Mining socio-political and socio-economic signals from social media content

Vasileios Lampos

Department of Computer Science
University College London

@lampos | lampos.net

Summer School on
"*Big Data & Networks in Social Sciences*"
University of Warwick, Sept. 21-23, 2016

Structure of the presentation

1. **Introductory remarks**

2. **Collective inference tasks**
   — Mining emotions
   — Modelling voting intention

3. **Personalised inference tasks**
   — Occupational class
   — Income
   — Socioeconomic status

4. **Concluding remarks**

Context and motivation

the Internet, the ***World Wide Web***, connectivity

numerous ***web products*** feeding from user activity

user-generated content, publicly available, esp. on social media platforms (e.g. Twitter)

large-scale digitised data, '***Big Data***', 'Data Science'

How can we use online user-generated content to enhance our understanding about our world?

Context and motivation

the Internet, the ***World Wide Web***, connectivity

numerous ***web products*** feeding from user activity

user-generated content, publicly available, esp. on social media platforms (e.g. Twitter)

large-scale digitised data, '***Big Data***', 'Data Science'

How can we use online user-generated content to enhance our understanding about our world?

About Twitter

And what about the statistical significance of the computed statistical significance? #inception_in_statistics

Reply Delete Favorite

RT if you love Justin Bieber. Delete ur account if you don't.

Reply Retweet Favorite

50	1
RETWEETS	FAVORITE

Why do I feel so happy today hihi. Bedtimeeee, good night. Yey thank You Lord for everything. Answered prayer ♥

Reply Retweet Favorite

i think i have the flu but i still look fabulous

Reply Retweet Favorite

About Twitter

And what about the statistical significance of

- **140 characters** per published status (***tweet***)
- users can **follow** and **be followed**
- embedded usage of **topics** (using #hashtags)
- user **interaction** (re-tweets, @mentions, likes)
- **real-time** nature
- **biased demographics** (13-15% of UK's population, age bias etc.)
- information is **noisy** and **not always accurate**

i think i have the flu but i still look fabulous

Reply Retweet Favorite

Inferring collective information from user-generated content

- mood / emotions
- voting intention

Lampos (Ph.D. Thesis, 2012)
Lansdall-Welfare, Lampos & Cristianini (WWW 2012)
Lampos, Preotiuc-Pietro & Cohn (ACL 2013)

Emotion taxonomies and quantification

> WordNet Affect
> Linguistic Inquiry and Word Count (LIWC)

(*Strapparava & Valitutti, 2004*; *Pennebaker et al., 2001, 2007*)

'Emotional' **keywords**, representing

+ **anger**, *e.g. angry, irritate*
+ **fear**, *e.g. fearful, afraid*
+ **joy**, *e.g. cheerful, enthusiastic*
+ **sadness**, *e.g. depressed, gloomy*
+ *plus other emotions*

Simply — *but maybe not good enough!* — we compute the **mean keyword frequency score** per emotion

Emotion taxonomies and quantification

> WordNet Affect
> Linguistic Inquiry and Word Count (LIWC)

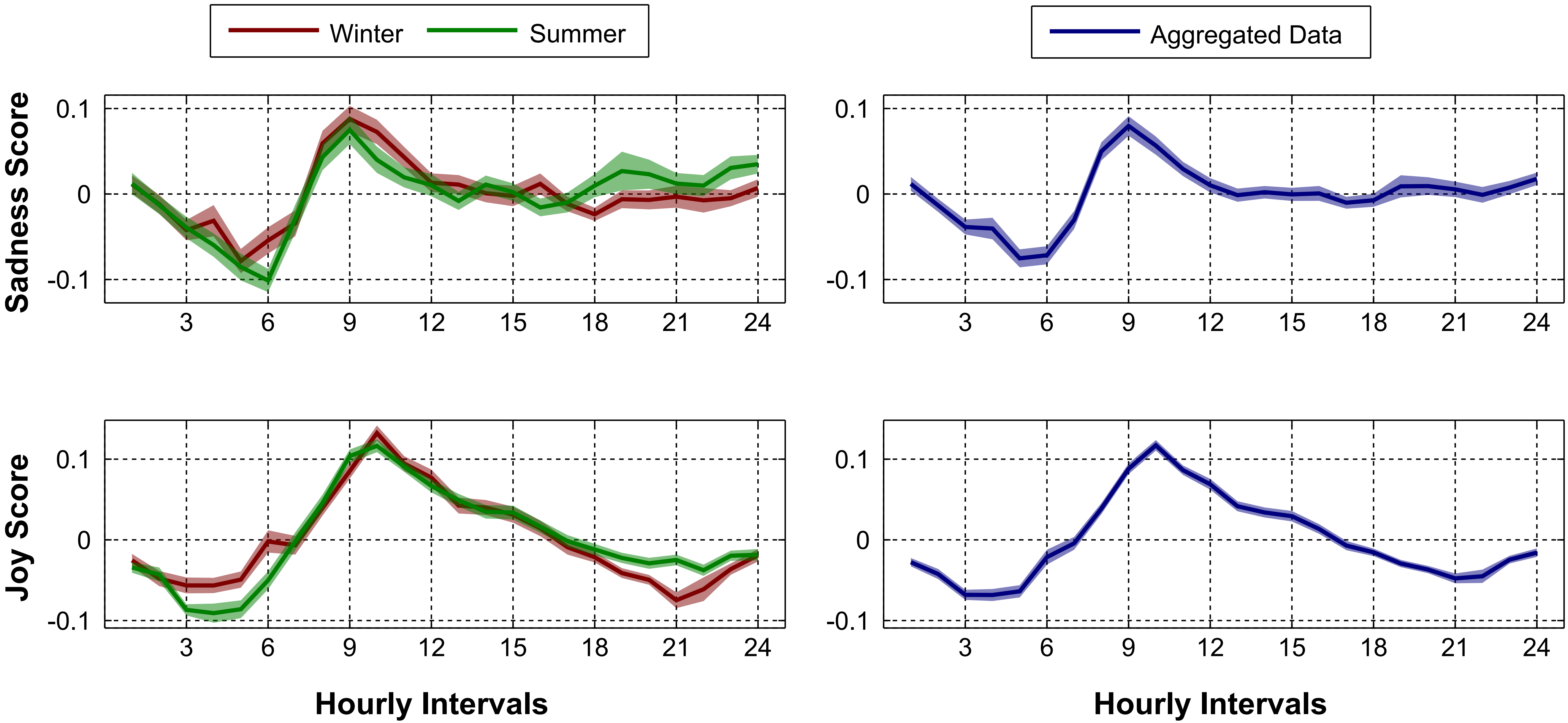
(*Strapparava & Valitutti, 2004*; *Pennebaker et al., 2001, 2007*)

'Emotional' **keywords**, representing

+ **anger**, *e.g. angry, irritate*
+ **fear**, *e.g. fearful, afraid*
+ **joy**, *e.g. cheerful, enthusiastic*
+ **sadness**, *e.g. depressed, gloomy*
+ *plus other emotions*

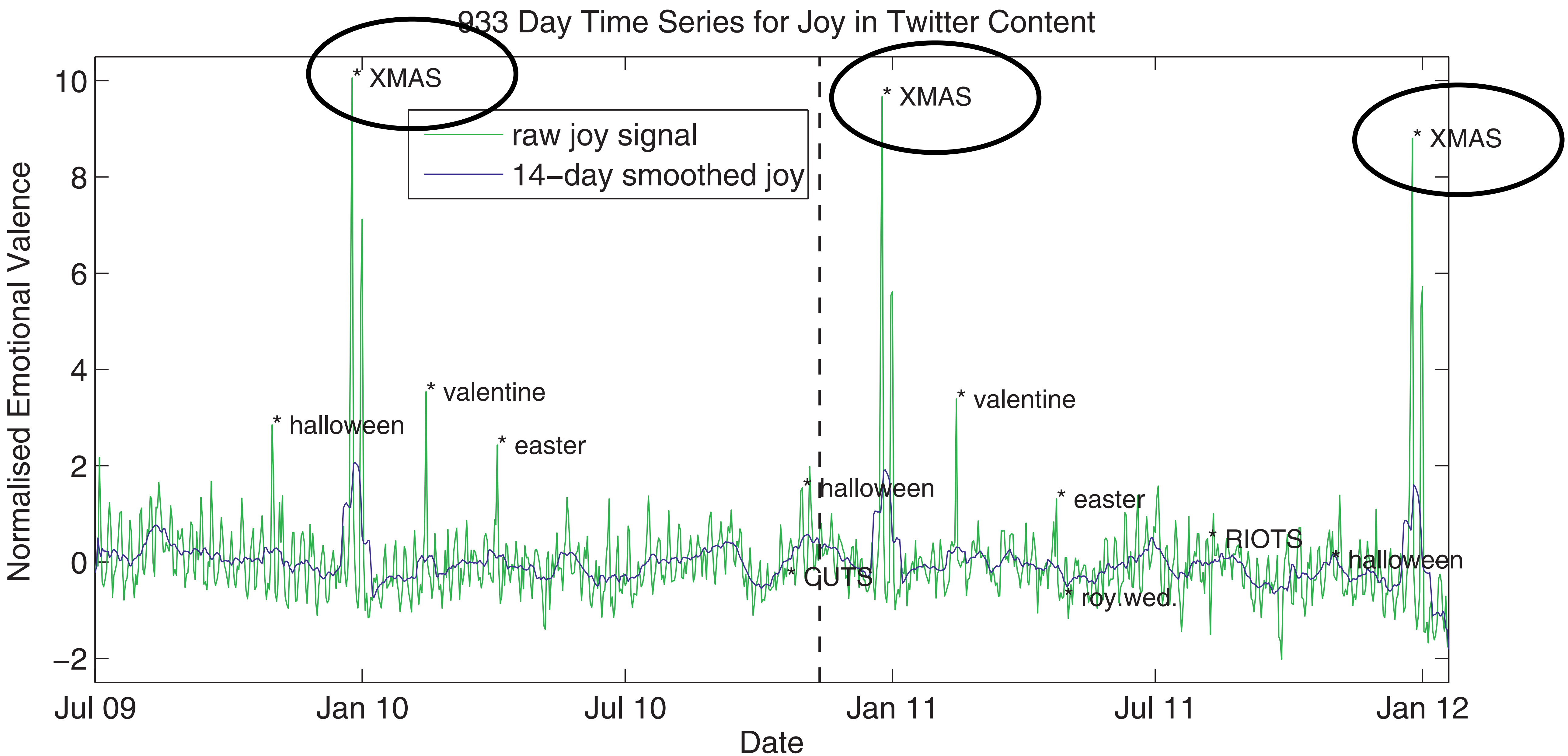
Simply — *but maybe not good enough!* — we compute the **mean keyword frequency score** per emotion

Circadian emotion patterns from Twitter (UK)



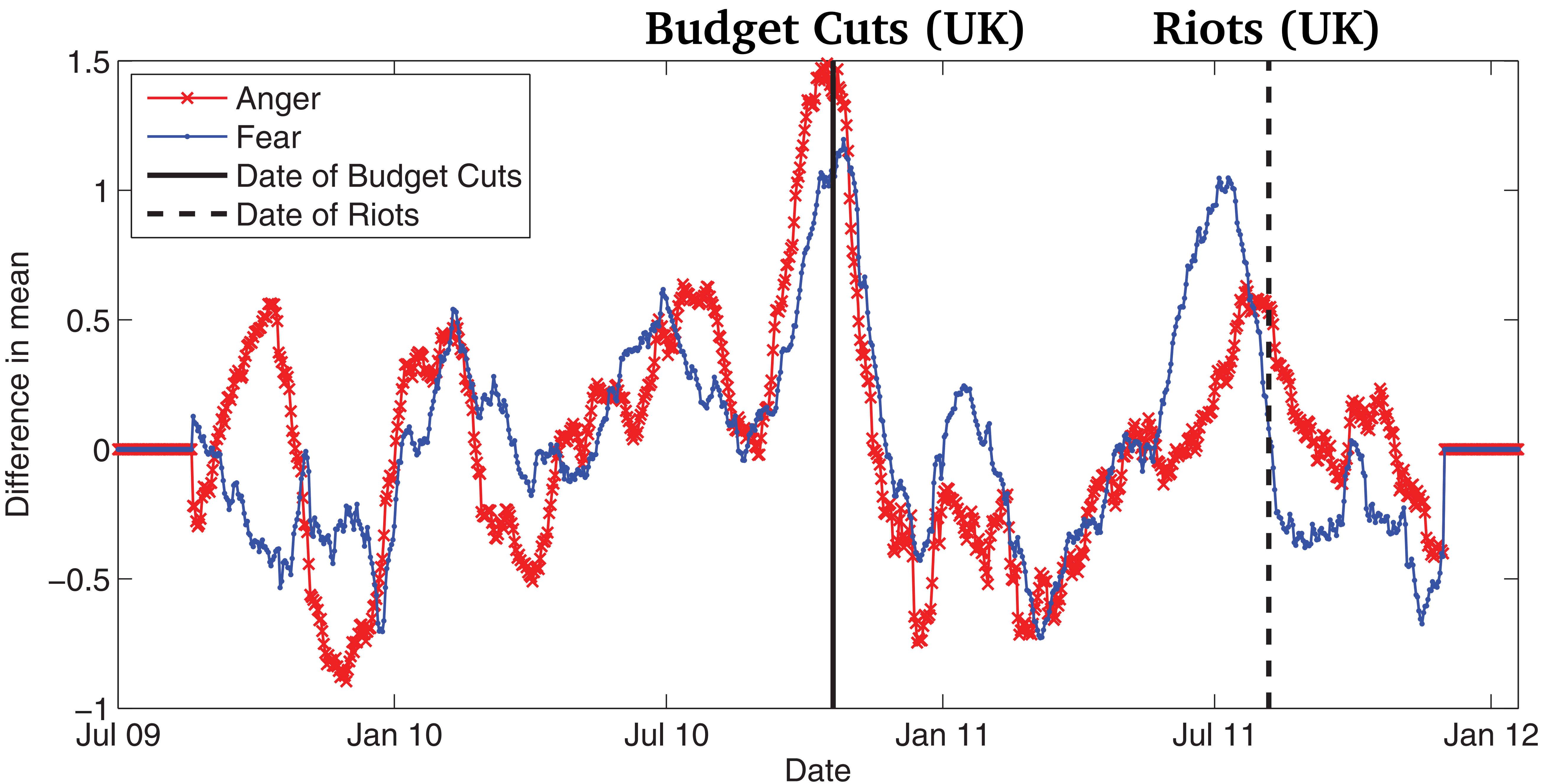
24h emotion patterns for 'joy' and 'sadness' for summer and winter with 95% confidence intervals

'Joy' time series based on Twitter (UK)



Clear peaking pattern during XMAS or other annual celebrations (Valentine's Day, Easter)

Recession, riots, and Twitter emotions (UK)



Difference in mean mood score 50 days prior and after each date; **peaks** indicate **increase in mood change**

Inferring voting intention — Data sets

United Kingdom

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

Austria

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

Regularised text regression

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

$$f(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{w} + \beta$$

Elastic Net (*Zou & Hastie, 2005*)

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta - \sum_{j=1}^{m} x_{ij} w_j \right)^2 + \lambda_1 \underbrace{\sum_{j=1}^{m} |w_j|}_{\textbf{L1-norm}} + \lambda_2 \underbrace{\sum_{j=1}^{m} w_j^2}_{\textbf{L2-norm}} \right\}$$

Regularised text regression

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

$$f(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}} \mathbf{w} + \beta$$

Elastic Net (*Zou & Hastie, 2005*)

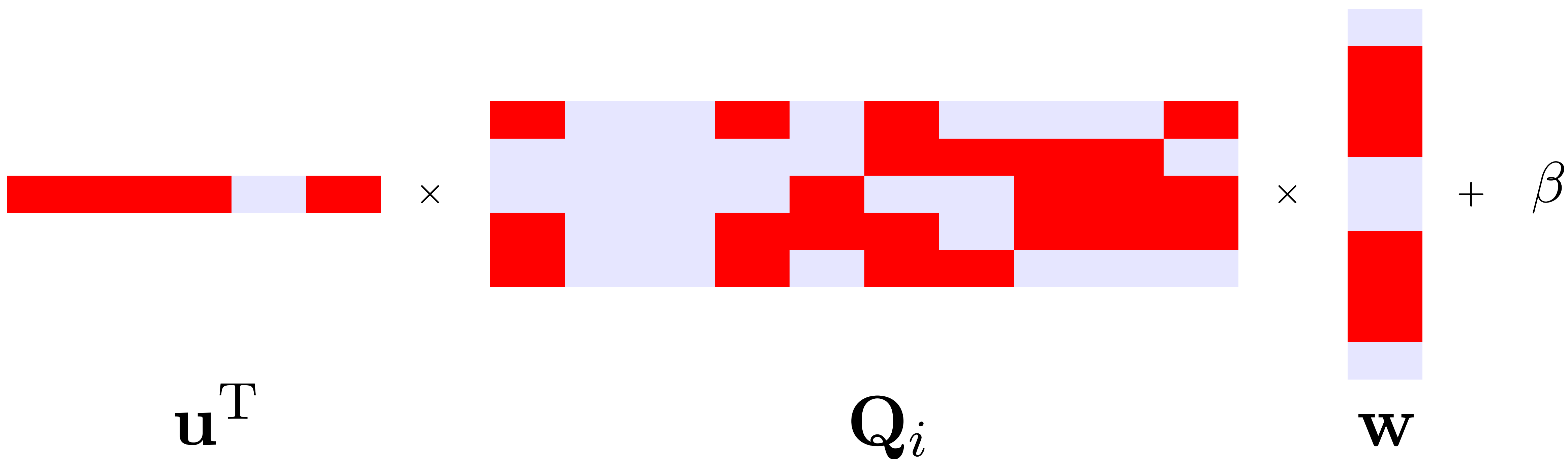
$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta - \sum_{j=1}^{m} x_{ij} w_j \right)^2 + \lambda_1 \sum_{j=1}^{m} |w_j| + \lambda_2 \sum_{j=1}^{m} w_j^2 \right\}$$

L1-norm **L2-norm**

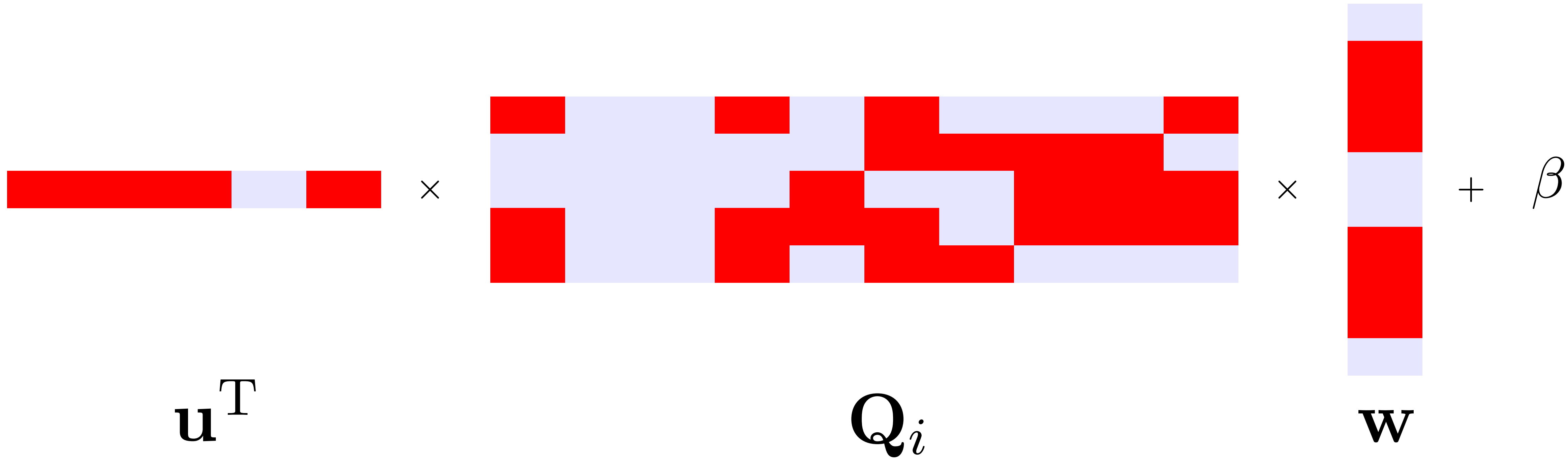
Bilinear (users+text) regularised regression

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

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



Bilinear elastic net (BEN)



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

where

$$\psi(\mathbf{x}, \lambda_1, \lambda_2) = \lambda_1 \|\mathbf{x}\|_{\ell_1} + \lambda_2 \|\mathbf{x}\|_{\ell_2}^2$$

Training bilinear elastic net (BEN)

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

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***)

Training bilinear elastic net (BEN)

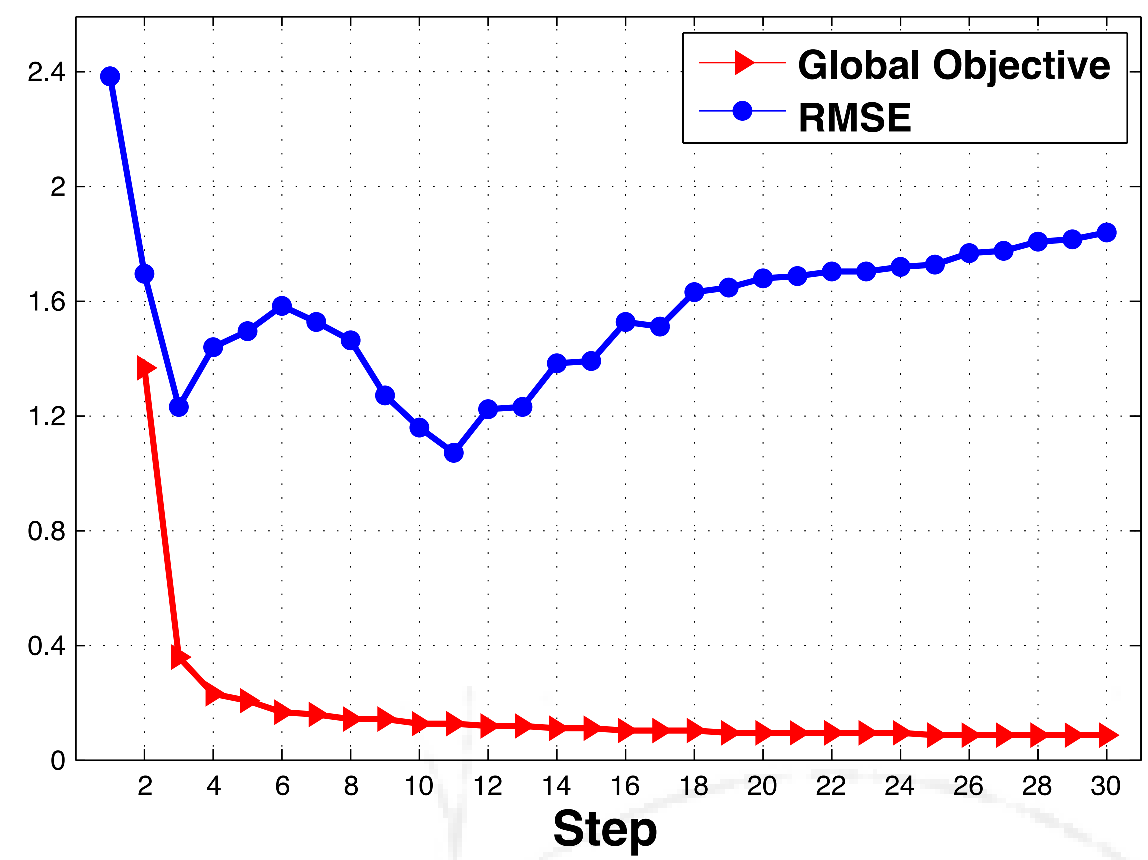
$$\underset{\mathbf{u},\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\mathbf{u}^{\mathrm{T}} \mathbf{Q}_i \mathbf{w} + \beta - y_i \right)^2 + \psi(\mathbf{u}, \theta_{\mathbf{u}}) + \psi(\mathbf{w}, \theta_{\mathbf{w}}) \right\}$$

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***)

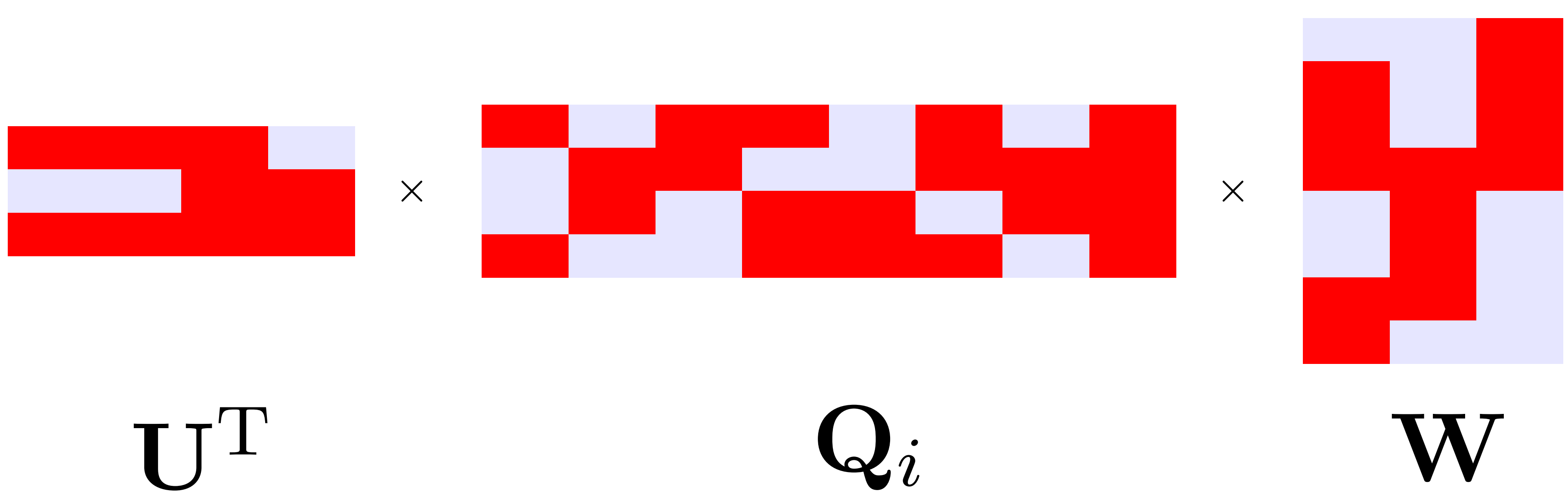
Corresponding prediction error on held out data (***blue***)



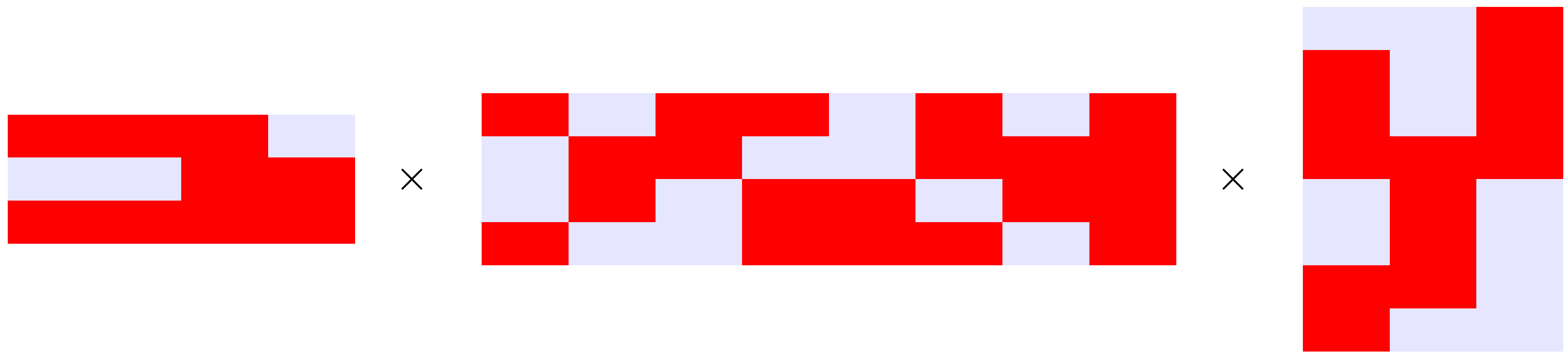
Bilinear and multi-task regression

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

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



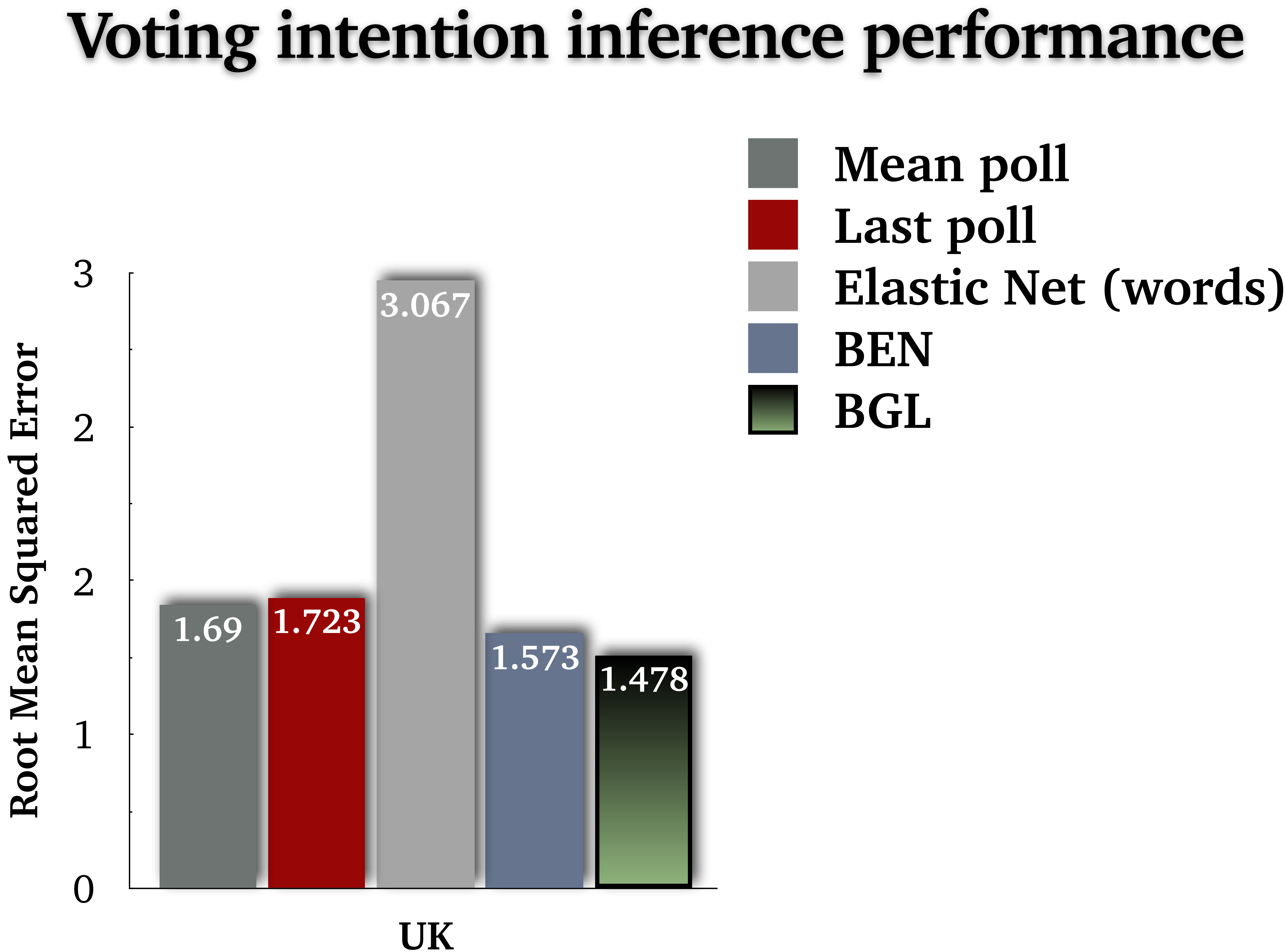
Bilinear Group $L_{2,1}$ (BGL)



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

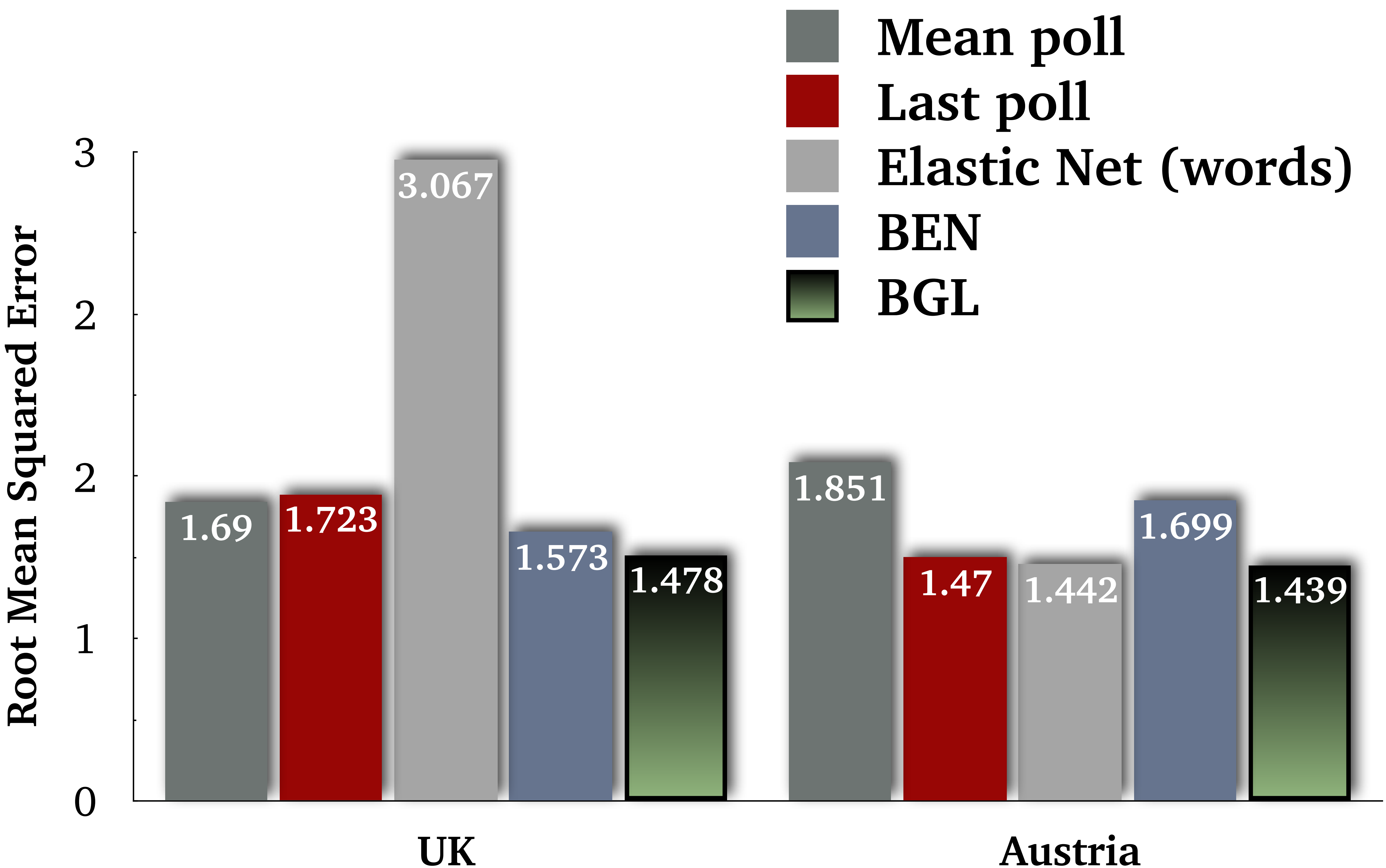
$$\underset{\mathbf{U},\mathbf{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\mathbf{u}^{\mathrm{T}} \mathbf{Q}_i \mathbf{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\mathbf{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\mathbf{W}_j\|_2 \right\}$$

+ a nonzero weighted feature (user or word) is encouraged to be nonzero **for all tasks**, but with potentially different weights
+ intuitive for **political preference inference**

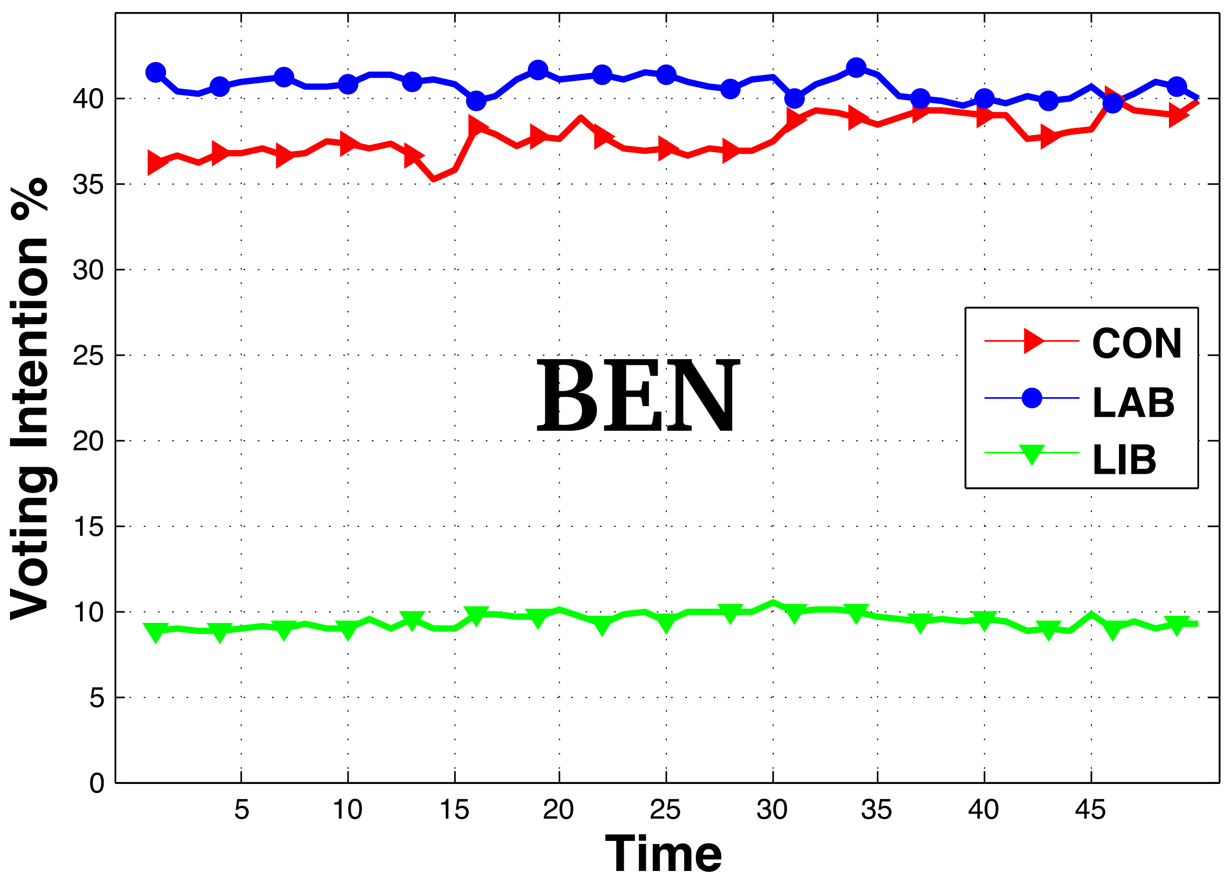


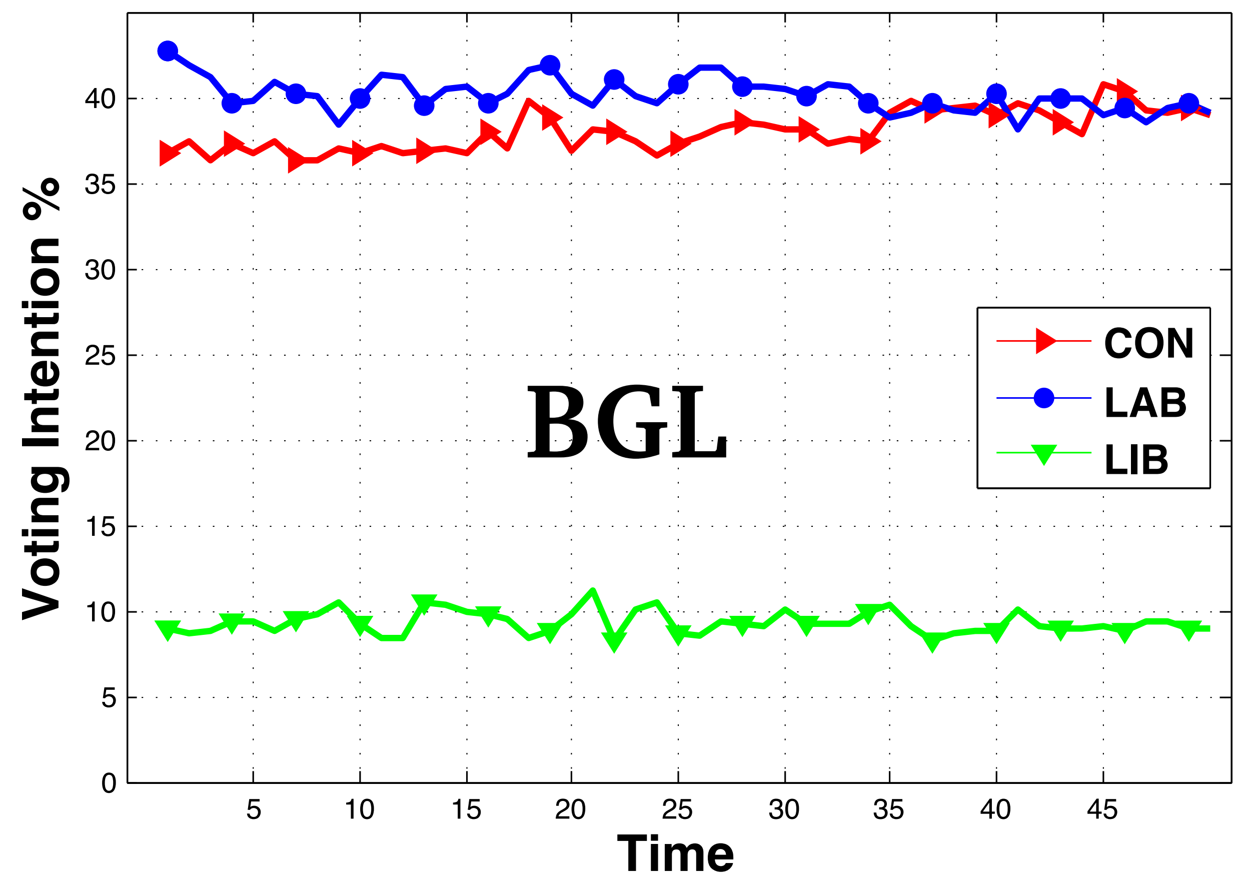
Voting intention inference performance
Mean poll
Last poll
Elastic Net (words)
BEN
BGL
3
2
2
1
0
Root Mean Squared Error
1.69
1.723
3.067
1.573
1.478
UK

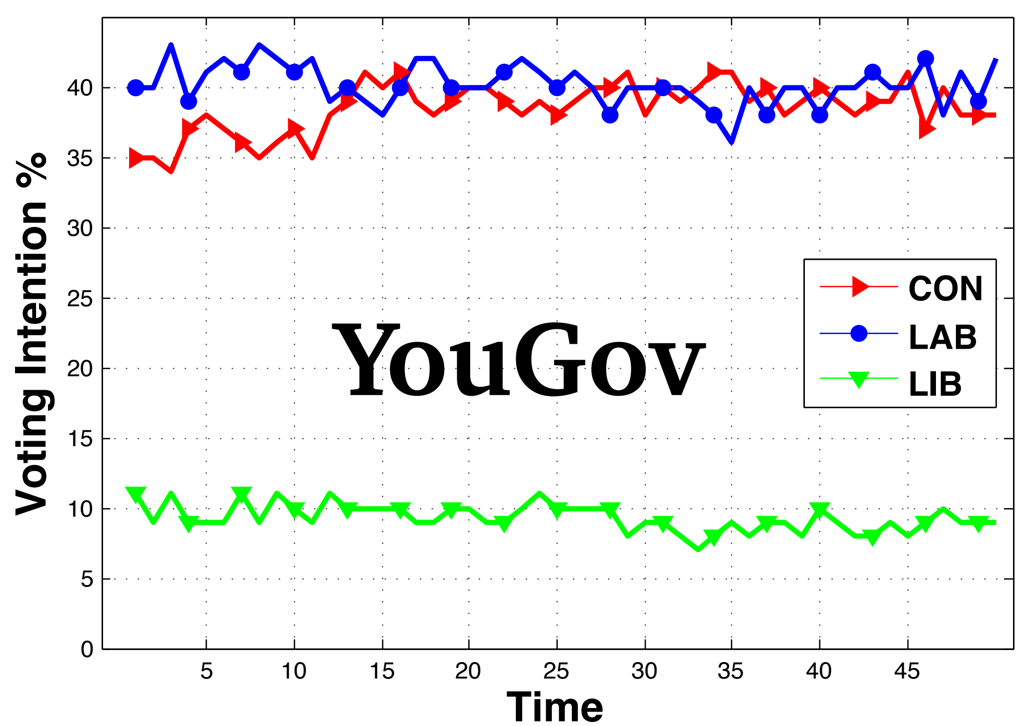
Voting intention inference performance



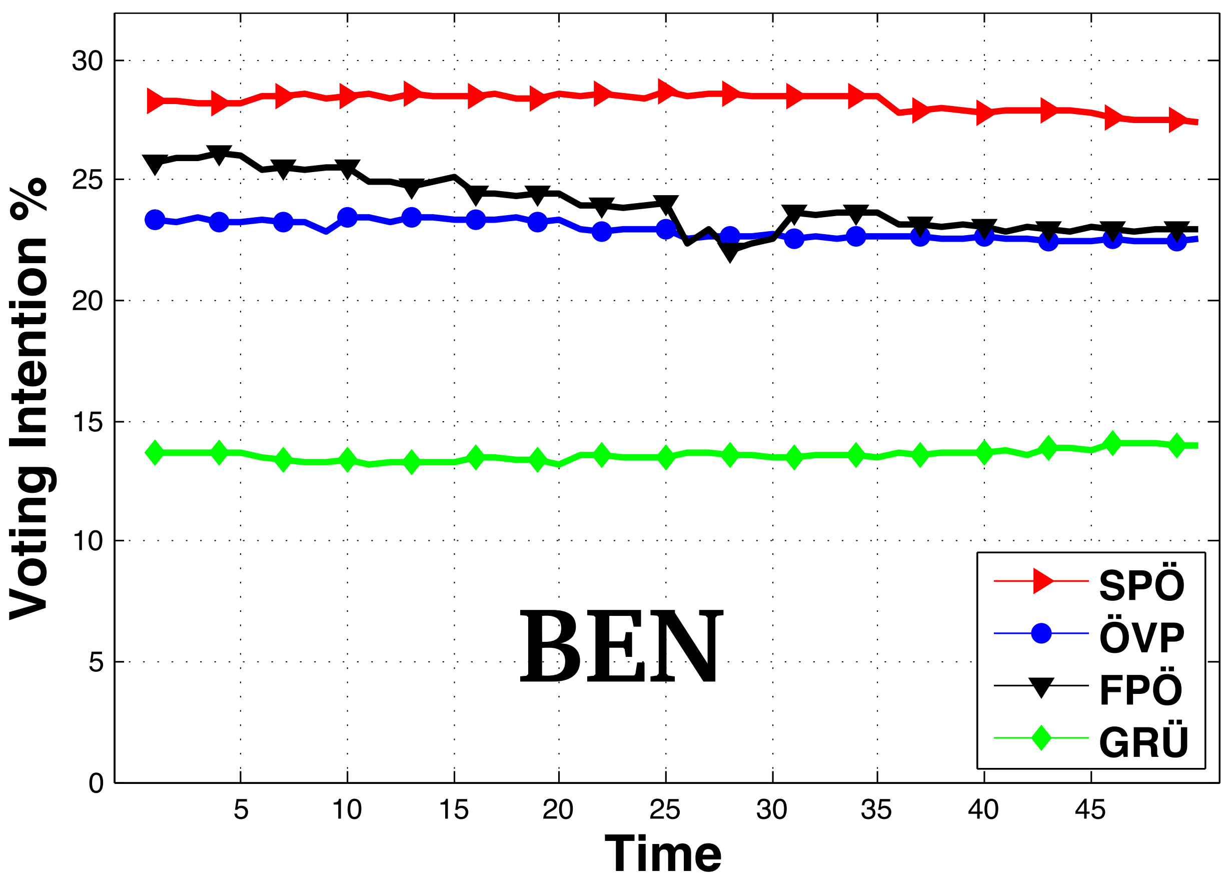
Voting intention comparative plots

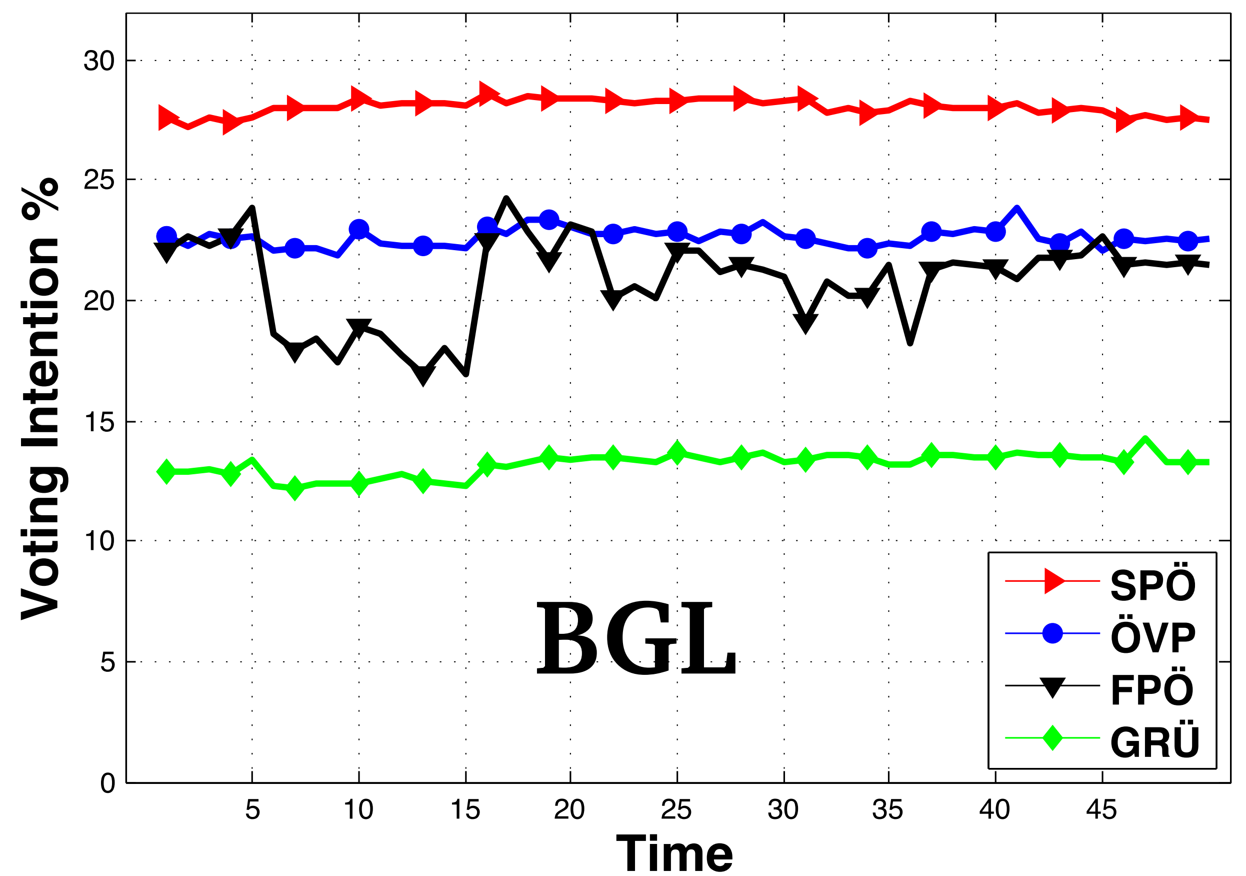


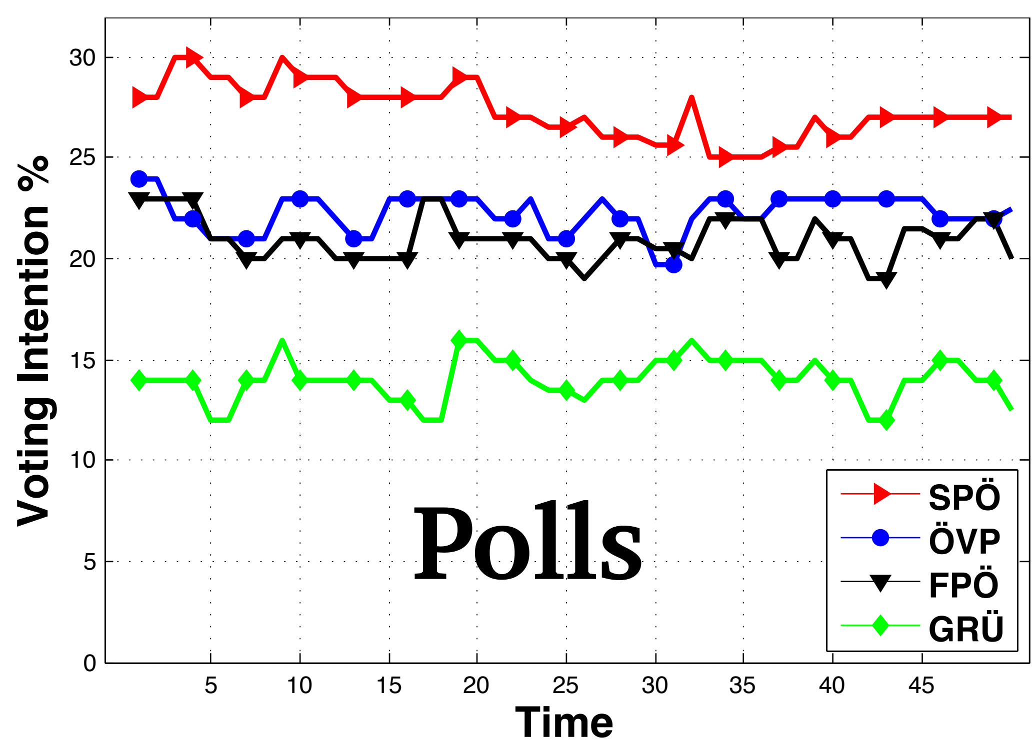




Voting intention comparative plots







Qualitative insights

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**	Normal 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

Inferring user-level information from user-generated content

- occupational class
- income
- socio-economic status (SES)

Preotiuc-Pietro, Lampos & Aletras (ACL 2015)
Preotiuc-Pietro, Volkova, Lampos, Bachrach & Aletras (PLOS ONE, 2015)
Lampos, Aletras, Geyti, Zou & Cox (ECIR 2016)

Linguistic expression and demographics

"Socioeconomic variables are influencing language use."

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

- **Validate this hypothesis** on a broader, larger data set using social media
- **Applications**
  - research, as in computational social science, health, and psychology
  - commercial

Standard Occupational Classification (SOC)

Major Group 1 (**C1**): Managers, Directors and Senior Officials
    Sub-major Group 11: Corporate Managers and Directors
        Minor Group 111: Chief Executives and Senior Officials
            Unit Group 1115: Chief Executives and Senior Officials
            •Job: chief executive, bank manager
            Unit Group 1116: Elected Officers and Representatives
        Minor Group 112: Production Managers and Directors
        Minor Group 113: Functional Managers and Directors
        Minor Group 115: Financial Institution Managers and Directors
        Minor Group 116: Managers and Directors in Transport and Logistics
        Minor Group 117: Senior Officers in Protective Services
        Minor Group 118: Health and Social Services Managers and Directors
        Minor Group 119: Managers and Directors in Retail and Wholesale
    Sub-major Group 12: Other Managers and Proprietors
Major Group (**C2**): Professional Occupations
            •Job: mechanical engineer, pediatrist
Major Group (**C3**): Associate Professional and Technical Occupations
            •Job: system administrator, dispensing optician
Major Group (**C4**): Administrative and Secretarial Occupations
            •Job: legal clerk, company secretary
Major Group (**C5**): Skilled Trades Occupations
            •Job: electrical fitter, tailor
Major Group (**C6**): Caring, Leisure and Other Service Occupations
            •Job: nursery assistant, hairdresser
Major Group (**C7**): Sales and Customer Service Occupations
            •Job: sales assistant, telephonist
Major Group (**C8**): Process, Plant and Machine Operatives
            •Job: factory worker, van driver
Major Group (**C9**): Elementary Occupations
            •Job: shelf stacker, bartender

provided by the Office for National Statistics (UK)

9 major groups

25 sub-major groups

90 minor groups

369 unit groups

Standard Occupational Classification (SOC)

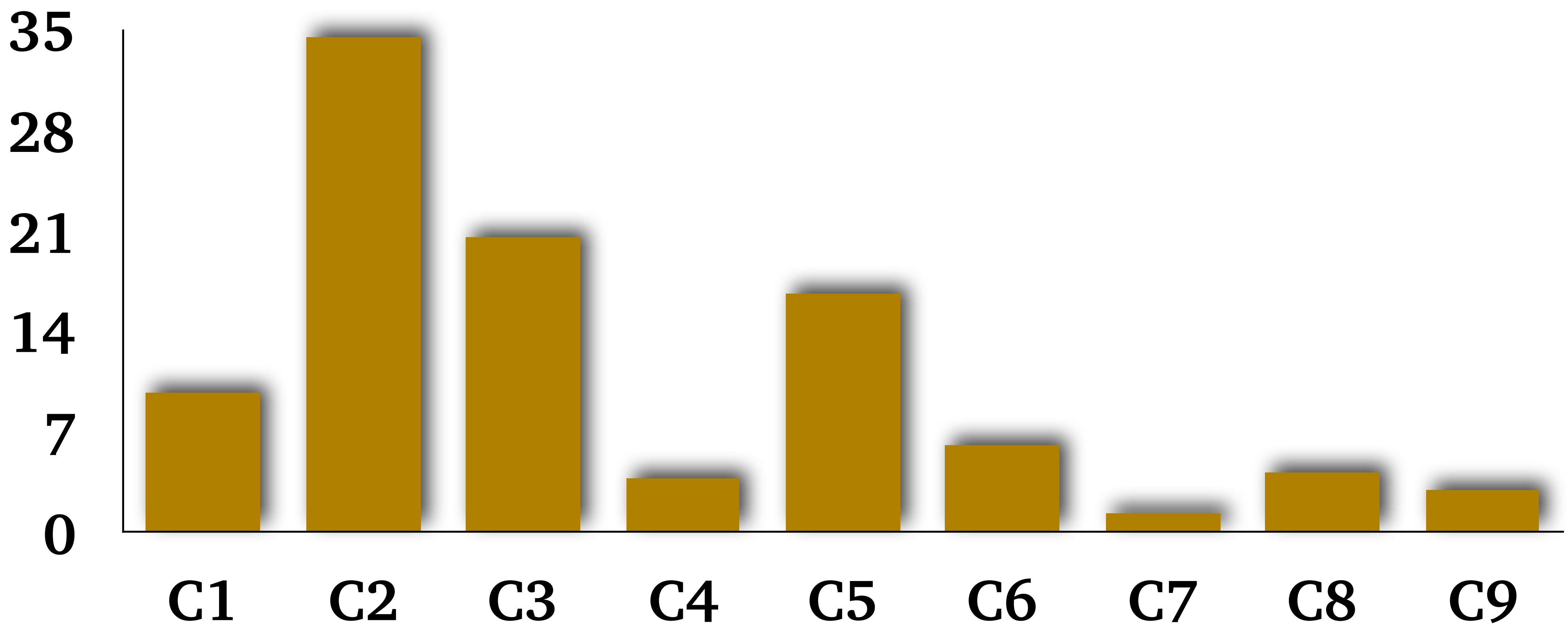
The 9 major occupational classes (C1-9)

C1 — Managers, Directors & Senior Officials
(*chief executive, bank manager*)
C2 — Professional Occupations (*postdoc, pediatrist*)
C3 — Associate Professional & Technical
(*system administrator, dispensing optician*)
C4 — Administrative & Secretarial (*legal clerk, secretary*)
C5 — Skilled Trades (*electrical fitter, tailor*)
C6 — Caring, Leisure, Other Service
(*nursery assistant, hairdresser*)
C7 — Sales & Customer Service (*sales assistant, telephonist*)
C8 — Process, Plant and Machine Operatives
(*factory worker, van driver*)
C9 — Elementary (*shelf stacker, bartender*)

Forming a Twitter user data set

+ **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



Twitter user attributes (*18 in total*)

number of
— followers
— friends
— followers/friends (ratio)
— times listed
— tweets
— favourites (likes)
— unique @-mentions
— tweets/day (avg.)
— retweets/tweet (avg.)

proportion of
— retweets done
— non duplicate tweets
— retweeted tweets
— hashtags
— tweets with hashtags
— tweets with @-mentions
— @-replies
— tweets with links
— tweets in English

Similarly to our paper for user impact estimation (*Lampos et al., 2014*)

Twitter user discussion topics (I)

Topics — Word clusters (#: 30, 50, 100, **200**)

+ ***SVD*** on the graph laplacian of the word by word similarity matrix using ***normalised PMI***, *i.e.* a form of spectral clustering (*Bouma, 2009*; *von Luxburg, 2007*)

+ ***Word2vec*** (skip-gram with negative sampling) to learn word embeddings; pairwise ***cosine similarity*** on the embeddings to derive a word by word similarity matrix; then spectral clustering on the similarity matrix (*Mikolov et al., 2013*)

Twitter user discussion topics (II)

Topic	Most central words; *Most frequent words*
Arts	archival, stencil, canvas, minimalist; *art, design, print*
Health	chemotherapy, diagnosis, disease; *risk, cancer, mental, stress*
Beauty Care	exfoliating, cleanser, hydrating; *beauty, natural, dry, skin*
Higher Education	undergraduate, doctoral, academic, students, curriculum; *students, research, board, student, college, education, library*
Football	bardsley, etherington, gallas; *van, foster, cole, winger*
Corporate	consortium, institutional, firm's; *patent, industry, reports*
Elongated Words	yaaayy, wooooo, woooo, yayyyyy, yaaaaay, yayayaya, yayy; *wait, till, til, yay, ahhh, hoo, woo, woot, whoop, woohoo*
Politics	religious, colonialism, christianity, judaism, persecution, fascism, marxism; *human, culture, justice, religion, democracy*

A few words about Gaussian Processes

Say $\boldsymbol{x} \in \mathbb{R}^d$ and we want to learn $f : \mathbb{R}^d \to \mathbb{R}$

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

mean function drawn on inputs

covariance function (kernel) drawn on pairs of inputs

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

Why do we use Gaussian Processes?

+ Kernelised, models nonlinearities
+ Interpretability (**A**uto**R**elevance **D**etermination)
+ Performance

(*Rasmussen & Williams, 2006*)

More information about Gaussian Processes

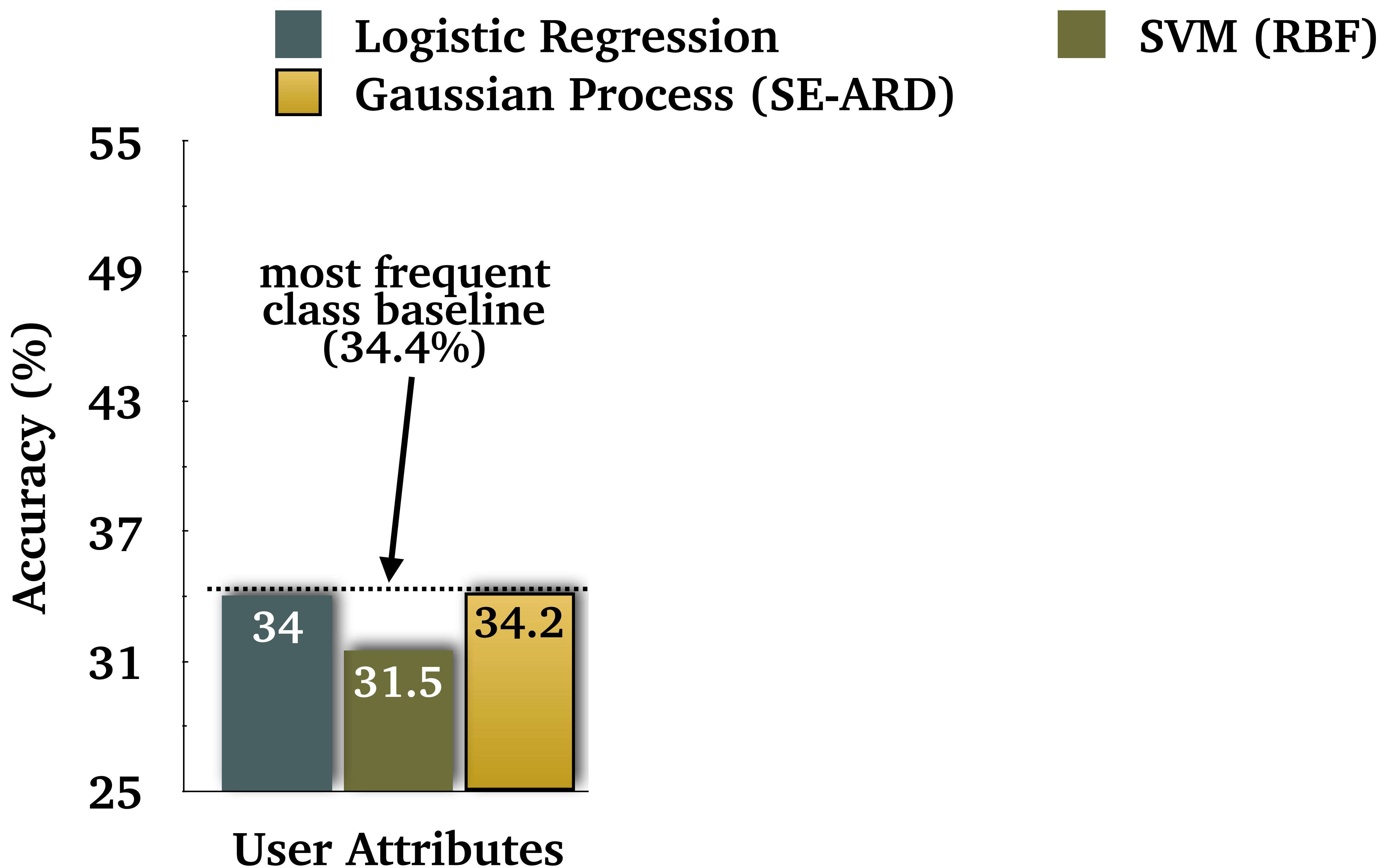
- Book: "*Gaussian Processes for Machine Learning*"
  http://www.gaussianprocess.org/gpml/
- Video-lecture: "*Gaussian Process Basics*"
  http://videolectures.net/gpip06_mackay_gpb/
- Tutorial tailored to statistical NLP tasks: "*Gaussian Processes for Natural Language Processing*"
  http://people.eng.unimelb.edu.au/tcohn/tutorial.html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

Gaussian Process classifier

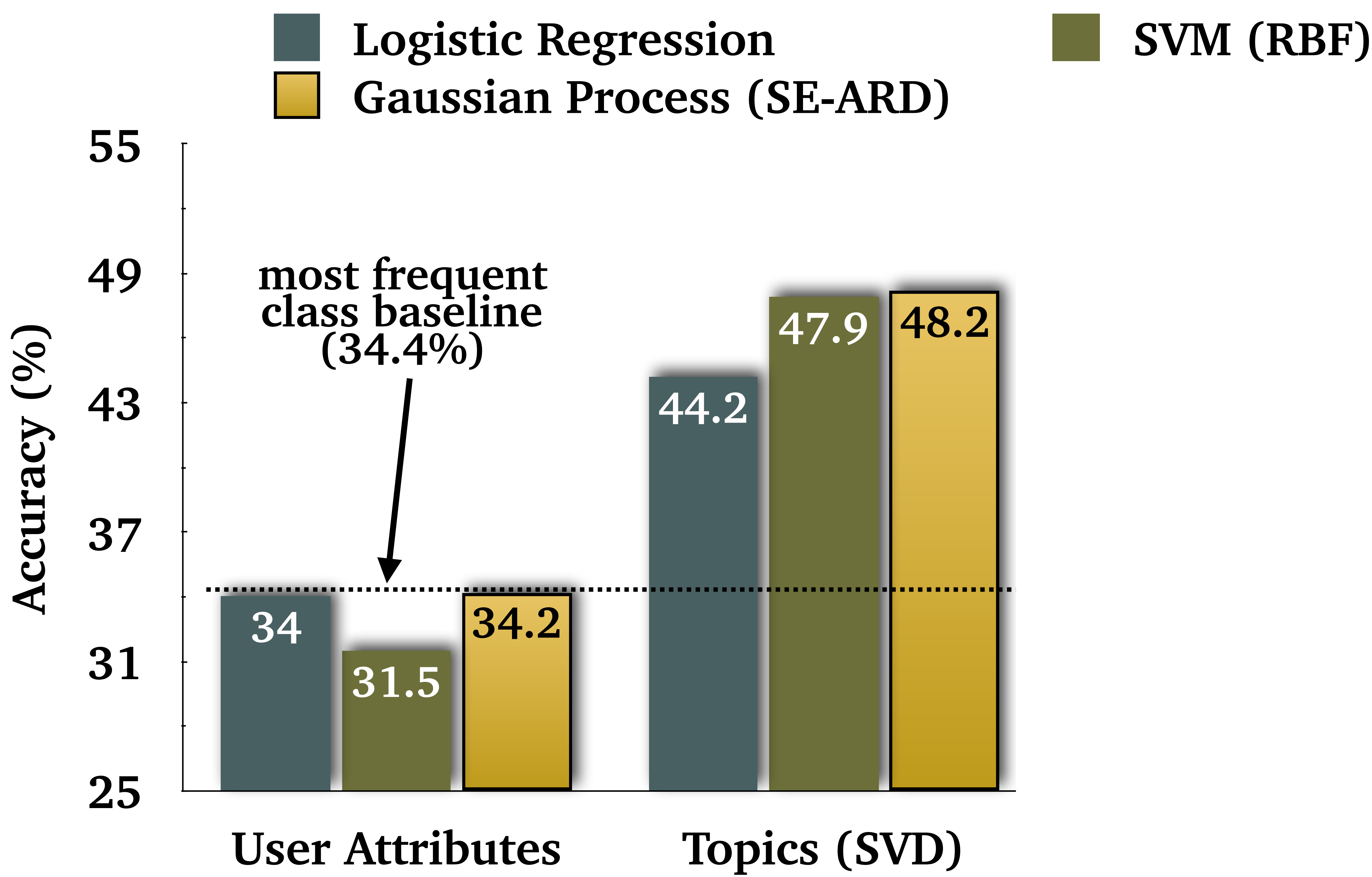
$$k_{\mathrm{ard}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left[\sum_i^d -\frac{(x_i - x_i')^2}{2l_i^2}\right]$$

- Squared-exponential ARD covariance function: determines (quantify) the relevancy of each user feature, *i.e.* **the relevance of feature *i* is inversely proportional to the length-scale hyper-parameter l_i**
- **9-class classification** using one vs. all
- GP hyper-parameter learning with **Expectation Propagation**
- Inference using **FITC** (500 inducing points)

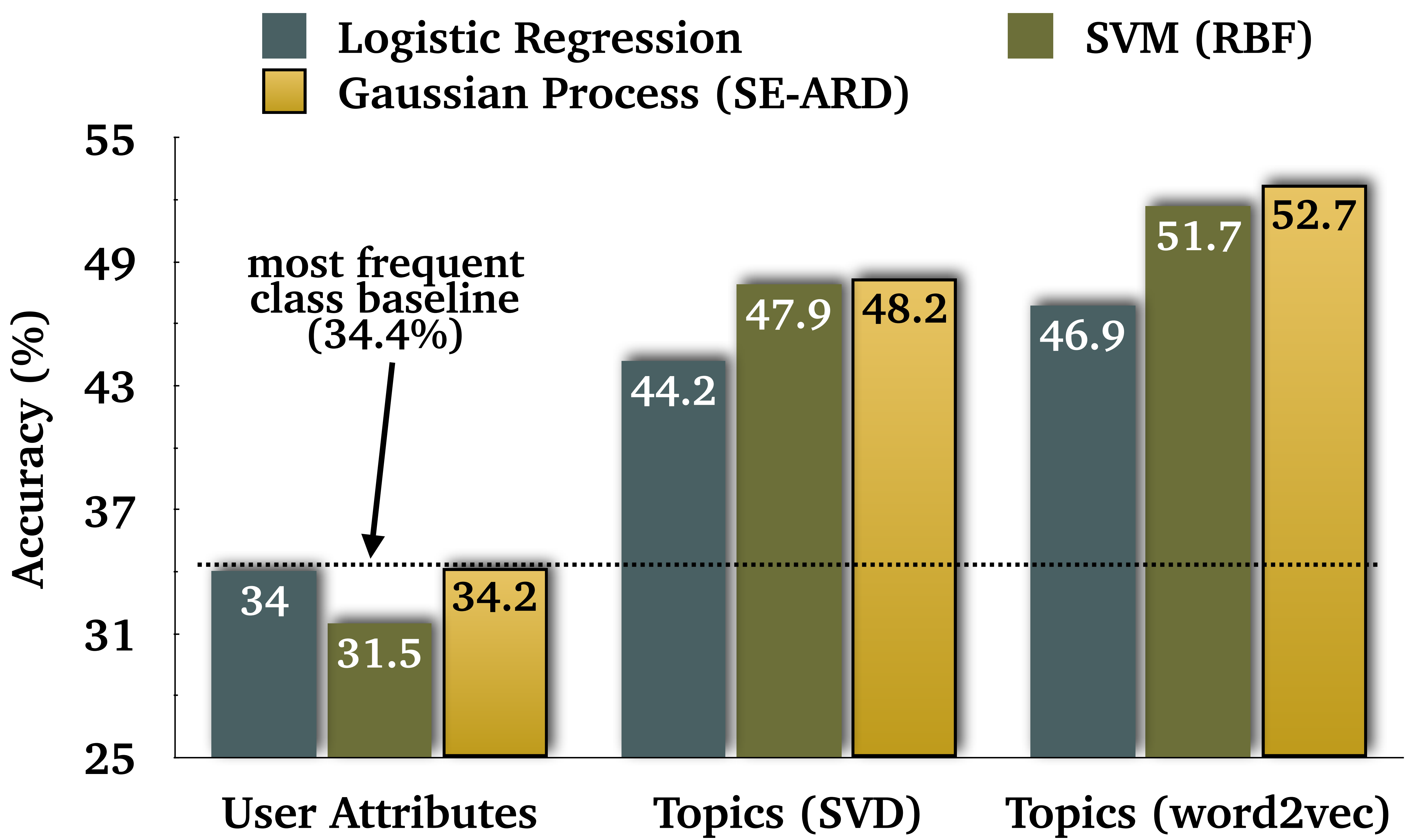
Occupation classification performance



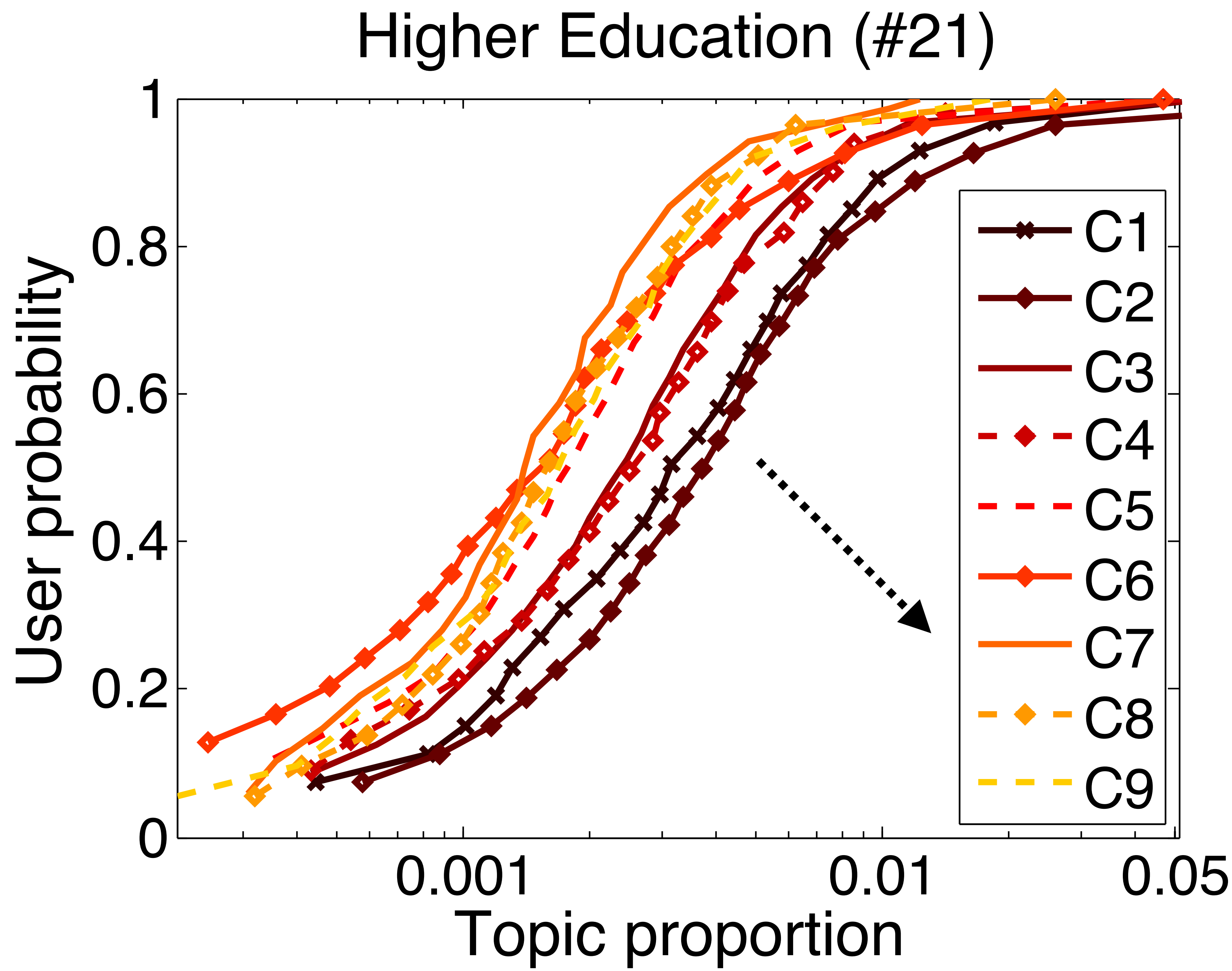
Occupation classification performance



Occupation classification performance

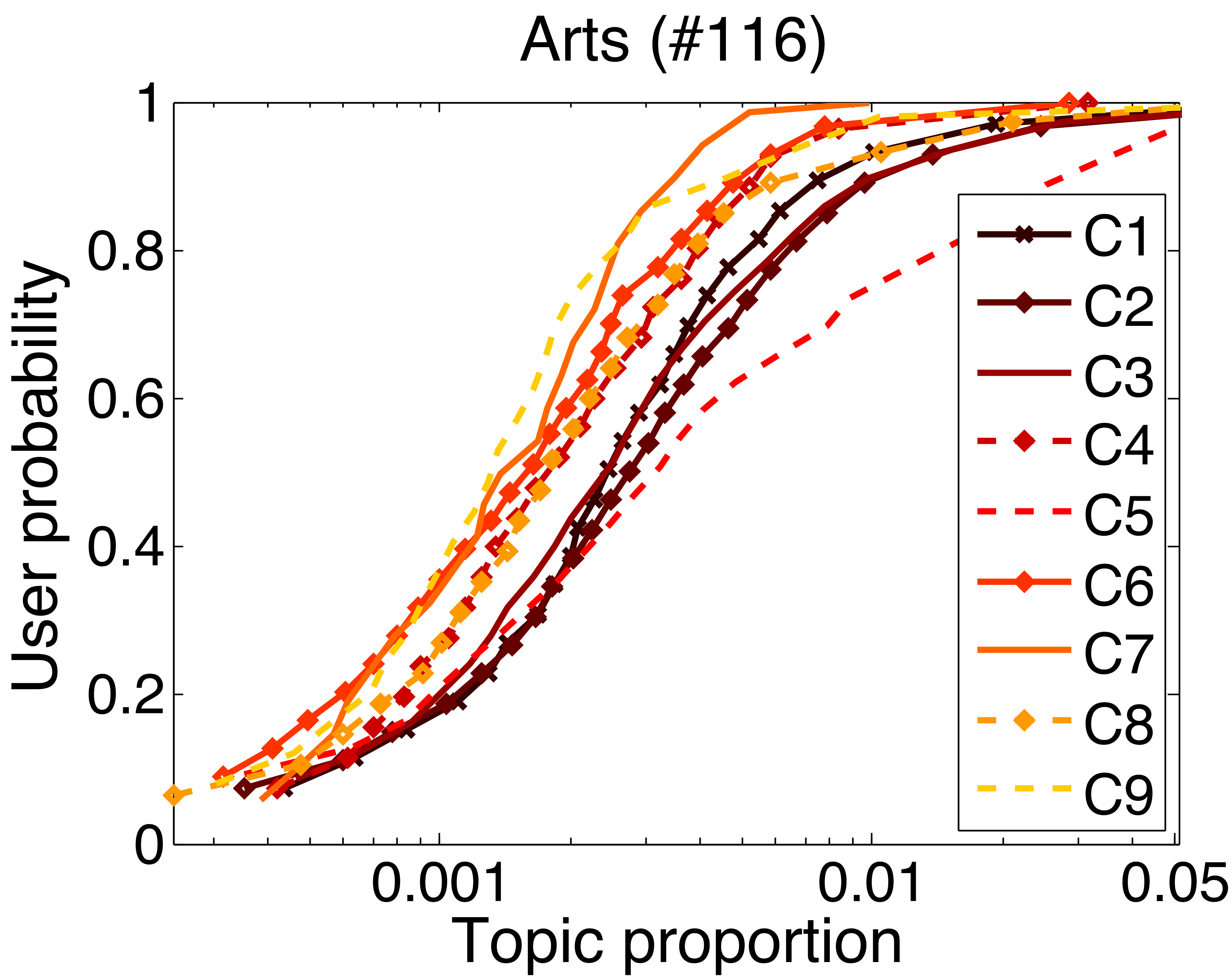


Occupation classification insights (I)



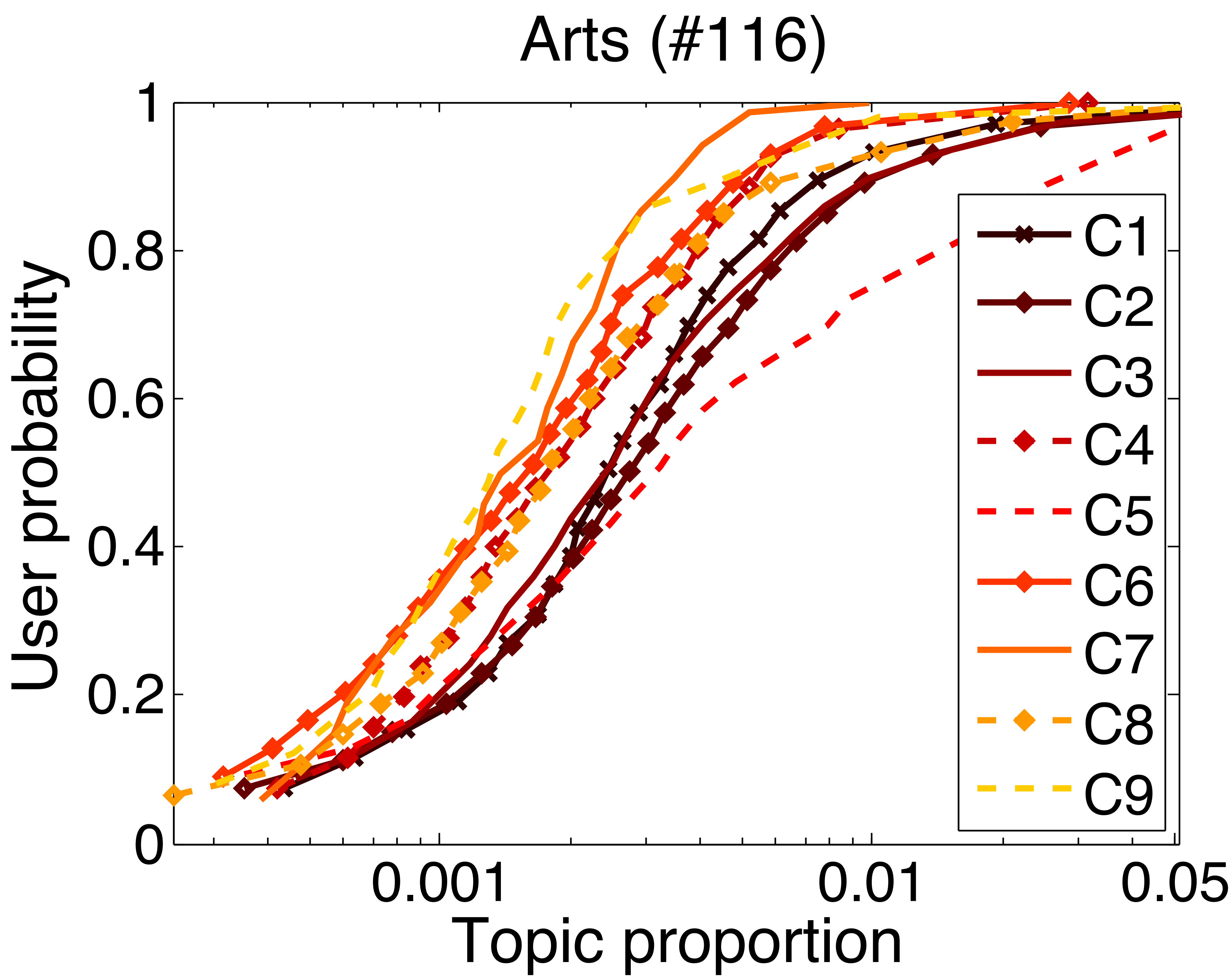
CDF of the topic "**Higher Education**": Topic **more prevalent in the upper classes** (C2, which includes education professionals, and C1), and less so in the lower classes

Occupation classification insights (II)



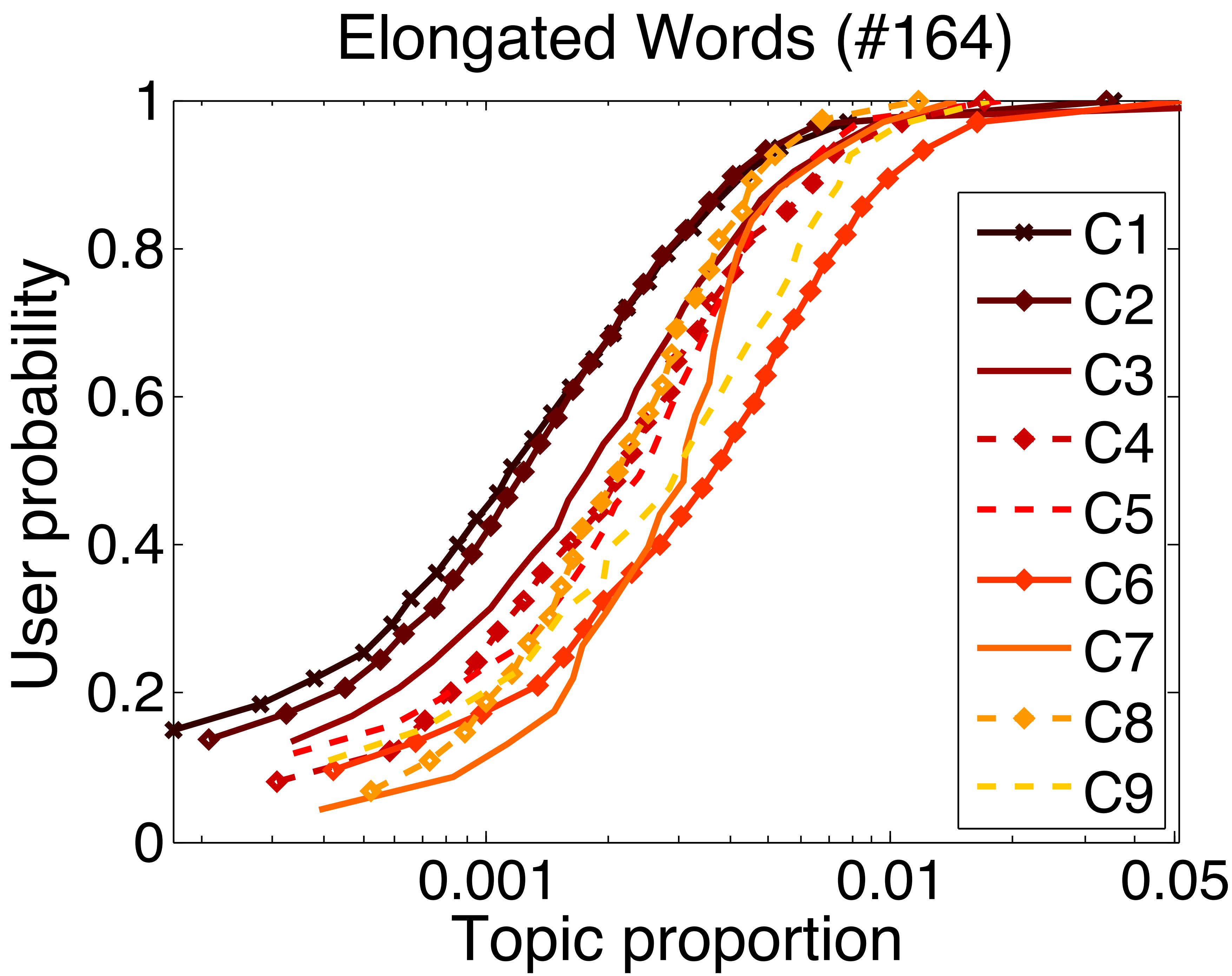
CDF of the topic "**Arts**": Topic **more prevalent in C5** (which includes artists) and **the upper classes**

Occupation classification insights (II)



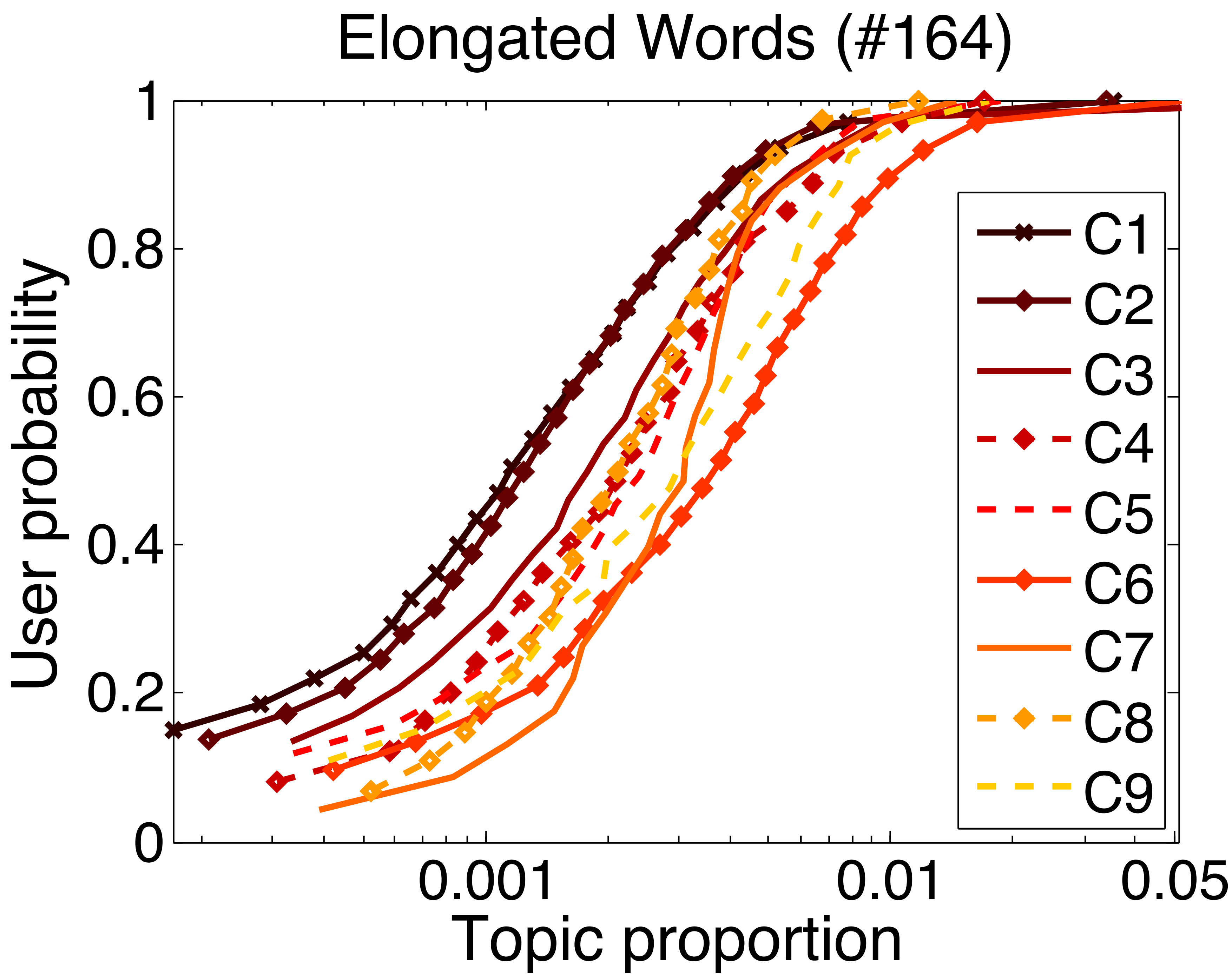
CDF of the topic "**Arts**": Topic **more prevalent in C5** (which includes artists) and **the upper classes**

Occupation classification insights (III)



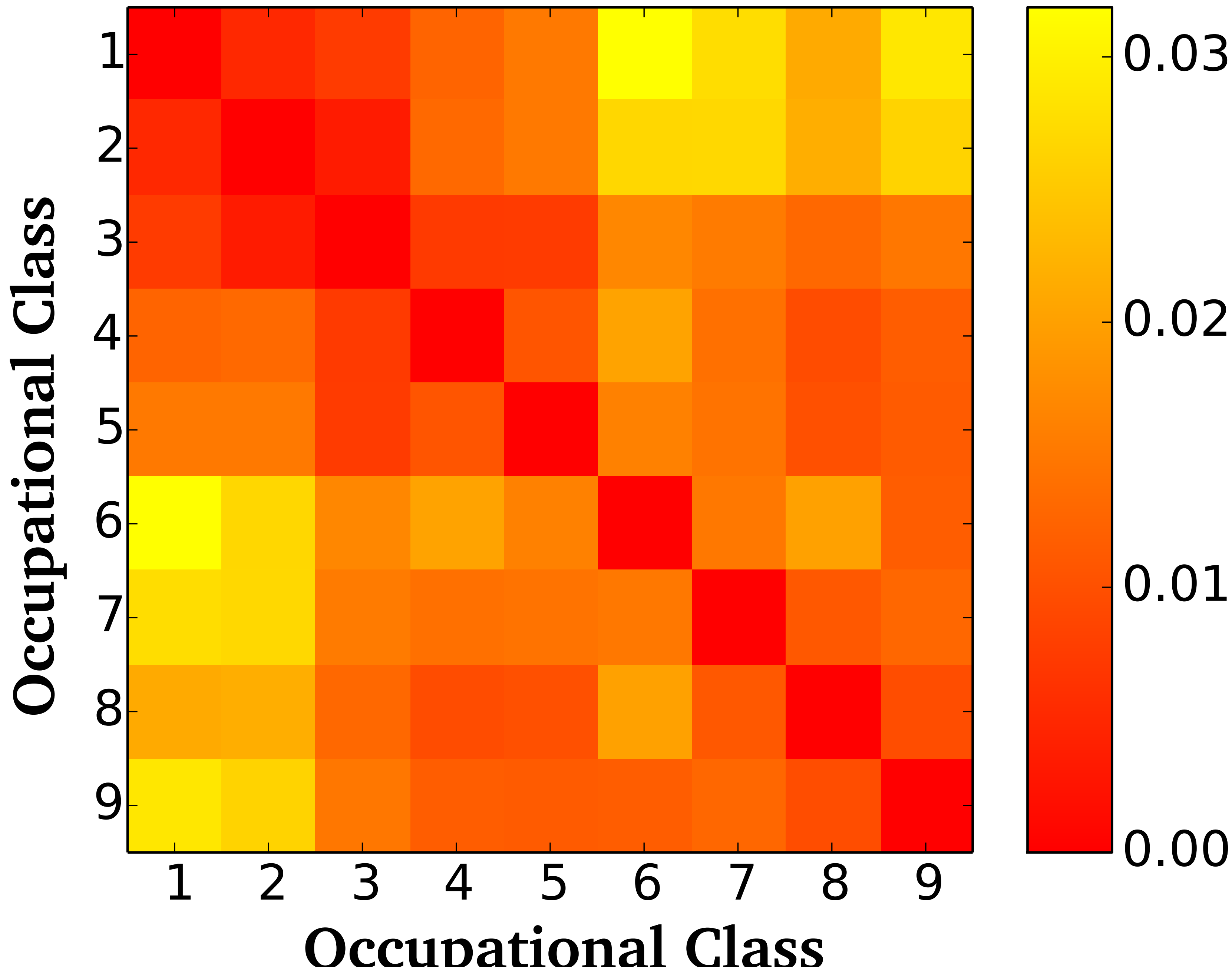
CDF of the topic "Elongated Words": Topic more prevalent in the lower classes, and less so in the upper classes

Occupation classification insights (III)



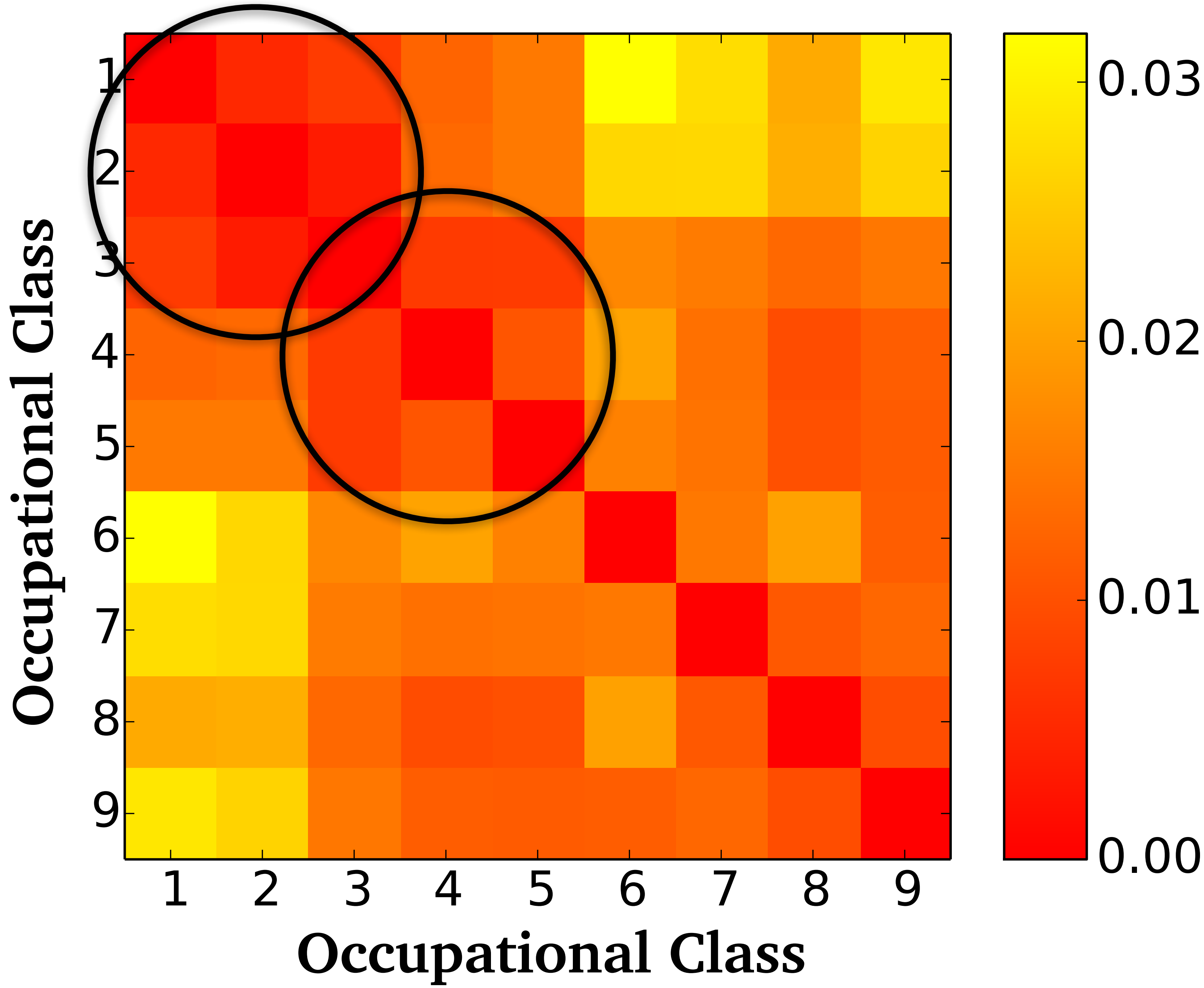
CDF of the topic "**Elongated Words**": Topic **more prevalent in the lower classes**, and less so in the upper classes

Occupation classification insights (IV)



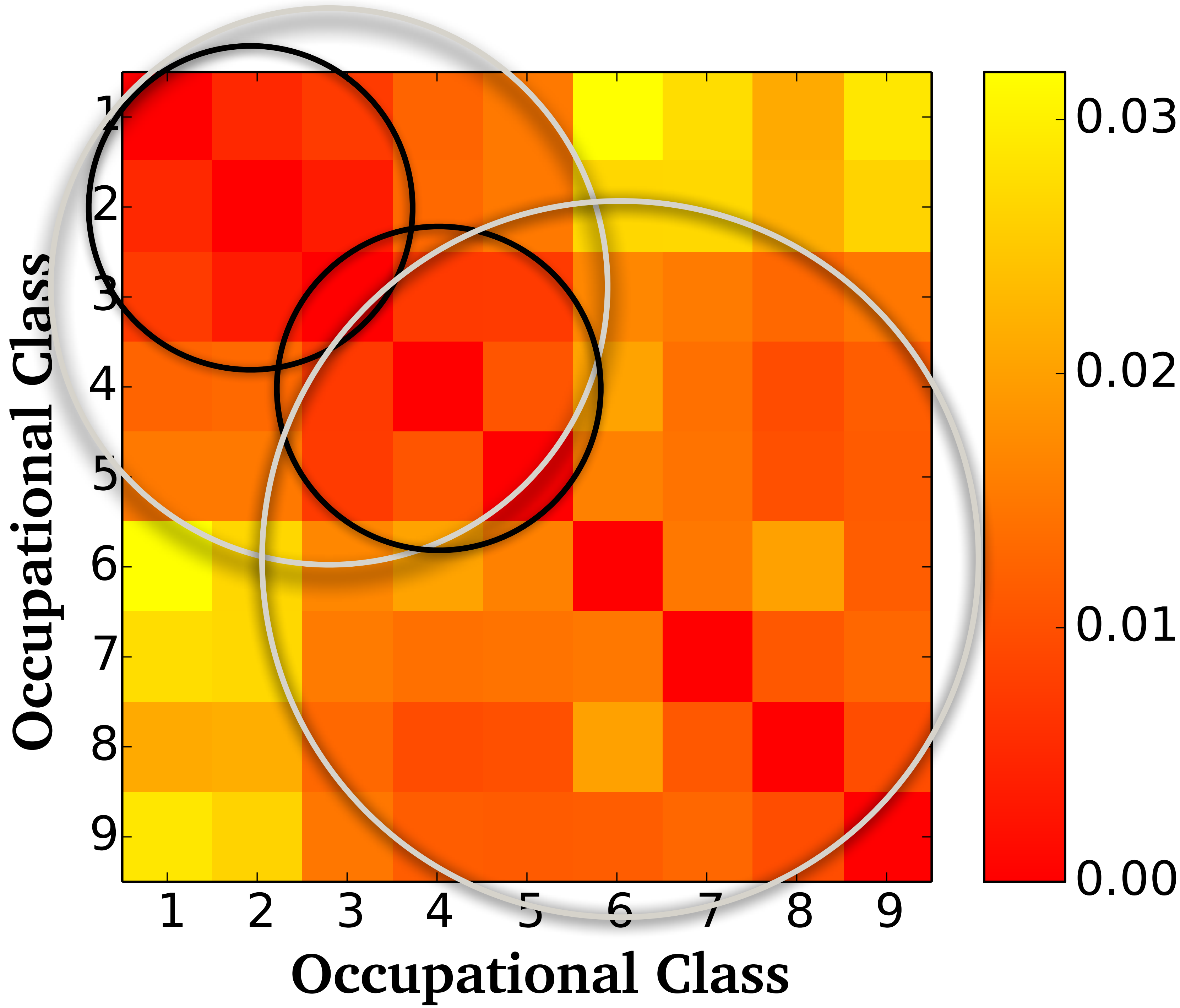
Topic distribution distance (*Jensen-Shannon divergence*)
for the different occupational classes (1-9)

Occupation classification insights (IV)



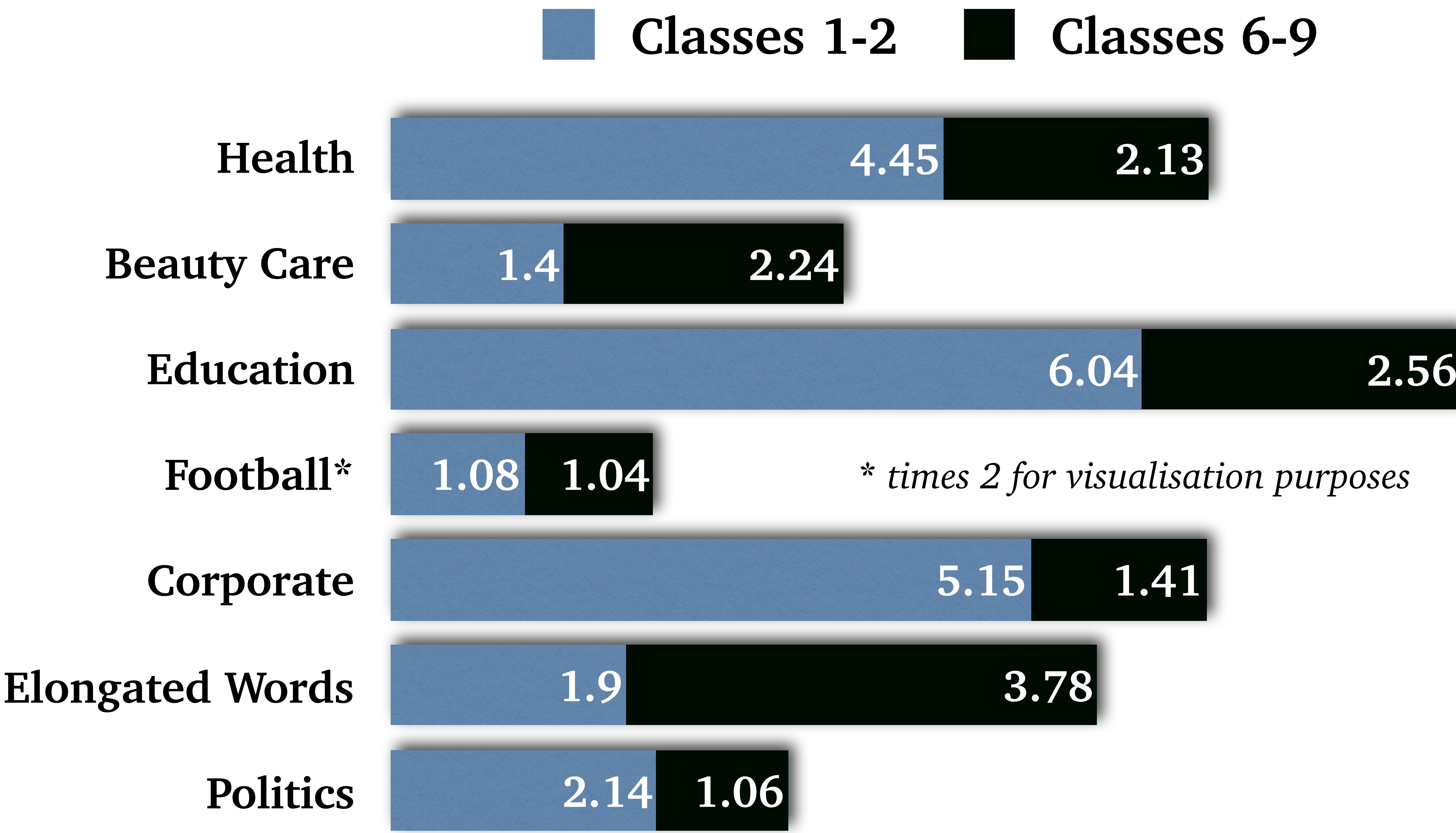
Topic distribution distance (*Jensen-Shannon divergence*)
for the different occupational classes (1-9)

Occupation classification insights (IV)



Topic distribution distance (*Jensen-Shannon divergence*)
for the different occupational classes (1-9)

Occupation classification insights (V)



Topic scores for occupational class supersets

Additional 'perceived' user features

+ Previously used features: **Profile** features, **Shallow** profile features, and **Topics**

+ Based on the work of *Volkova et al. (2015)*, we also incorporated:
  > Inferred Psycho-**Demo**graphic 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.

Defining the user income regression task

Group 112: Production Managers and Directors (50,952 GBP/year)	
•Job titles: engineering manager, managing director, production manager, construction manager, quarry manager, operations manager	
Group 241: Conservation and Environment Professionals (53,679 GBP/year)	
•Job titles: conservation officer, ecologist, energy conservation officer, heritage manager, marine conservationist, energy manager, environmental consultant, environmental engineer, environmental protection officer, environmental scientist, landfill engineer	
Group 312: Draughtspersons and Related Architectural Technicians (29,167 GBP/year)	
•Job titles: architectural assistant, architectural, technician, construction planner, planning enforcement officer, cartographer, draughtsman, CAD operator	
Group 411: Administrative Occupations: Government and Related Organisations (20,373 GBP/year)	
•Job titles: administrative assistant, civil servant, government clerk, revenue officer, benefits assistant, trade union official, research association secretary	
Group 541: Textiles and Garments Trades (18,986 GBP/year)	
•Job titles: knitter, weaver, carpet weaver, curtain maker, upholsterer, curtain fitter, cobbler, leather worker, shoe machinist, shoe repairer, hosiery cutter, dressmaker, fabric cutter, tailor, tailoress, clothing manufacturer, embroiderer, hand sewer, sail maker, upholstery cutter	
Group 622: Hairdressers and Related Services (10,793 GBP/year)	
•Job titles: barber, colourist, hair stylist, hairdresser, beautician, beauty therapist, nail technician, tattooist	
Group 713: Sales Supervisors (18,383 GBP/year)	
•Job titles: sales supervisor, section manager, shop supervisor, retail supervisor, retail team leader	
Group 813: Assemblers and Routine Operatives (22,491 GBP/year)	
•Job titles: assembler, line operator, solderer, quality assurance inspector, quality auditor, quality controller, quality inspector, test engineer, weightbridge operator, type technician	
Group 913: Elementary Process Plant Occupations (17,902 GBP/year)	
•Job titles: factory cleaner, hygene operator, industrial cleaner, paint filler, packaging operator, material handler, packer	

Same Twitter data set as in the job classification task

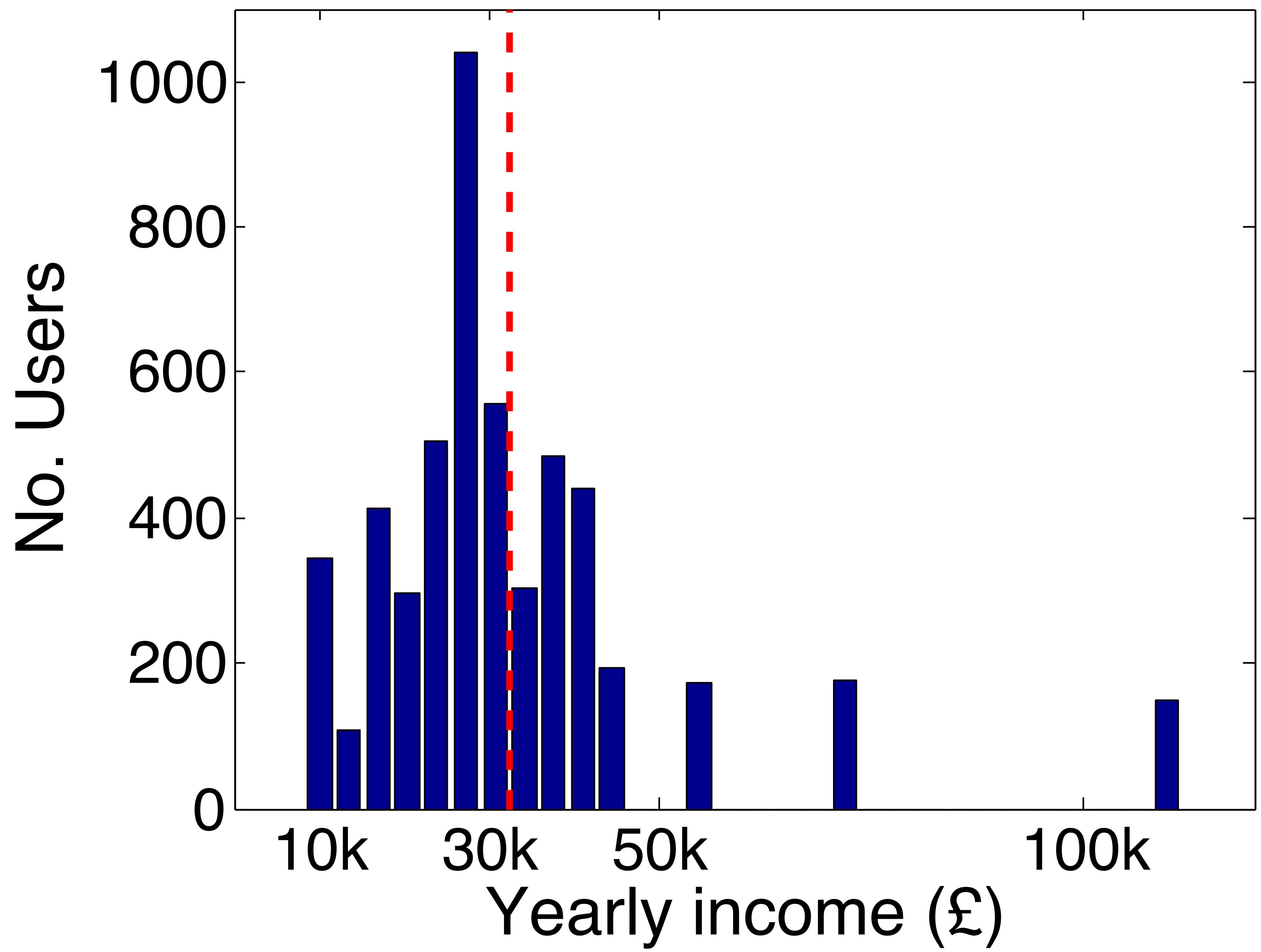
Use an income mapping from SOC to create real-valued target data for the regression task

User income regression: data

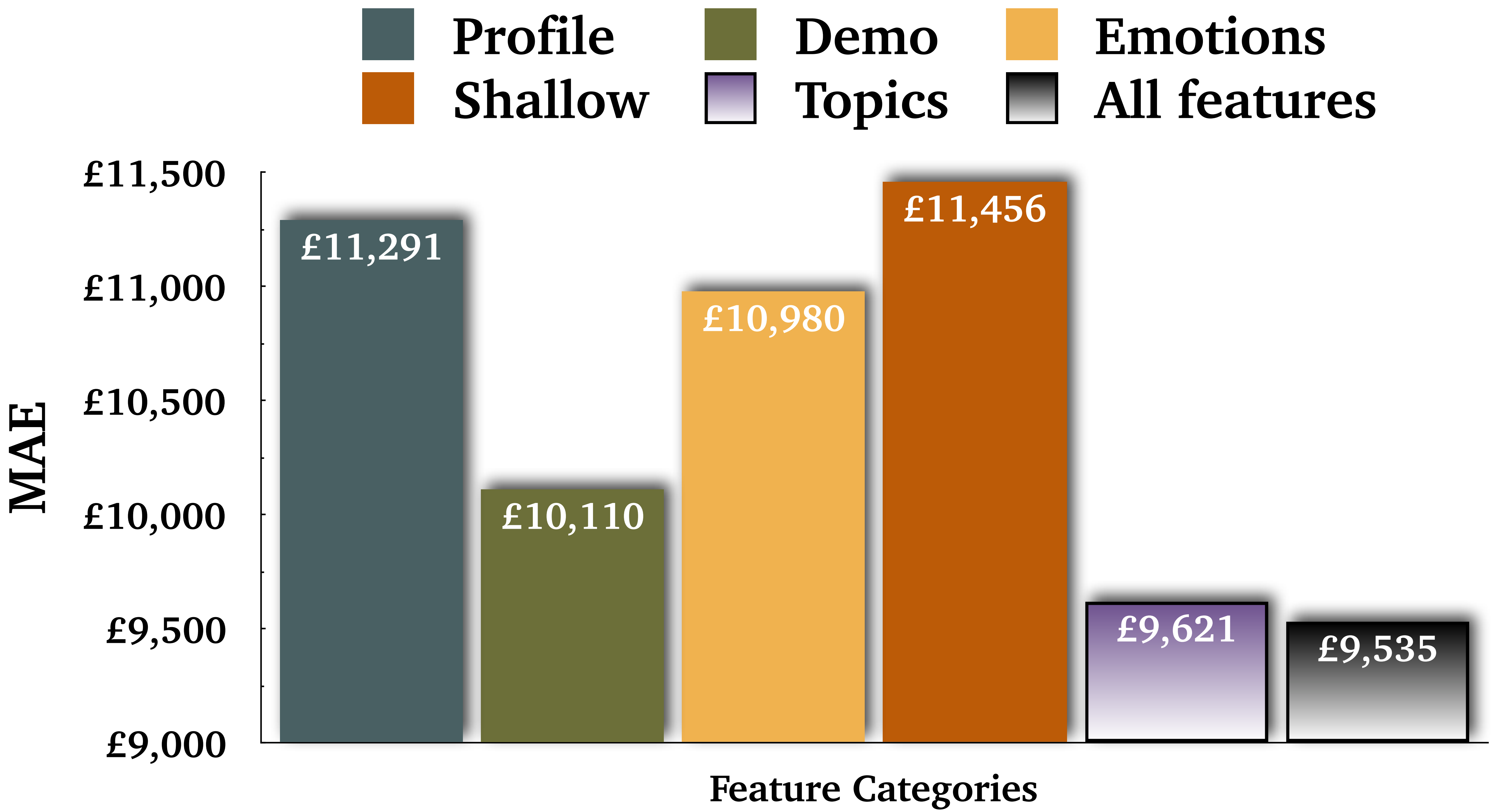
+ **5,191** Twitter users mapped to their occupations, then mapped to an average income in GBP (£) using the *SOC* taxonomy

+ ~11 million tweets

+ **Download the data**

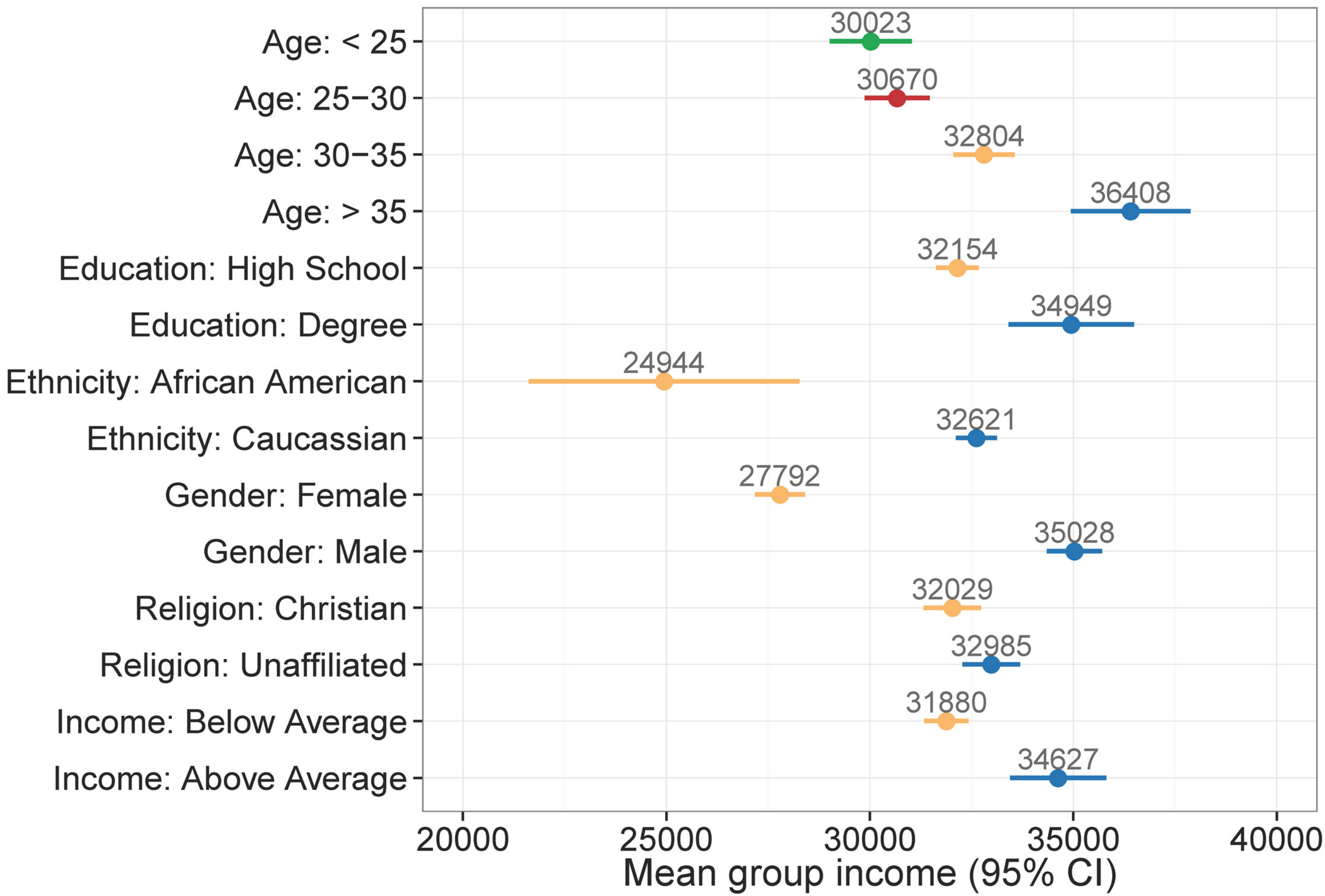


User income regression performance



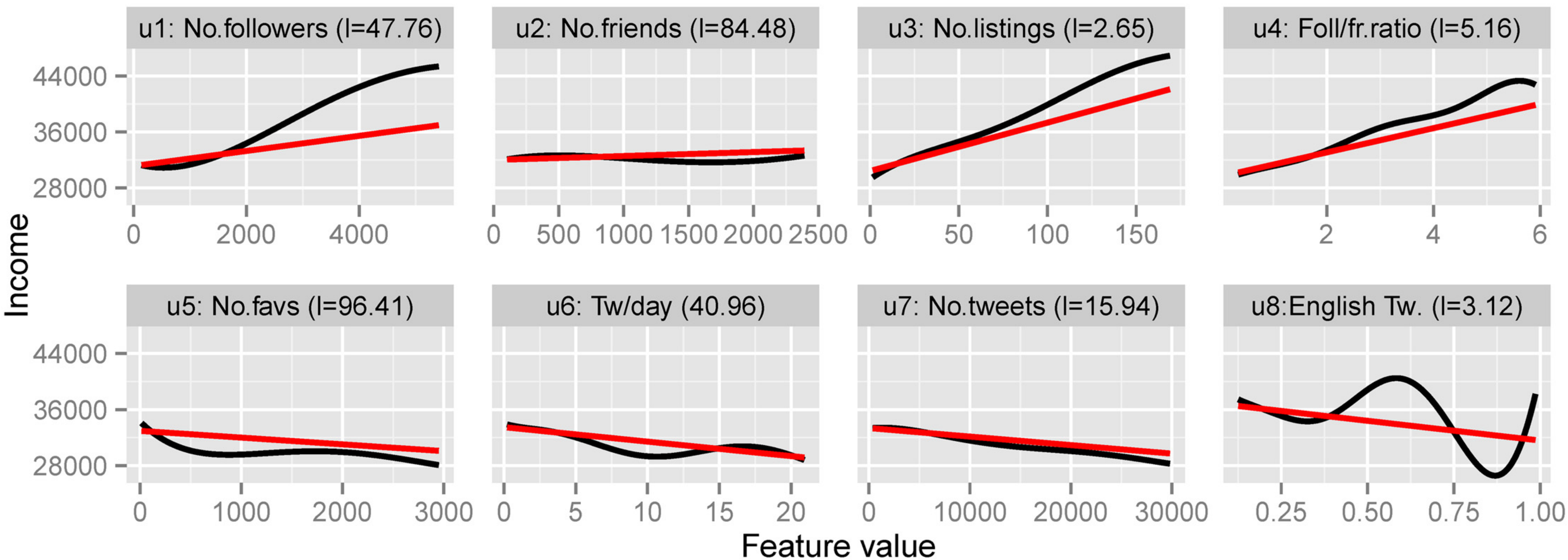
Income inference error (Mean Absolute Error) using GP regression or a linear ensemble for all features

User income regression insights (I)



User income regression insights (II)

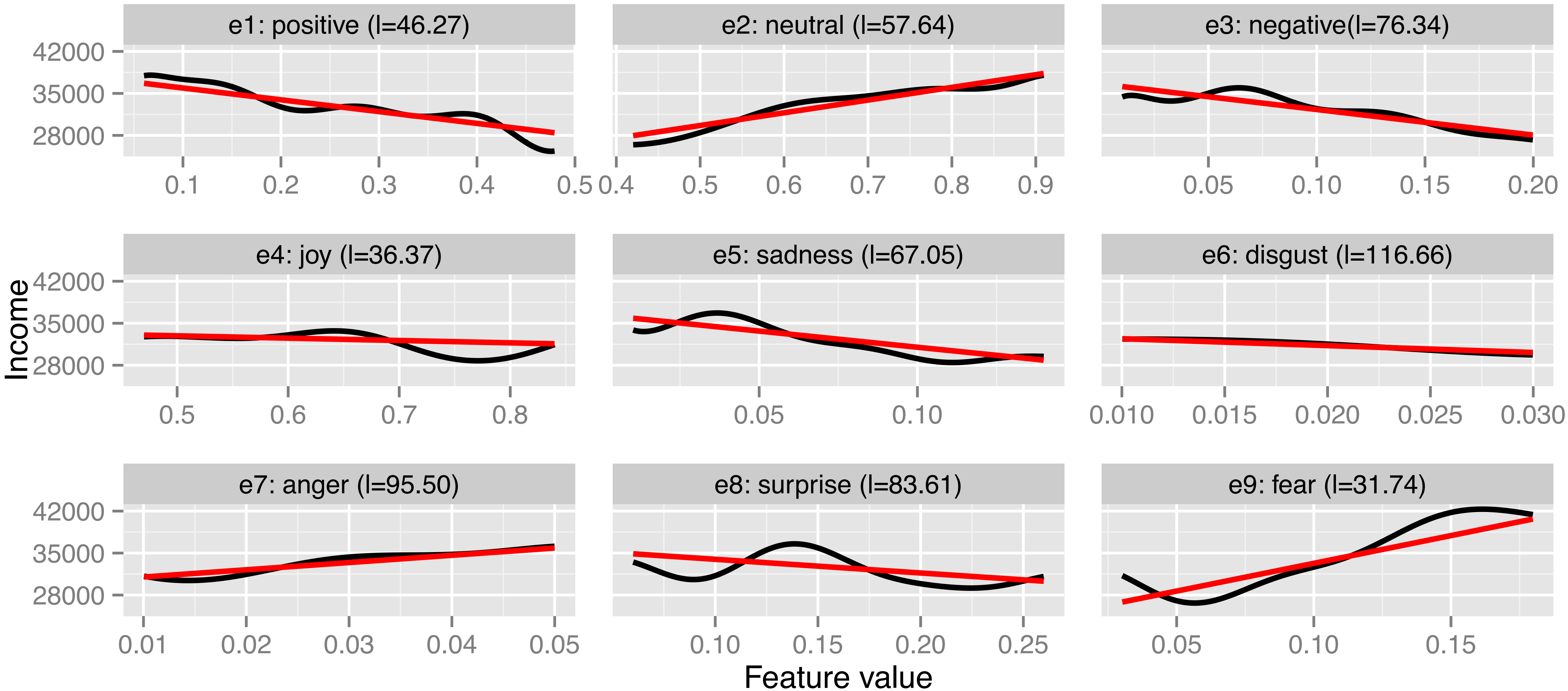
Relating income and user attributes



Linear vs **GP** fit

User income regression insights (III)

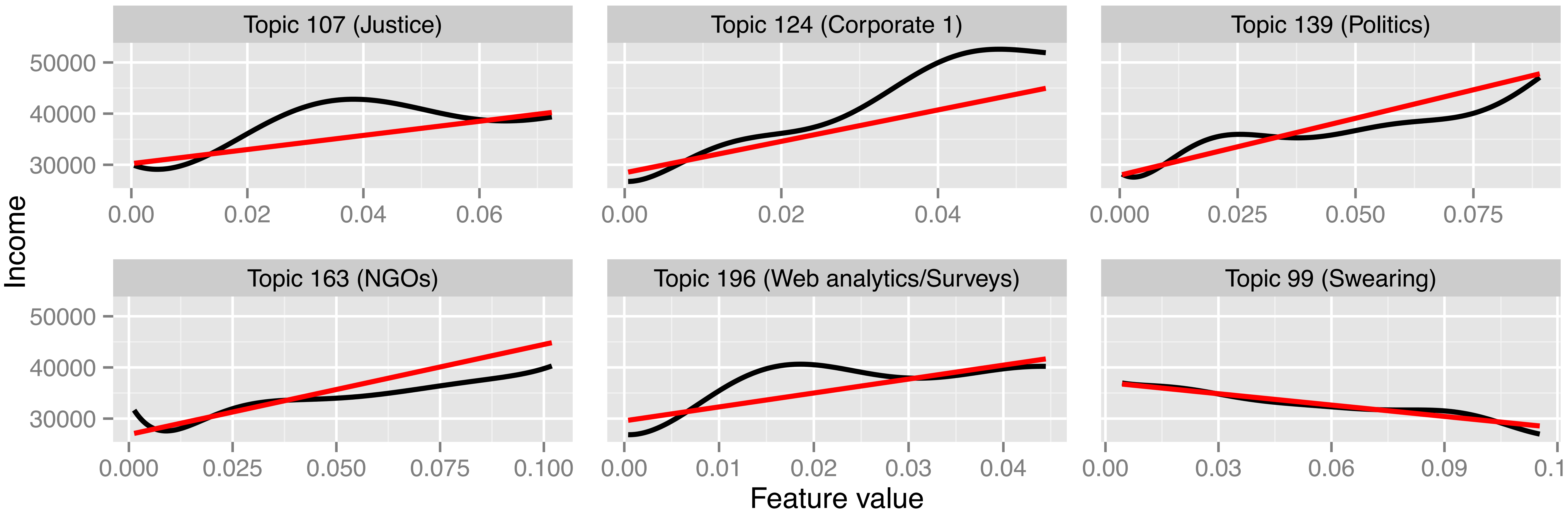
Relating income and emotion



Linear vs **GP** fit

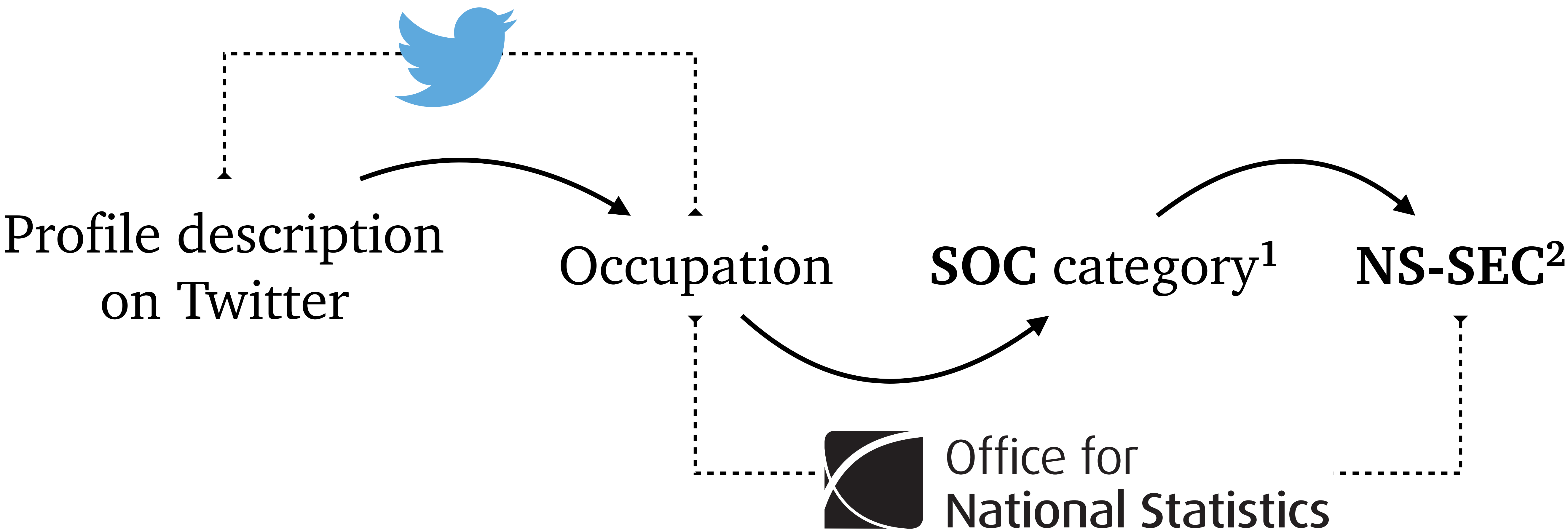
User income regression insights (IV)

Relating income and topics of discussion



Linear vs **GP** fit

Defining a user SES classification task



1. **S**tandard **O**ccupational **C**lassification job groups
2. **N**ational **S**tatistics **S**ocio-**E**conomic **C**lassification: Map from the job groups in the SOC to a socioeconomic status (SES): *upper*, *middle* or *lower*

UK Twitter user data set for SES classification

+ **1,342** UK Twitter user profiles
+ 2 million tweets
+ Date interval: Feb. 1, 2014 to March 21, 2015
+ Labelled with a **socioeconomic status** (SES), using the occupational class proxy from SOC and NS-SEC: ***upper***, ***middle***, or ***lower***
+ 1,291 **user features** following the previous paradigms, *i.e.* quantifying behaviour, impact, profile info, text in tweets and topics from tweets
+ **Download the data set**

SES classification performance

3-class 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%

middle & lower merged

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

… using a Gaussian Process classifier

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

Conclusions — Mining socio-political and socio-economic signals from social media

- collective emotion
- voting intention
- occupational class
- income
- socio-economic status

Further thoughts

+ **User-generated content** is a **valuable asset**
+ **Nonlinear models** tend to perform better given the multimodality of the feature space
+ **Deeper representations** of text tend to improve performance
+ **Qualitative analysis** is important
  > Evaluation
  > Interesting insights

Some of the future research challenges

+ Work closer with **domain experts**
+ Better understanding of online media **biases**, *e.g.* demographics, external influence etc.
+ **Generalisation**, defining **limitations**, more rigorous **evaluation** frameworks
+ Methodological improvements
+ **Ethical concerns**

Acknowledgements

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

Nikolaos Aletras (*Amazon*)
Yoram Bachrach (*Microsoft Research*)
Trevor Cohn (*University of Melbourne*)
Ingemar J. Cox (*UCL*)
Nello Cristianini (*University of Bristol*)
Daniel Preotiuc-Pietro (*Penn*)
Thomas Lansdall-Welfare (*University of Bristol*)
Svitlana Volkova (*PNNL*)
Bin Zou (*UCL*)

Currently **funded by**

i-sense

Thank you!

Any questions?

Slides can be downloaded from
lampos.net/talks

@lampos | lampos.net

References

Argyriou, Evgeniou & Pontil. ***Convex Multi-Task Feature Learning*** (Machine Learning, 2008)

Bernstein. ***Language and social class*** (Br J Sociol, 1960)

Bouma. ***Normalized (pointwise) mutual information in collocation extraction*** (GSCL, 2009)

Labov. ***The Social Stratification of English in New York City*** (Cambridge Univ Press, 1972; 2006, 2nd ed.)

Lampos. ***Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods*** (Ph.D. Thesis, University of Bristol, 2012)

Lampos, Aletras, Geyti, Zou & Cox. ***Inferring the Socioeconomic Status of Social Media Users based on Behaviour and Language*** (ECIR, 2016)

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)

Lansdall-Welfare, Lampos & Cristianini. ***Effects of the Recession on Public Mood in the UK*** (WWW, 2012)

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)

Pennebaker, Booth & Francis. ***Linguistic Inquiry and Word Count: LIWC2007*** (Tech. Report, 2001, 2007)

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)

Strapparava & Valitutti. ***WordNet-Affect: An affective extension of WordNet***. LREC, 2004.

Volkova, Bachrach, Armstrong & Sharma. ***Inferring Latent User Properties from Texts Published in Social Media*** (AAAI, 2015)

von Luxburg. ***A tutorial on spectral clustering*** (Stat Comput, 2007)

Zou & Hastie. ***Regularization and variable selection via the elastic net*** (J R Stat Soc Series B Stat Methodol, 2005)