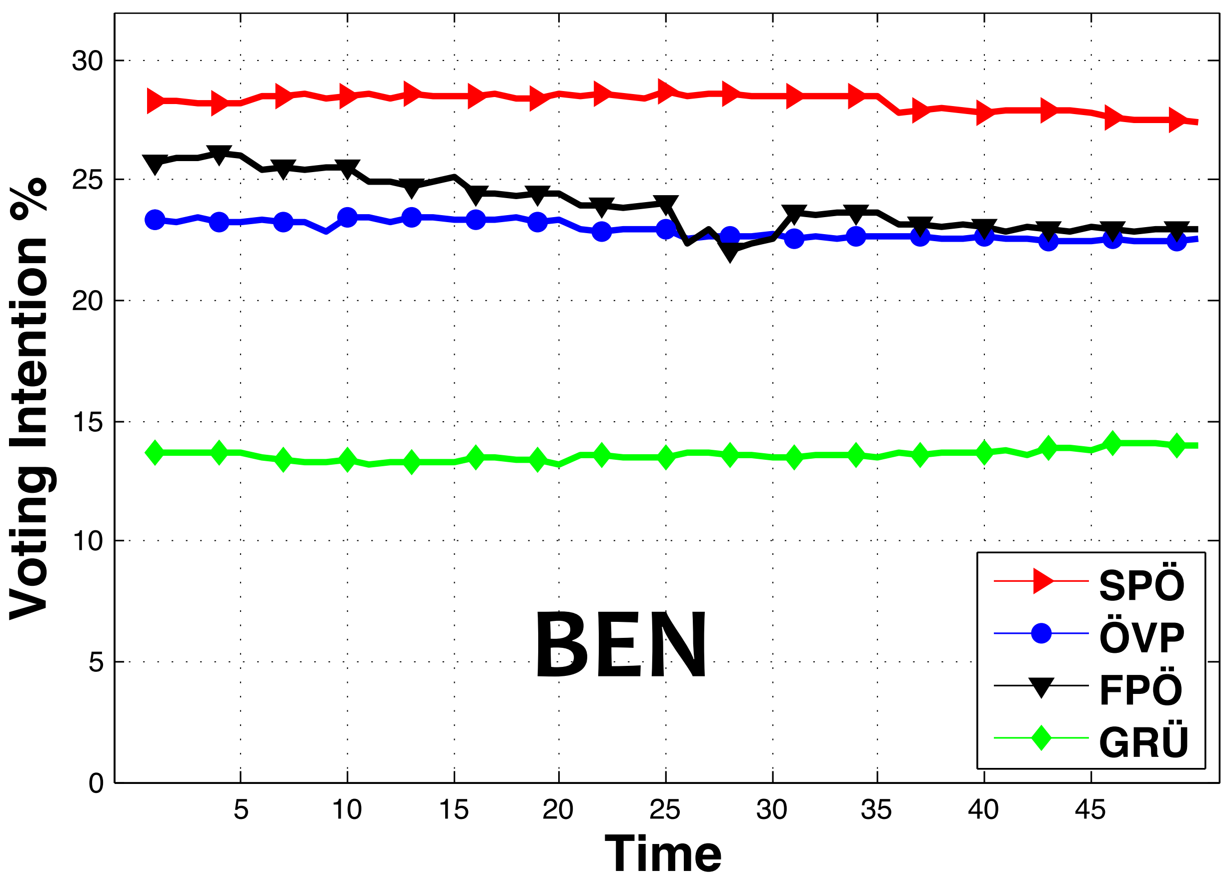
Inferring voting intention from Twitter: UK

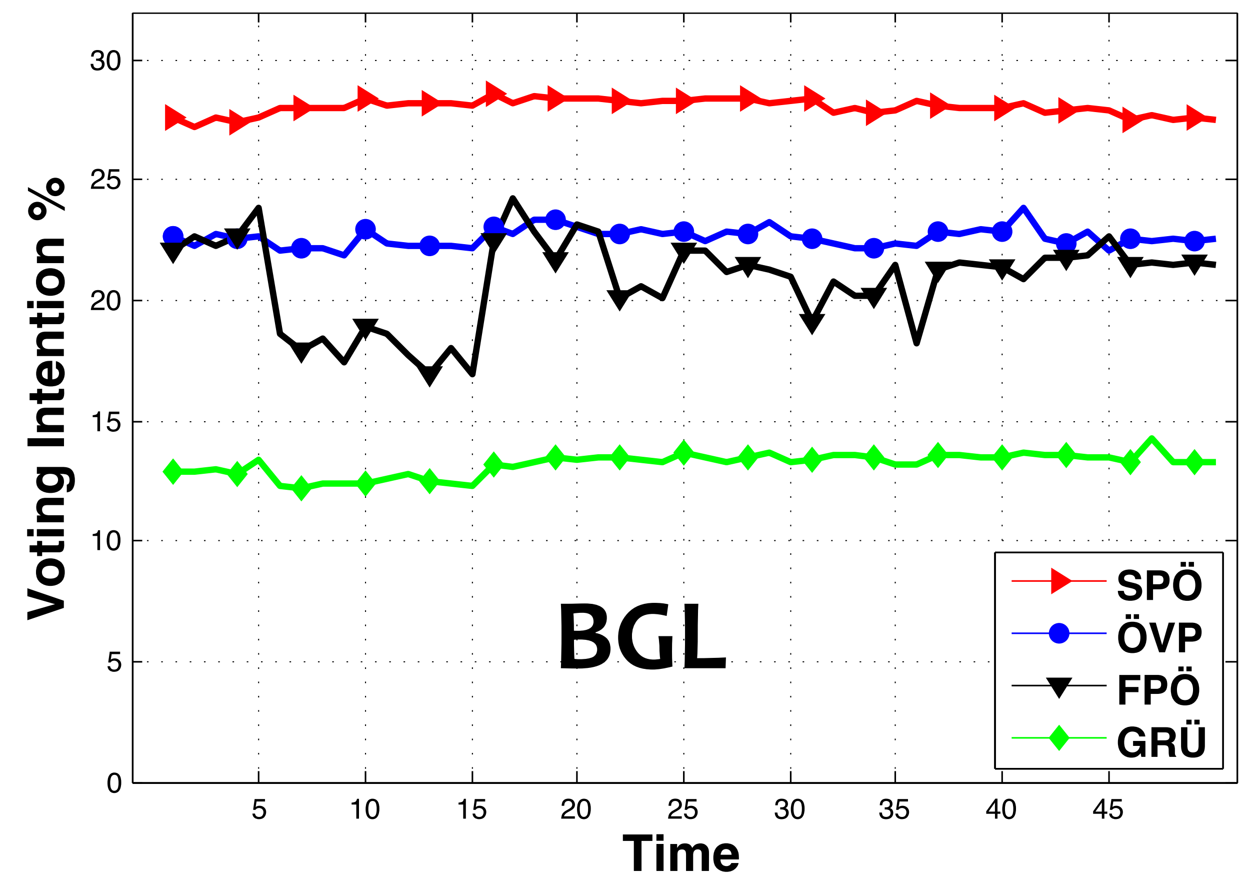


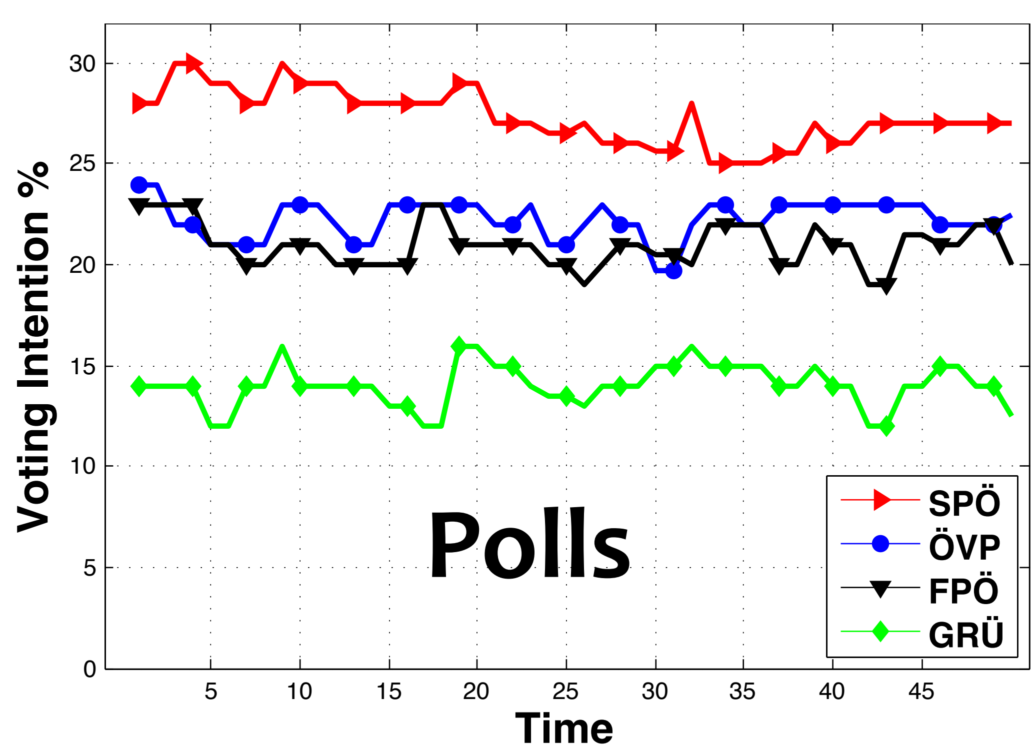




Inferring voting intention from Twitter: Austria



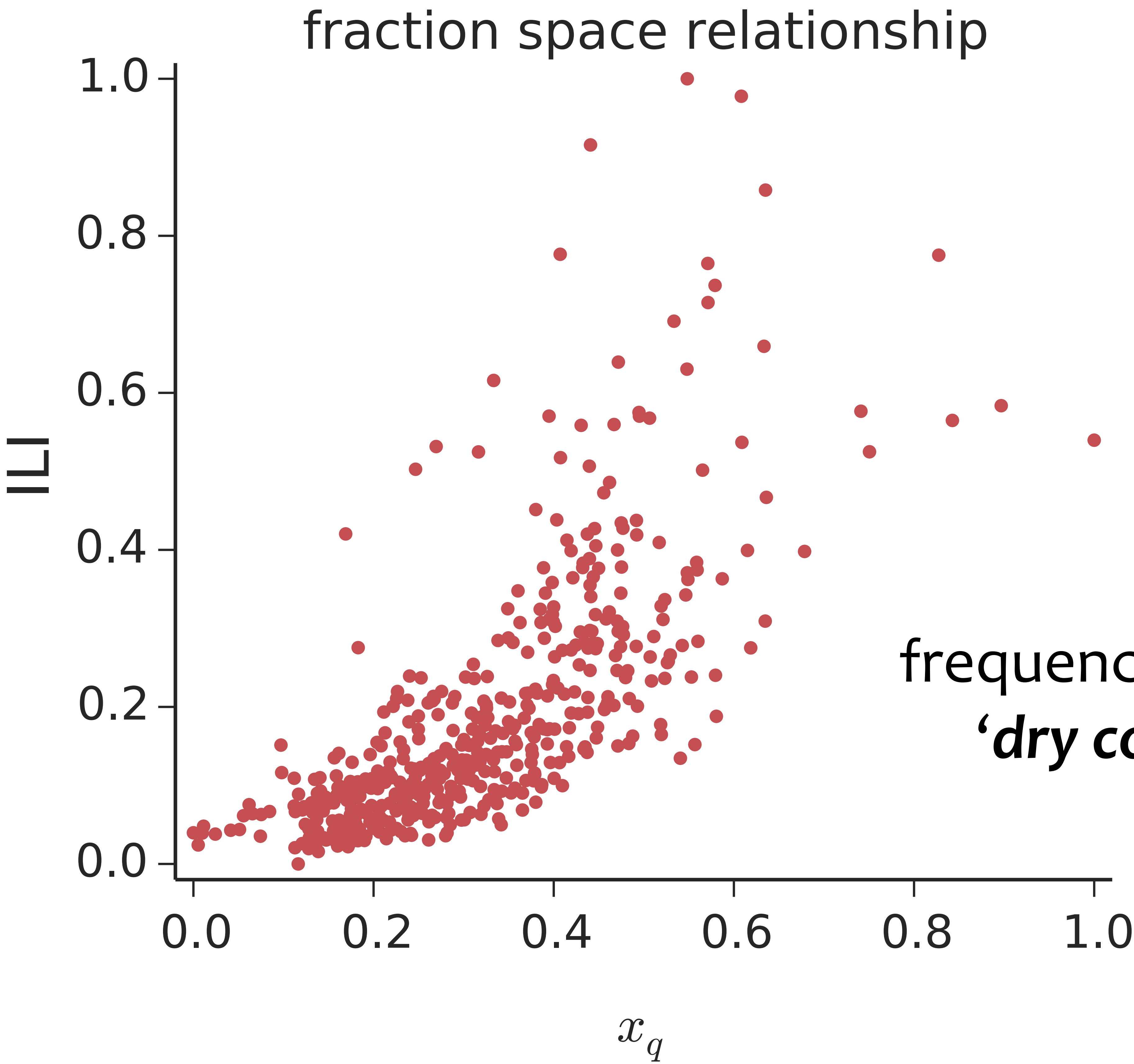




Inferring voting intention from Twitter: Qualitative outcomes

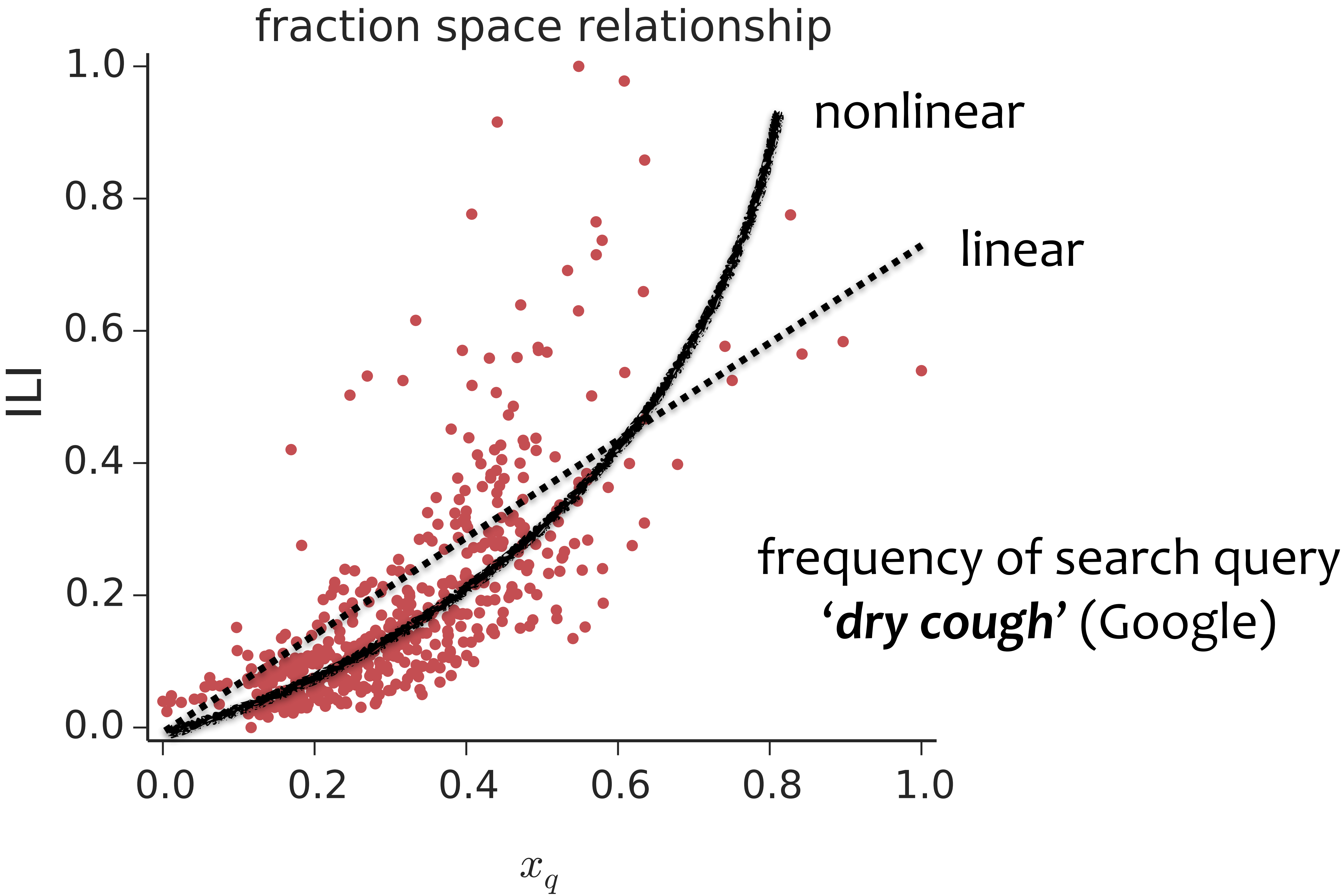
Party	Tweet	Score	User type
SPÖ *centre*	*Inflation rate in Austria slightly down in July from 2.2 to 2.1%. Accommodation, Water, Energy more expensive.*	**0.745**	Journalist
ÖVP *centre right*	*Can really recommend the book "Res Publica" by Johannes #Voggenhuber! Food for thought and so on #Europe #Democracy*	**-2.323**	User
FPÖ ***far right***	*Campaign of the Viennese SPO on "Living together" plays right into the hands of right-wing populists*	**-3.44**	Human rights
GRÜ *centre left*	*Protest songs against the closing-down of the bachelor course of International Development: <link> #ID_remains #UniBurns #UniRage*	**1.45**	Student Union

Nonlinearities in the data (1)



frequency of search query
'dry cough' (Google)

Nonlinearities in the data (2)



Gaussian Processes (GPs)

Based on d-dimensional input data $\boldsymbol{x} \in \mathbb{R}^d$

we want to learn a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$

$$\boxed{f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))}$$

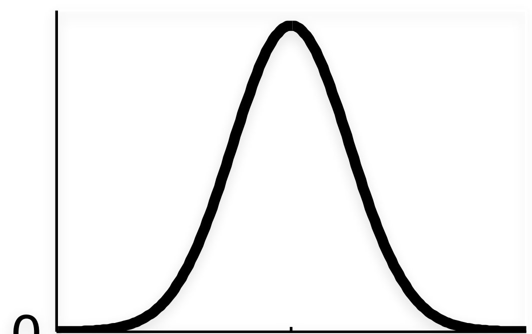
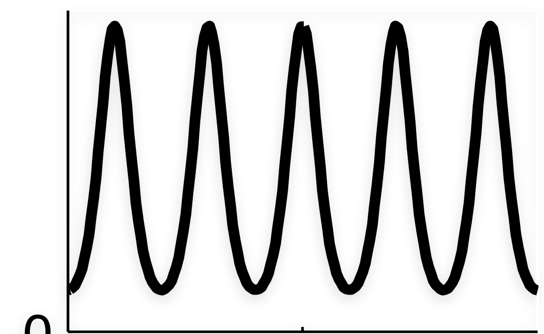
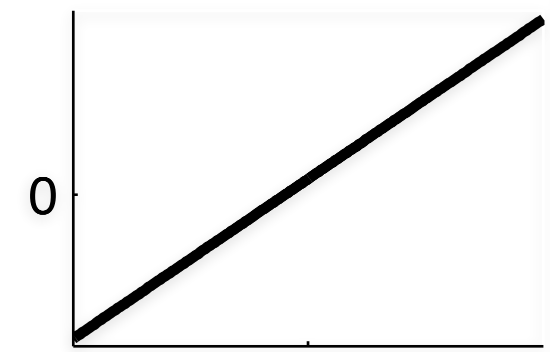
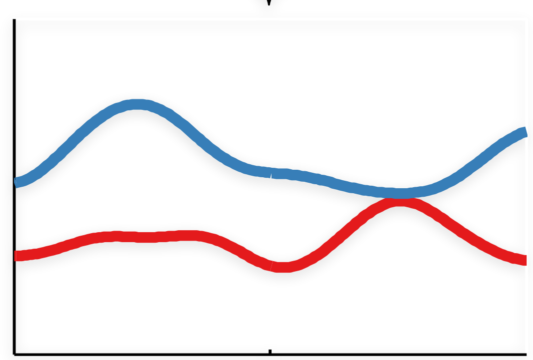
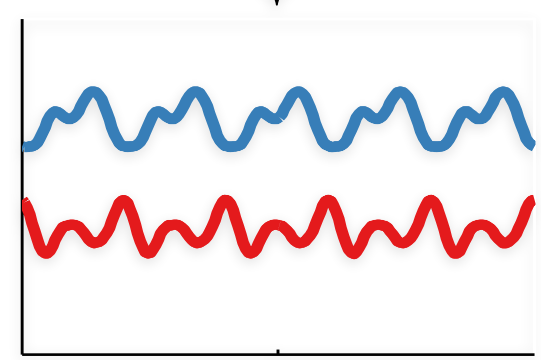
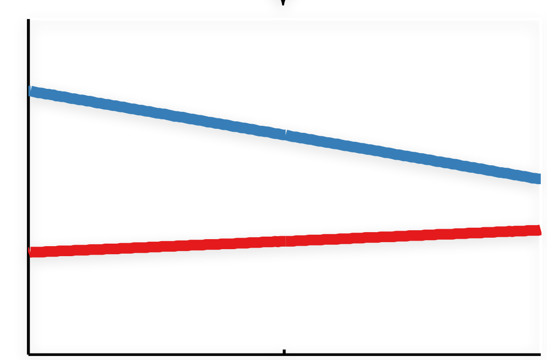
mean function
drawn on inputs

covariance function (or kernel)
drawn on pairs of inputs

<u>Formally</u>: *Sets of random variables any finite number of which have a **multivariate Gaussian distribution***

(*Rasmussen & Williams, 2006*)

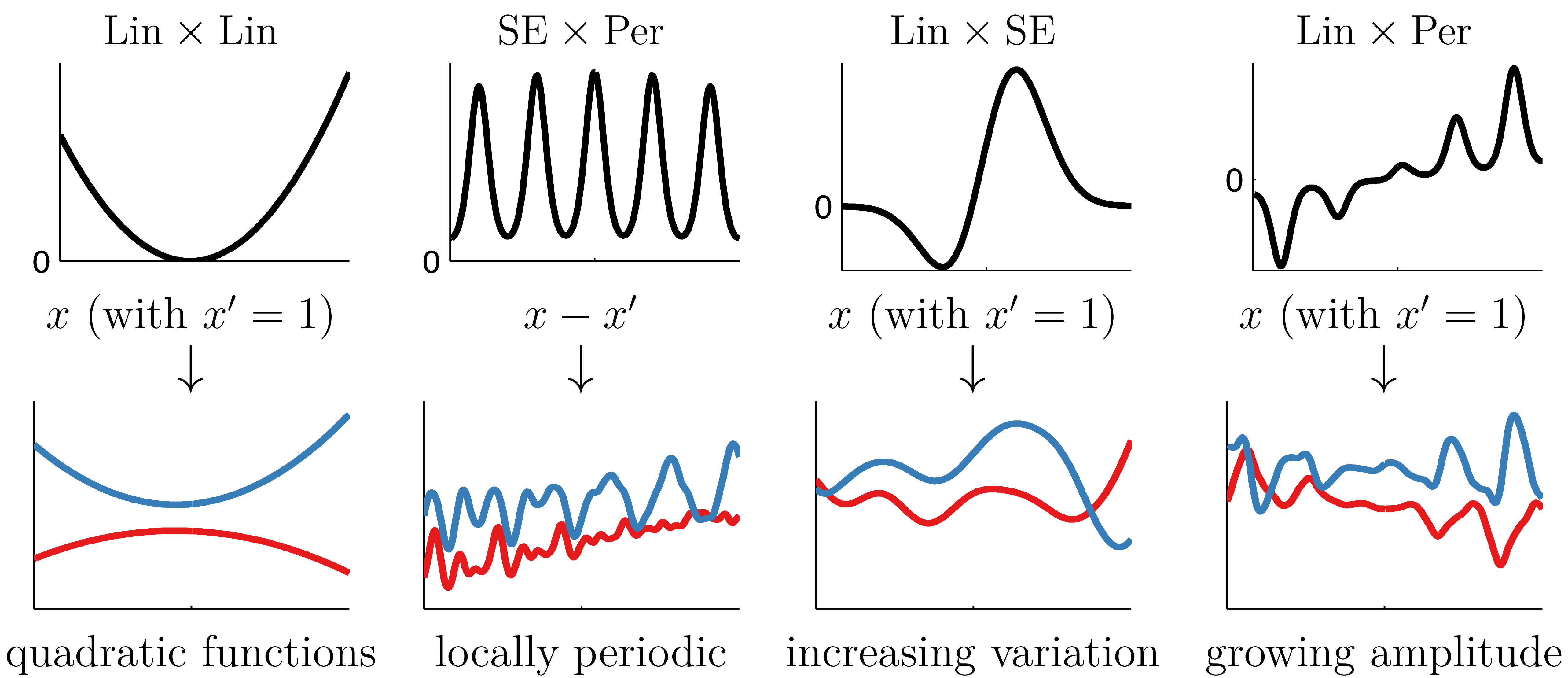
Common covariance functions (kernels)

Kernel name:	Squared-exp (SE)	Periodic (Per)	Linear (Lin)
$k(x, x') =$	$\sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$	$\sigma_f^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{x-x'}{p}\right)\right)$	$\sigma_f^2(x-c)(x'-c)$
Plot of $k(x, x')$:	 $x - x'$	 $x - x'$	 x (with $x' = 1$)
	$\downarrow$	$\downarrow$	$\downarrow$
Functions $f(x)$ sampled from GP prior:	 x	 x	 x
Type of structure:	local variation	repeating structure	linear functions

(Duvenaud, 2014)

Combining kernels in a GP

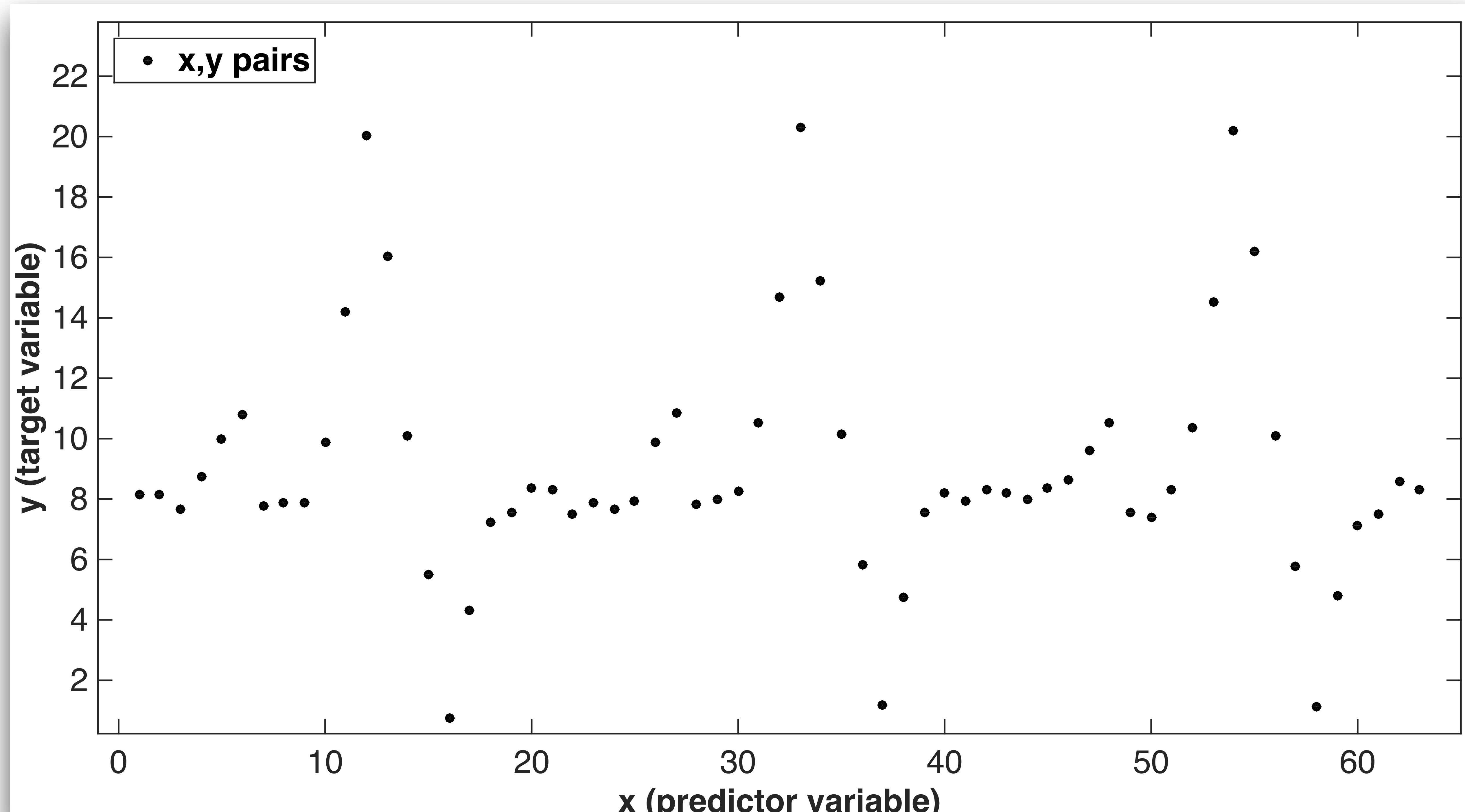
it is possible to **add** or **multiply** kernels
(among other operations)



(*Duvenaud, 2014*)

GPs for regression: A toy example (1)

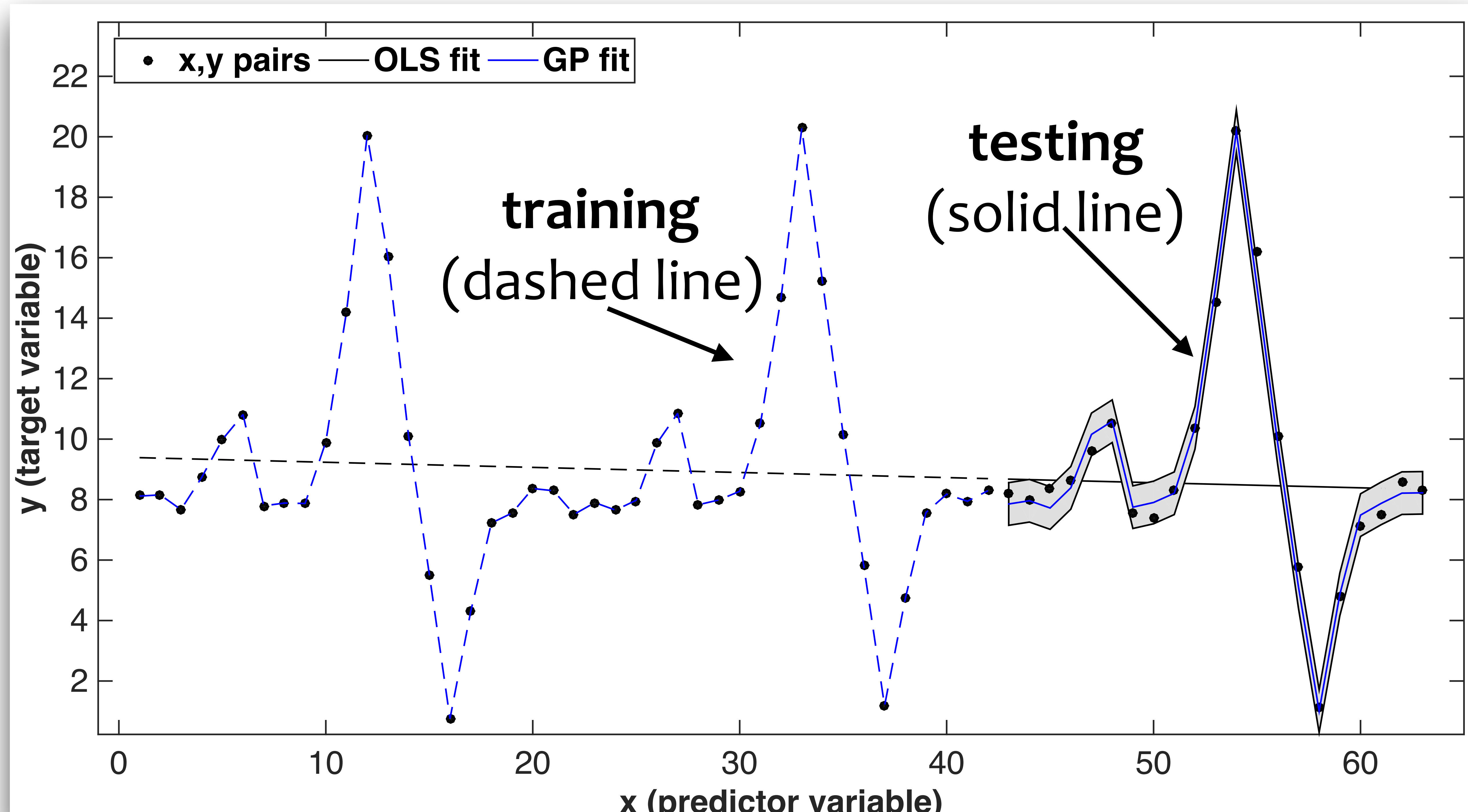
take some (x,y) pairs with some obvious *nonlinear* underlying structure



GPs for regression: A toy example (2)

Addition of 2 GP kernels:

periodic + squared exponential + noise



More information about GPs

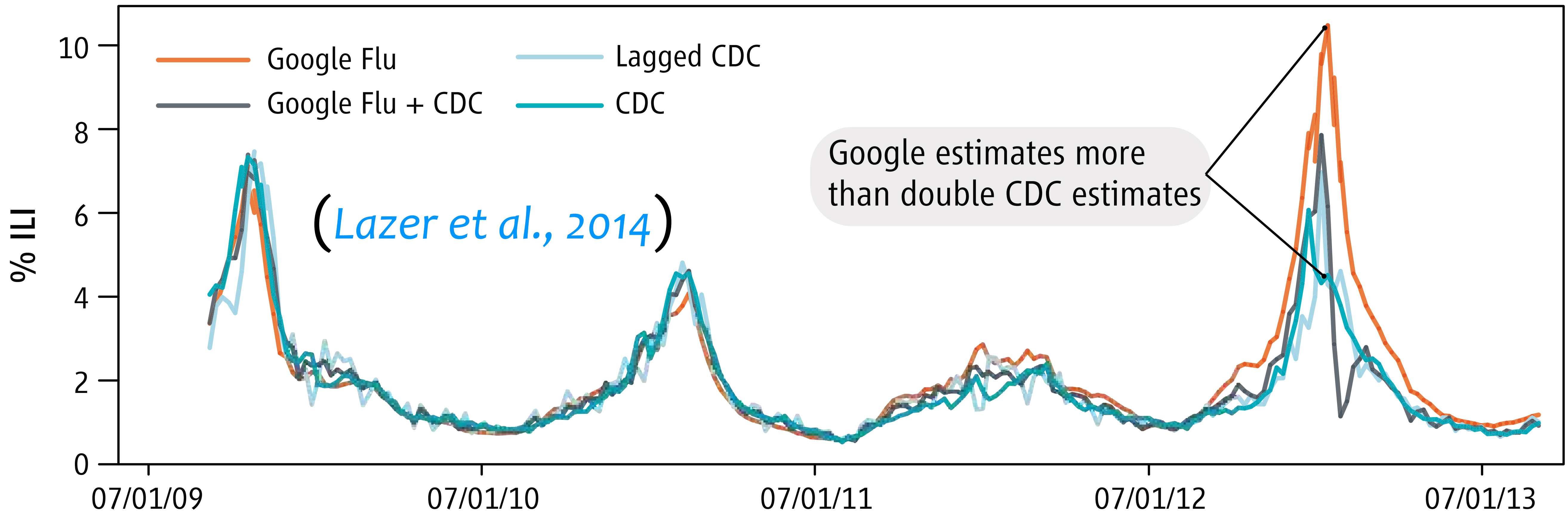
+ Book — "*Gaussian Processes for Machine Learning*"
  http://www.gaussianprocess.org/gpml/
+ Tutorial — "*Gaussian Processes for Natural Language Processing*"
  http://people.eng.unimelb.edu.au/tcohn/tutorial.html
+ Video-lecture — "*Gaussian Process Basics*"
  http://videolectures.net/gpip06_mackay_gpb/
+ Software I — *GPML* for Octave or MATLAB
  http://www.gaussianprocess.org/gpml/code
+ Software II — *GPy* for Python
  http://sheffieldml.github.io/GPy/

Google Flu Trends: The idea

Can we **turn search query information** (statistics) to estimates about the **rate of *influenza-like illness*** in the real-world population?

Google Flu Trends: Failure

$$\boldsymbol{logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon}$$ (Ginsberg et al., 2009)



The estimates of the online Google Flu Trends tool were approx. ***two times larger*** *than the ones from the CDC in 2012/13*

Google Flu Trends: Hypotheses for failure

+ ‘**Big Data**’ are not always good enough; may not always capture the target signal properly
+ The estimates were based on a rather **simplistic model**
+ The model was OK, but some **spurious search queries** invalidated the ILI inferences, e.g. ‘flu symptoms’
+ **Media hype** about the topic of ‘flu’ significantly increased the search query volume from people that were just seeking information (non patients)
+ **Side note**: *CDC’s estimates are* ***not*** *necessarily the ground truth*; they can also go wrong sometimes, although we generally assume that they are a good representation of the real signal

Google Flu Trends revised: *Data* (1)

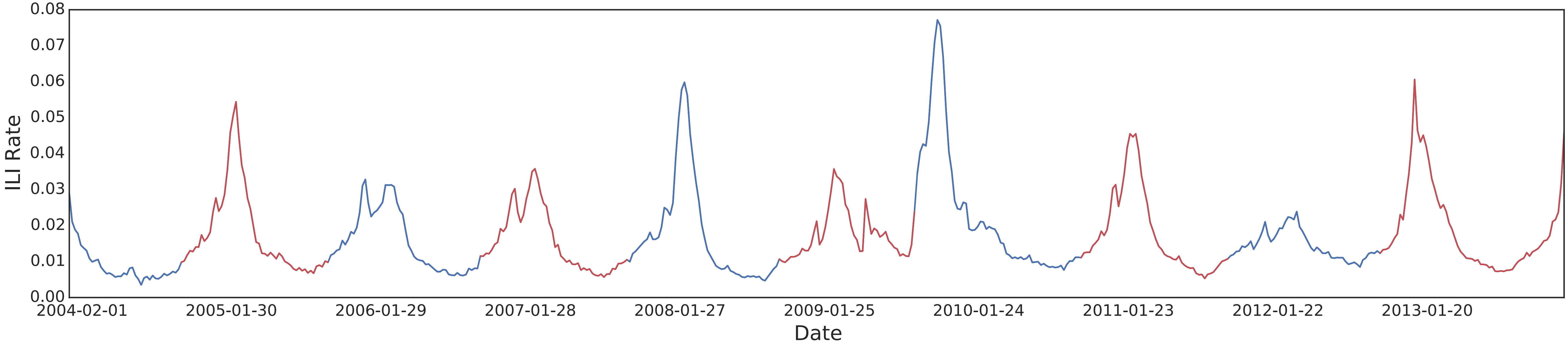
Google **search query logs**

> geo-located in **US** regions
> from 4 Jan. 2004 to 28 Dec. 2013 (521 weeks, ~**decade**)
> filtered by a *very* relaxed health-topic classifier
> intersection among frequently occurring search queries in all US regions
> weekly frequencies of **49,708 queries** (# of features)
> all data have been anonymised and aggregated

plus **corresponding ILI rates from the CDC**

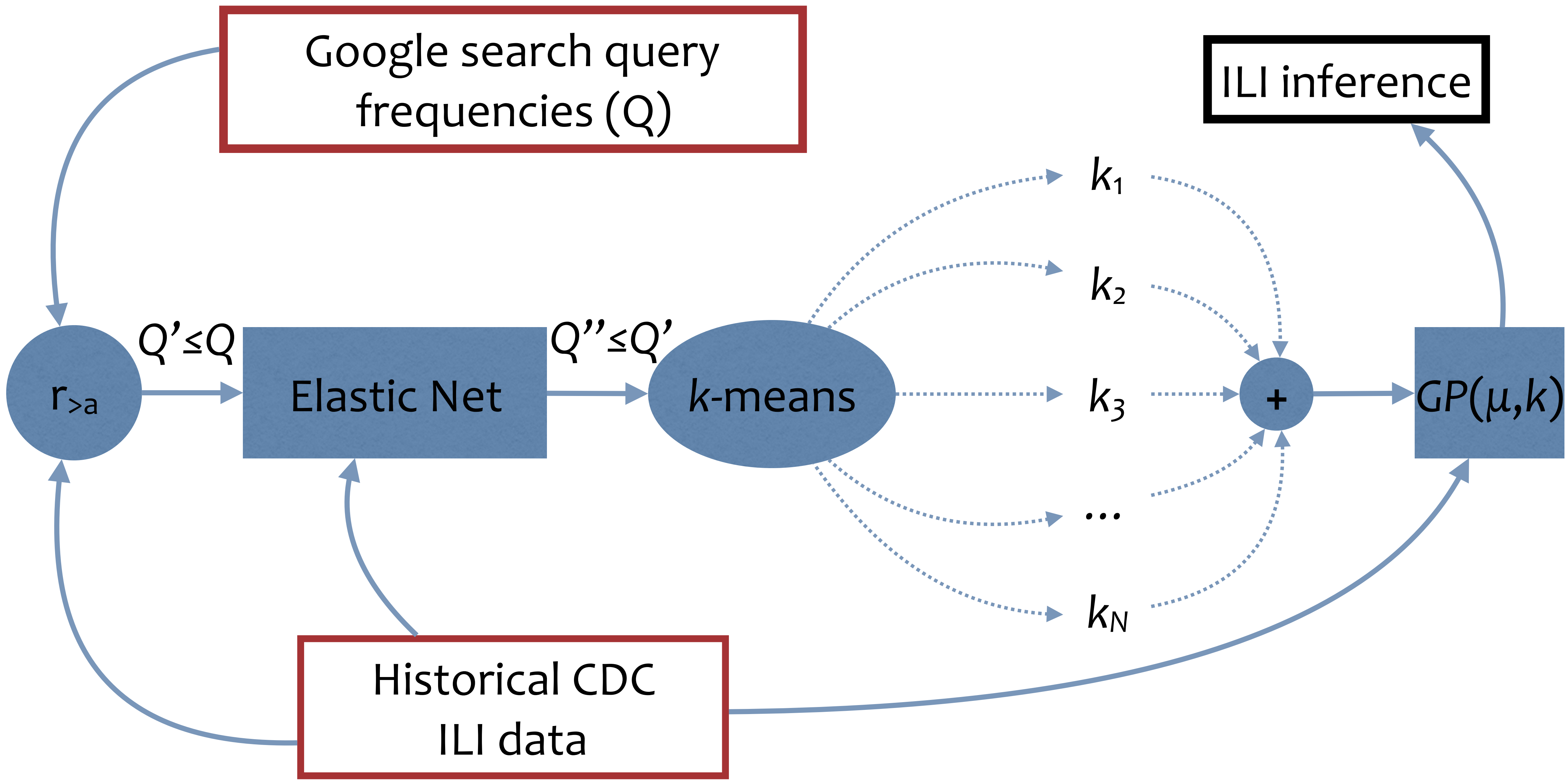
Google Flu Trends revised: *Data* (2)

Corresponding ILI rates from the CDC



different colouring per flu season

Google Flu Trends revised: *Methods* (1)



(*Lampos, Miller, Crossan & Stefansen, 2015*)

Google Flu Trends revised: *Methods* (2)

1. Keep search queries with $r \geq 0.5$ (*reduces the amount of irrelevant queries*)

2. Apply the previous model (**GFT**) to get a baseline performance estimate

3. Apply **elastic net** to select a subset of search queries and compute another baseline

4. Group the selected queries into $N = 10$ **clusters** using k-means to account for their different semantics

5. Use a different **GP covariance function** on top of each query cluster to explore non-linearities

Google Flu Trends revised: *Methods* (3)

$$\boxed{k(\mathbf{x}, \mathbf{x}') = \left(\sum_{i=1}^{C} k_{\mathrm{SE}}(\mathbf{c}_i, \mathbf{c}_i')\right) + \sigma_{\mathrm{n}}^2 \cdot \delta(\mathbf{x}, \mathbf{x}')}$$

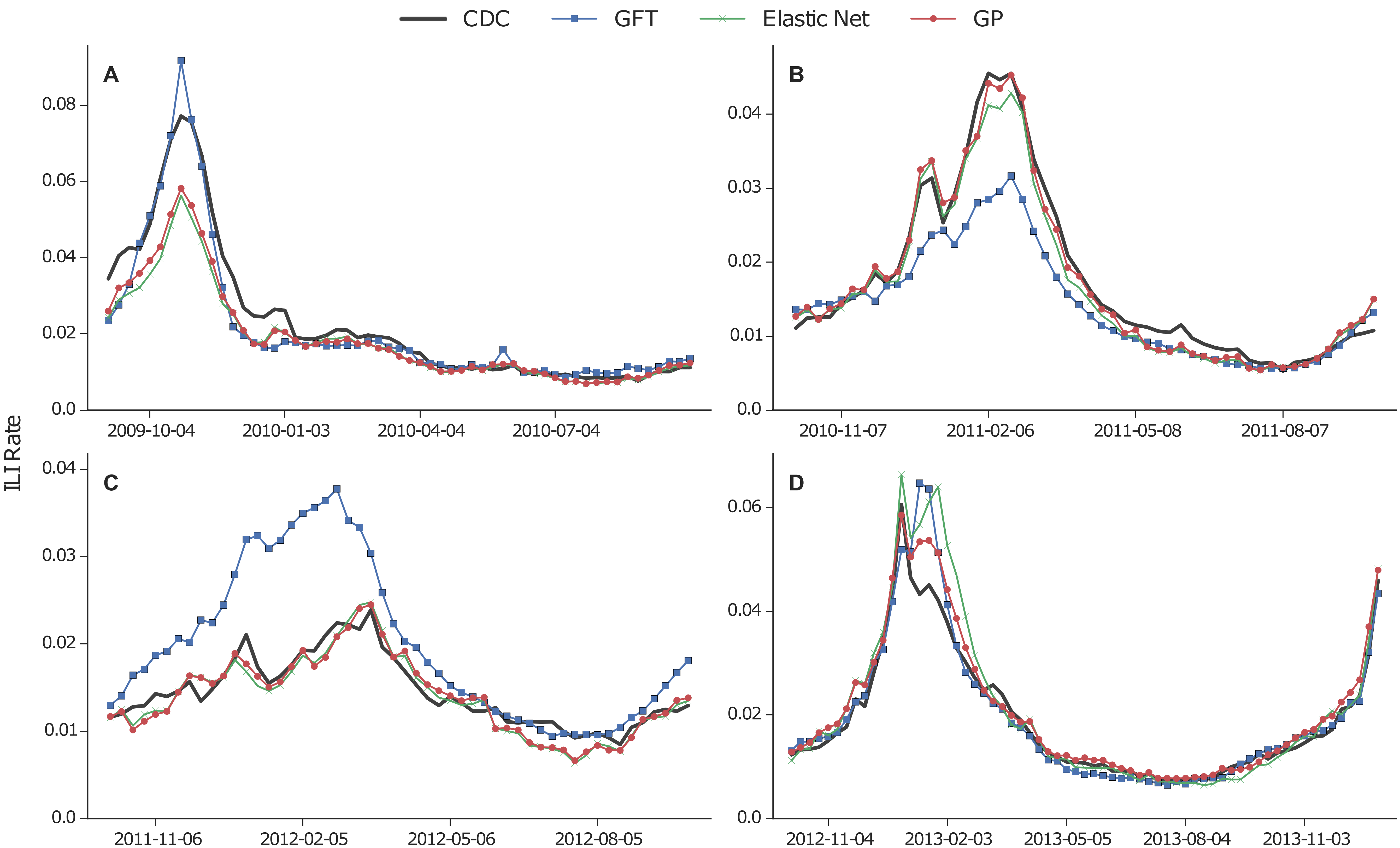
\+ **protect a model from radical changes** in the frequency of single queries that are not representative of a cluster

\+ model the **contribution of various thematic concepts** (captured by different clusters) to the final prediction

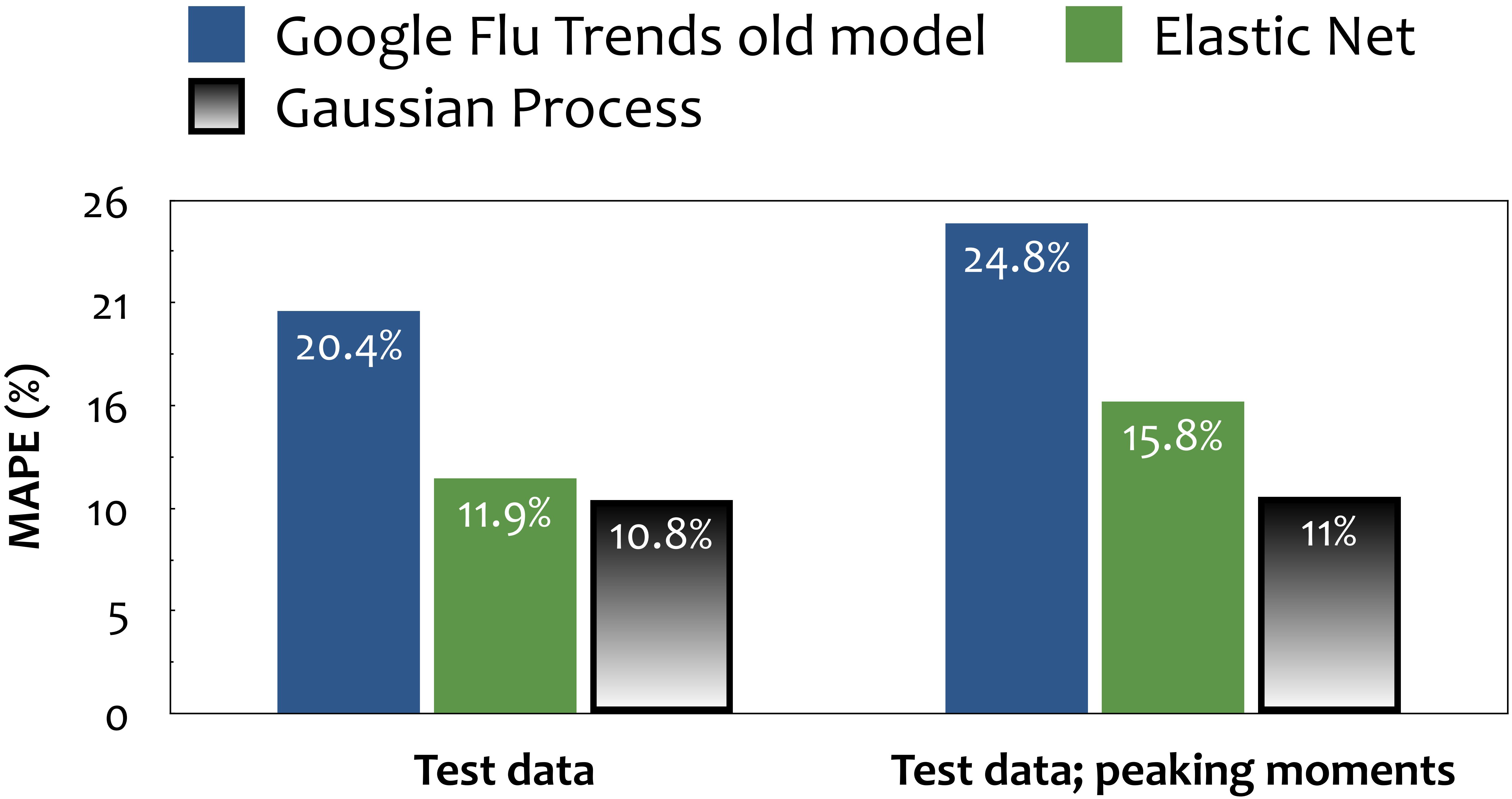
\+ learning a sum of lower-dimensional functions: significantly smaller input space, much **easier learning task,** fewer samples required, more statistical traction obtained

\- imposes the assumption that the relationship between queries in separate clusters provides no information about ILI (*reasonable trade-off*)

Google Flu Trends revised: *Results* (1)



Google Flu Trends revised: *Results* (2)



Mean absolute percentage (%) of error (MAPE) in flu rate estimates during a 5-year period (2008-2013)

Google Flu Trends revised: *Results* (3)

impact of automatically selected queries in
a flu estimate during the ***over-predictions***

previous GFT model

'rsv' —	25%
'flu symptoms' —	18%
'benzonatate' —	6%
'symptoms of pneumonia' —	6%
'upper respiratory infection' —	4%

Google Flu Trends revised: *Methods* (4)

Auto-regressive moving average with exogenous inputs (***ARMAX***)

$$\boxed{y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{D} w_i h_{t,i} + \varepsilon_t}$$

AR component

Moving average component

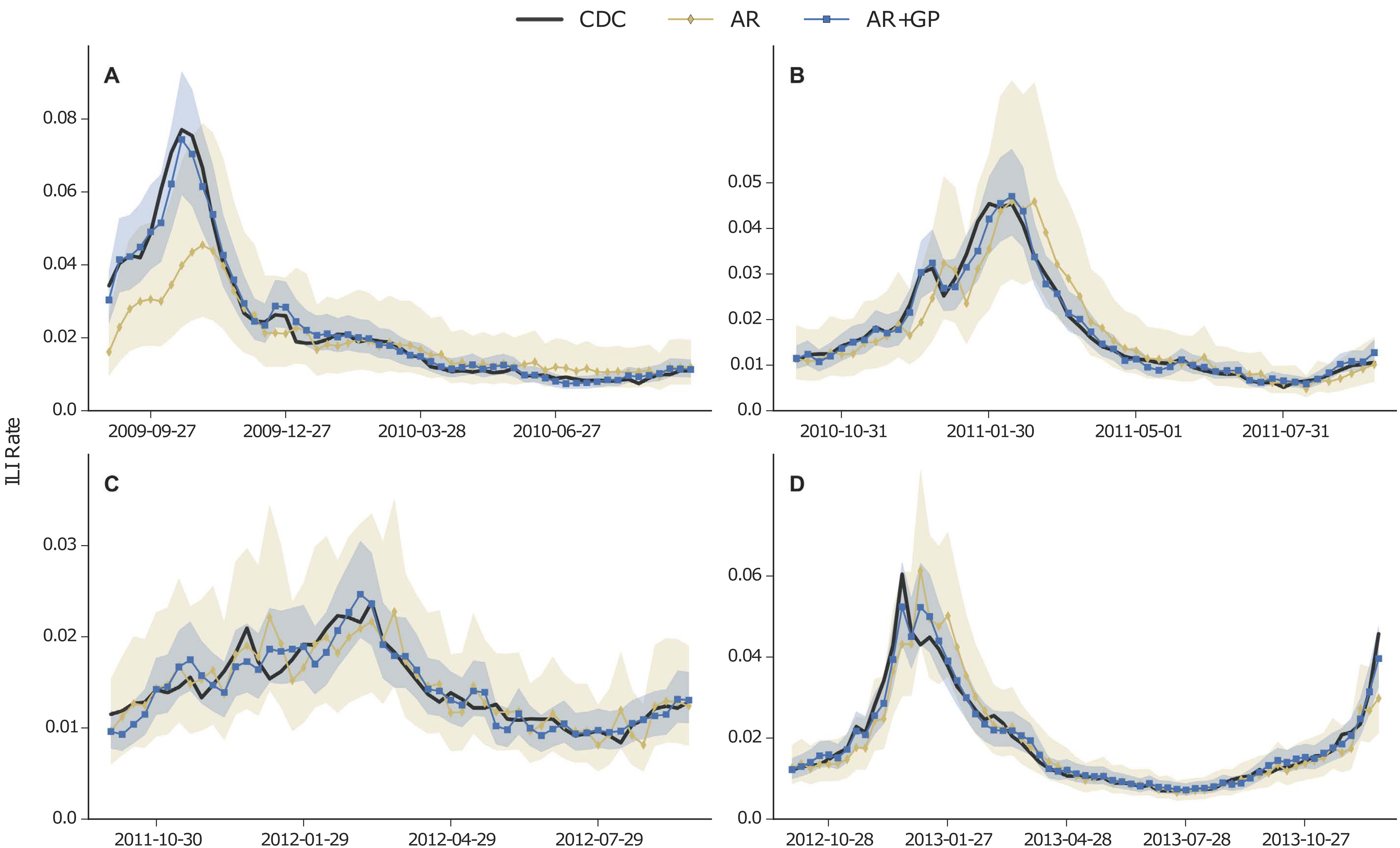
Exogenous input

Seasonal ARMAX

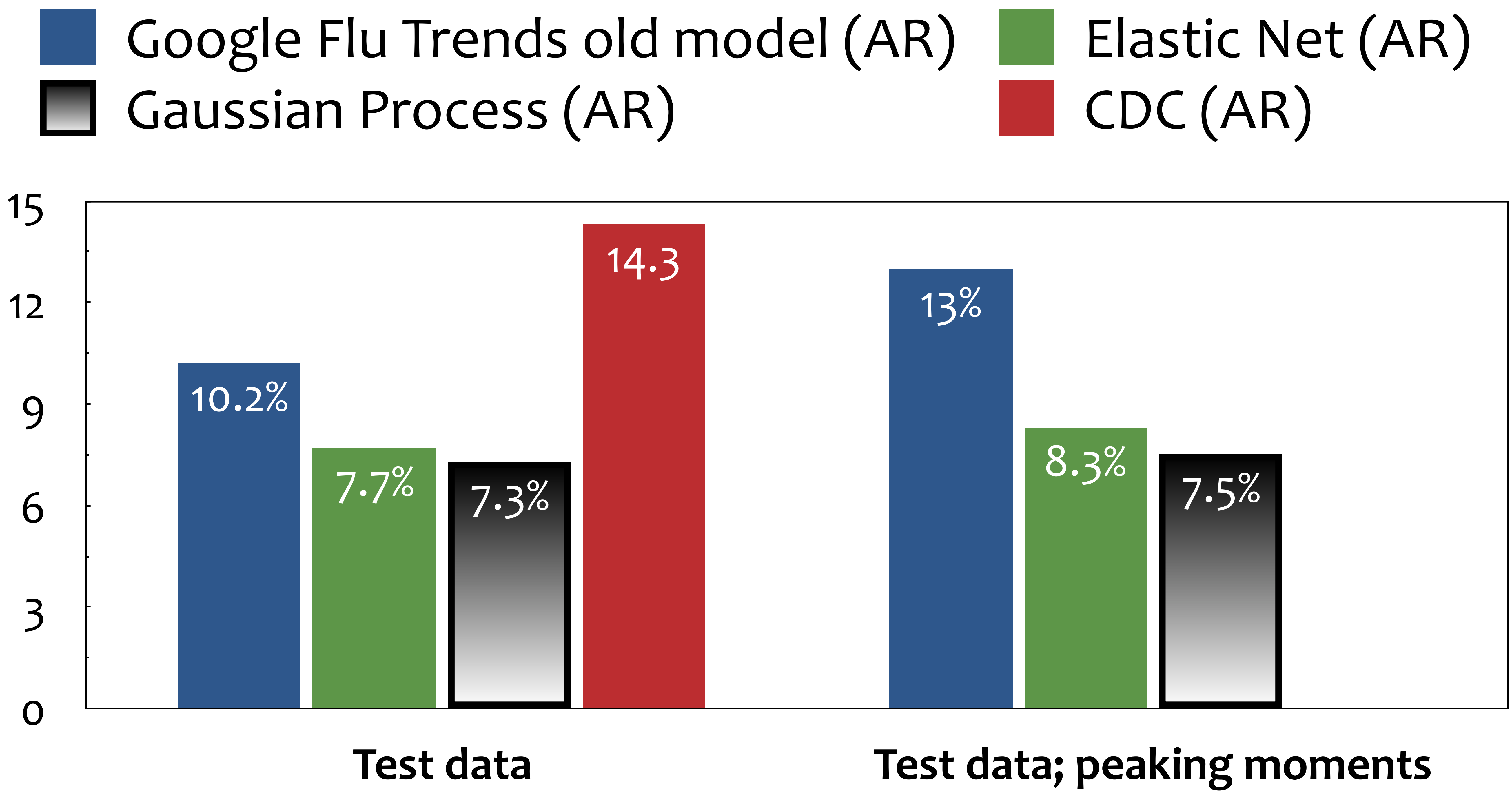
$$y_t = \underbrace{\sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{J} \omega_i y_{t-52-i}}_{\text{AR and seasonal AR}} +$$

$$\underbrace{\sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{K} \nu_i \epsilon_{t-52-i}}_{\text{MA and seasonal MA}} + \underbrace{\sum_{i=1}^{D} w_i h_{t,i}}_{\text{regression}} + \epsilon_t$$

Google Flu Trends revised: *Results* (4)



Google Flu Trends revised: *Results* (5)



MAPE (%) in flu rate autoregressive (AR) estimates during a 4-year period (2009-2013)

Personalised inference tasks using social media content

Lampos, Aletras, Preotiuc-Pietro & Cohn, 2014;
Preotiuc-Pietro, Lampos & Aletras, 2015;
Preotiuc-Pietro, Volkova, Lampos, Bachrach & Aletras, 2015;
Lampos, Aletras, Geyti, Zou & Cox, 2015

Occupational class inference: Motivation

"Socioeconomic variables are influencing language use."

(*Bernstein, 1960*; *Labov, 1972/2006*)

+ Validate this hypothesis on a broader, larger data set using social media (*Twitter*)
+ Downstream applications
  > research (social science & other domains)
  > commercial
+ Proxy for additional user attributes, *e.g.* income and socioeconomic status

(*Preotiuc-Pietro, Lampos & Aletras, 2015*)

Occupational class inference: SOC 2010

Standard Occupational Classification (**SOC**)

C1 — Managers, Directors & Senior Officials
e.g. *chief executive, bank manager*

C2 — Professional Occupations (*e.g. mechanical engineer, pediatrist*)

C3 — Associate Professional & Technical
e.g. *system administrator, dispensing optician*

C4 — Administrative & Secretarial (e.g. *legal clerk, secretary*)

C5 — Skilled Trades (e.g. *electrical fitter, tailor*)

C6 — Caring, Leisure, Other Service
e.g. *nursery assistant, hairdresser*

C7 — Sales & Customer Service (e.g. *sales assistant, telephonist*)

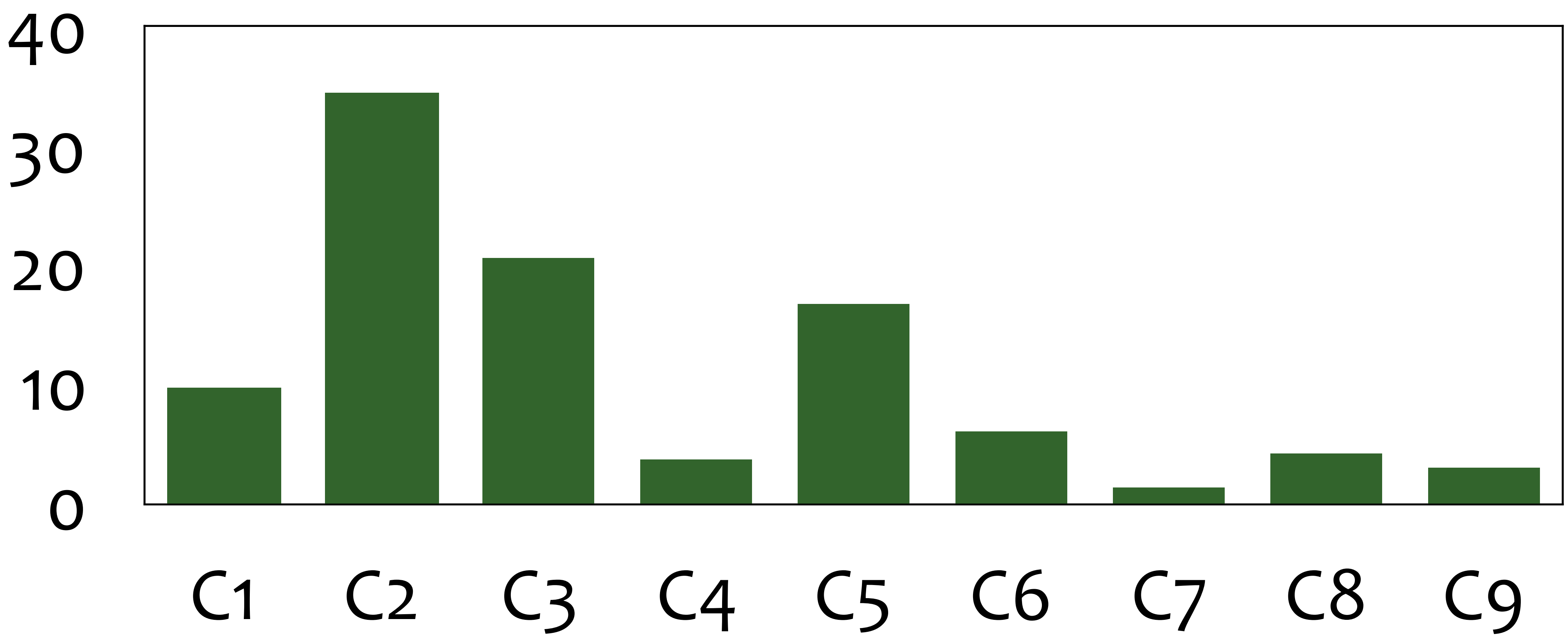
C8 — Process, Plant and Machine Operatives
e.g. *factory worker, van driver*

C9 — Elementary (e.g. *shelf stacker, bartender*)

Occupational class inference: Data

- **5,191** Twitter users mapped to their occupations, then mapped to one of the 9 SOC categories
- 10 million tweets
- **Download the data set**

% of users per SOC category



Occupational class inference: Features

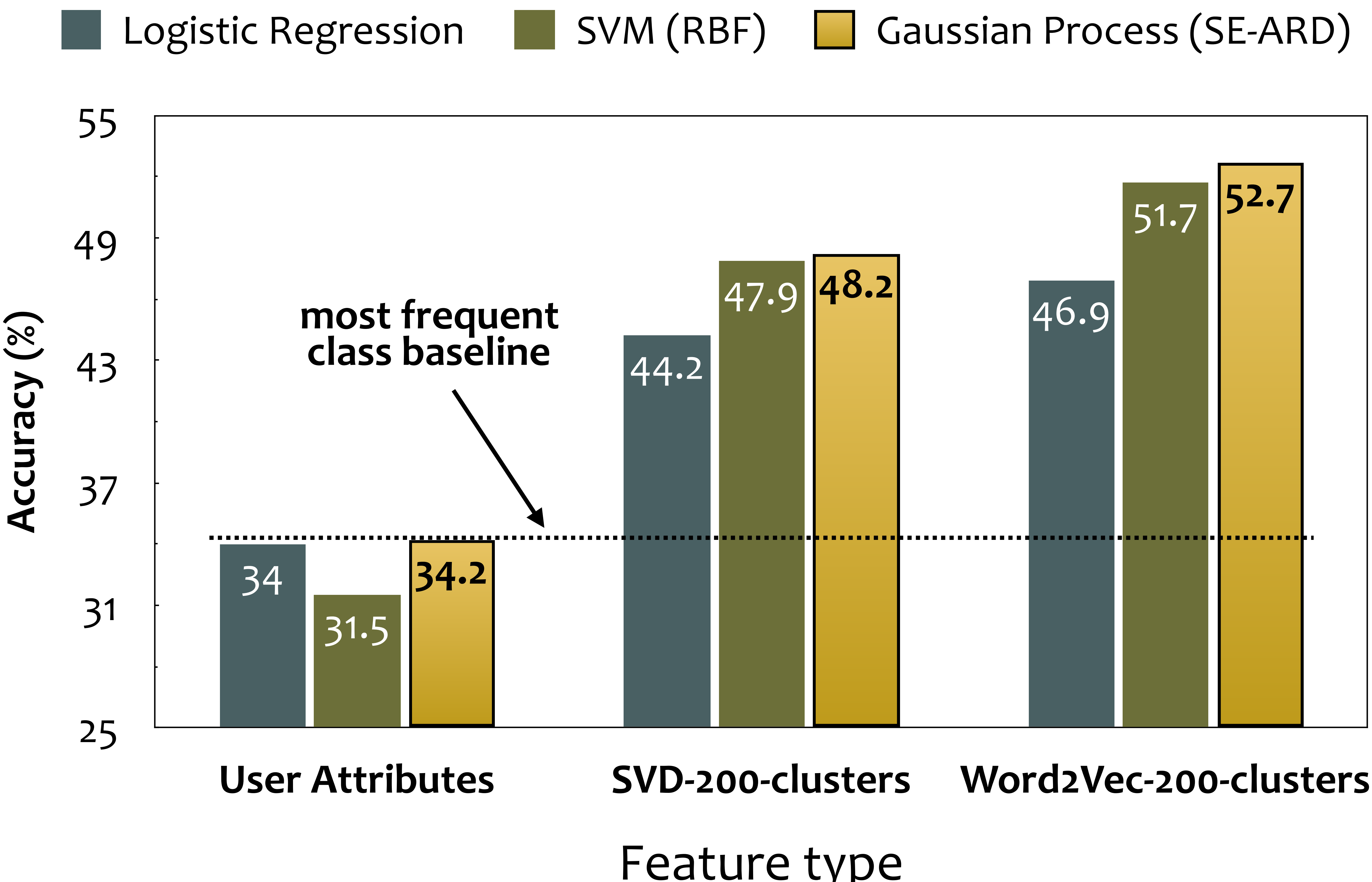
User attributes (18)

+ number of followers, friends, listings, follower/friend ratio, favourites, tweets, retweets, hashtags, @-mentions, @-replies, links and so on

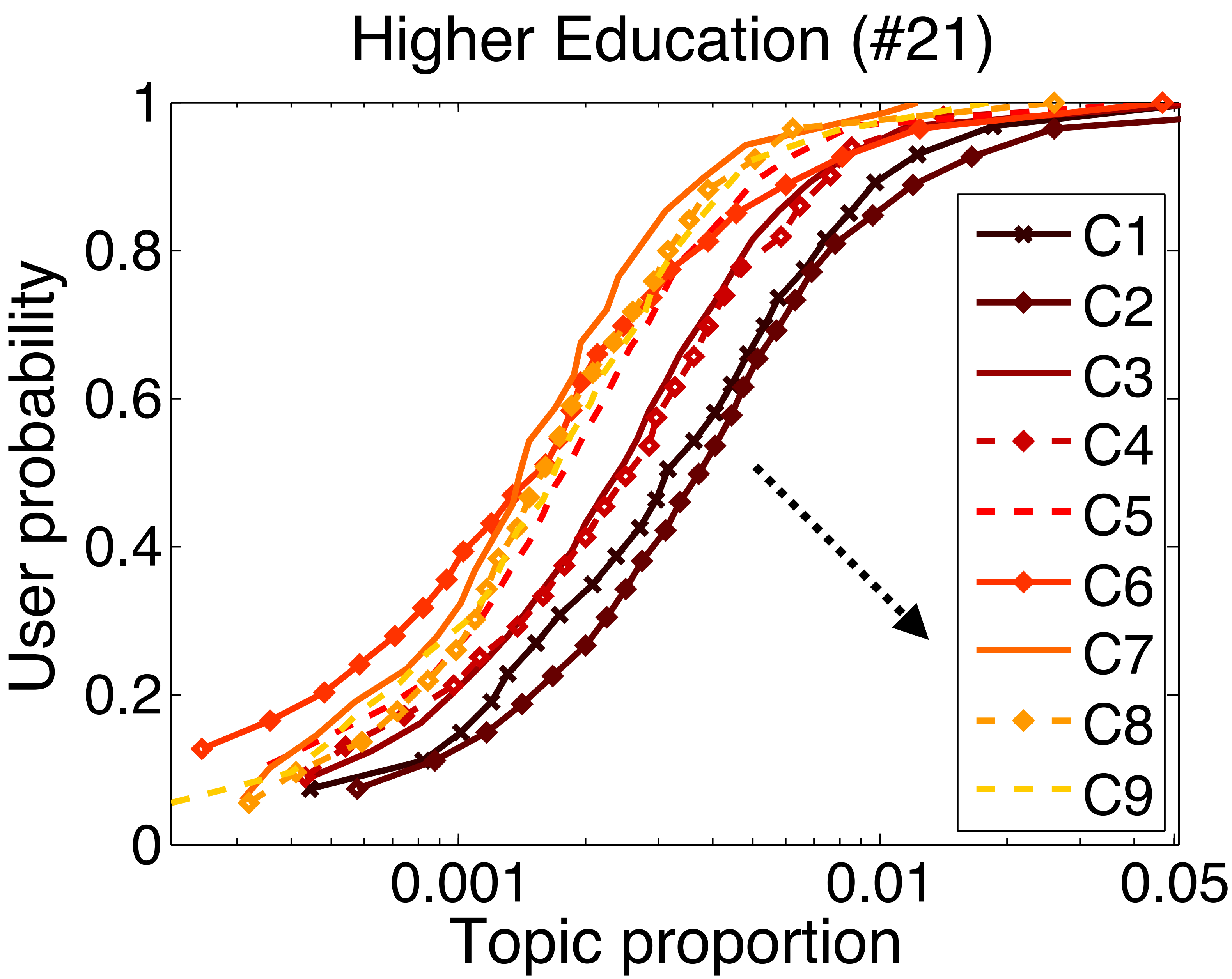
Topics — Word clusters (200)

+ ***SVD*** on the graph laplacian of the word x word similarity matrix using normalised PMI, i.e. a form of spectral clustering (*Bouma, 2009*; *von Luxburg, 2007*)
+ Skip-gram model with negative sampling to learn word embeddings (***Word2Vec***); pairwise cosine similarity on the embeddings to derive a word x word similarity matrix; then spectral clustering on the similarity matrix (*Mikolov et al., 2013*)

Occupational class inference: Performance

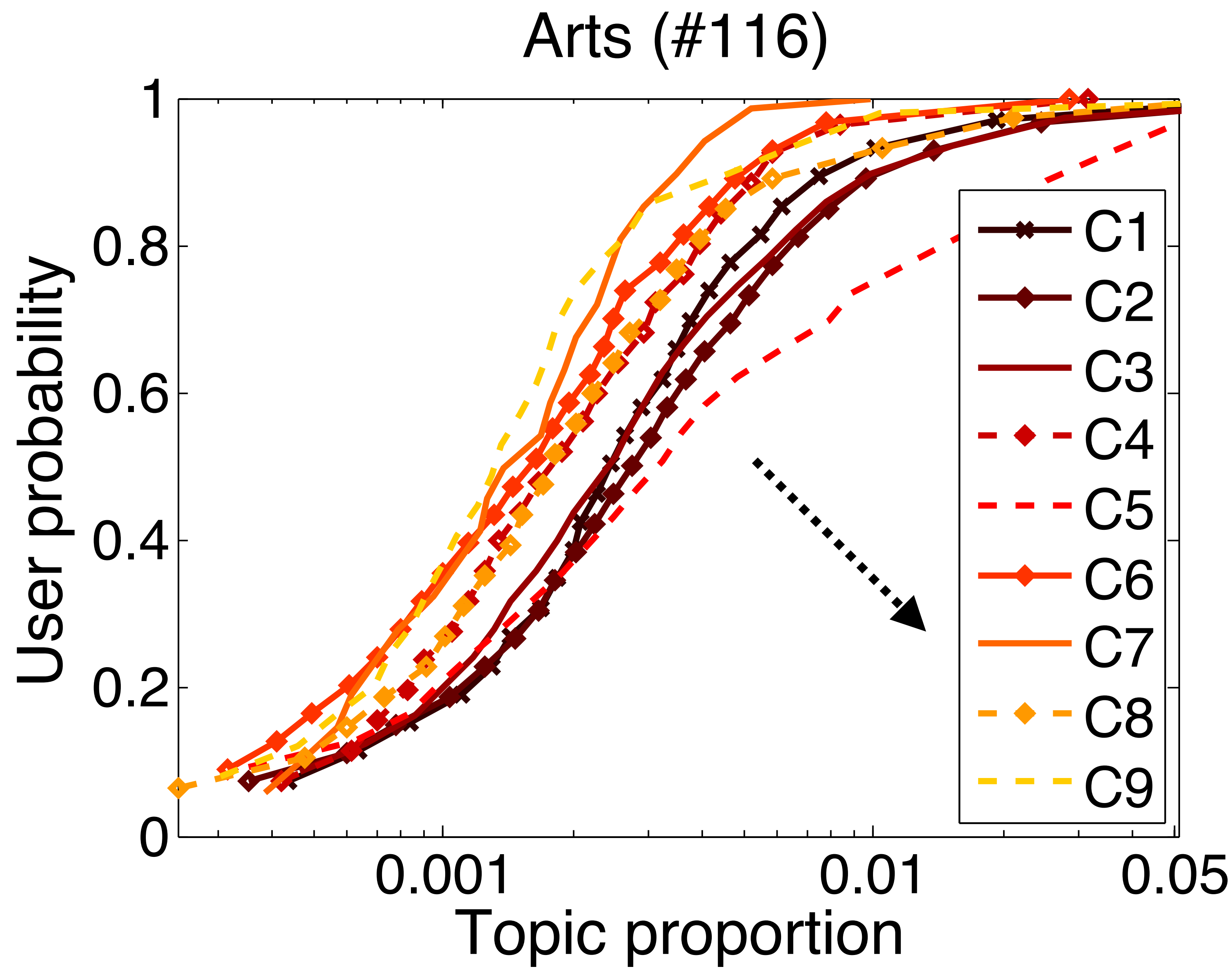


Occupational class inference: Topic CDFs (1)



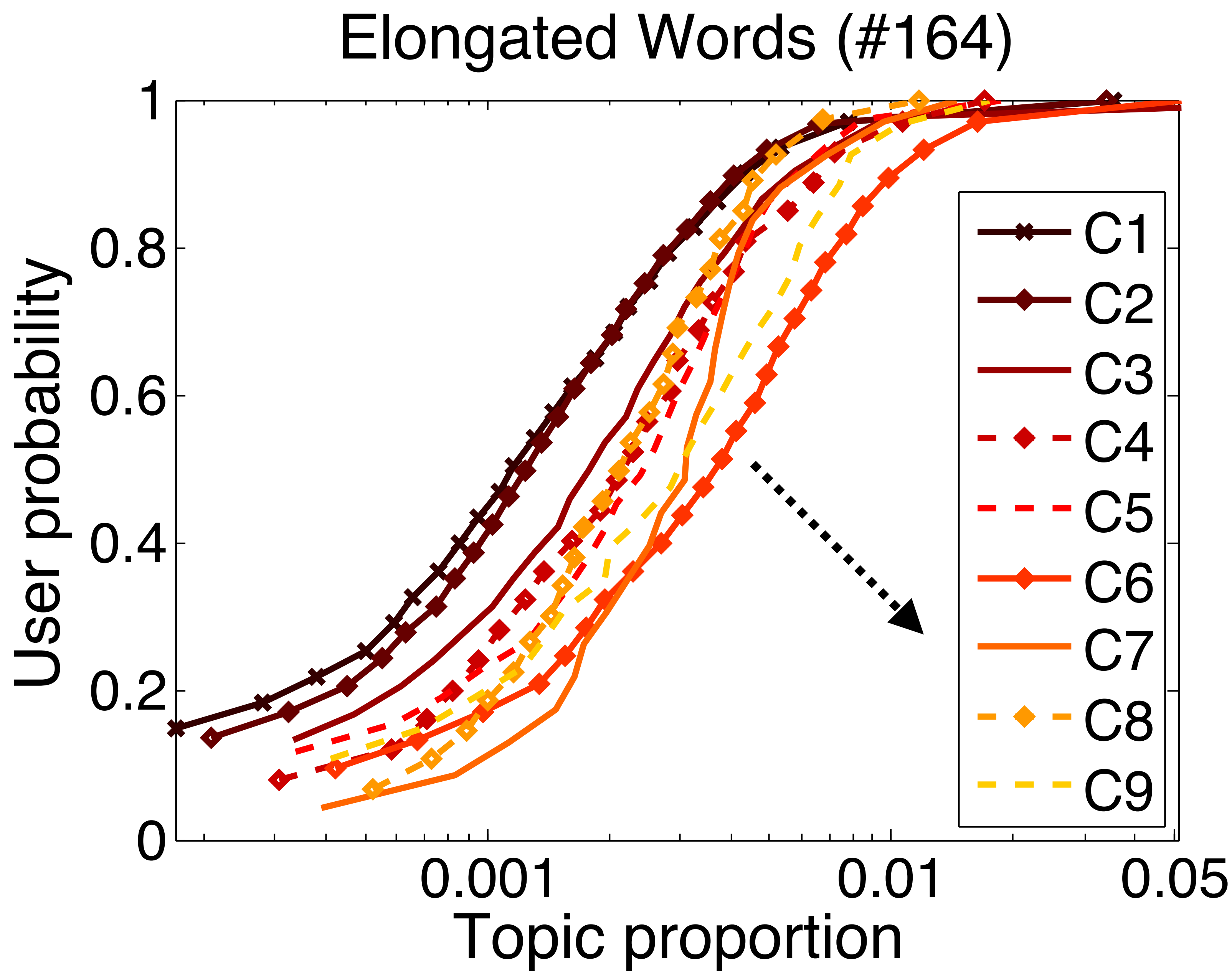
*Topic more **prevalent** in a class (C1-C9), if the line leans closer to the **bottom-right corner** of the plot*

Occupational class inference: Topic CDFs (2)



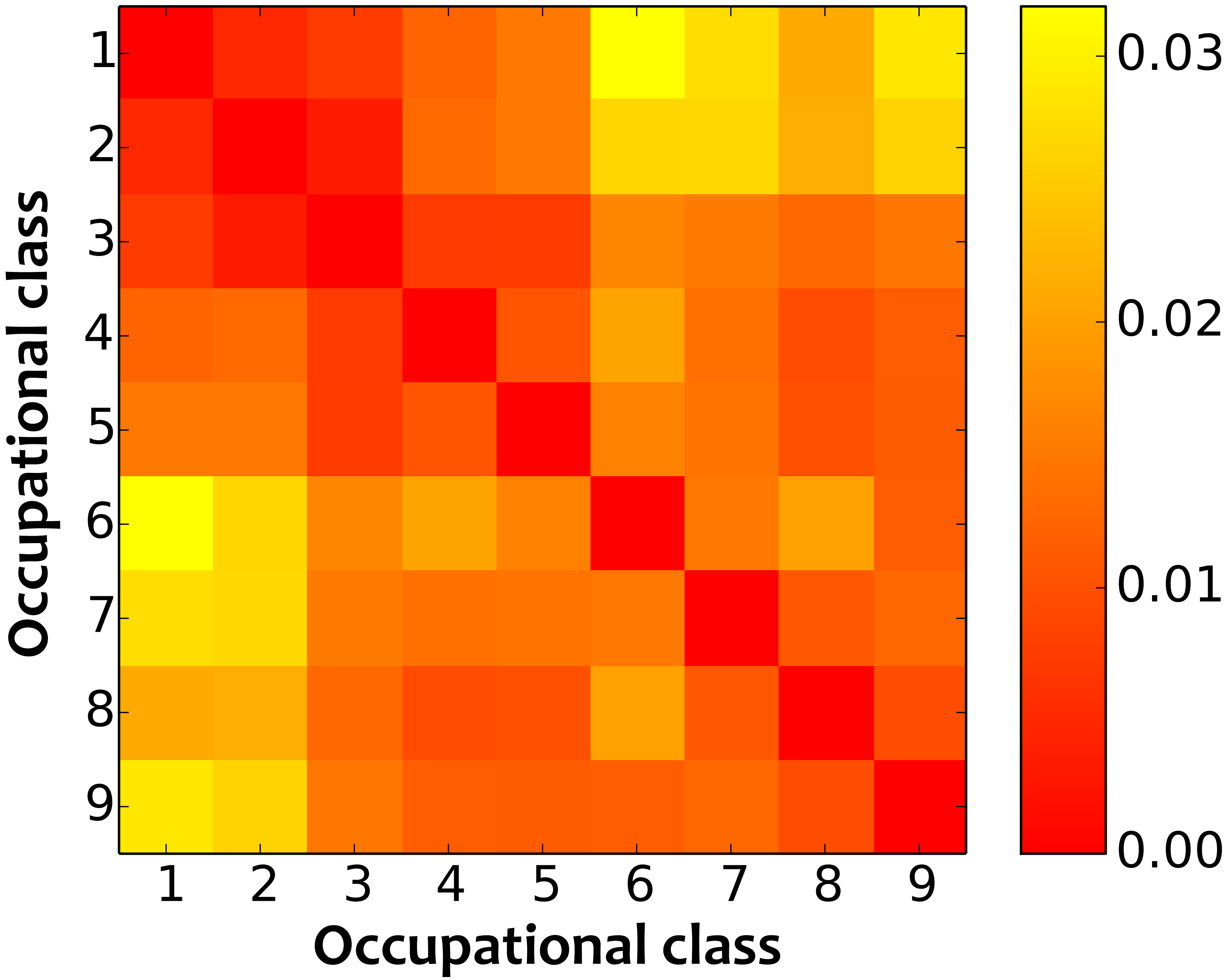
*Topic more **prevalent** in a class (C1-C9), if the line leans closer to the **bottom-right corner** of the plot*

Occupational class inference: Topic CDFs (3)



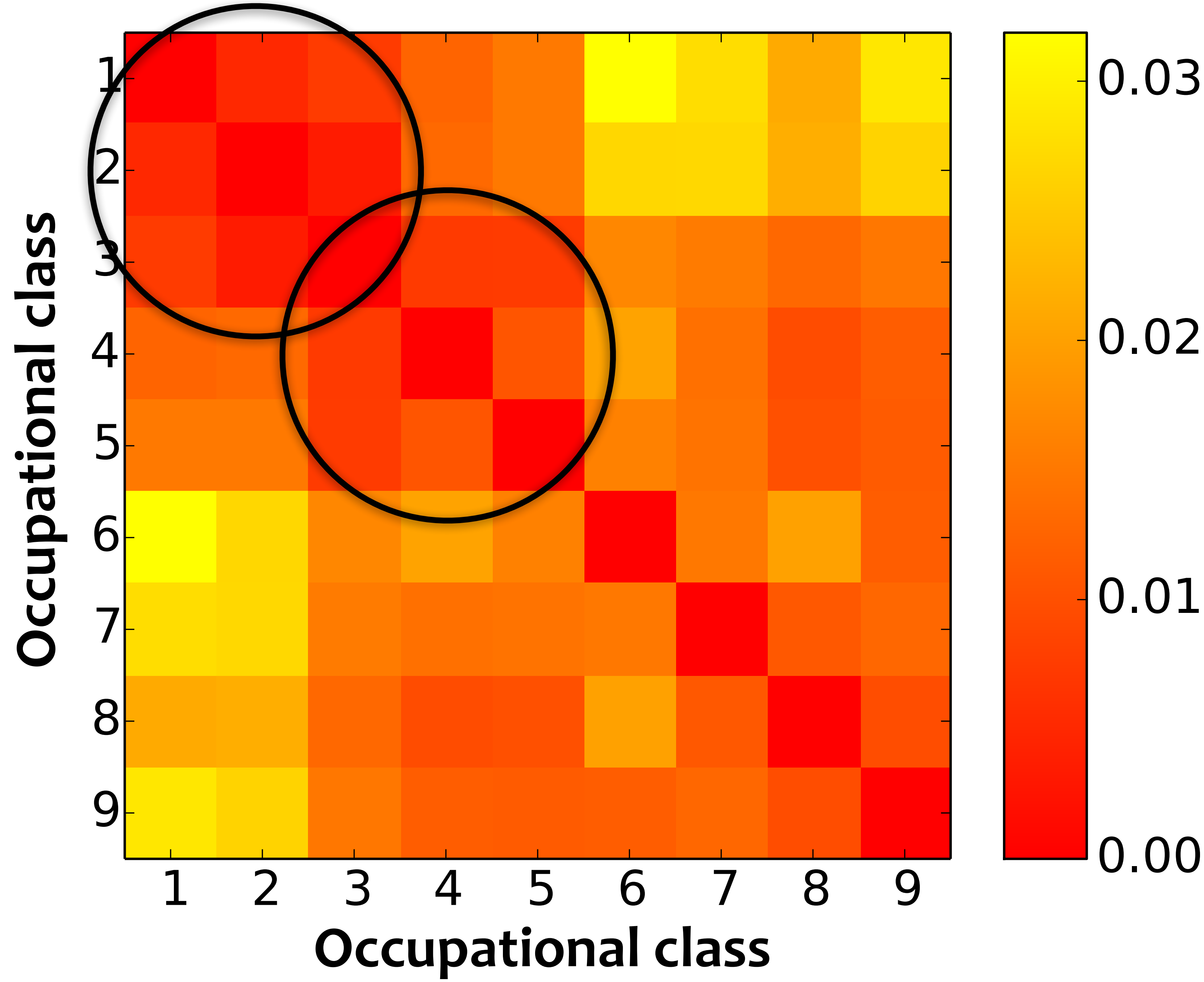
*Topic more **prevalent** in a class (C1-C9), if the line leans closer to the **bottom-right corner** of the plot*

Occupational class inference: Topic similarity



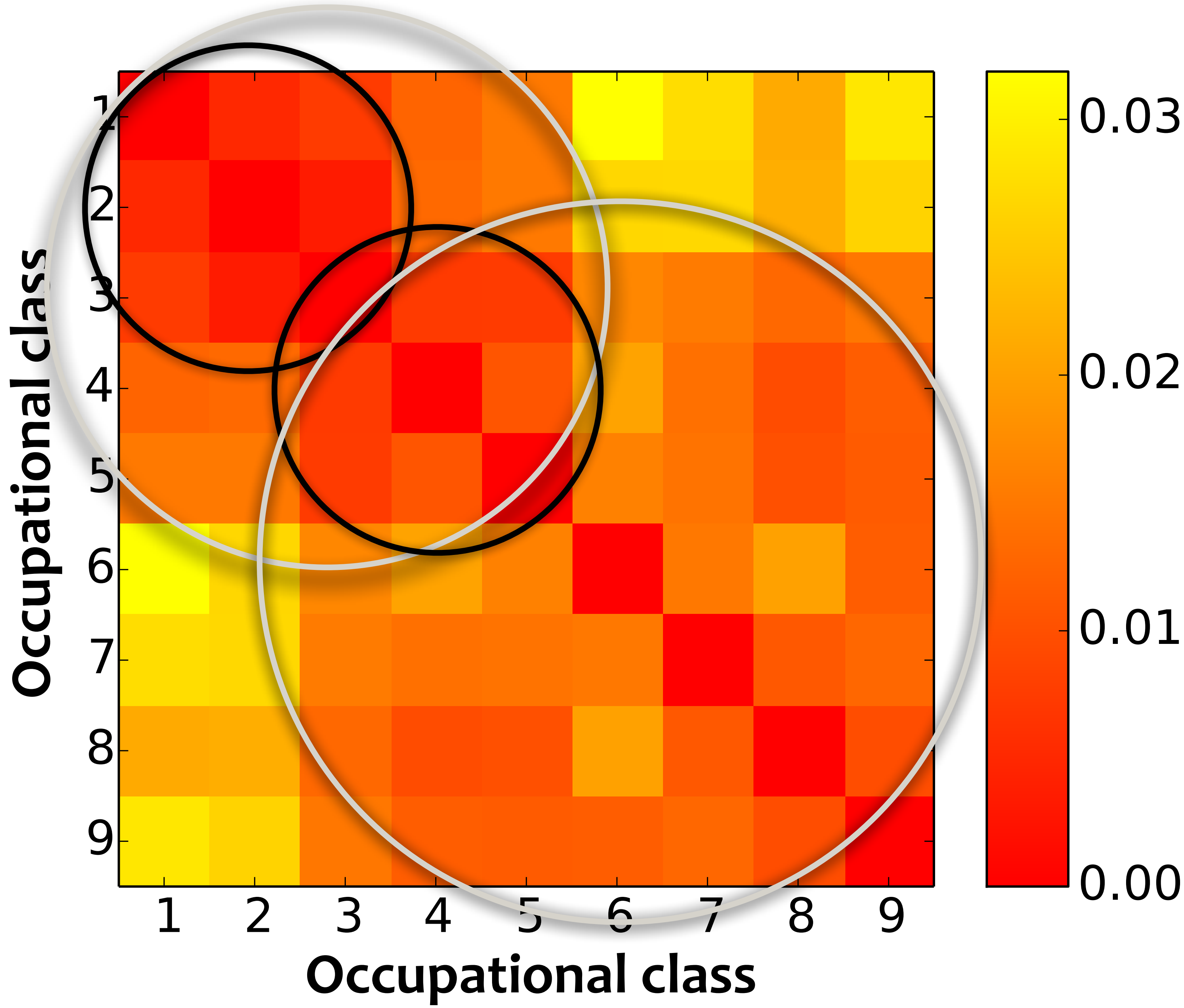
Topic distribution distance (*Jensen-Shannon divergence*)
for the different occupational classes

Occupational class inference: Topic similarity



Topic distribution distance (*Jensen-Shannon divergence*)
for the different occupational classes

Occupational class inference: Topic similarity

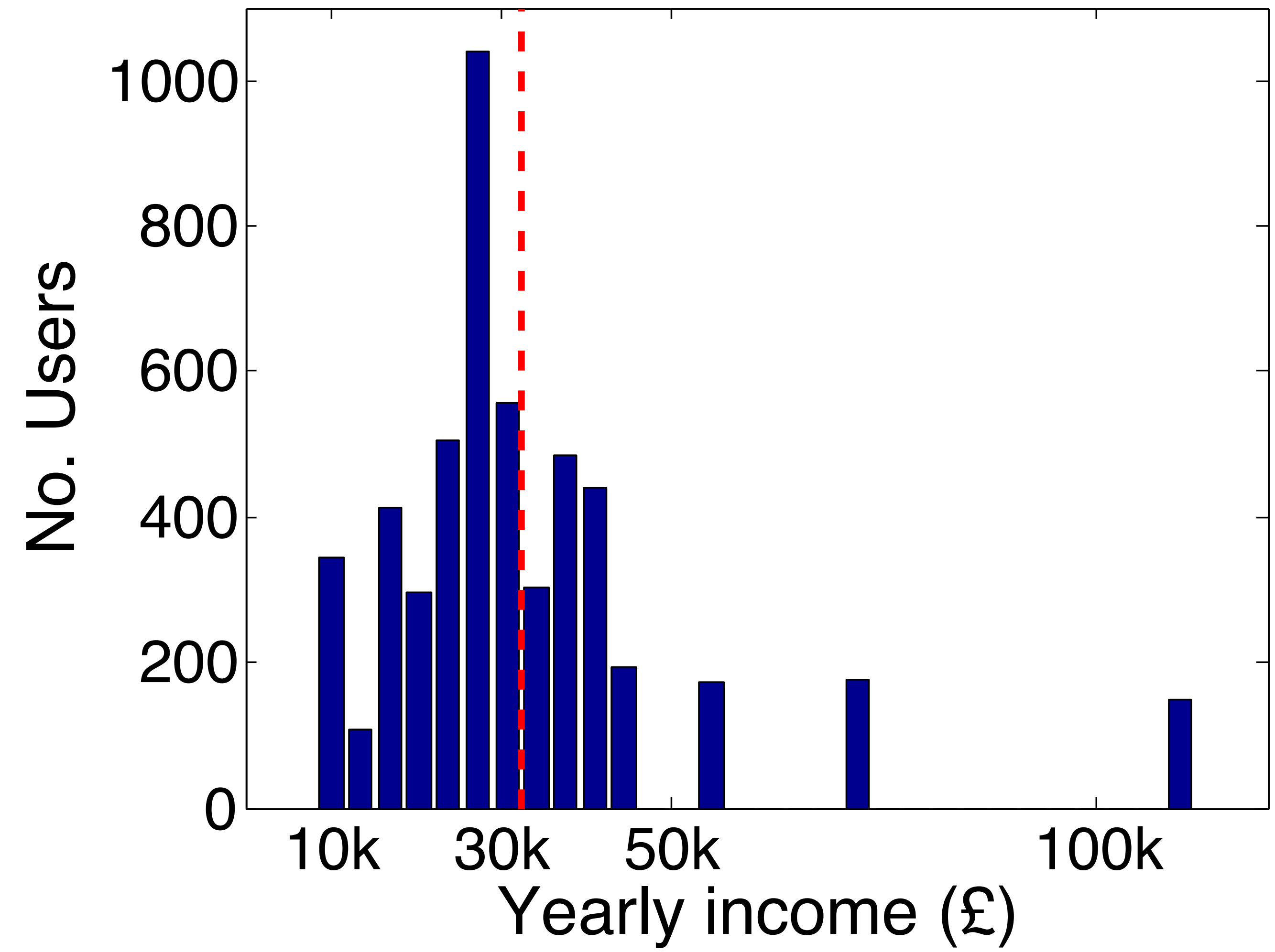


Topic distribution distance (*Jensen-Shannon divergence*) for the different occupational classes

Income inference: Data

+ **5,191** Twitter users (same as in the previous study) mapped to their occupations, then mapped to an average income in GBP (£) using the *SOC* taxonomy
+ approx. 11 million tweets
+ **Download the data set**

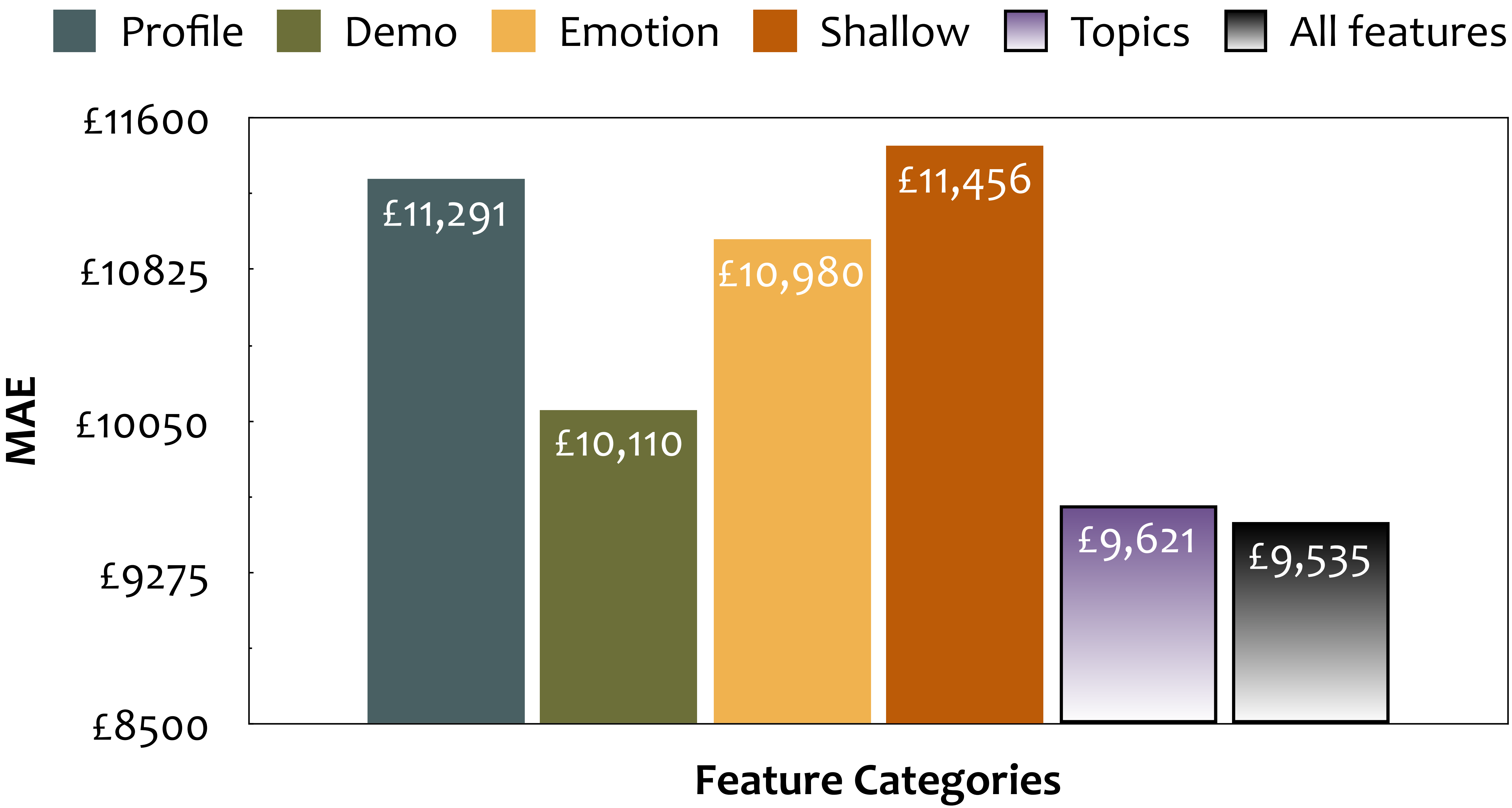
(*Preotiuc-Pietro, Volkova, Lampos, Bachrach & Aletras, 2015*)



Income inference: Features

+ **Profile** (8)
  e.g. #followers, #followees, times listed etc.
+ **Shallow textual features** (10)
  e.g. proportion of hashtags, @-replies, @-mentions etc.
+ **Inferred (perceived) psycho-demographic features** (15)
  e.g. gender, age, education level, religion, life satisfaction, excitement, anxiety etc.
+ **Emotions** (9)
  e.g. positive / negative sentiment, joy, anger, fear, disgust, sadness, surprise etc.
+ **Word clusters — Topics of discussion** (200)
  based on word embeddings and by applying spectral clustering

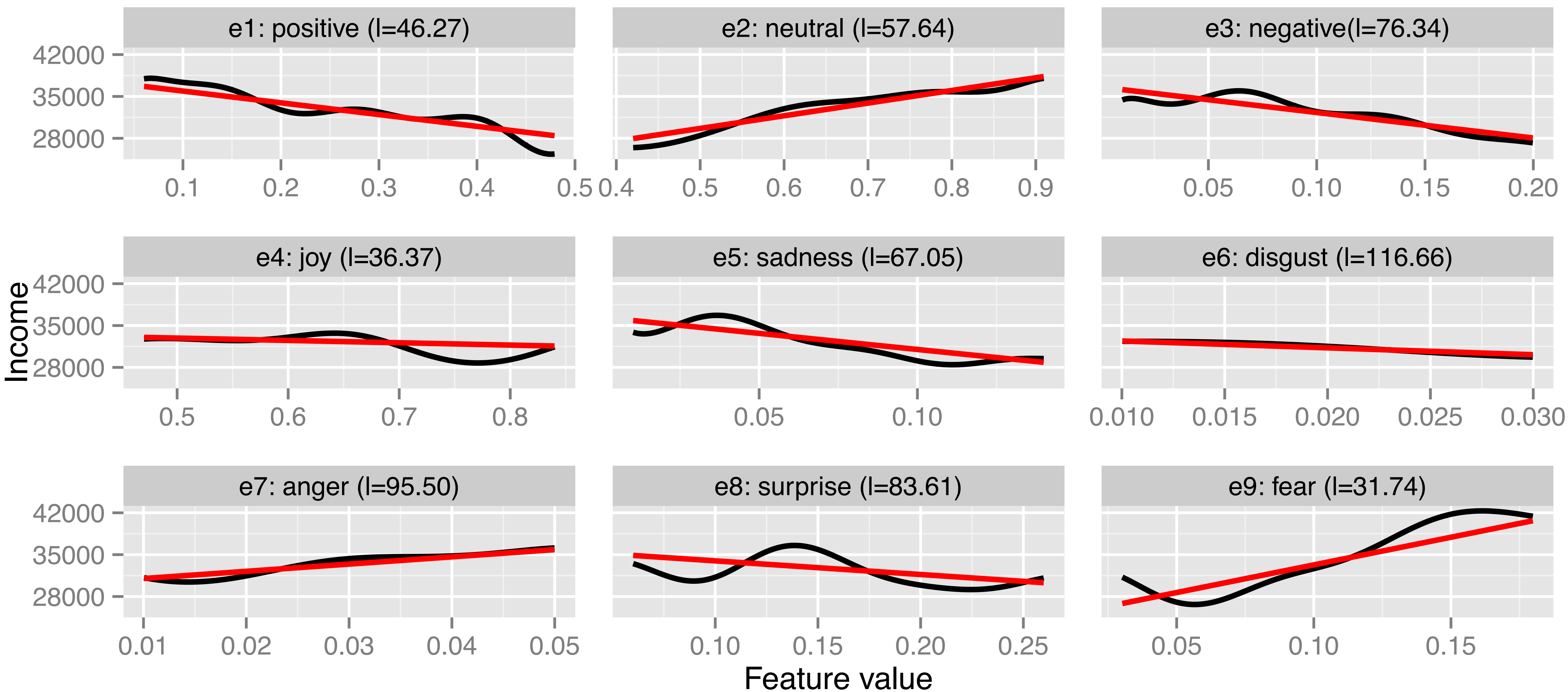
Income inference: Performance



Income inference error (Mean Absolute Error) using GP regression or a linear ensemble for all features

Income inference: Qualitative analysis (1)

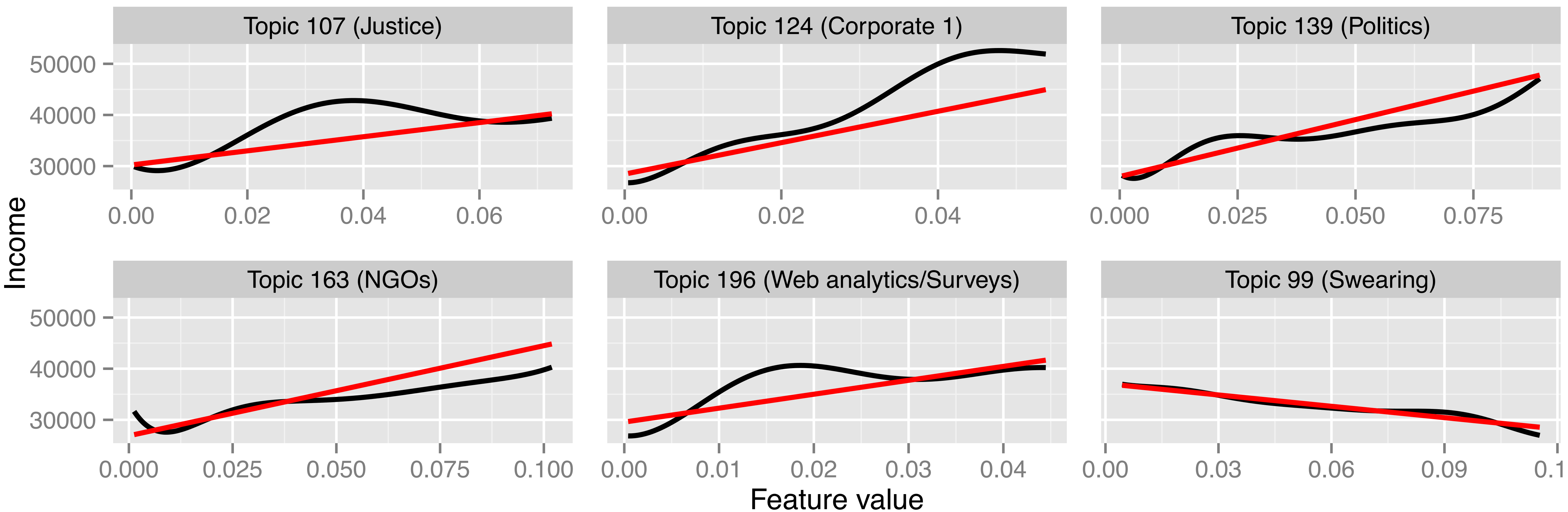
Relating income and emotion



Linear vs **GP** fit

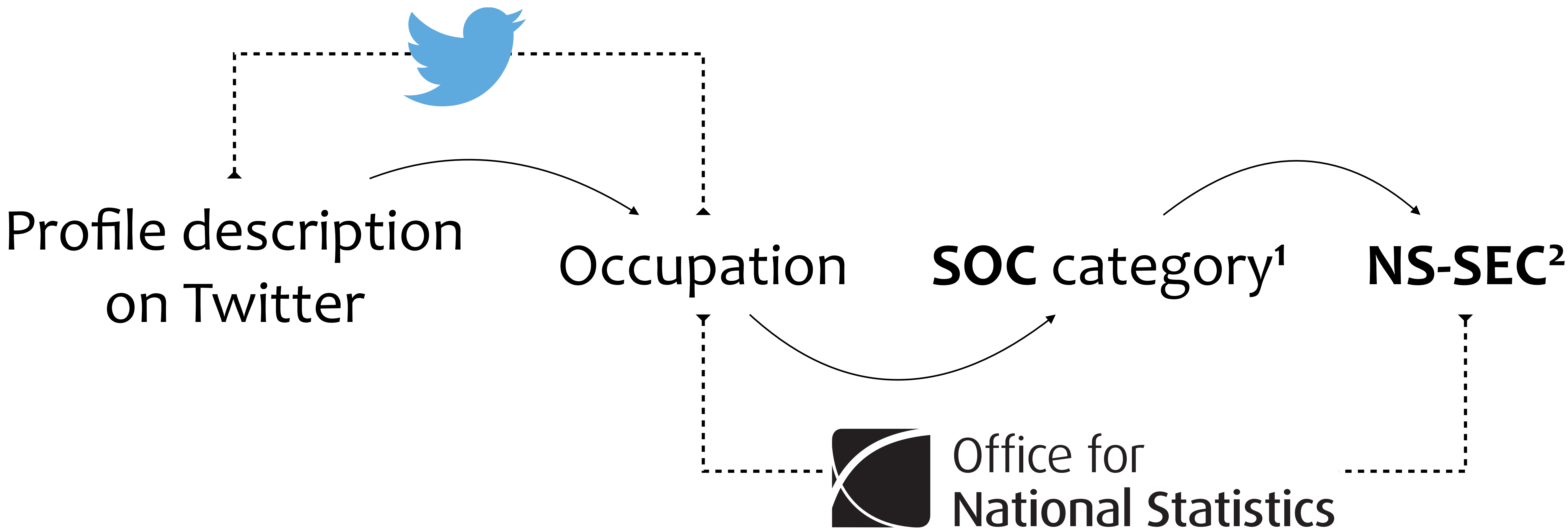
Income inference: Qualitative analysis (2)

Relating income and topics of discussion



Linear vs **GP** fit

Inferring the socioeconomic status: Task



1. **S**tandard **O**ccupational **C**lassification: 369 job groupings
2. **N**ational **S**tatistics **S**ocio-**E**conomic **C**lassification: Map from the job groupings in SOC to a socioeconomic status, *i.e.* {*upper*, *middle* or *lower*}

Inferring the socioeconomic status: Data & Features

- **1,342** Twitter user profiles
  distinct data set from the previous works
- 2 million tweets
- Date interval: Feb. 1, 2014 to March 21, 2015
- Each user has a **socioeconomic status (SES) label:**
  {*upper*, *middle*, *lower*}
- **Download the data set**

1,291 features representing
user behaviour (4), *biographical / profile information* (523), *text in the tweets* (560), *topics of discussion* (200), and *impact on the platform* (4)

Inferring the socioeconomic status: Results

Confusion matrices for the 3- and 2-way classification

	T1	T2	T3	P
O1	606	84	53	81.6%
O2	49	186	45	66.4%
O3	55	48	216	67.7%
R	854%	58.5%	68.8%	75.1%

	T1	T2	P
O1	584	115	83.5%
O2	126	517	80.4%
R	82.3%	81.8%	82.0%

Classification performance (using a GP classifier)

Classification	Accuracy (%)	Precision (%)	Recall (%)	F1
2-way	82.05 (2.4)	82.2 (2.4)	81.97 (2.6)	.821 (.03)
3-way	75.09 (3.3)	72.04 (4.4)	70.76 (5.7)	.714 (.05)

Characterising user impact: Task & Data

$$\boxed{\mathrm{S}(\phi_{\mathrm{in}}, \phi_{\mathrm{out}}, \phi_{\lambda}) = \ln\left(\frac{(\phi_{\lambda}+\theta)(\phi_{\mathrm{in}}+\theta)^2}{\phi_{\mathrm{out}}+\theta}\right)}$$

$$(\phi_{\mathrm{in}}^2/\phi_{\mathrm{out}}) = (\phi_{\mathrm{in}} - \phi_{\mathrm{out}}) \times (\phi_{\mathrm{in}}/\phi_{\mathrm{out}}) + \phi_{\mathrm{in}}$$

ϕ_{in} —> number of followers

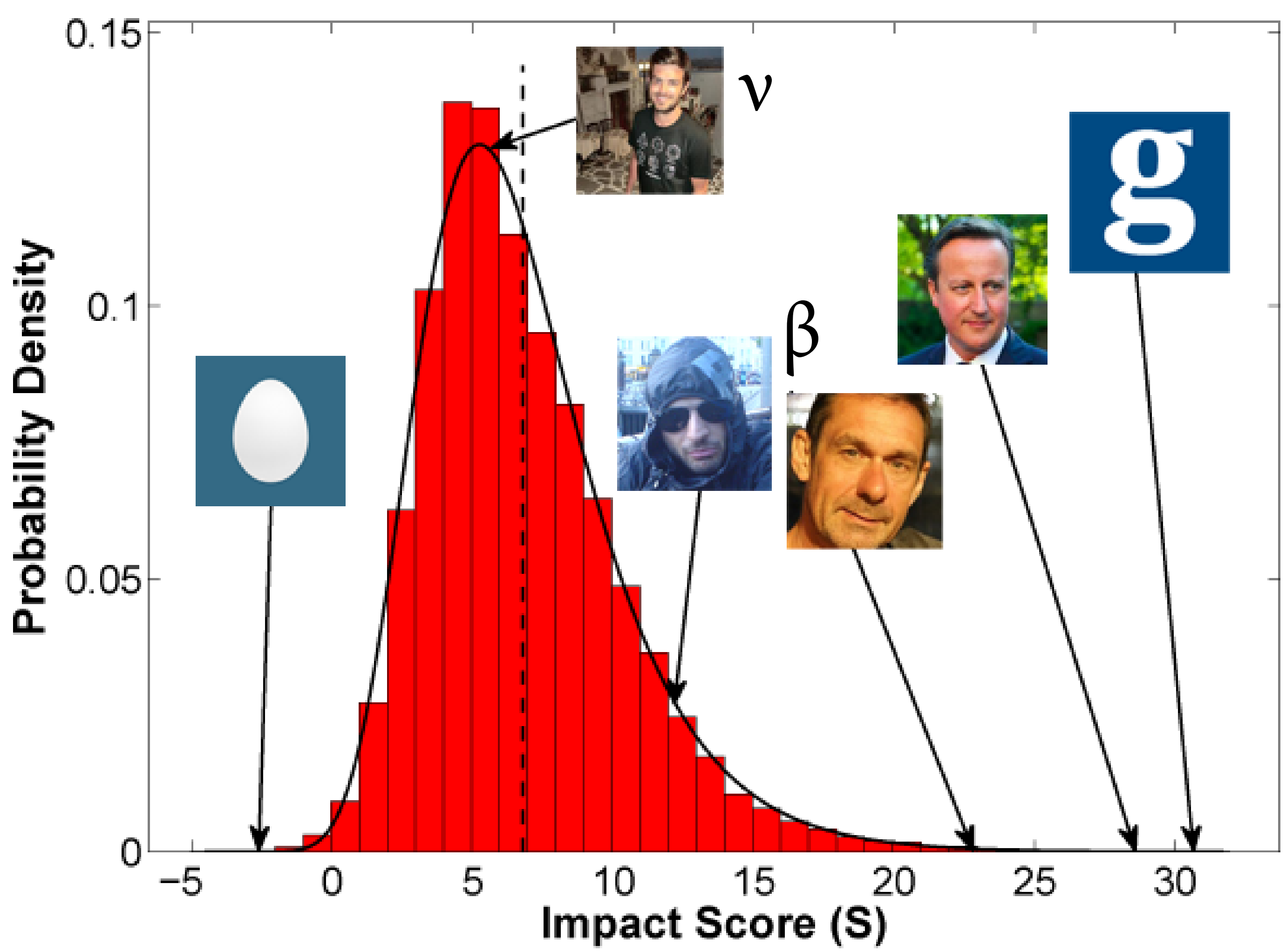
ϕ_{out} —> number of followees

ϕ_{λ} —> number of times listed

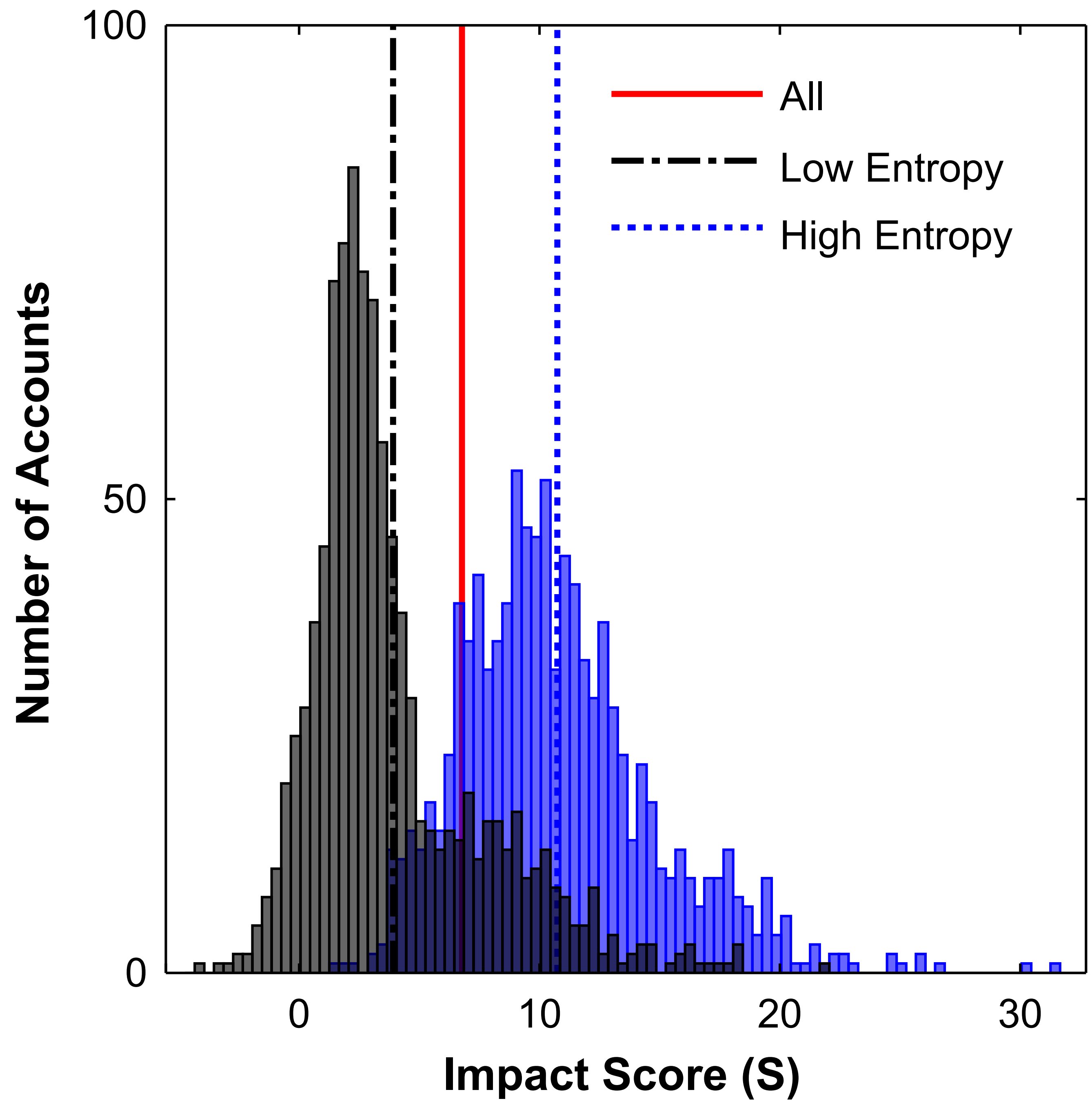
$\theta = 1$ —> logarithm is applied on a positive number

β Vasileios Lampos ~ *@lampos*
ν Nikolaos Aletras ~ *@nikaletras*

40K Twitter accounts (UK) considered

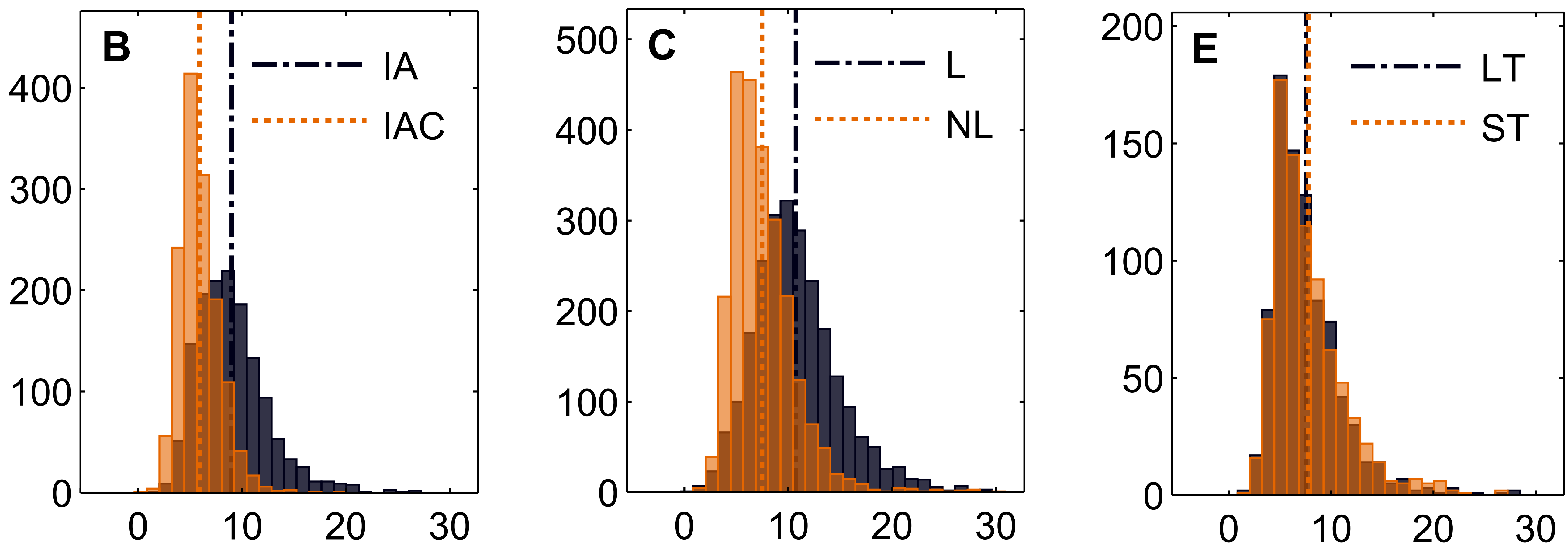


Characterising user impact: Topic entropy



On average, the **higher** the user ***impact score***, the **higher** the ***topic entropy***

Characterising user impact: Use case scenarios



Interactive (**IA**) vs. clique interactive (**IAC**)

Links (**L**) vs. very few links (**NL**)

Light topics (**LT**) vs. more 'serious' topics (**ST**)

Impact distribution under user behaviour scenarios

Concluding remarks

+ **User-generated content** is a ***valuable asset***
  > improve health surveillance tasks
  > mine collective knowledge
  > infer user characteristics
  > *numerous other tasks*

+ ***Nonlinear models*** tend to perform better given the multimodality of the feature space

+ ***Deep representations*** of text tend to improve performance (*better representations*)

+ ***Qualitative analysis*** is important
  > Evaluation
  > Interesting insights

Future research challenges

+ Interdisciplinary research tasks require to work closer with ***domain experts***
+ Understand better the ***biases*** in the online media (demographics, information propagation, external influence etc.)
+ Attack more interesting (*usually more complex*) questions, attempt to ***generalise*** findings, identify and define ***limitations***
+ Conduct more rigorous ***evaluation***
+ Improve on existing methods (‘***deeper***’ understandings & interpretations)
+ ***Ethical concerns***

Acknowledgements

All **collaborators** (*alphabetical order*)
in research mentioned today

Nikolaos Aletras (*Amazon*), **Yoram Bachrach** (*Microsoft Research*), **Trevor Cohn** (*Univ. of Melbourne*), **Ingemar J. Cox** (*UCL & Univ. of Copenhagen*), **Nello Cristianini** (*Univ. of Bristol*), **Steve Crossan** (*Google*), **Jens K. Geyti** (*UCL*), **Andrew C. Miller** (*Harvard Univ.*), **Daniel Preotiuc-Pietro** (*Penn*), **Christian Stefansen** (*Google*), **Sviltana Volkova** (*PNNL*), **Bin Zou** (*UCL*)

Currently **funded by**

i-sense

Thank you.
Any questions?

Slides can be downloaded from
lampos.net/talks-posters

References

Argyriou, Evgeniou & Pontil. ***Convex Multi-Task Feature Learning*** (Machine Learning, 2008)
Bach. ***Bolasso: Model Consistent Lasso Estimation through the Bootstrap*** (ICML, 2008)
Bernstein. ***Language and social class*** (Br J Sociol, 1960)
Bouma. ***Normalized (pointwise) mutual information in collocation extraction*** (GSCL, 2009)
David Duvenaud. ***Automatic Model Construction with Gaussian Processes*** (Ph.D. Thesis, Univ of Cambridge, 2014)
Ginsberg et al. ***Detecting influenza epidemics using search engine query data*** (Nature, 2009)
Hastie, Tibshirani & Friedman. ***The Elements of Statistical Learning*** (Springer, 2009)
Labov. ***The Social Stratification of English in New York City*** (Cambridge Univ Press, 1972; 2006, 2nd ed.)
Lampos & Cristianini. ***Nowcasting Events from the Social Web with Statistical Learning*** (ACM TIST, 2012)
Lampos, Aletras, Geyti, Zou & Cox. ***Inferring the Socioeconomic Status of Social Media Users based on Behaviour and Language*** (ECIR, 2016)
Lampos, Miller, Crossan & Stefansen. ***Advances in nowcasting influenza-like illness rates using search query logs*** (Nature Sci Rep, 2015)
Lampos, Preotiuc-Pietro, Aletras & Cohn. ***Predicting and Characterising User Impact on Twitter*** (EACL, 2014)
Lampos, Preotiuc-Pietro & Cohn. ***A user-centric of voting intention from Social Media*** (ACL, 2013)
Lazer, Kennedy, King and Vespignani. ***The Parable of Google Flu: Traps in Big Data Analysis*** (Science, 2014)
Mairal, Jenatton, Obozinski & Bach. ***Network Flow Algorithms for Structured Sparsity*** (NIPS, 2010)
Mikolov, Chen, Corrado & Dean. ***Efficient estimation of word representations in vector space*** (ICLR, 2013)
Preotiuc-Pietro, Lampos & Aletras. ***An analysis of the user occupational class through Twitter content*** (ACL, 2015)
Preotiuc-Pietro, Volkova, Lampos, Bachrach & Aletras. ***Studying User Income through Language, Behaviour and Affect in Social Media*** (PLoS ONE, 2015)
Rasmussen & Williams. ***Gaussian Processes for Machine Learning*** (MIT Press, 2006)
Tibshirani. ***Regression shrinkage and selection via the lasso*** (J R Stat Soc Series B Stat Methodol, 1996)
von Luxburg. ***A tutorial on spectral clustering*** (Stat Comput, 2007)
Zhao & Yu. ***On model selection consistency of lasso*** (JMLR, 2006)
Zou & Hastie. ***Regularization and variable selection via the elastic net*** (J R Stat Soc Series B Stat Methodol, 2005)