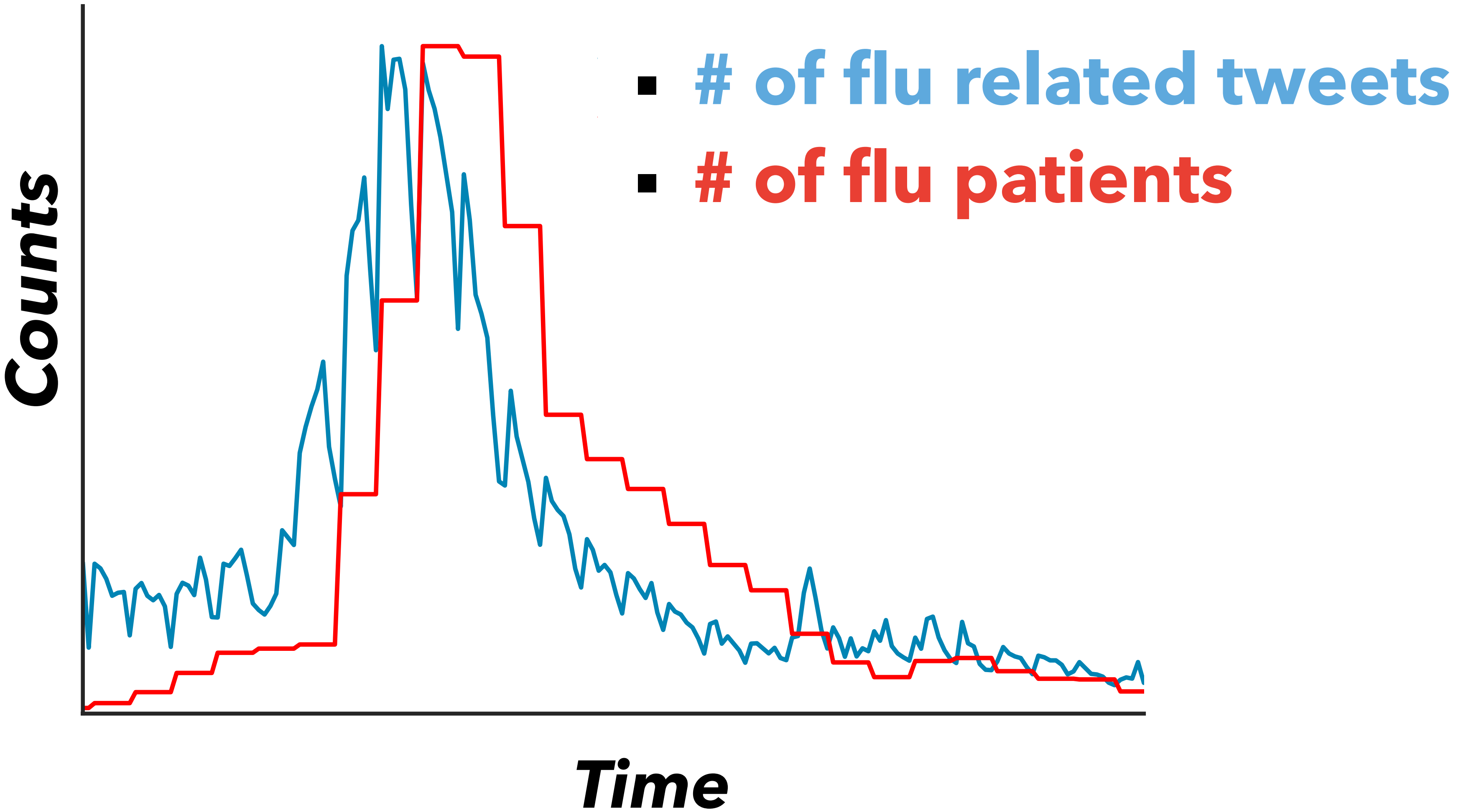
Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction

Hayate ISO, Shoko WAKAMIYA, Eiji ARAMAKI

Social Computing Laboratory NAIST since 2015 SOCIOCOM.jp NAIST®

Twitter for Public health

- Many users **tweet** when they caught a **disease**
- **# of tweets** is in proportion to **# of flu patients**



Noise included in tweets

Website:

By patients:

By healthy people:

Influencer
@not_influenza

For more information about bird flu *link*

High Fever
@flu_patient

I got a flu… I couldn't do anymore…

Healthy person
@organic

I've never caught a flu

Injection lover
@prevention

I got a flu shot yesterday

Noise included in tweets

Website:

By patients:

Only counts this type of tweets

By healthy people:

Influencer
@not_influenza

For more information about bird flu *link*

High Fever
@flu_patient

I got a flu… I couldn’t do anymore…

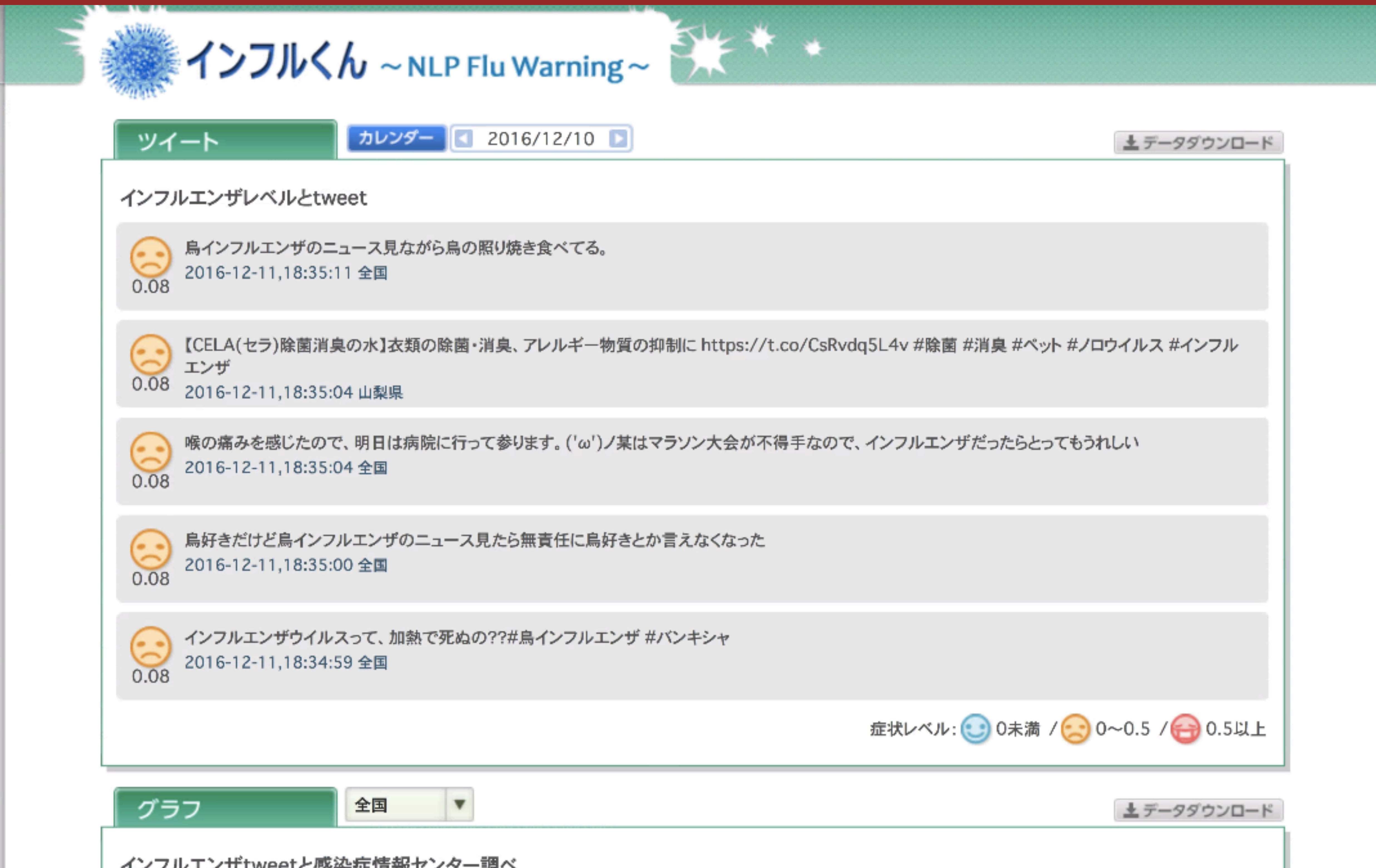
Healthy person
@organic

I’ve never caught a flu

Injection lover
@prevention

I got a flu shot yesterday

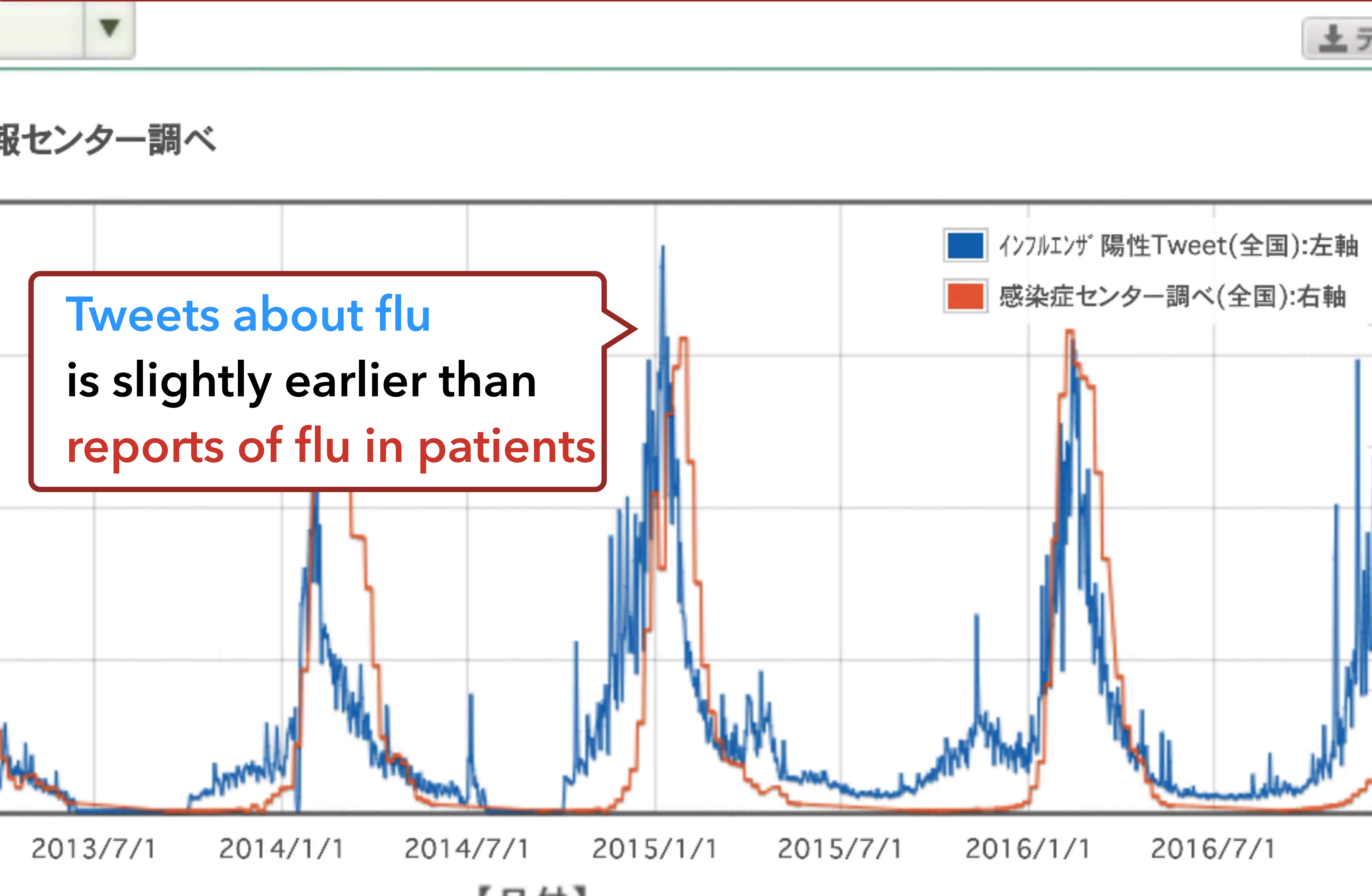
Our lab runs flu surveillance system



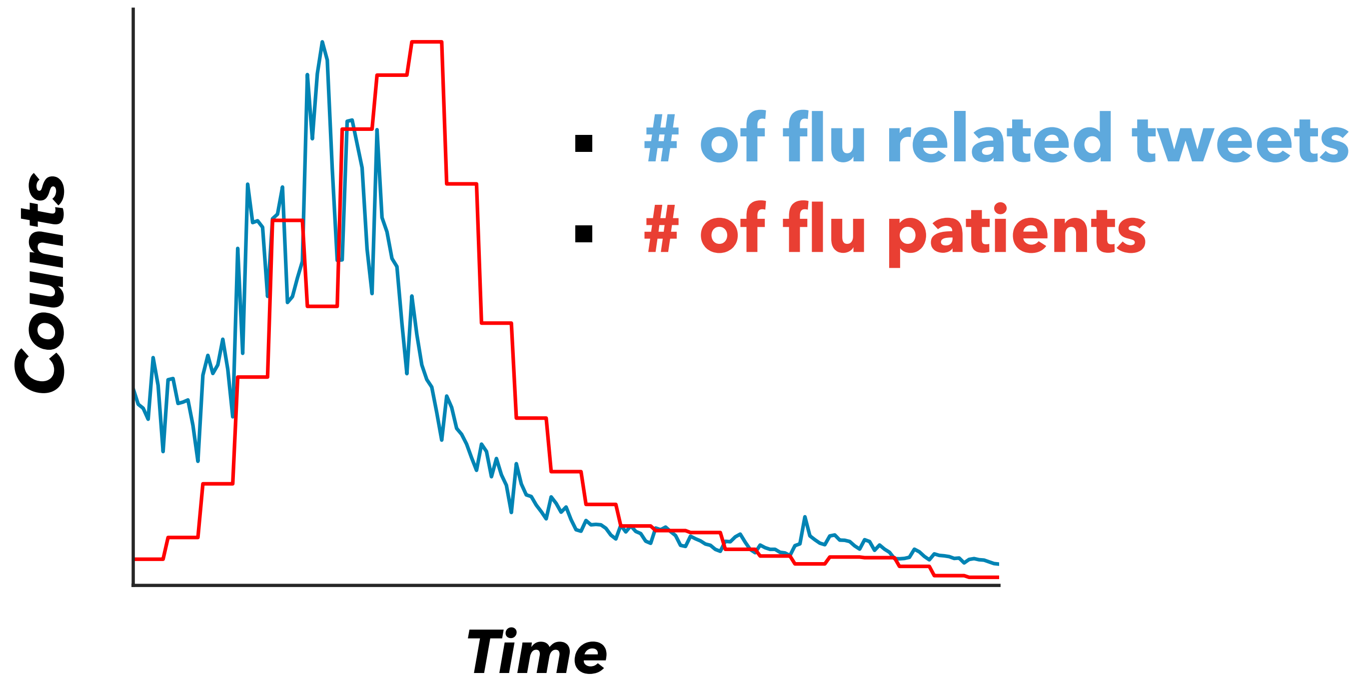
Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter." *In Proc of EMNLP 2011*. http://mednlp.jp/influ_map/

Similarity between Tweets and Patients

Tweets about flu
is slightly earlier than
reports of flu in patients

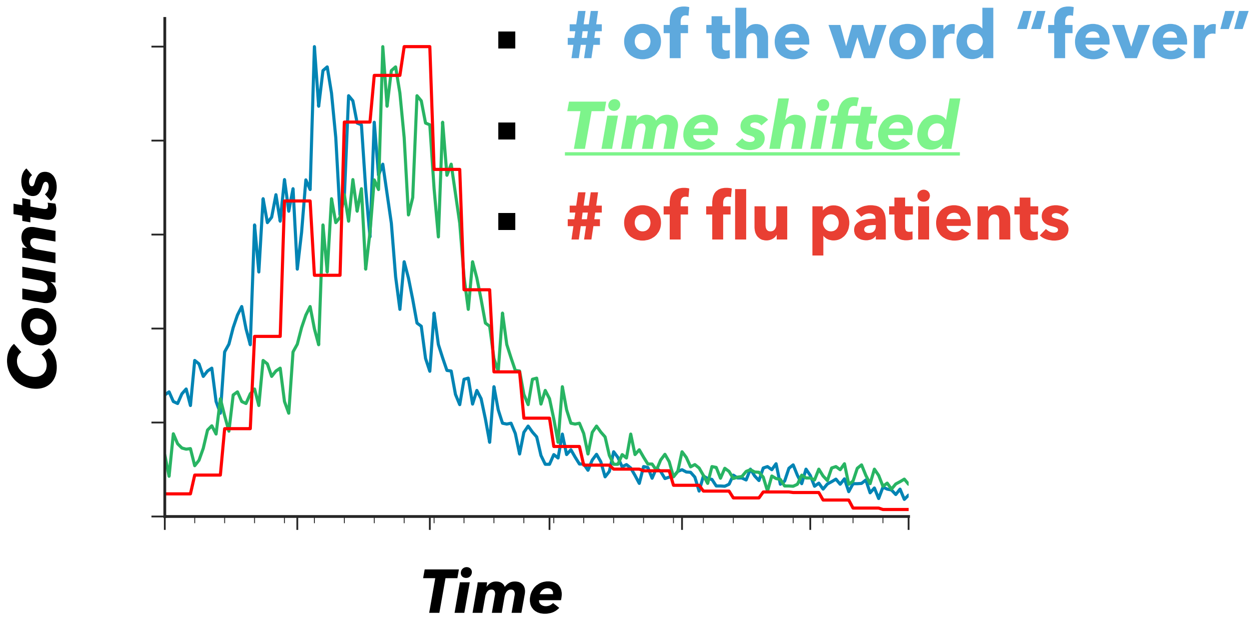


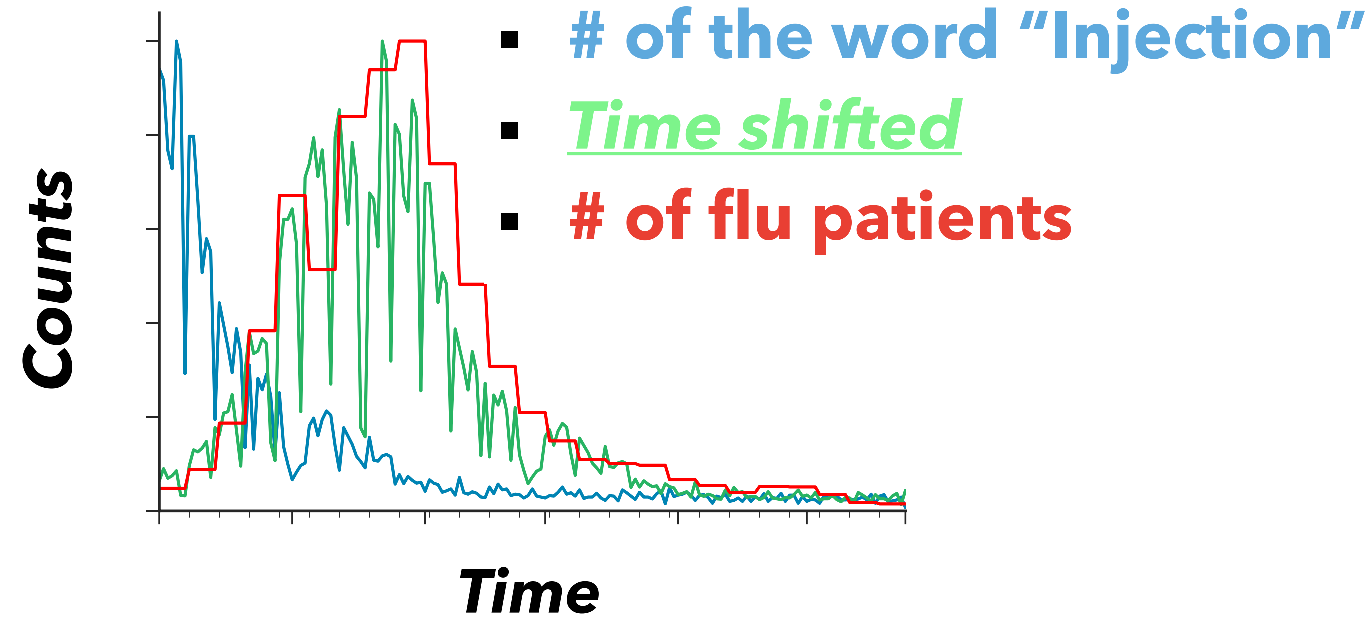
Each word has a specific time-lag



The word *"Fever"*
16 days time lag

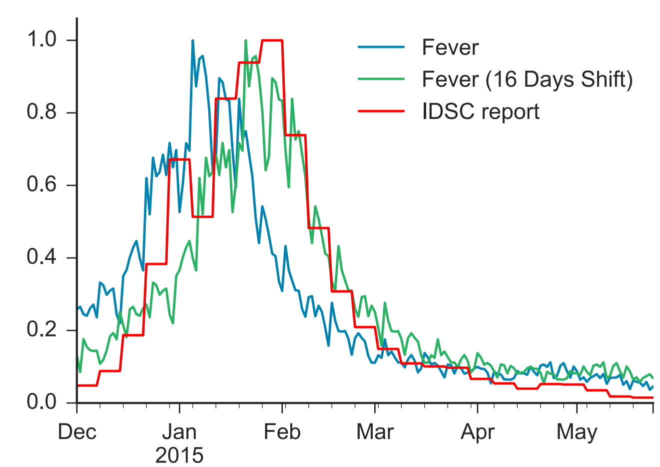
The word *"Injection"*
55 days time lag

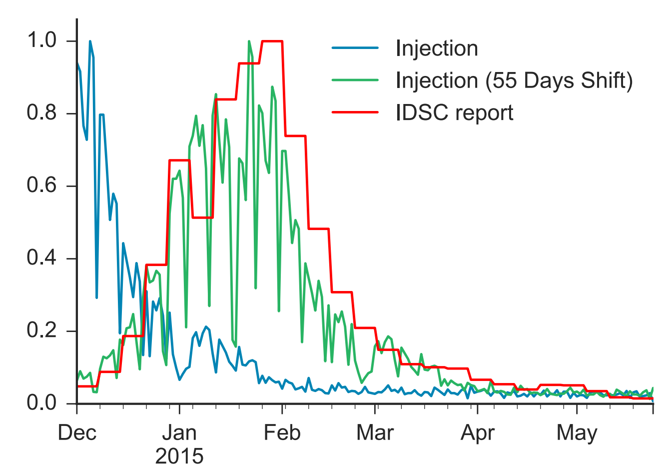




What is Forecasting Words?

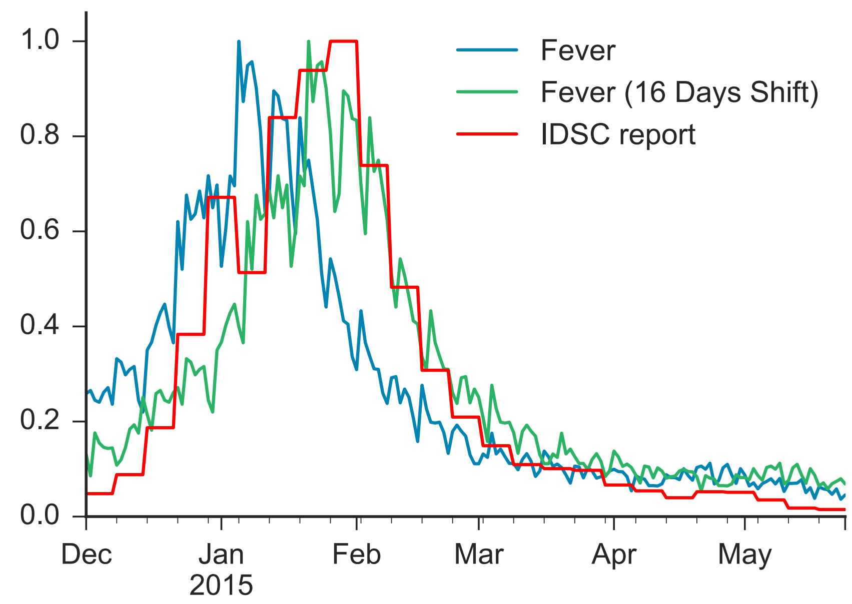
- Twitter tends to be an **early indicator** of actual condition
- We observed that each word has a specific **time lag** with actual condition
- Our objective: **more flexible modeling**
  - Estimate **time-difference**
  - Extend **future forecasting** model



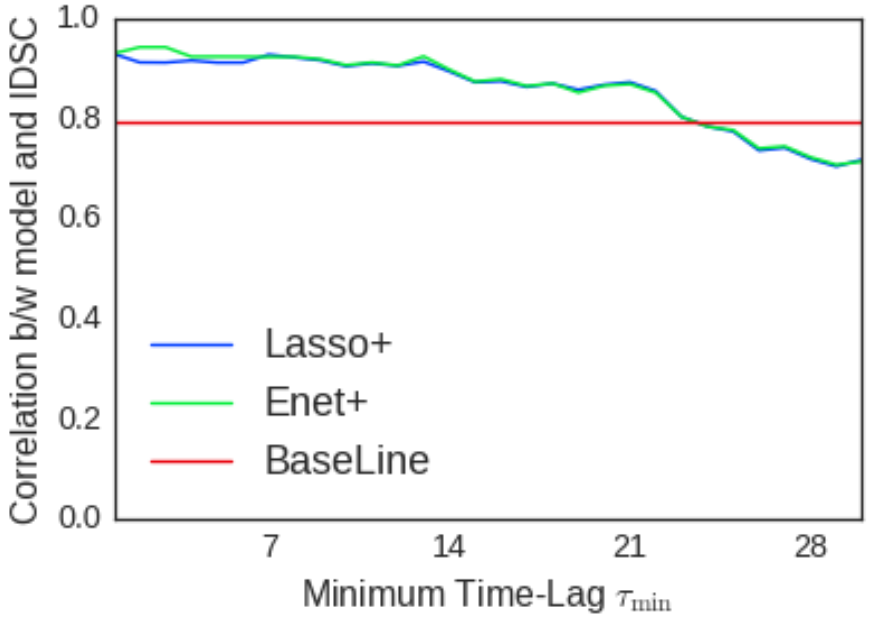


Outline

Data



Time shift: Nowcasting



Time shift: Forecasting

Outline

Data

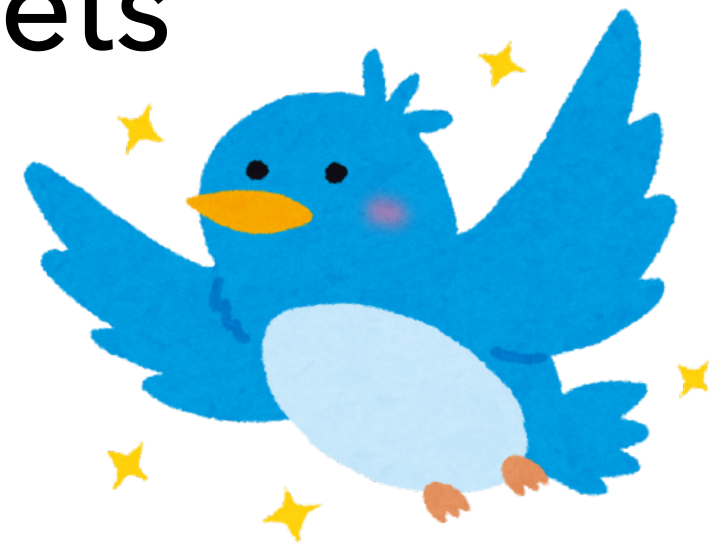
Time shift: Nowcasting

Time shift: Forecasting

Training data: Twitter Corpus

- **Query**: The word **"flu"** in Japanese (INFLU / I-N-FU-RU/)
- **Period**: Aug 2012 ~ Jan 2016

  (3 year 5 month)
- **Size of corpus**: 7.7 Million tweets

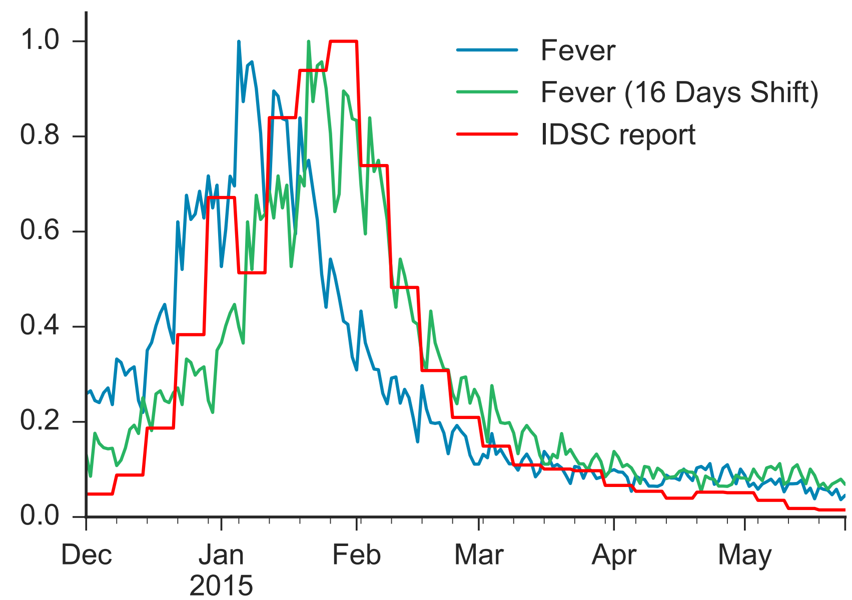


Gold standard: IDSC reports

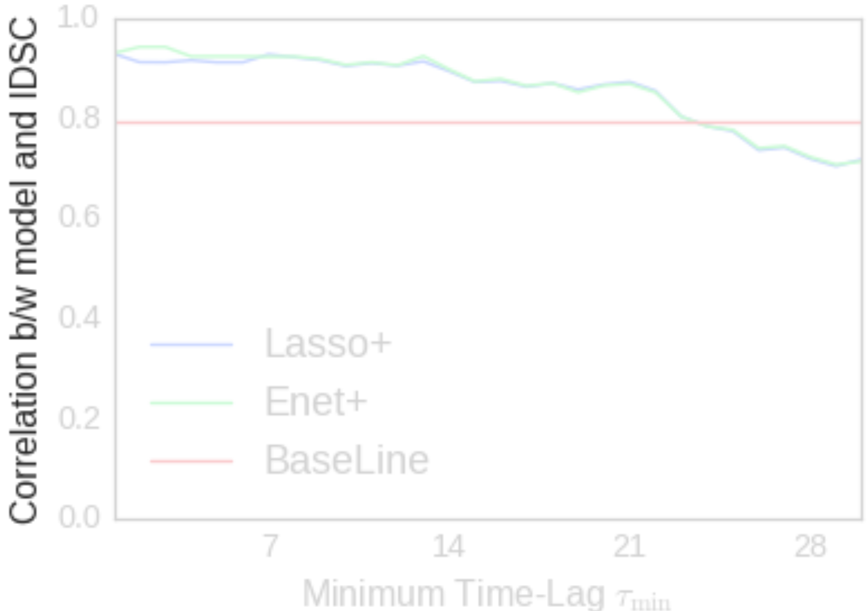
- Infectious Disease Surveillance Center (IDSC) reports # of flu patients **once a week**
- They gather the number of flu patients during the period of epidemic
- We split IDSC reports into three seasons as follows:
  - Season 1: Dec 1, 2012 ~ May 31, 2013
  - Season 2: Dec 1, 2013 ~ May 31, 2014
  - Season 3: Dec 1, 2014 ~ May 24, 2014

Outline

Data



Time shift: Nowcasting



Time shift: Forecasting

Time lag measure: Cross Correlation

- Cross Correlation is used to search for the most suitable time shift width for each word frequency as between **# of tweets τ days before** and **# of actual patients**

$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))^2 \sum_{t=1}^{T} (y^{(t)} - \bar{y})^2}},$$

where $\bar{x}_v(\tau) = \sum_{t=1}^{T} x_v^{(t-\tau)} / T$

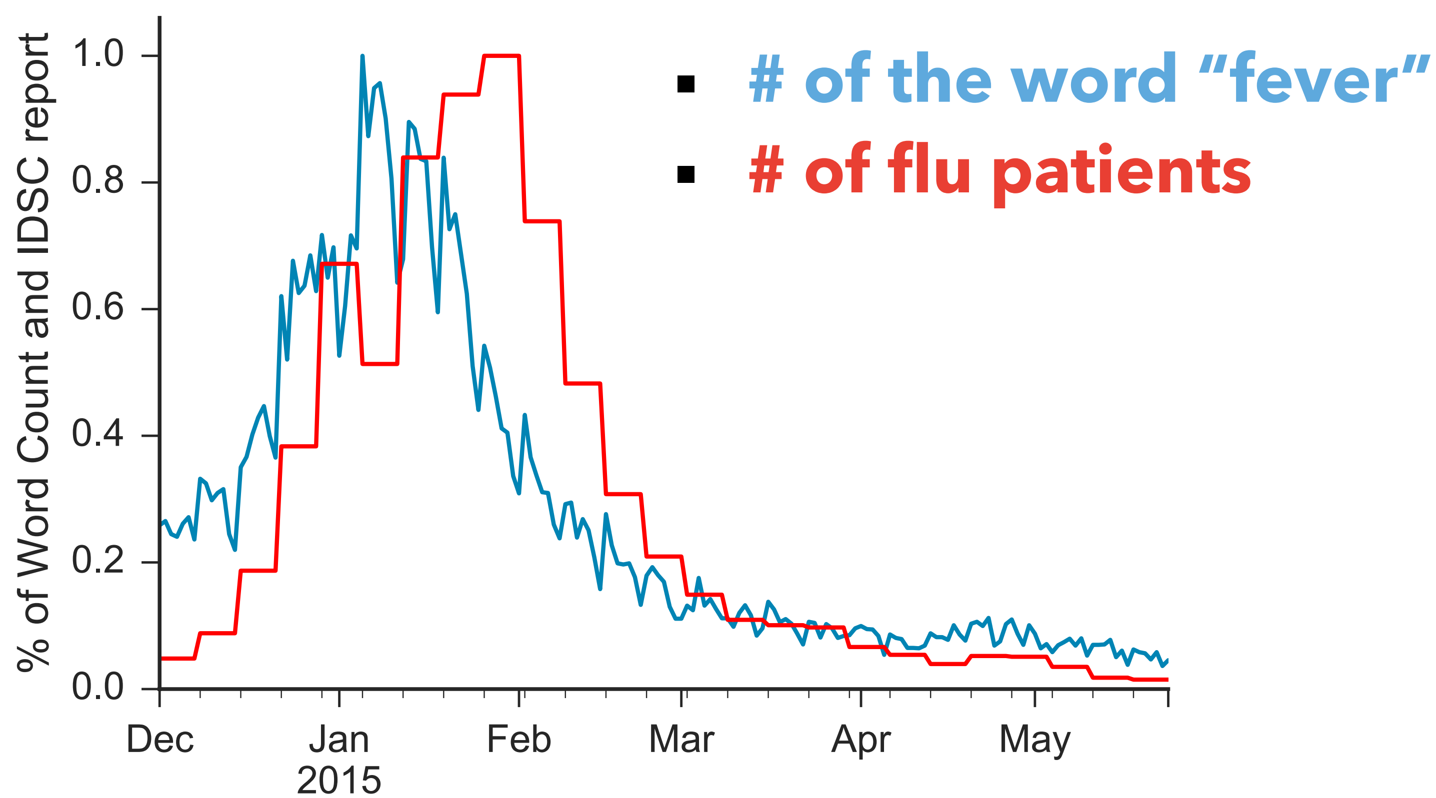
※ The cross correlation is exactly the same as the *Pearson's correlation* when ***τ*** **= 0**.

Motivating examples

- Cross Correlation *r*:

$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))^2 \sum_{t=1}^{T} (y^{(t)} - \bar{y})^2}},$$

- When τ = 0, *r* is 0.75 B/T tweet and IDSC reports

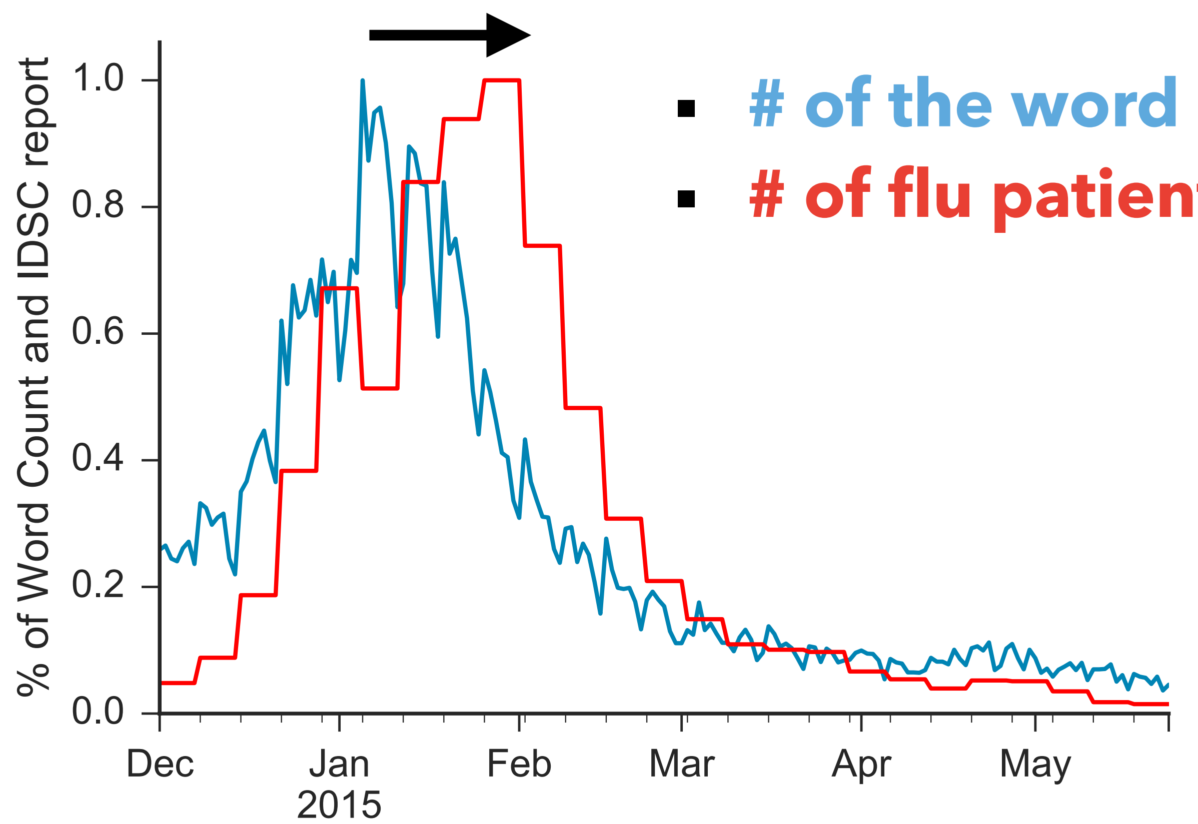


Motivating examples

- Cross Correlation *r*:

$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^{T} (x_v^{(t-\tau)} - \bar{x}_v(\tau))^2 \sum_{t=1}^{T} (y^{(t)} - \bar{y})^2}},$$

- When τ increases, word counts moves to right side:



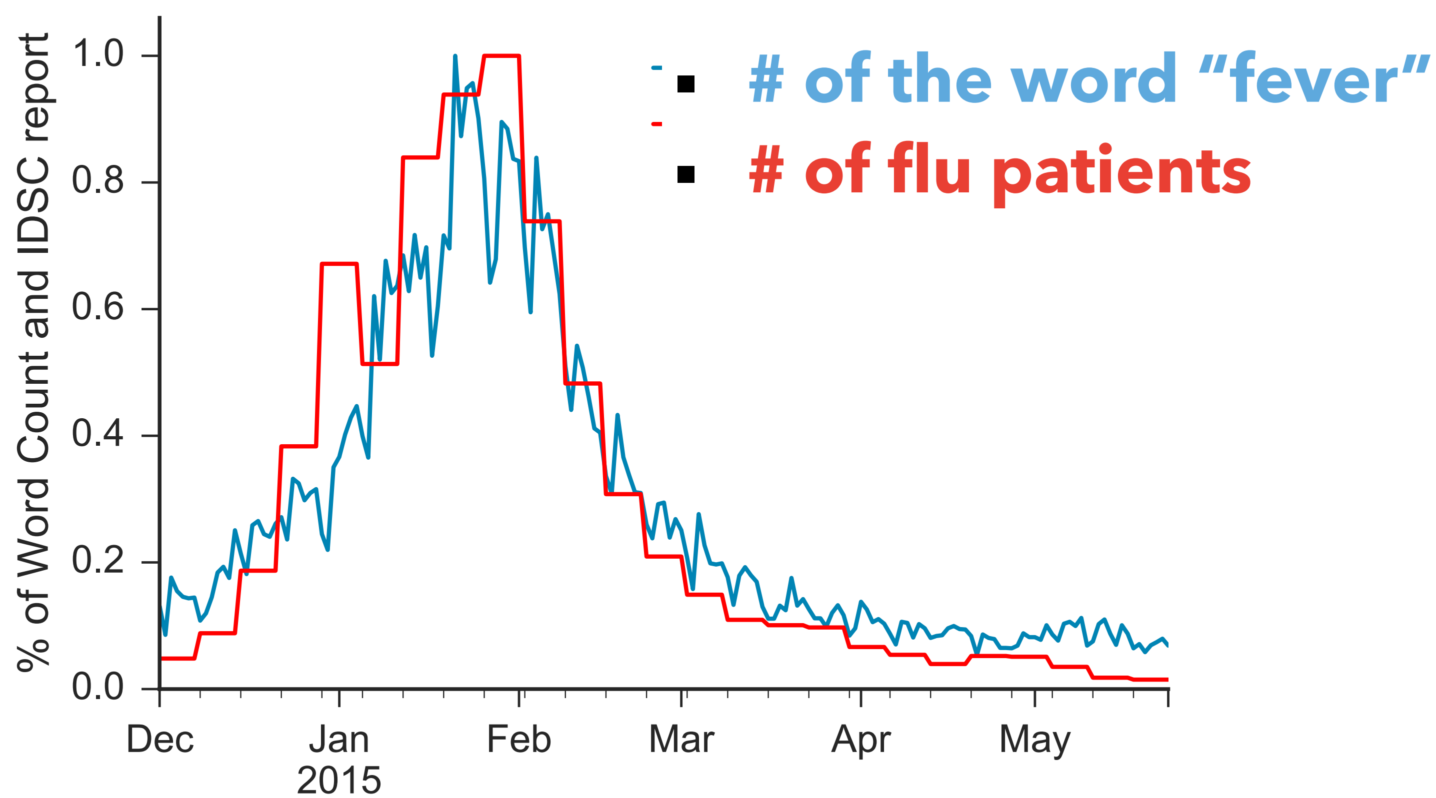
- **# of the word "fever"**
- **# of flu patients**

Motivating examples

- Cross Correlation *r*:

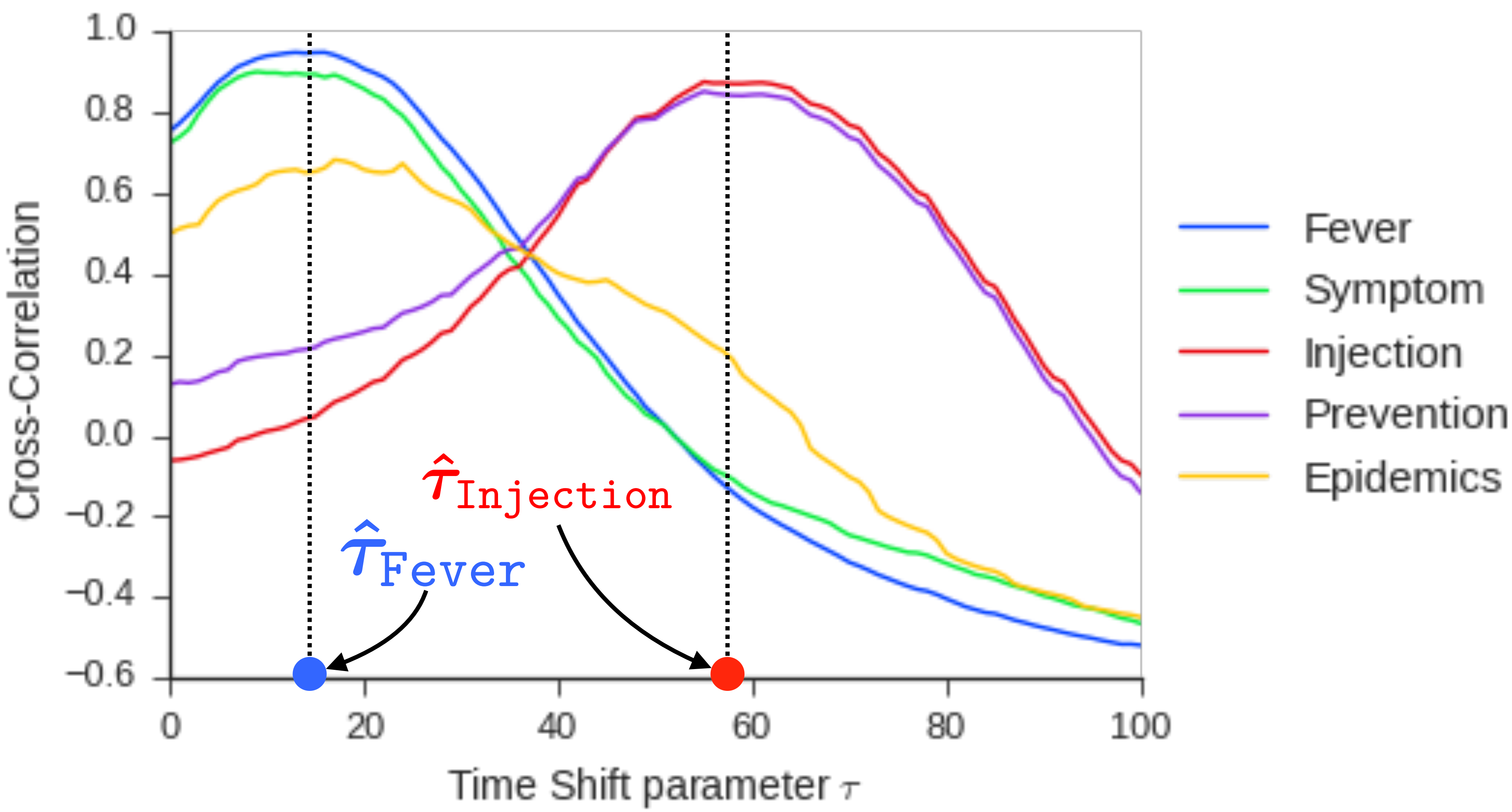
$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v(\tau))(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v(\tau))^2 \sum_{t=1}^{T}(y^{(t)} - \bar{y})^2}},$$

- When τ = 16, *r* is 0.95 B/T tweet and IDSC reports



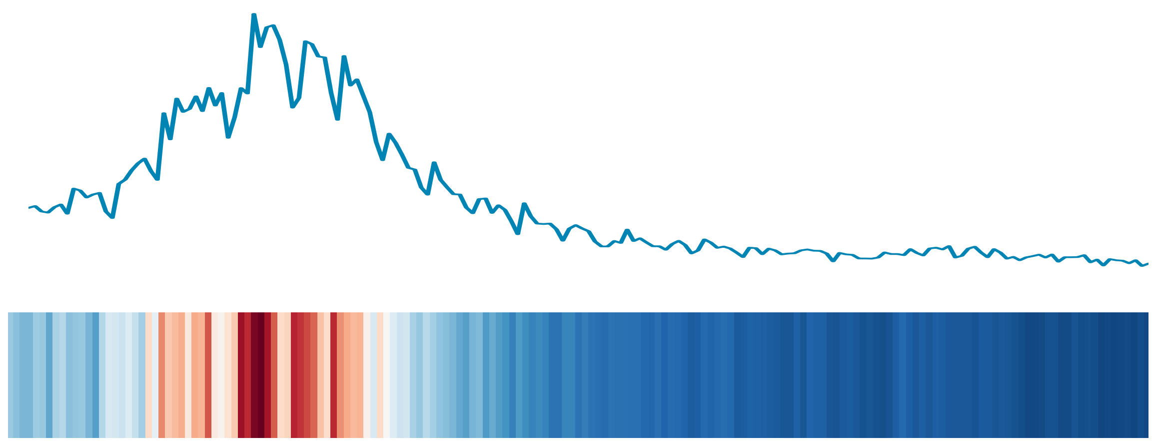
Estimate optimal time-lag

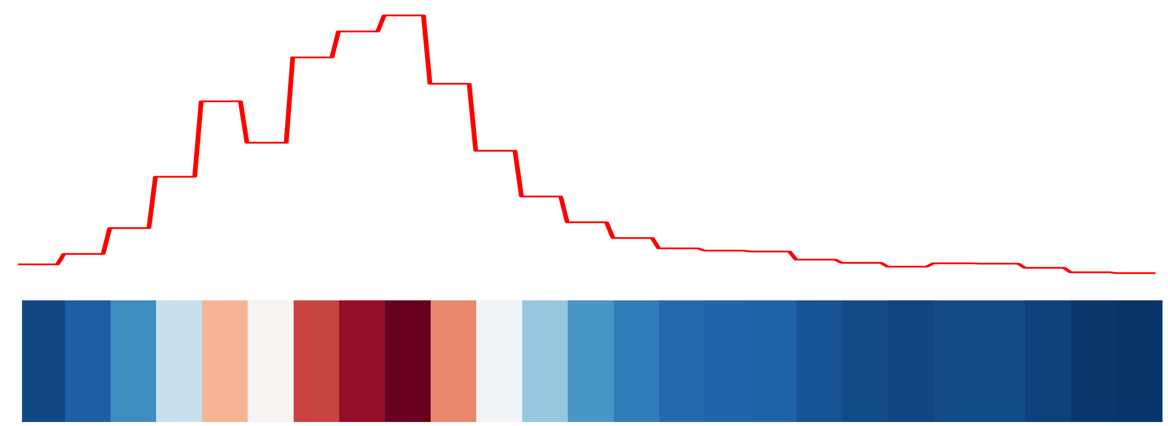
- We define optimal time-lag τ̂ by **maximizing** the **cross correlation**



$$\hat{\tau}_v = \arg\max_{\tau} r_{x_v, y}(\tau)$$

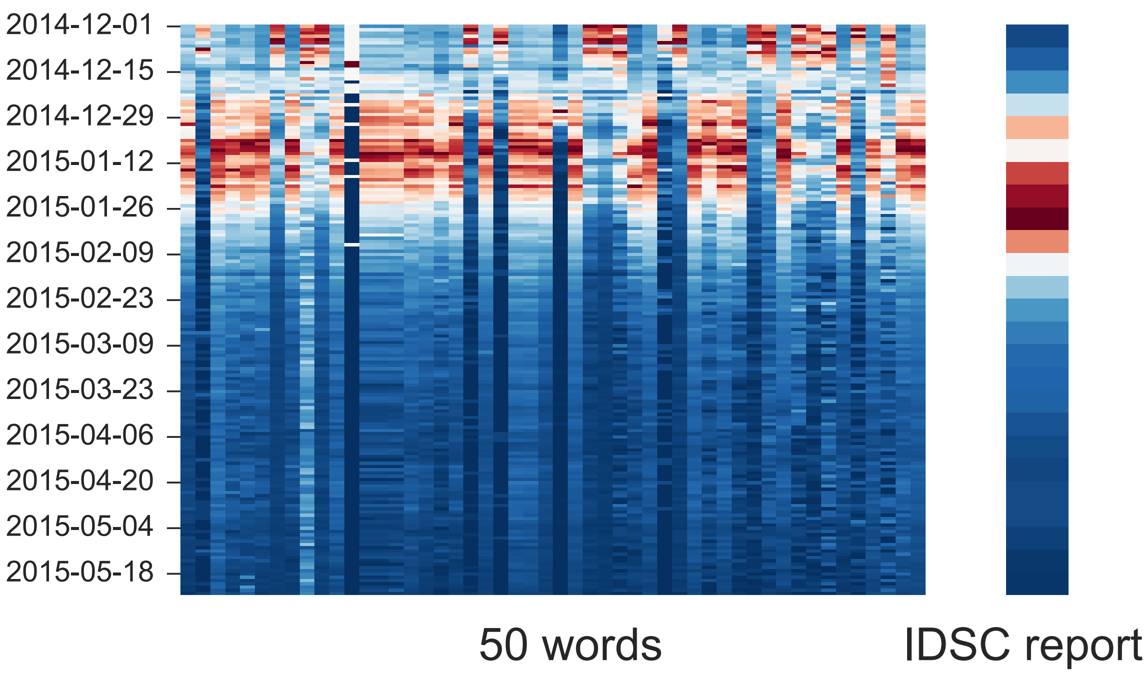
Heatmap representation of Matrix





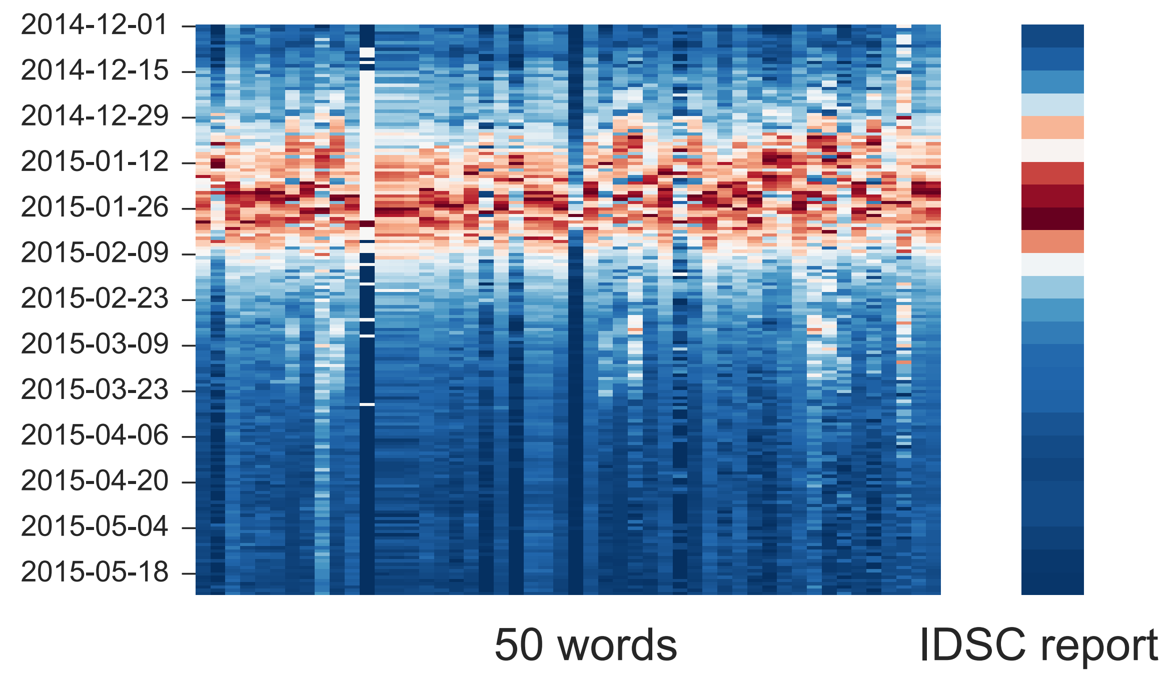
Raw word counts

# of patients



X y

Apply time-shift



X y

Effectiveness of time shift

- Regression for **<u>nowcasting</u> with applying time-shift or not:**
  - Lasso (Tibshirani, 1994)
  - Elastic-Net (Zou and Hastie, 2005)
- The searching range of time shift τ is in [0, ..., 60]

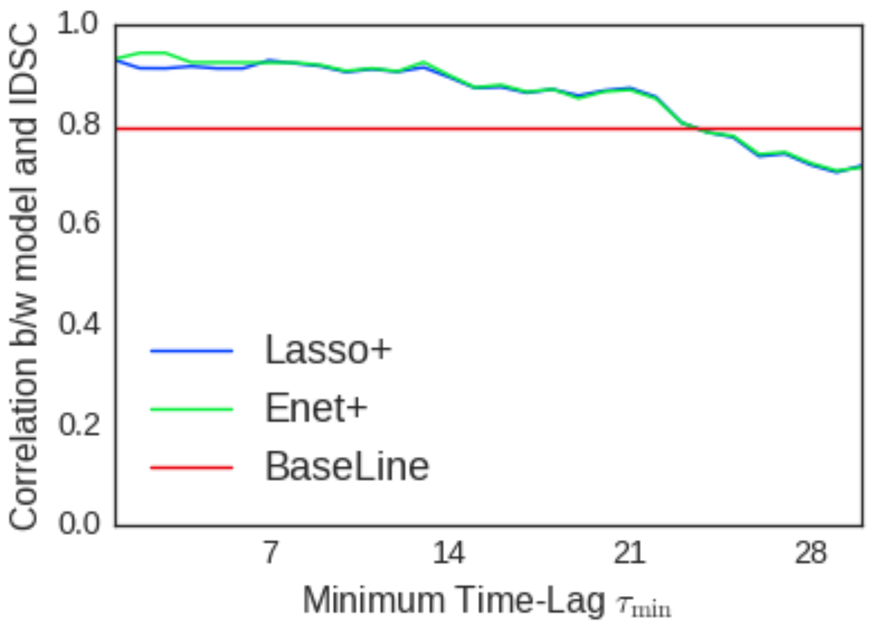
	Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
	Test	Season 1		Season 2		Season 3		
with time-shift	Lasso+	**0.952**	0.907	0.951	0.888	0.955	**0.963**	**0.936**
with time-shift	ENet+	0.944	0.898	**0.960**	0.878	**0.967**	0.959	0.934
without time-shift	Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.825
without time-shift	ENet	0.900	**0.927**	0.809	**0.914**	0.792	0.805	0.858

※ Higher is better

Outline

Data

Time shift:
Nowcasting

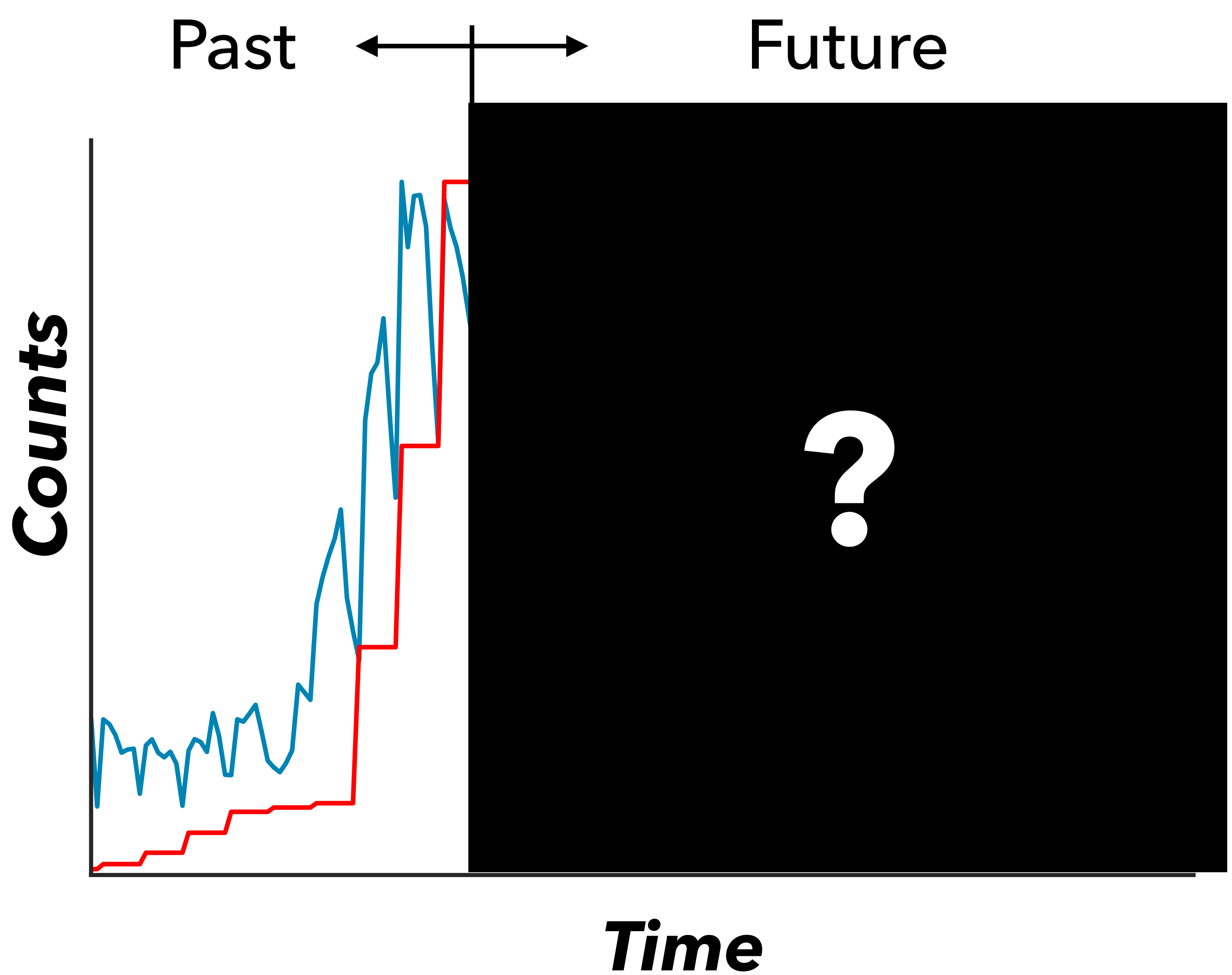


Time shift:
Forecasting

Limitation

- To estimate specific day of the **epidemic** through Twitter, we need to gather **same day's tweet**

- How to predict ***future disease outbreaking***?

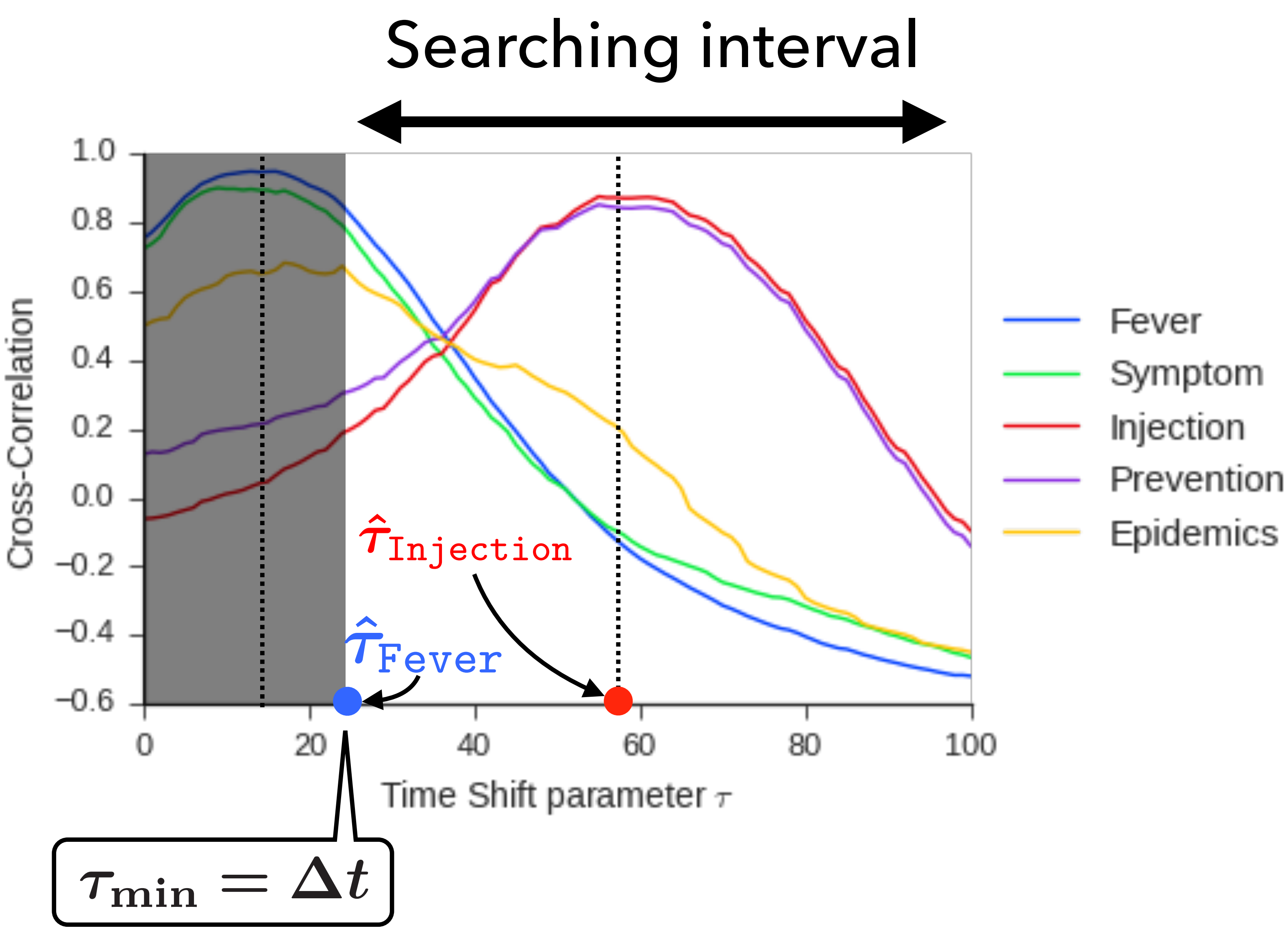


- **# of flu related tweets**
- **# of flu patients**

Restrict time shift estimation

- In order to forecast ***Δt*** days future epidemics,

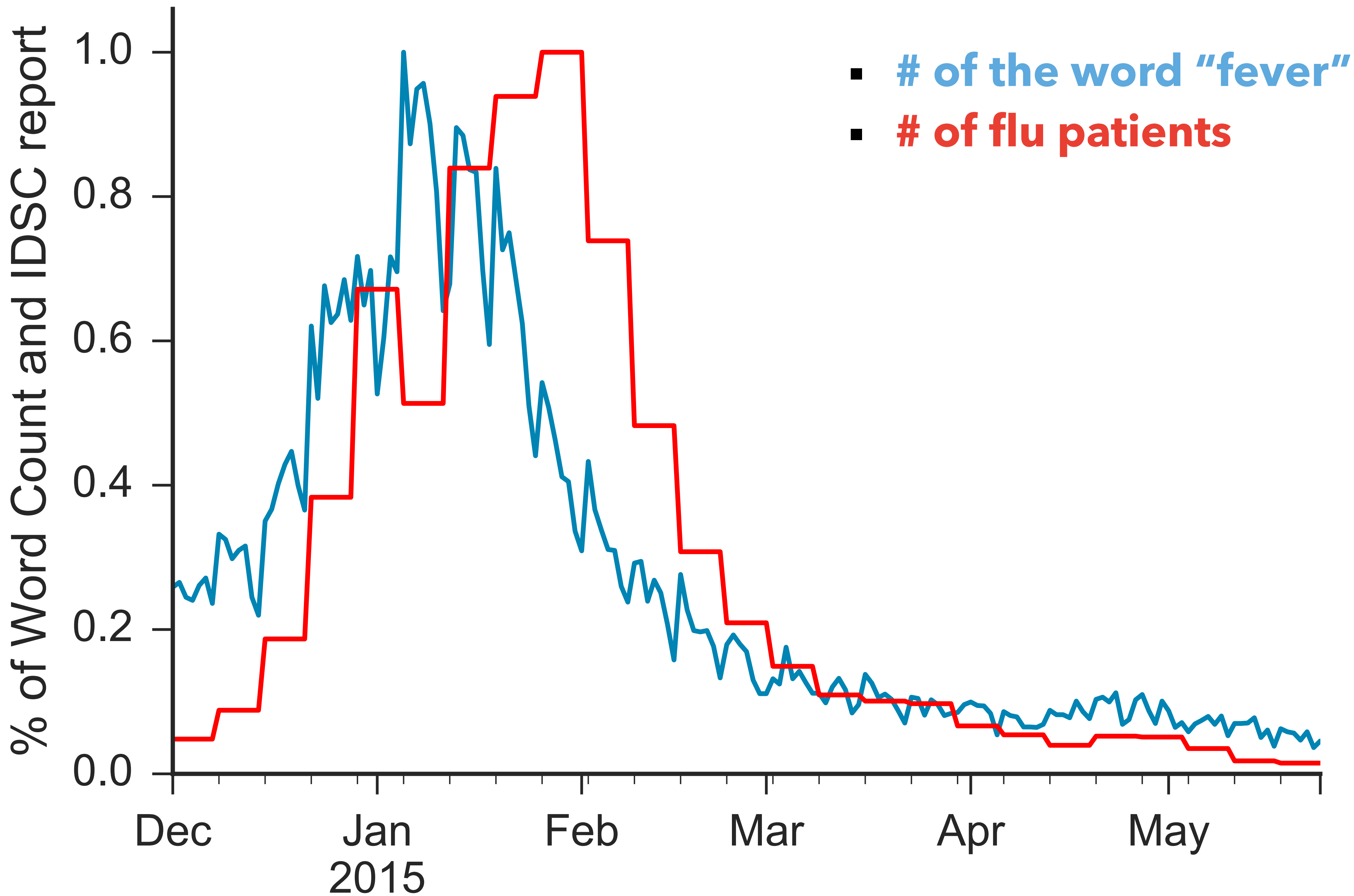
  we restrict searching interval of time shift at least ***Δt*** days



$$\hat{\tau}_v = \underset{\tau \in \{\tau_{\min}, \dots\} \subset \mathbb{N}}{\arg\max}\, r_{x_v, y}(\tau)$$

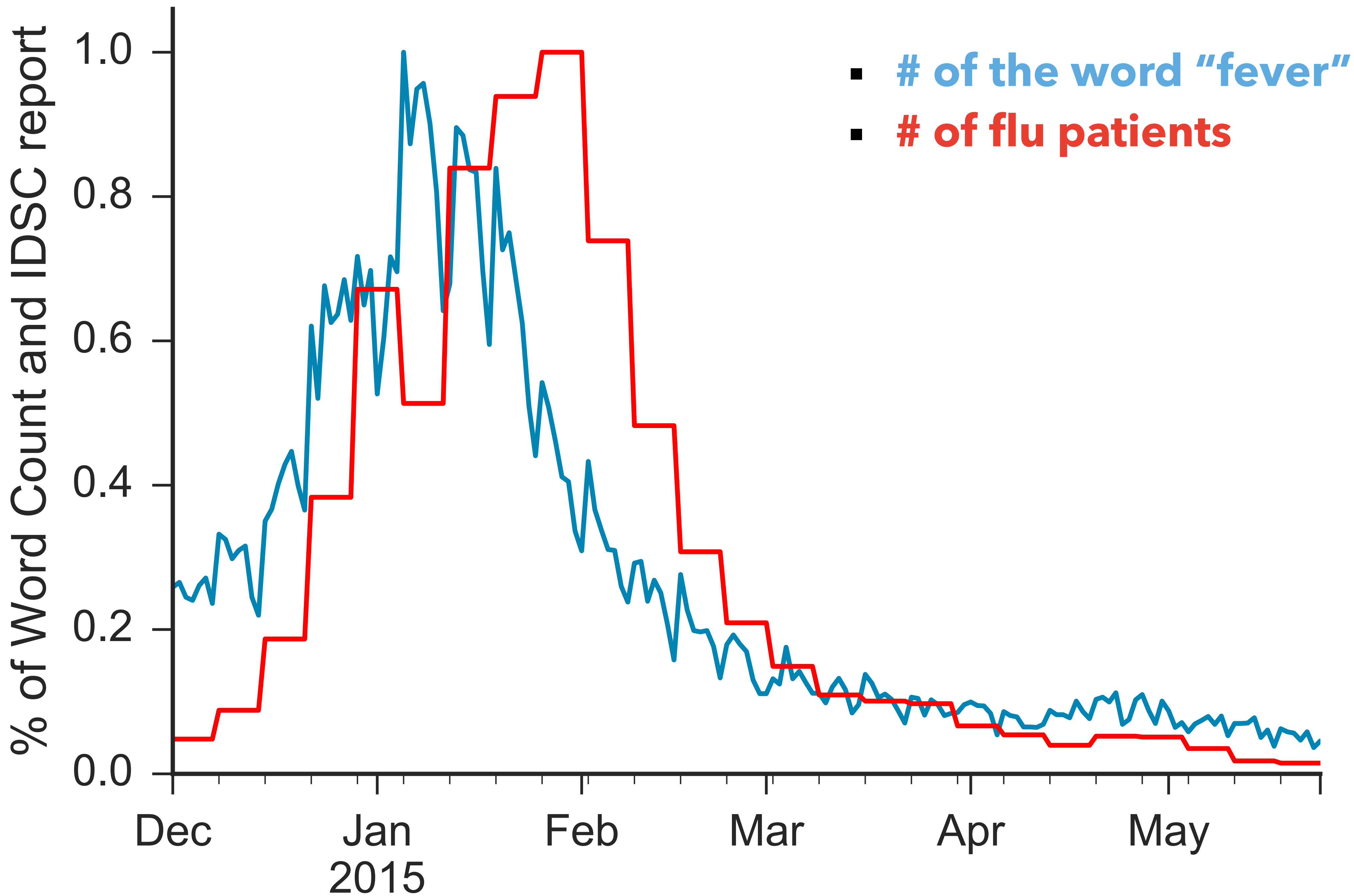
Motivating example

- Nowcasting case: $\tau \in [0, \tau_{max}]$



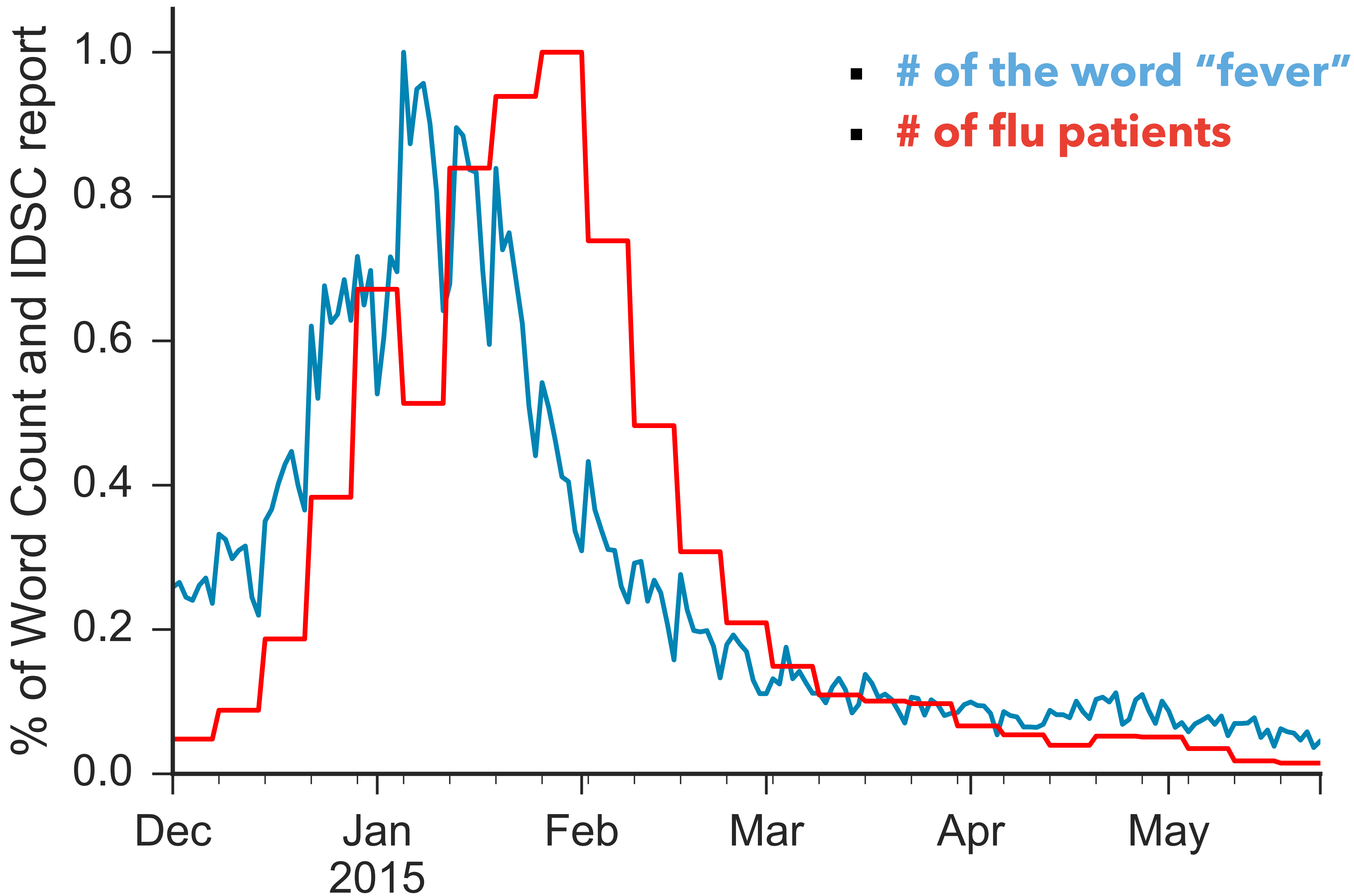
Motivating example

- Forecasting case (10 days future): $\tau \in [10, \tau_{max}]$



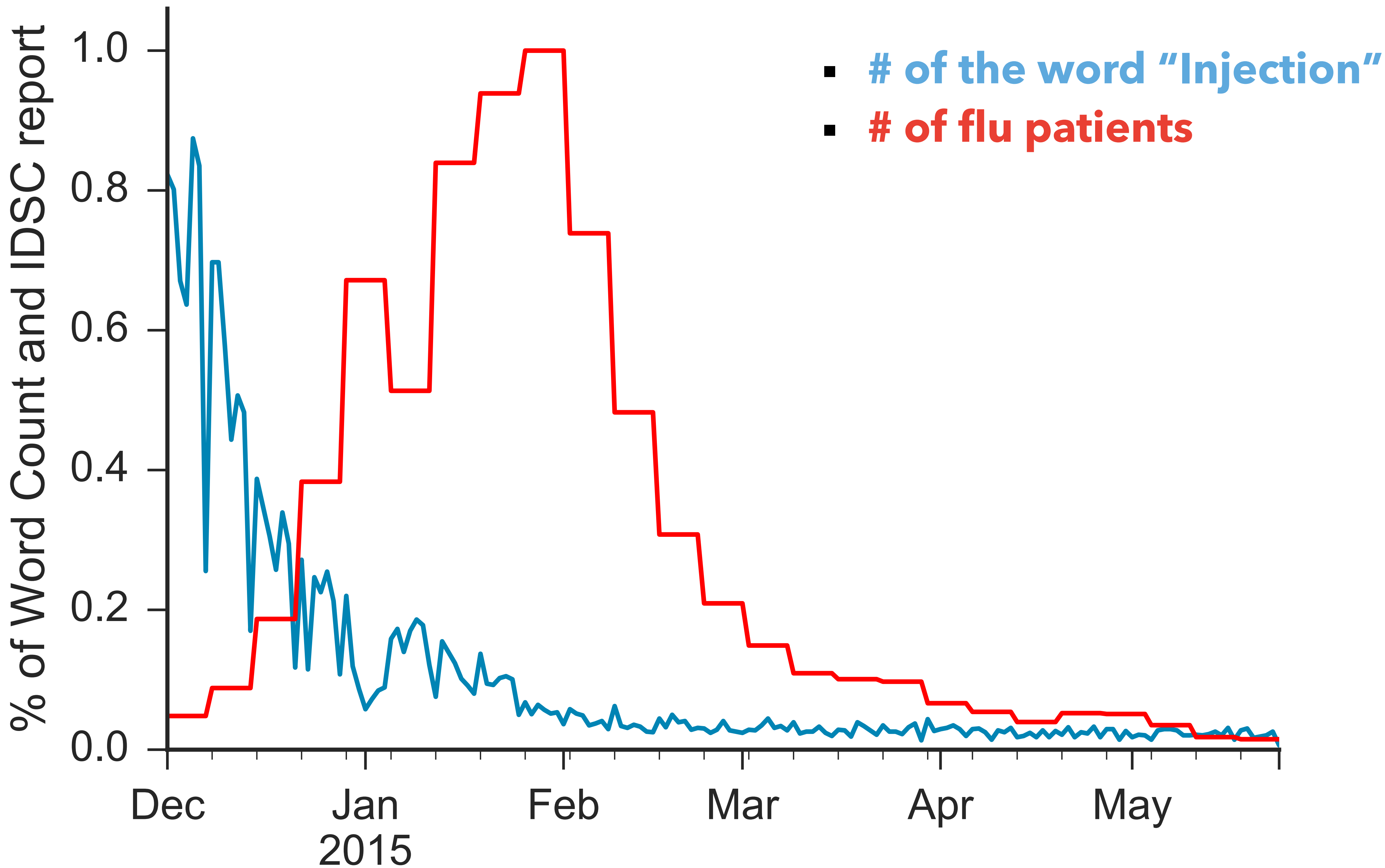
Motivating example

- Forecasting case (30 days future): $\tau \in [30, \tau_{max}]$



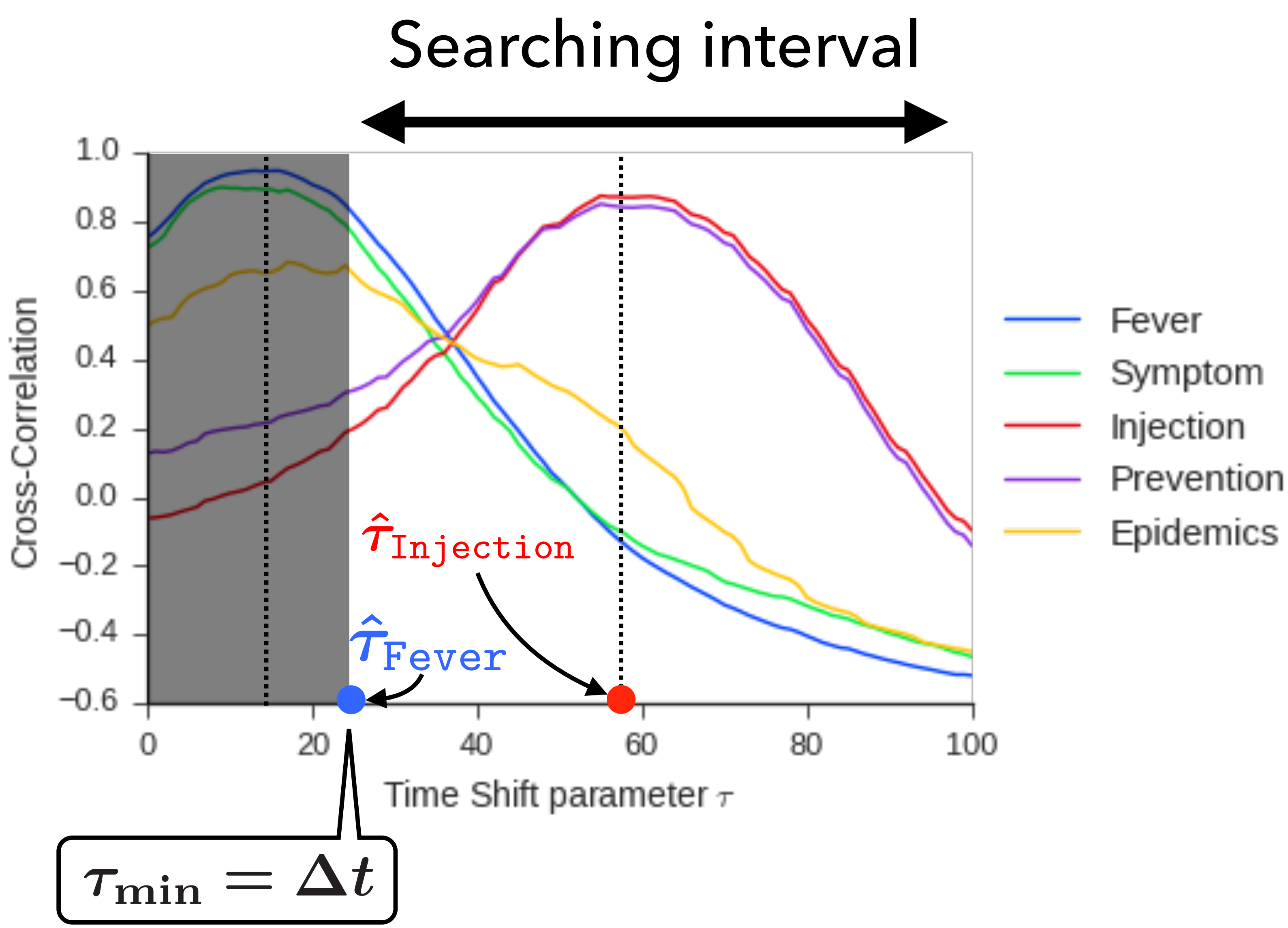
Motivating example

- Forecasting case (30 days future): $\tau \in [30, \tau_{max}]$



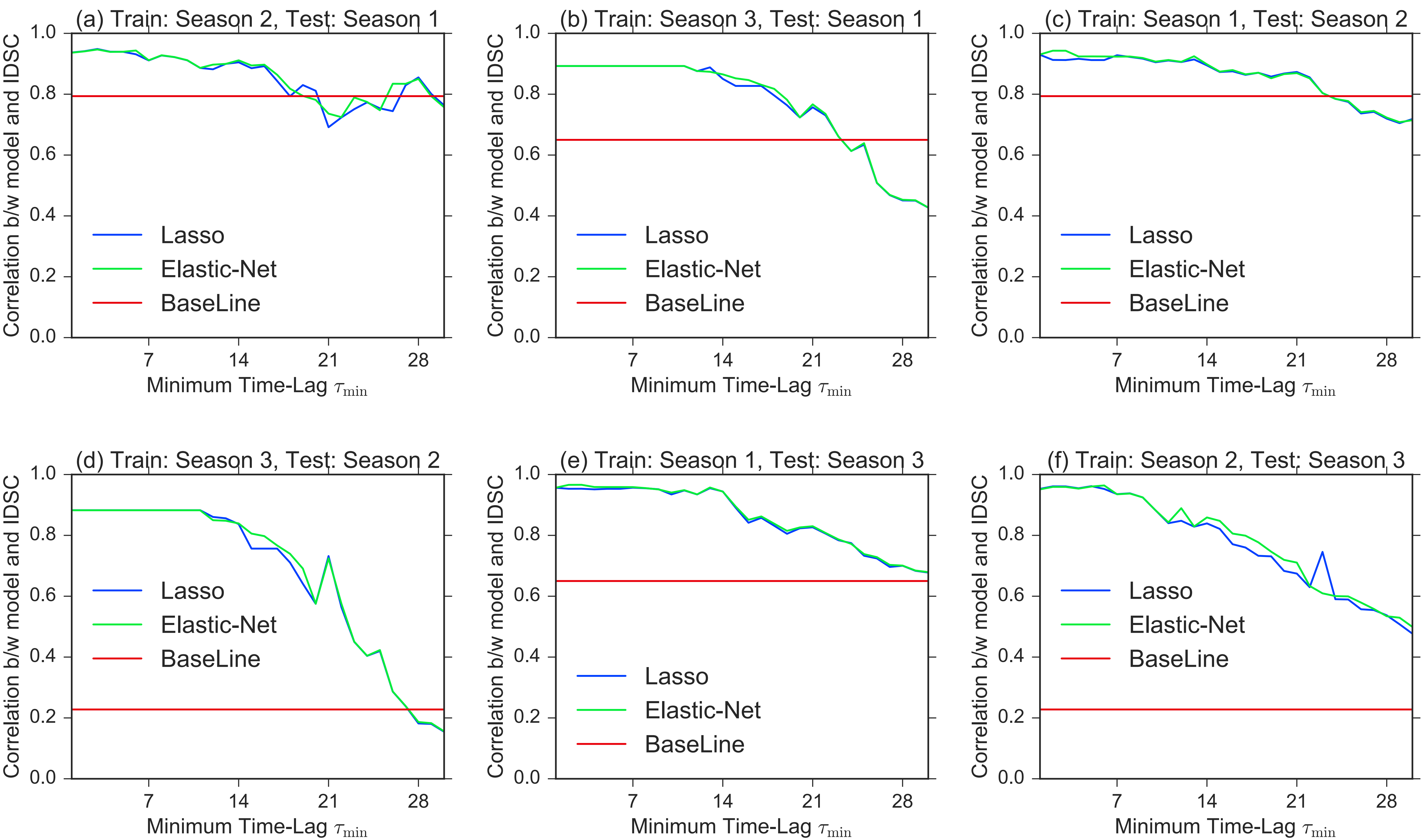
Forecasting Modeling

- In each ***Δt***, we search optimal time shift for all words.
- Estimate model by Lasso & ENet using these features.



$$\hat{\tau}_v = \mathop{\arg\max}_{\tau \in \{\tau_{\min}, \dots\} \subset \mathbb{N}} r_{x_v, y}(\tau)$$

Our model beyonds baseline



- BaseLine: $\hat{y}_{\mathtt{test}}(t) = y_{\mathtt{train}}(t)$

※ **Higher is better**

Summary

- We discovered the time difference between twitter and actual phenomena.
- We proposed but handling such difference to improve the nowcasting performance and extend for forecasting model.
- Our method is widely applicable for other time series data which has time-lag between response and predictors.

Code and Data available at http://sociocom.jp/~iso/forecastword