Flu Detector - Tracking Epidemics on Twitter

Vasileios Lampos, Tijl De Bie and Nello Cristianini

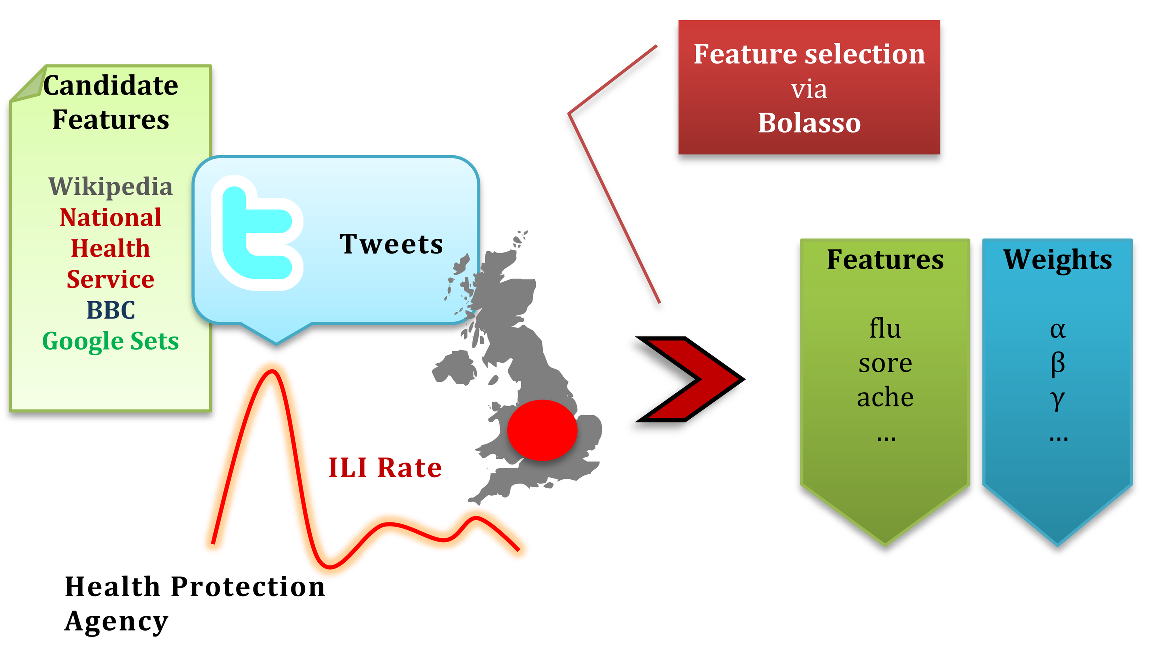
Intelligent Systems Laboratory, University of Bristol, UK

intelligentSystems.bristol.ac.uk

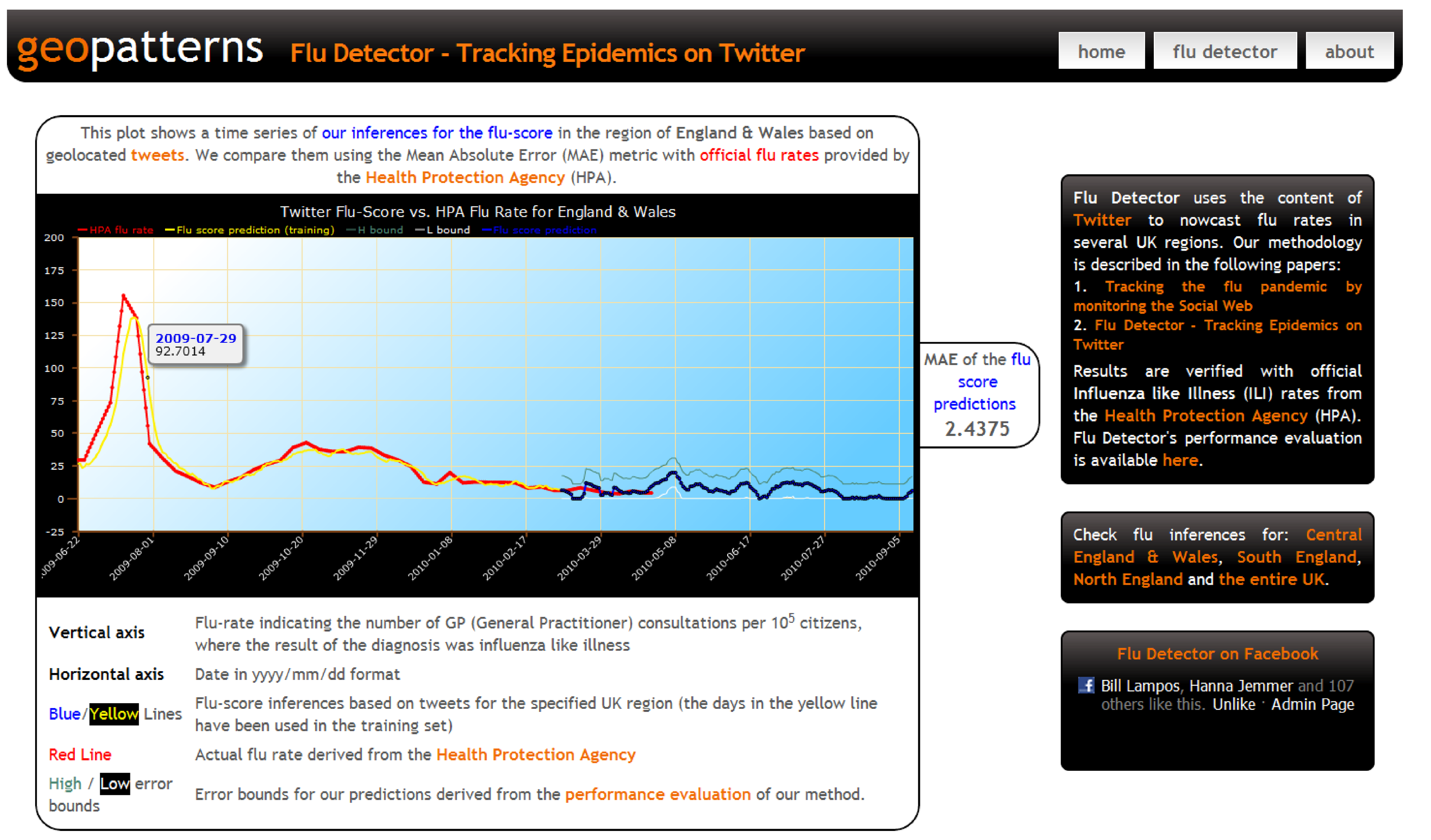
✓I. What Is This All About?

Flu Detector is a tool with a web interface for nowcasting the prevalence of **Influenza-like Illness** (ILI) in several UK regions using the contents of **Twitter**. We automatically select a set of representative flu-words (markers, features) via **Bolasso** and learn their weights by applying linear LS regression. Ground truth is acquired from the **Health Protection Agency** (HPA). Flu Detector applies and extends the findings of [3].

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



✓II. Web Interface & Data



Flu Detector makes flu-score inferences for **Central England & Wales** (r_1), **South England** (r_2) and **North England** (r_3) as well as for some unions of them (updated on a daily basis).

For the experimental purposes of this work, from 22/06/2009 to 28/03/2010 we were collecting:

▶ a daily average of 200,000 **tweets** geolocated in the 49 most populated UK's urban centres

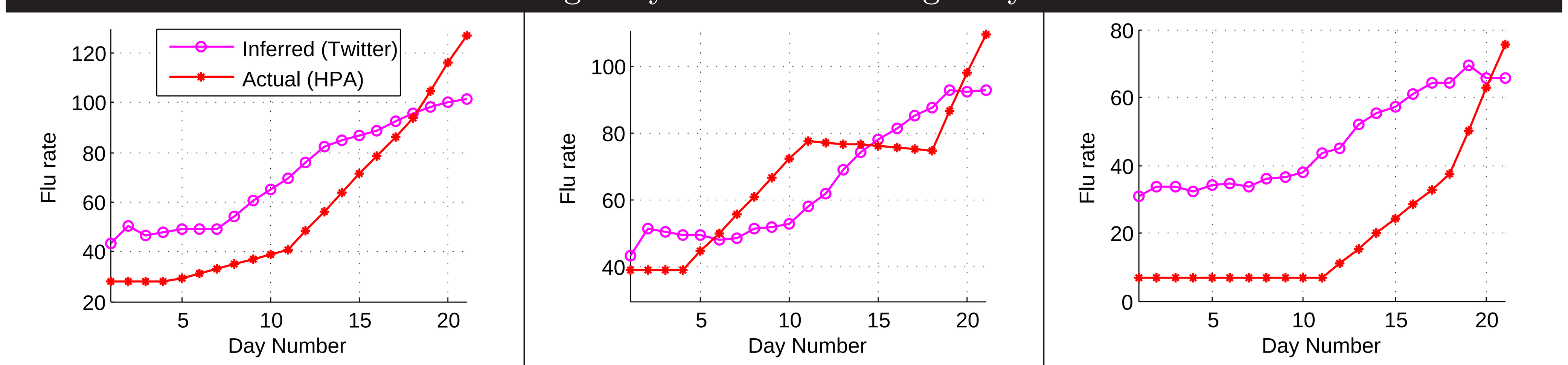
▶ weekly **reports from HPA** for the same regions based on data gathered by the Royal College of General Practitioners (GP) expressing the number of GP consultations per 10^5 citizens, where the the diagnosis result was ILI. Each weekly rate is expanded over a 7-day period and the expanded ground truth time series are smoothed with a 7-point moving average.

✓V. Validation

Performance is evaluated by computing the Mean Absolute Error ($\mathcal{MAE}$) between the inferred and the target values. When the ground truth signal is clearly present, we additionally compute its linear correlation coefficient (ρ) with the inferences.

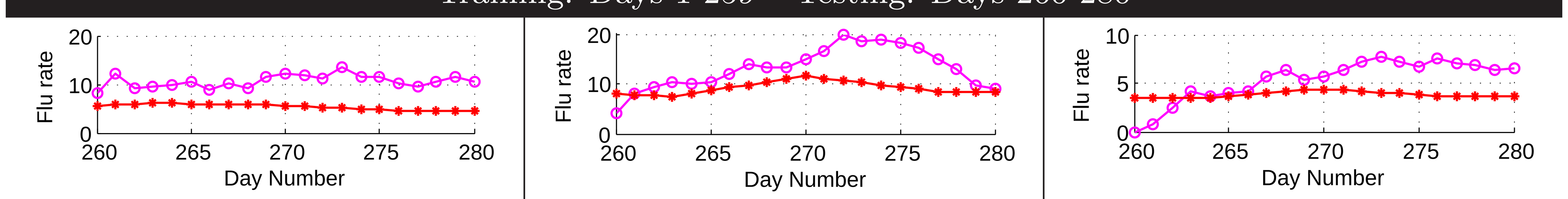
C. England & Wales	S. England	N. England

Training: Days 22-280 – Testing: Days 1-21



C. England & Wales	S. England	N. England
$\mathcal{MAE} = 18.34$	$\mathcal{MAE} = 9.38$	$\mathcal{MAE} = 27.29$
$\rho = 0.94$, $p = 3.5e-10$	$\rho = 0.84$, $p = 2.2e-6$	$\rho = 0.87$, $p = 4e-7$

Training: Days 1-259 – Testing: Days 260-280



C. England & Wales	S. England	N. England
$\mathcal{MAE} = 5.27$	$\mathcal{MAE} = 4.26$	$\mathcal{MAE} = 2.18$

✓III. Notation-Definitions

▶ Set of **candidate markers**: $\mathcal{C} = \{c_i\}$, $i \in [1, \theta]$

▶ Their respective **weights**: $\mathcal{W} = \{w_i\}$, $i \in [1, \theta]$

▶ Set of **tweets**: $\mathcal{T} = \{t_u\}$, $u \in [1, k]$

▶ **Regions**: $\mathcal{R} = \{r_j\}$, $j \in [1, 3]$

▶ A subset selection of $\mathcal{C}$ is denoted with $\mathcal{C}^{(s)}$

▶ Function for forming vector space representations:

$$g(t_u, c_i) = \begin{cases} 1 & \text{if } c_i \text{ appears in } t_u \\ 0 & \text{otherwise} \end{cases}$$

▶ **Unweighted flu-subscore** of a marker c_i in a set of tweets $\mathcal{T}$:

$$f_{\mathcal{C}}(\mathcal{T}, c_i) = \sum_u g(t_u, c_i)/k$$

▶ **Flu-score** of a set of tweets $\mathcal{T}$:

$$f_{\mathcal{S}}(\mathcal{T}, \mathcal{W}, \mathcal{C}) = \sum_u \sum_i w_i \times g(t_u, c_i)/k$$

✓IV. Methodology

▶ **Form a pool of** $\theta = 2675$ **candidate markers** (or features) using several influenza related web references (Wikipedia, NHS, BBC, Google Sets). The majority of the candidate markers is not directly related to flu.

▶ **Compute** their daily, unweighted **flu-subscores** $f_{\mathcal{C}}(\mathcal{T}_r, c_i)$ for a region r given $\mathcal{T}_r$, the set of tweets for region r.
For a day d, Twitter's regional flu-score is represented as a vector

$$\mathcal{F}_{d,r} = [f_{\mathcal{C}}(\mathcal{T}_r, c_1) \ \dots \ f_{\mathcal{C}}(\mathcal{T}_r, c_\theta)]^T.$$

For a region r and a period of ℓ days, we form an $\ell \times \theta$ array with the time series of the flu-subscores for all candidate markers:

$$\mathcal{X}_{\ell,r} = [\mathcal{F}_{1,r} \ \dots \ \mathcal{F}_{\ell,r}]^T.$$

▶ **HPA**'s flu rates for region r and the same period of ℓ days are denoted by vector y_r.

▶ **Bolasso** [2] is applied for extracting a consistent set of markers with respect to the ground truth. Internally, Bolasso uses LASSO method for performing regression with L1-regularisation [1]. LASSO is formulated as the following optimisation problem:

$$\min_w \quad \|\mathcal{X}_{\ell,r} w - y_r\|_2^2$$
$$\text{s.t.} \quad \|w\|_1 \leq t,$$

where vector w is guaranteed to be a sparse solution and t is the regularisation parameter. A soft version of Bolasso is used, *i.e.* we select the markers that have non zero weights in $s = 65\%$ to 75% of the bootstraps. The selected $h \leq \theta$ markers are denoted with $c_i^{(s)}$, $i \in [1, h]$ and the corresponding $\ell \times h$ array of their flu-subscores time series with $\mathcal{X}_{\ell,r}^{(s)}$.

▶ Finally, we perform linear LS **regression** to learn the weights ($w^{(s)}$) of the selected markers.

$$\min_{w_s} \|\mathcal{X}_{\ell,r}^{(s)} w^{(s)} - y_r\|_2^2$$

Acknowledgements

We would like to thank **Twitter Inc.** for making its data publicly available. This work is partially supported by European Commission through the **PASCAL2** NoE (FP7-216866). V. Lampos is supported by **EPSRC** (DTA/SB1826) and **NOKIA Research**. N. Cristianini is supported by a **Royal Society Wolfson Merit Award**.

References

[1] R. Tibshirani: Regression shrinkage and selection via the lasso. In Journal of the Royal Statistical Society 58B, 267–288 (1996).

[2] F.R. Bach: Bolasso: model consistent Lasso estimation through the bootstrap. ICML 25, 33–40 (2008).

[3] V. Lampos and N. Cristianini: Tracking the flu pandemic by monitoring the Social Web. 2nd IAPR Workshop on CIP, 411–416 (2010).