



UCL



An introduction to ‘digital’ health surveillance from online user-generated content

Vasileios Lampos

*Department of Computer Science
University College London*

(March, 2016; @ DIKU)

... before we begin

- + Research Fellow, Department of Computer Science, University College London 
- + Funded by the  **i-sense** project
(*Early-Warning Sensing Systems for Infectious Diseases*)
- + Research interests
 - + Artificial Intelligence
 - + Statistical Natural Language Processing
 - + Online user-generated content mining
 - + Digital Health & Computational Social Science

Website: lampos.net | Twitter: [@lampos](https://twitter.com/lampos)

Structure of the lecture

1. Essentials on public health surveillance
2. The very basics of linear regression
3. **Using linear regression to develop Google Flu Trends**
4. More regression basics: regularised regression
5. **Using regularised linear regression to map Twitter data to estimates about influenza rates**
6. Further regression basics (incl. a nonlinear approach)
7. **Improving Google Flu Trends**
8. **Assessing the impact of a health intervention using Internet data**
9. Recap and concluding remarks

1. Essentials on public health surveillance

Public health surveillance

... is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice.

It can:

- + serve as an **early warning system** for impending public health emergencies
- + document the **impact of an intervention**, or track progress towards specified goals
- + monitor and clarify the epidemiology of health problems, to allow priorities to be set and to **inform public health policy and strategies**

http://www.who.int/topics/public_health_surveillance/en/

Examples of public health surveillance

Syndromic surveillance

using health data preceding a solid diagnosis to signal a potential outbreak
e.g. visits to general practitioners, hospitals, emergency call systems, school absenteeism, over-the-counter drug sales

Laboratory-based surveillance

laboratory-confirmed cases (laboratory testing and diagnosis)

Organisations

Centers for Disease Control and Prevention (CDC) in the US

European Centre for Disease Prevention and Control (ECDC) in the EU

Public Health England (PHE)

Staten Serum Institut (SSI) in Denmark

<http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>

Limitations of traditional health surveillance

- + Derivations are based on the subset of people that actively seek medical attention

non-adults or the elderly are responsible for the majority of doctor visits or hospital admissions

thus, these methods may not always be able to capture a disease outbreak emerging in the actual population

- + Infrastructure is required

i.e. a health surveillance system may not be applicable to under-developed parts of the world

- + Time delays

it may take days to process the records of general practitioners and hospitals and make an estimate about the rate of a disease

Digital health surveillance

also known as Info-veillance

Syndromic surveillance that utilises the online (web) contents (Eysenbach, 2006)

Examples of online user-generated content (**UGC**):

- + search engine query logs
 - + social media
 - + online fora, specialised email lists (e.g. medical)
-
- > Famous digital health surveillance example:
[Google Flu Trends](#) [[Link 2](#)]
 - > Infamous one (*and under development*): [Flu Detector](#)

Advantages of digital health surveillance

- + Online content can potentially access a larger and **more representative** part of the population (*or at least a complementary one*)
- + 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

Challenges in digital health surveillance

- + Online information is **noisy** and oftentimes **inaccurate**
- + Statistical natural language processing is not perfect, i.e. **word sense disambiguation** may not always be successful
- + Online behaviour and content may respond to other factors, such as **news media coverage**
- + **Evaluation** of outcomes (e.g. estimated disease rates or the impact of a health intervention) is hard

2. The very basics of (linear) regression

Broad definitions for regression

Regression

A statistical tool for investigating (*and estimating*) the relationship between variables. There is usually one dependent variable (y), the one we want to estimate, and one (or more) independent variables (x), also known as predictors or observations.

$$y \approx f(x, w)$$

Text regression

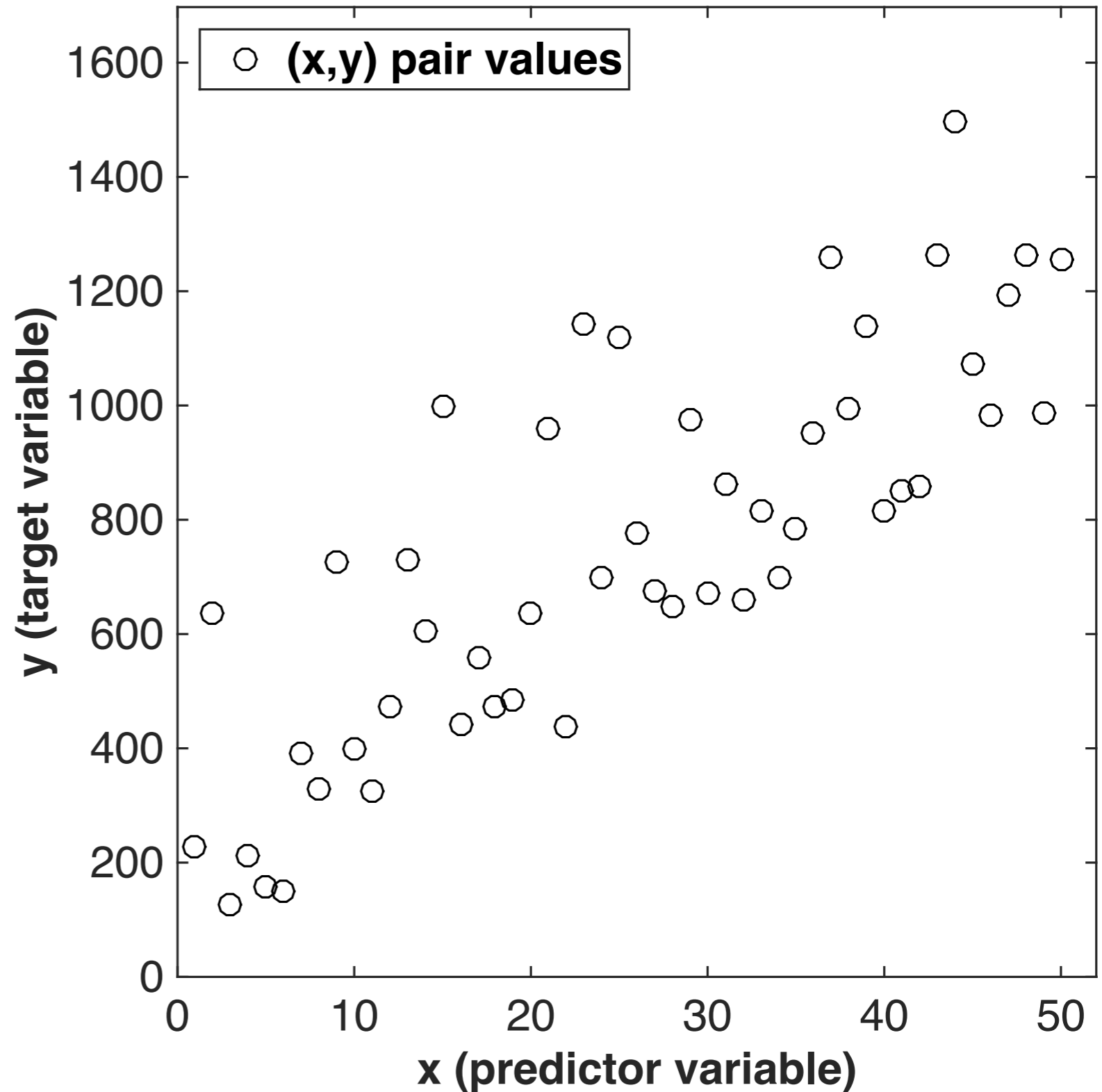
Regression, where the observed (input) variable is based on textual information

Regression: An example

y is the **target variable**
we want to estimate

by looking at the
observed variable x

*their relationship looks
like a **linear** one*



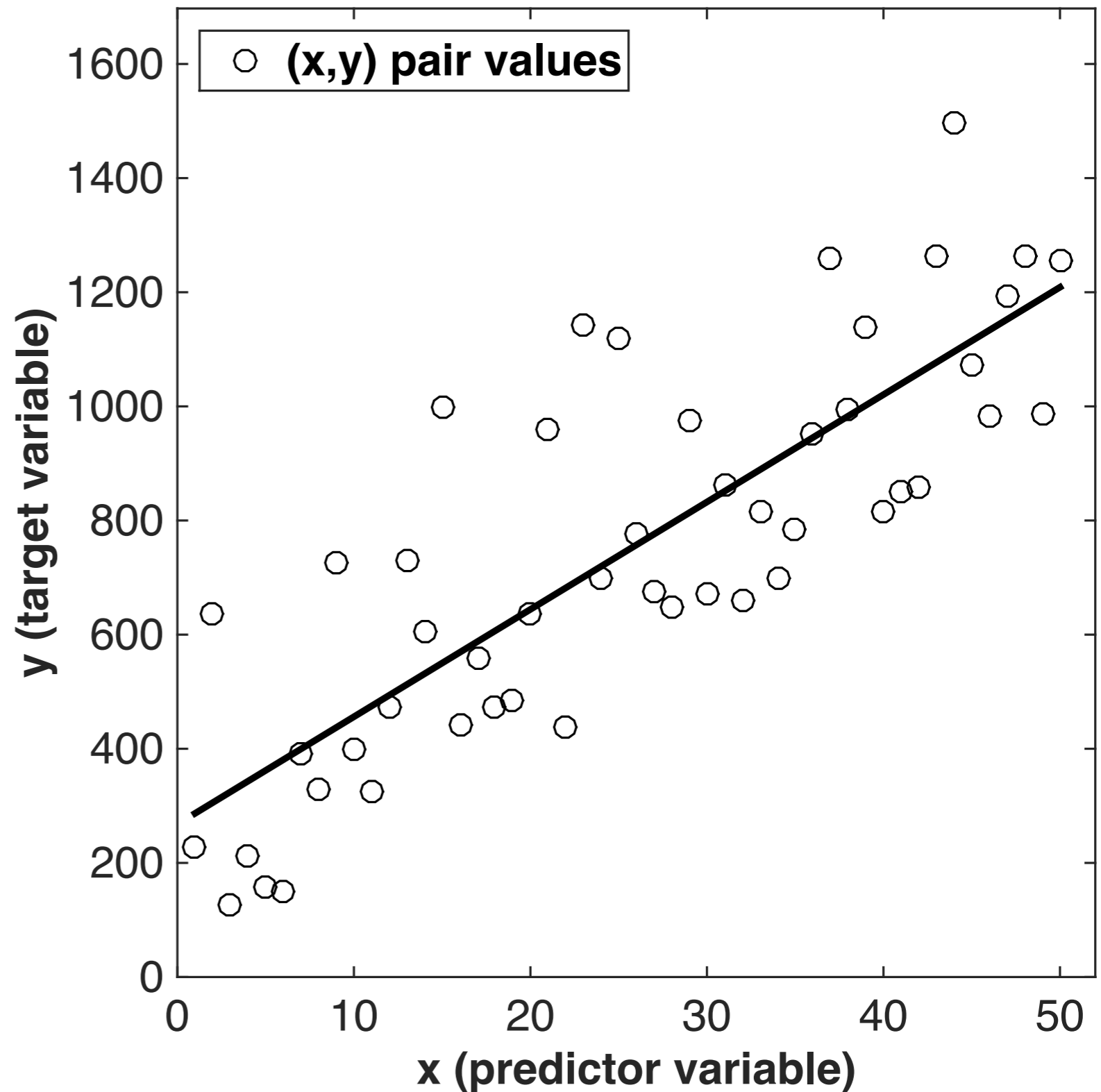
Regression: An example

which means that we can almost *accurately* (?) learn a weight w and a bias term β such that

$$y = w * x + \beta$$

In this case:

$$y_{OLS} = 18.817 * x + 267.922$$



Ordinary Least Squares (**OLS**) regression (1)

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

$$\operatorname{argmin}_{w, \beta} \sum_{i=1}^n \left(y_i - \beta - \sum_{j=1}^m x_{ij} w_j \right)^2$$

i.e. find the best values that minimise this function (summation)

Ordinary Least Squares (OLS) regression (2)

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

$$\operatorname{argmin}_{\mathbf{w}, \beta} \sum_{i=1}^n \left(y_i - \beta - \sum_{j=1}^m x_{ij} w_j \right)^2 \quad \text{or below in **matrix form**, i.e. using vectors and matrices instead of scalars}$$

$$\operatorname{argmin}_{\mathbf{w}_*} \|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2, \quad \text{where } \mathbf{X}_* = [\mathbf{X} \operatorname{diag}(\mathbf{I})]$$

$$\Rightarrow \mathbf{w}_* = \left(\mathbf{X}_*^T \mathbf{X}_* \right)^{-1} \mathbf{X}_*^T \mathbf{y}$$

Regression: How *good* is an inference? (1)

Target variable: $y = y_1, \dots, y_N$

Estimates: $\hat{y} = \hat{y}_1, \dots, \hat{y}_N$

Mean Squared Error:
$$\text{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{t=1}^N (\hat{y}_t - y_t)^2$$

Mean Absolute Error:
$$\text{MAE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{t=1}^N |\hat{y}_t - y_t|$$

Mean Absolute Percentage of Error:

$$\text{MAPE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{t=1}^N \left| \frac{\hat{y}_t - y_t}{y_t} \right|$$

Regression: How *good* is an inference? (2)

Target variable: $y = y_1, \dots, y_N$

Estimates: $\hat{y} = \hat{y}_1, \dots, \hat{y}_N$

Pearson (*linear*) correlation $\in [-1, 1]$

$$r = \frac{1}{N-1} \sum_{t=1}^N \left(\frac{y_t - \mu(y)}{\sigma(y)} \right) \left(\frac{\hat{y}_t - \mu(\hat{y})}{\sigma(\hat{y})} \right)$$

Note: Pearson correlation is not always indicative of performance (i.e. it can be occasionally misleading), but useful nonetheless

3. *Using OLS regression to map search query frequency to an influenza rate estimate — Google Flu Trends*

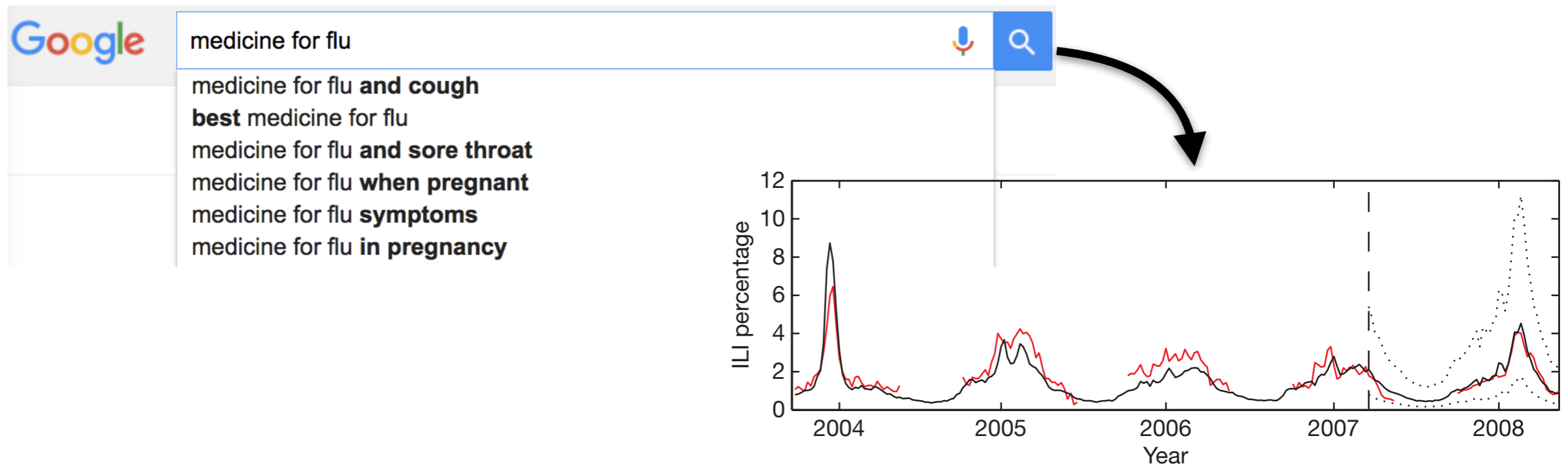
(Ginsberg et al., 2009)

Google Flu Trends: *the idea* (1)



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

Google Flu Trends: *the idea* (2)



Why is this an interesting task?

1. For all the reasons we mentioned already!
(*see the advantages of digital health surveillance*)
2. Plus, seasonal influenza epidemics are a major public health concern, i.e. causing **250,000 to 500,000 deaths worldwide per year.**

Google Flu Trends: *the data*

- + Search query logs (anonymised) between **2003** and **2008**
- + Weekly counts of **50 million queries** conducted by users located in the US and its 9 broad regions (formed by aggregations of member states)
- + Each query q **normalised** using the total number of searches conducted in the same weekly time interval (t) and location

$$\frac{\text{\#searches for } q \text{ in week } t}{\text{\#searches in week } t}$$

- + Model training and evaluation based on **CDC records**

Google Flu Trends: *the method* (1)

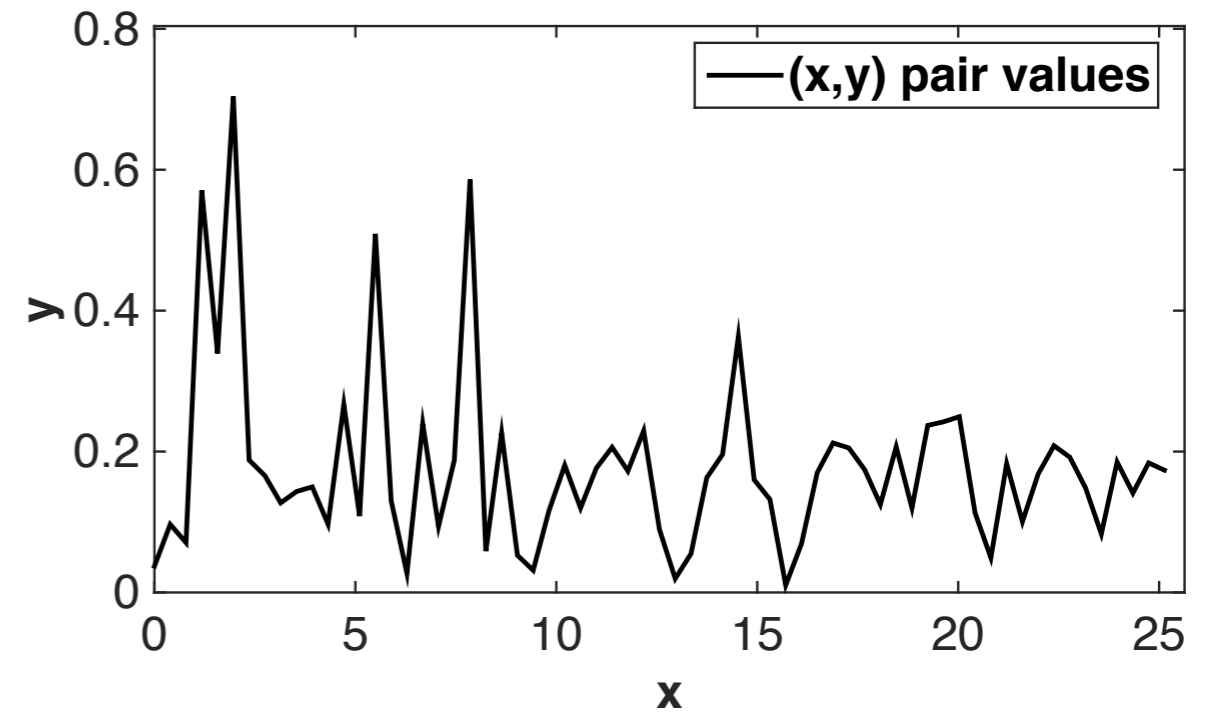
$$\mathit{logit}(P) = \beta_0 + \beta_1 \times \mathit{logit}(Q) + \varepsilon$$

- Q** Aggregate frequency of a set of search queries
- P** Percentage (*probability*) of doctor visits
- β_0** Regression bias term
- β_1** Regression weight (*one weight only*)
- ε** independent, zero-centered noise (*assumed*)

Google Flu Trends: *the method* (2)

the *logit* function

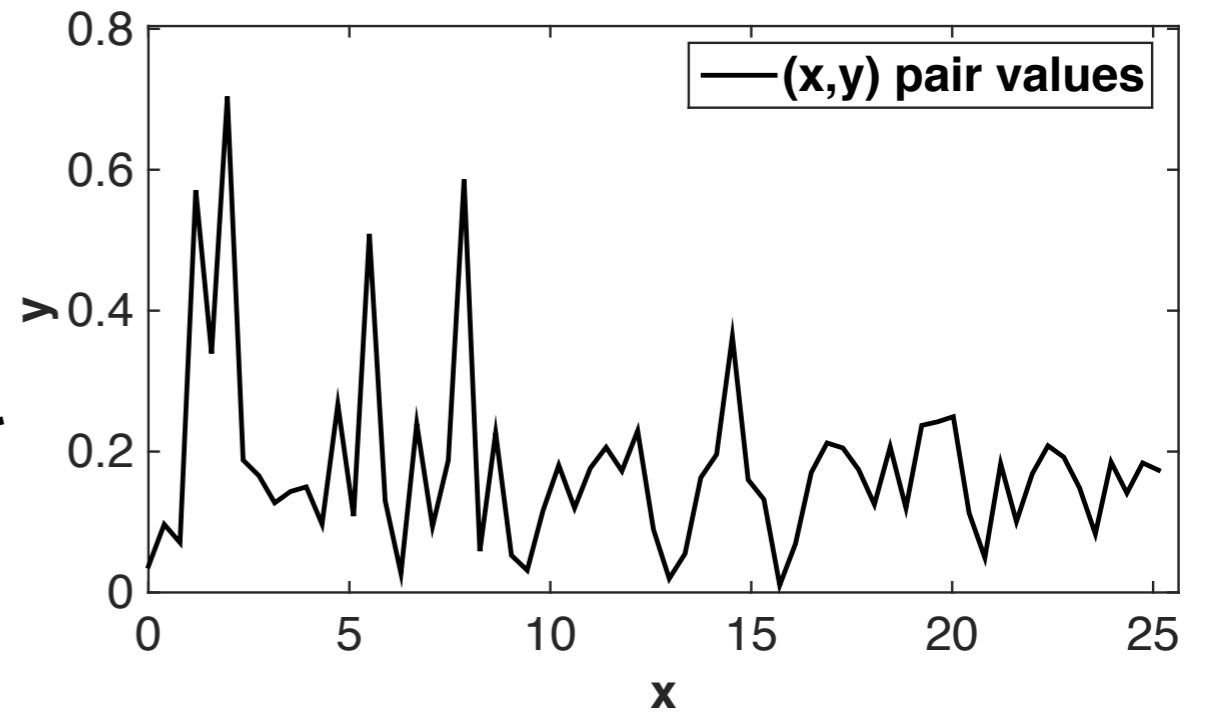
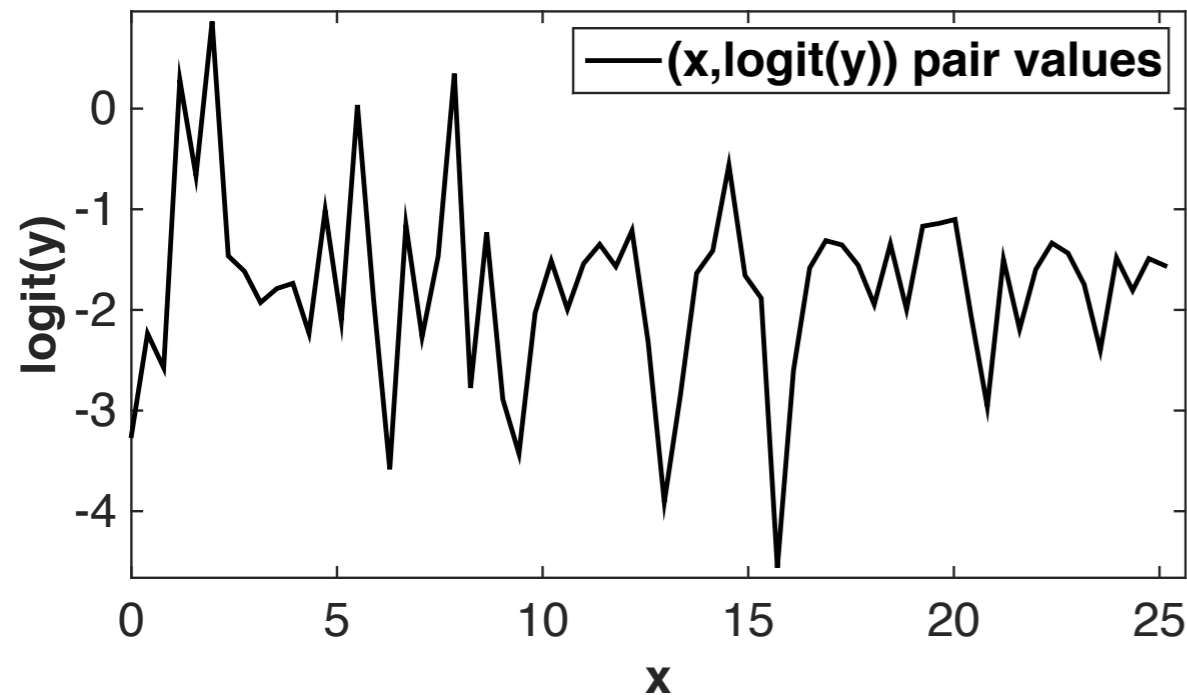
$$\text{logit}(\alpha) = \log(\alpha / (1 - \alpha))$$



Google Flu Trends: *the method* (2)

the *logit* function

$$\text{logit}(\alpha) = \log(\alpha / (1 - \alpha))$$



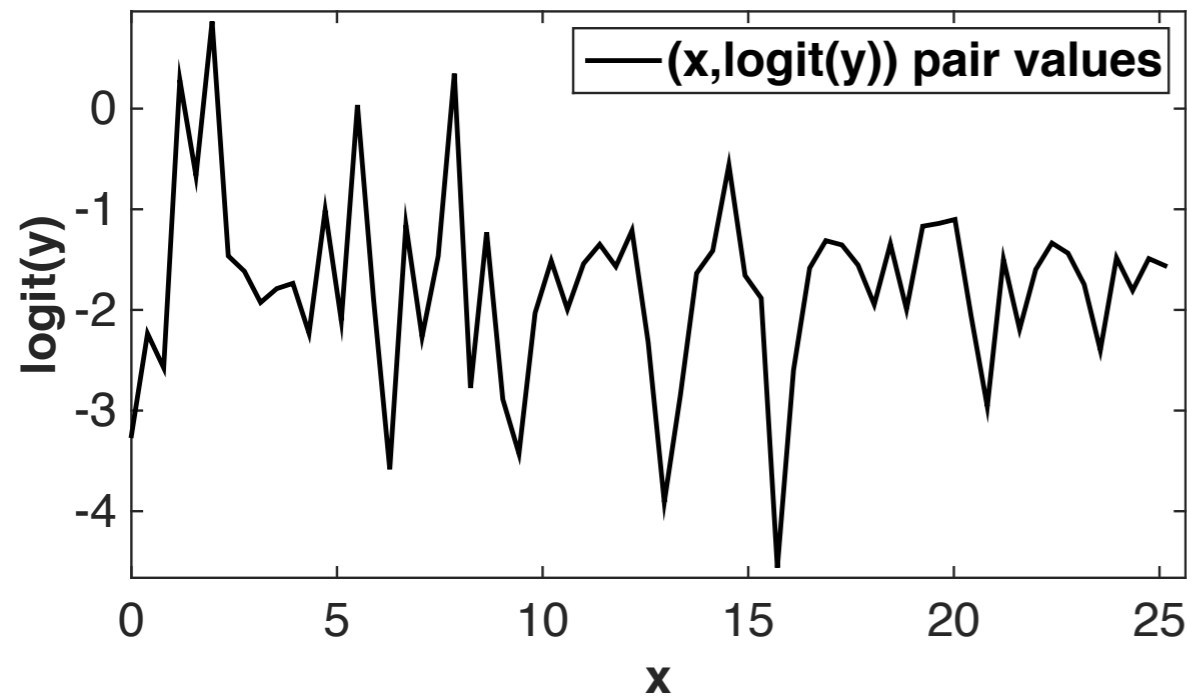
intermediate values are ‘squashed’

border values are ‘emphasised’

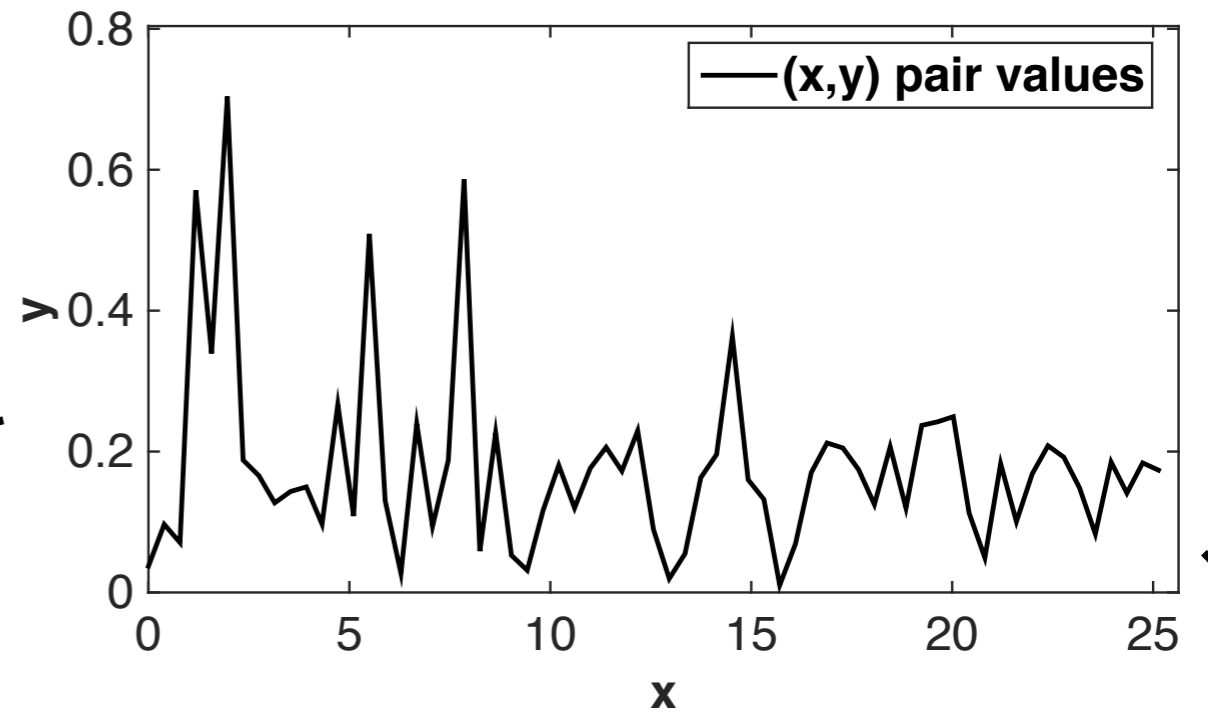
Google Flu Trends: *the method* (2)

the *logit* function

$$\text{logit}(\alpha) = \log(\alpha / (1 - \alpha))$$

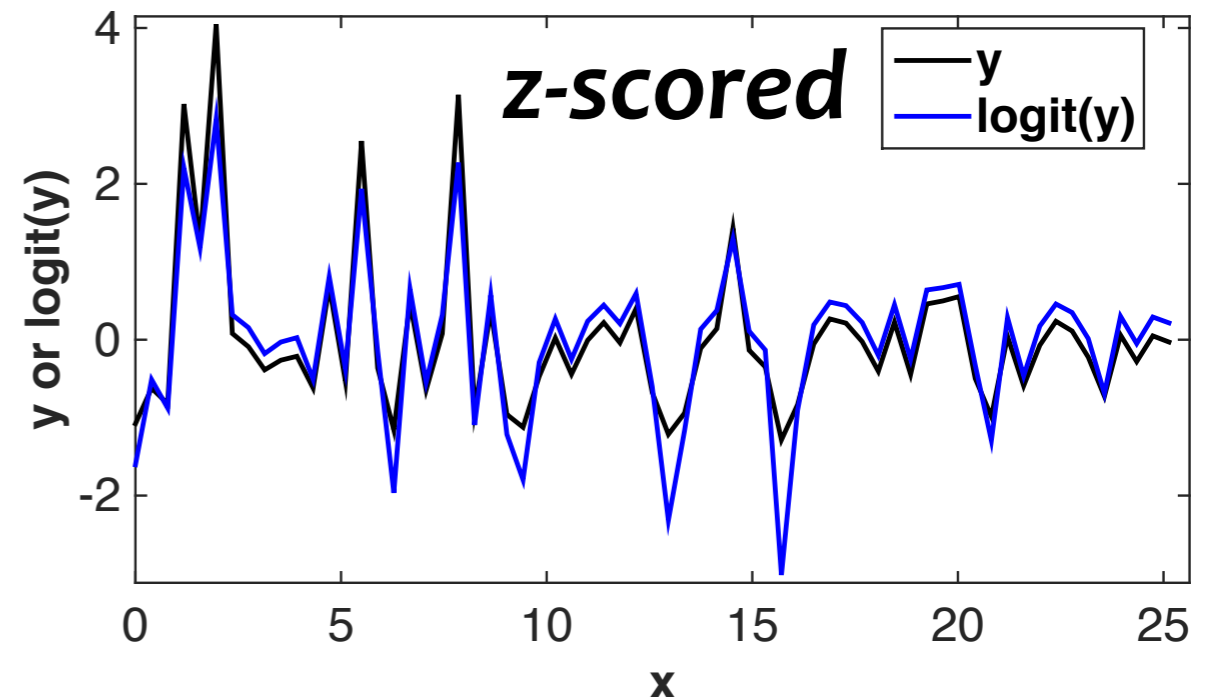


logit



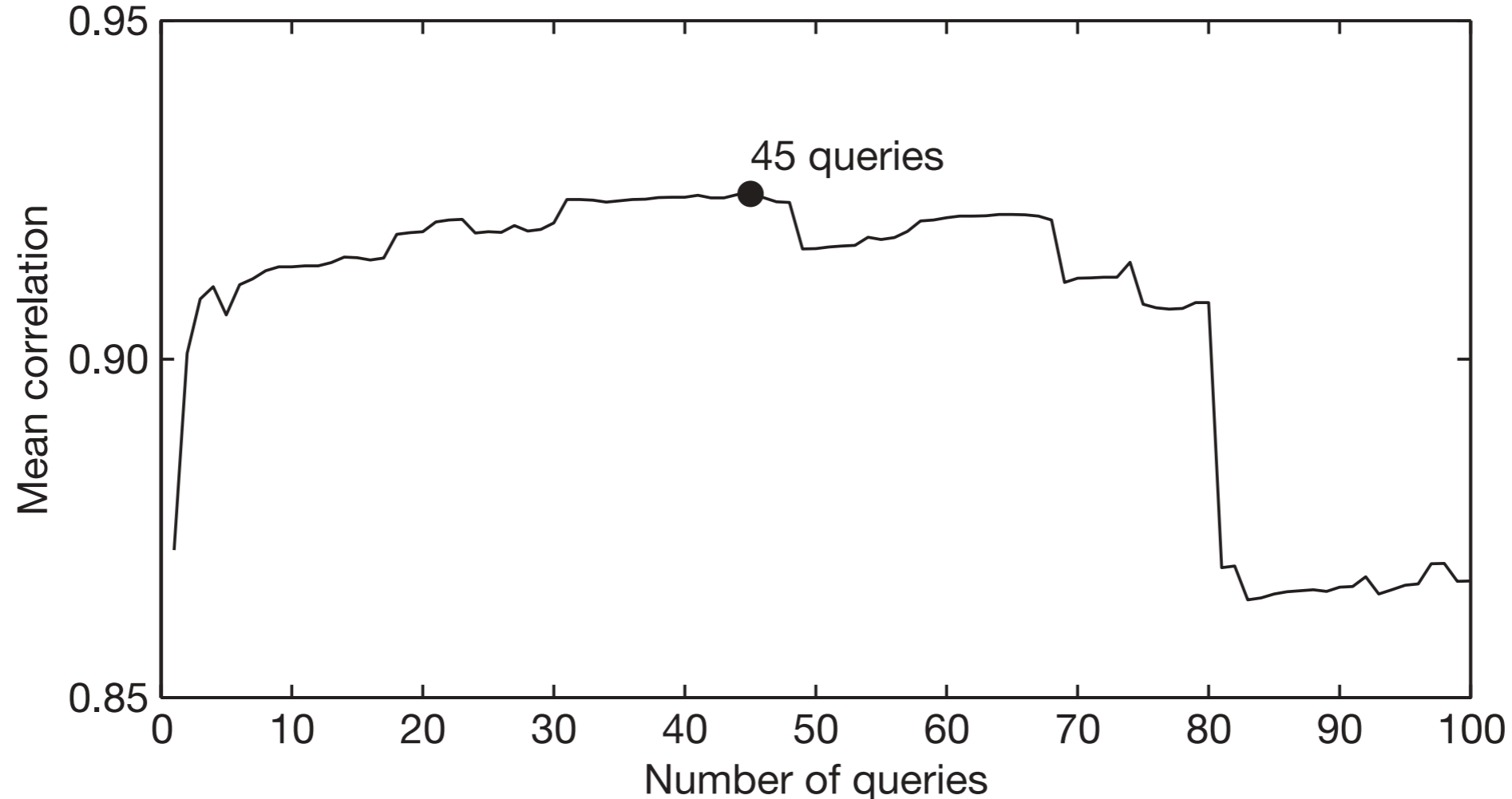
intermediate values are 'squashed'

border values are 'emphasised'



Google Flu Trends: *the method* (3)

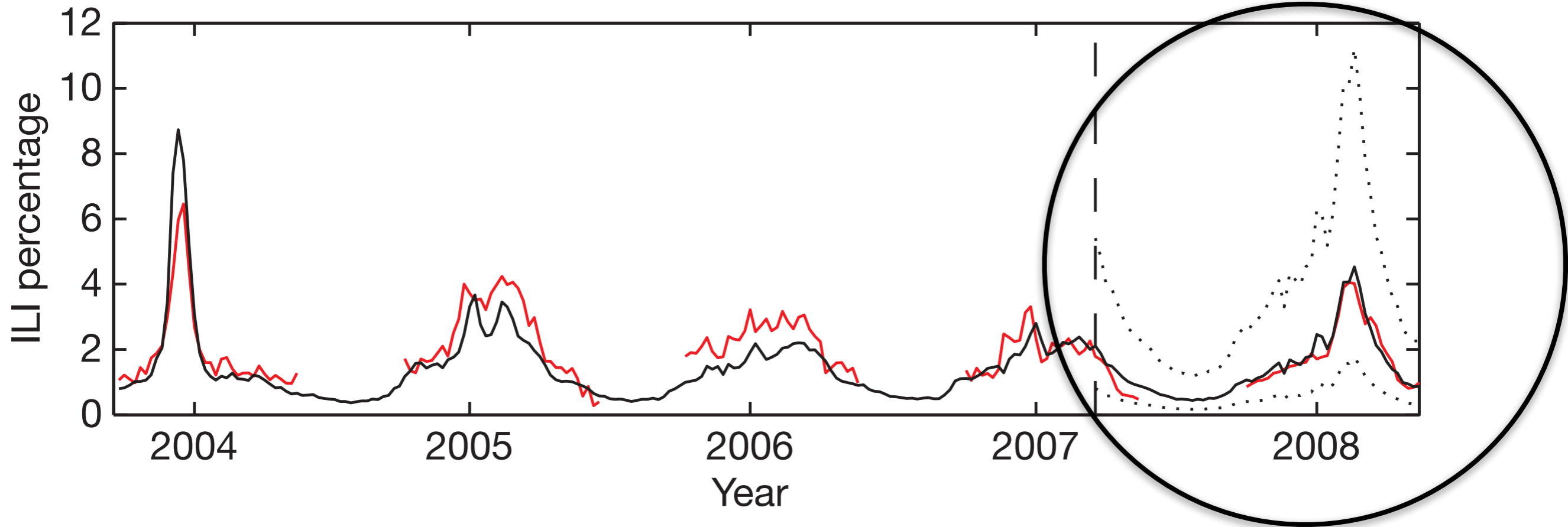
1. A flu rate estimation model is trained on each query (50 million) separately for the 9 US regions, i.e. 450 million models are trained
2. The N top performing queries (on average across the 9 regions) are identified; based on Pearson correlation (r) between inferences and CDC ILI rates
3. Starting from the best performing query and adding up one query each time, a new model is trained and evaluated



Google Flu Trends: *the results* (1)

Search query topic	Top 45 queries	
	<i>n</i>	Weighted
Influenza complication	11	18.15
Cold/flu remedy	8	5.05
General influenza symptoms	5	2.60
Term for influenza	4	3.74
Specific influenza symptom	4	2.54
Symptoms of an influenza complication	4	2.21
Antibiotic medication	3	6.23
General influenza remedies	2	0.18
Symptoms of a related disease	2	1.66
Antiviral medication	1	0.39
Related disease	1	6.66
Unrelated to influenza	0	0.00
Total	45	49.40

Google Flu Trends: *the results* (2)



Mean Pearson correlation (r)
for the 9 regions, $r = 0.97$

**4. *More regression basics:
Regularised regression***

Limitations of least squares regression

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

$$\operatorname{argmin}_{\mathbf{w}_*} \|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2, \text{ where } \mathbf{X}_* = [\mathbf{X} \operatorname{diag}(\mathbf{I})]$$

$$\Rightarrow \mathbf{w}_* = \left(\mathbf{X}_*^T \mathbf{X}_* \right)^{-1} \mathbf{X}_*^T \mathbf{y}$$

- May be **singular**, thus difficult to invert
- High dimensional models are **difficult to interpret**
- **Unsatisfactory prediction accuracy** (*estimates have large variance*)

Regularised regression: Ridge

$$\mathbf{w}_* = \underbrace{\left(\mathbf{X}_*^T \mathbf{X}_* + \lambda \mathbf{I} \right)}_{\text{non singular}}^{-1} \mathbf{X}_*^T \mathbf{y}$$

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

$$\text{or } \operatorname{argmin}_{\mathbf{w}_*} \left\{ \|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2 + \lambda \|\mathbf{w}\|_{\ell_2}^2 \right\}$$

also known as **L2-norm regularisation**

Pros and Cons of ridge regression

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

$$\text{or } \operatorname{argmin}_{\mathbf{w}_*} \left\{ \|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2 + \lambda \|\mathbf{w}\|_{\ell_2}^2 \right\}$$

- + size constraint on the weight coefficients (**regularisation**); resolves problems caused by *collinear* variables
- + less degrees of freedom; often **better** predictive accuracy **than OLS regression**
- does **not** perform **feature selection** (all coefficients are nonzero); performance could be improved

Regularised regression: Lasso (1)

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

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

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

also known as **L1-norm regularisation**

Pros and Cons of lasso

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

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

- **no closed form solution** (quadratic optimisation needed)
- + Least Angle Regression (**LARS**) algorithm explores the entire regularisation path, i.e. **all values for λ**
- + \mathbf{w} tends to be **sparse** enhancing both the **interpretability** of a model and providing (often) **better performance**
- if $m > n$, at most n variables can be selected, i.e. have a nonzero weight
- **collinear predictors** (high pair-wise correlation) may lead to **inconsistent models**

5. *Using lasso regression to map Twitter data to an influenza rate estimate*

(Lampos and Cristianini, 2010)

About Twitter (1)

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 RETWEETS	1 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 (2)

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

- > **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)
- > **information is noisy and not always accurate**

i think i have the flu but i still look fabulous

 Reply  Retweet  Favorite

Twitter 'Flu Trends': *the data*

Twitter

- > **27 million** tweets
- > from 22/06/2009 to 06/12/2009
- > geolocated in the **UK**
centred around 54 cities (10 Km radius)

Health surveillance data

- > **influenza-like illness (ILI) rates** from the Health Protection Agency (now Public Health England) and the Royal College of General Practitioners (RCGP)
- > expressing the **number of ILI doctor consultations per 100,000 citizens** in the population for various UK regions

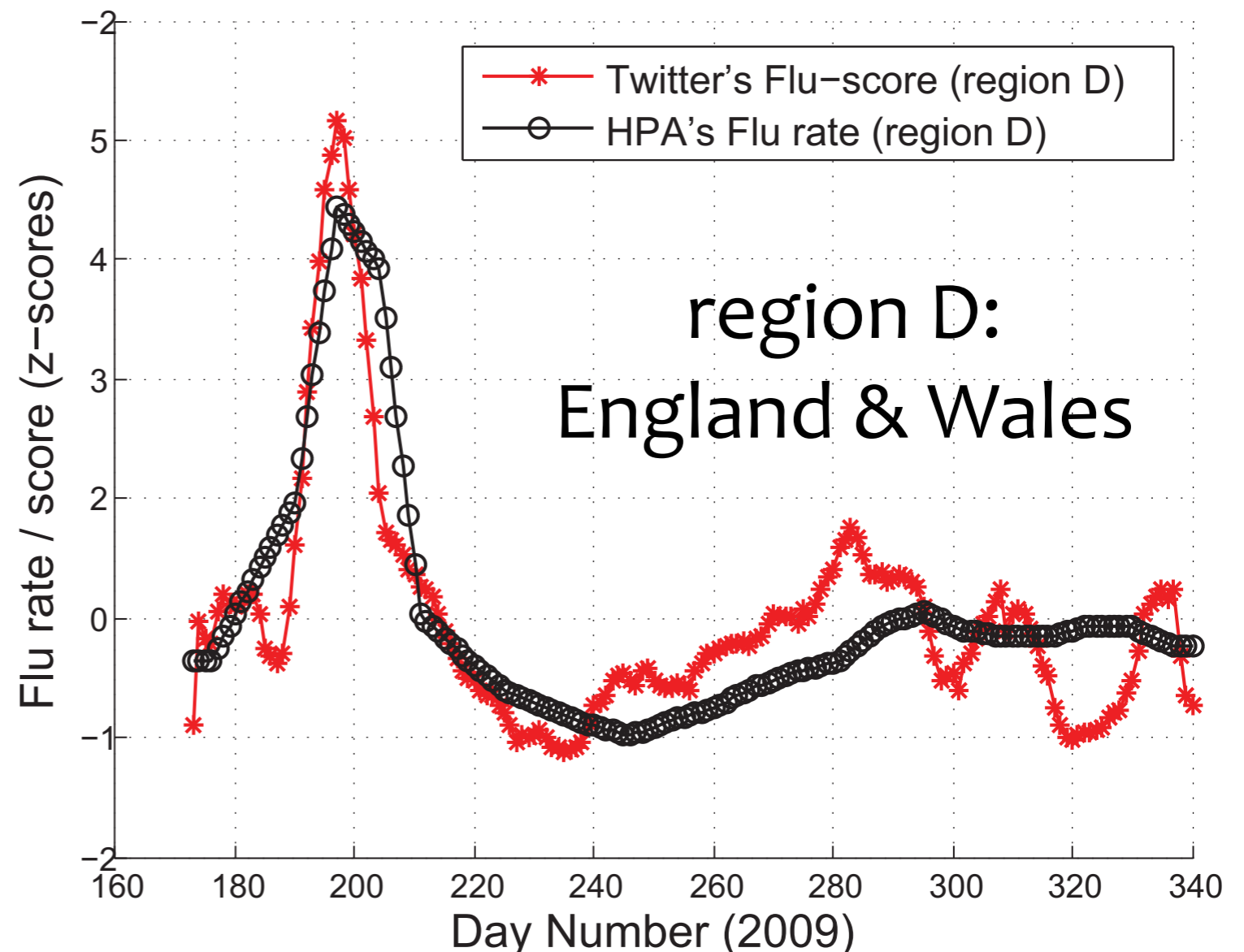
Twitter 'Flu Trends': *the methods* (1)

Is there a signal in the data?

List with **41 handpicked keywords** related to flu
e.g. 'fever', 'runny nose', 'sore throat', 'headache' etc.

Compute their
aggregate daily
frequency & compare
it to HPA records

$r = 0.82$ to 0.86



Twitter 'Flu Trends': *the methods* (2)

Can we improve the obtained correlations by learning a weight for each keyword using OLS regression? (Yes)

Train/Test	A	B	C	D	E	Avg.
A	-	0.8389	0.9605	0.9539	0.9723	0.9314
B	0.7669	-	0.8913	0.9487	0.8896	0.8741
C	0.8532	0.702	-	0.8887	0.9445	0.8471
D	0.8929	0.9183	0.9388	-	0.9749	0.9312
E	0.9274	0.8307	0.9204	0.9749	-	0.9134
Total Avg.						0.8915

Twitter ‘Flu Trends’: *the methods* (3)

What if we made our feature set (keywords) more rich?

- + Use Wikipedia pages about flu, Patient forums (*more informal language*), expert pages (e.g. NHS-based) to **expand our keywords to 1,560** (from 41)
- + Many related keywords — much more unrelated
- + **Stop word removal**, i.e. basic words that bear no particular meaning are removed, e.g. ‘and’, ‘a’, ‘the’, ‘they’
- + **Porter stemming** is applied to normalise word endings, e.g. both ‘happy’ and ‘happiness’ are converted to ‘happi’
- + Lasso regularised regression to select and weight subset of keywords for capturing an ILI rate

Twitter ‘Flu Trends’: *the results* (1)

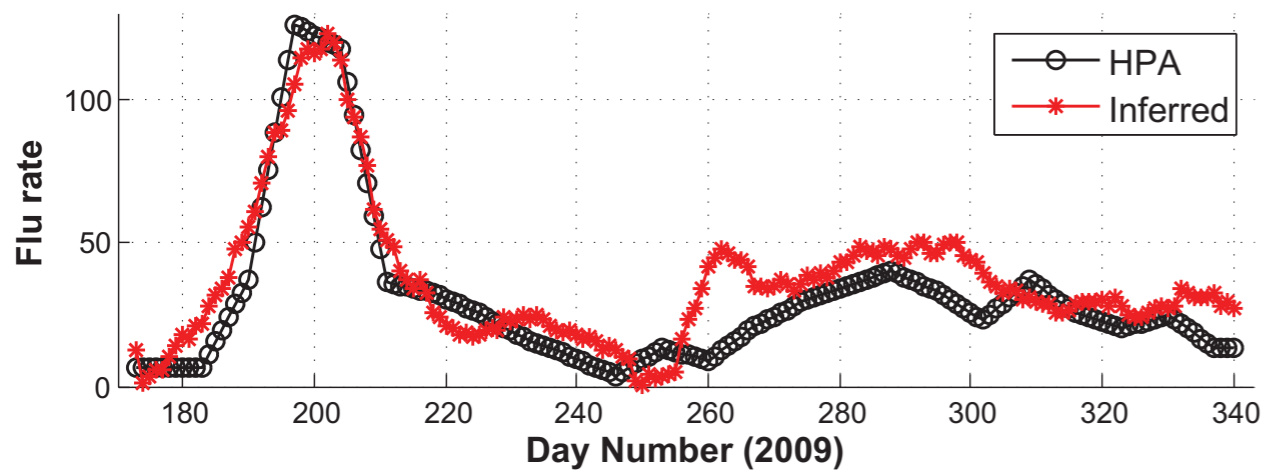
Lasso regression on the extended set of keywords seems to improve average performance ($r = 0.9256$)

Train/Validate	A	B	C	D	E
A	-	0.9594	0.9375	0.9348	0.9297
B	0.9455	-	0.9476	0.9267	0.9003
C	0.9154	0.9513	-	0.8188	0.908
D	0.9463	0.9459	0.9424	-	0.9337
E	0.8798	0.9506	0.9455	0.8935	-
				Total Avg.	0.9256

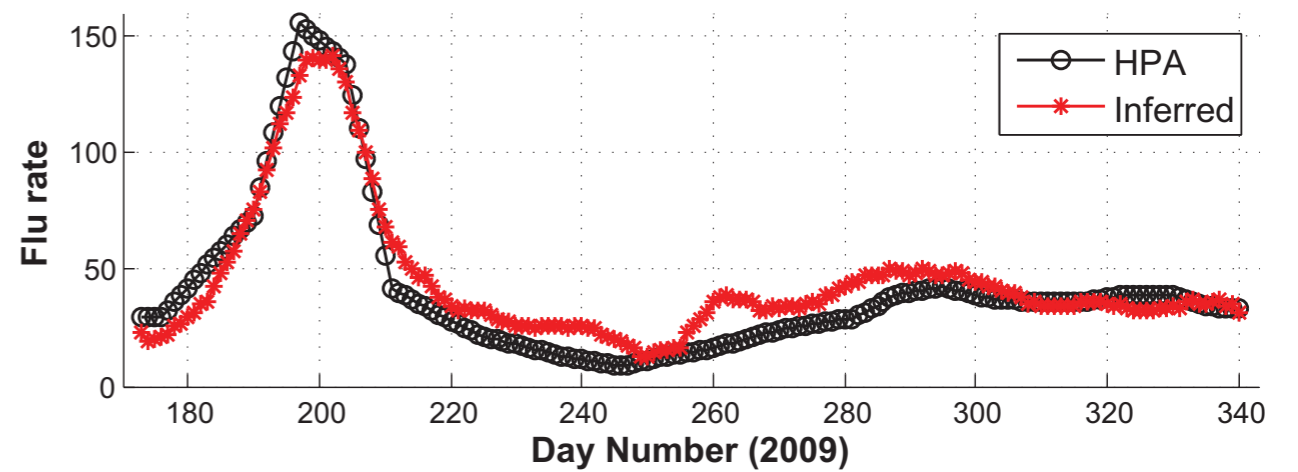
Train on one region, validate λ (*regularisation parameter*) on another, test performance on the remaining regions

Twitter 'Flu Trends': *the results* (2)

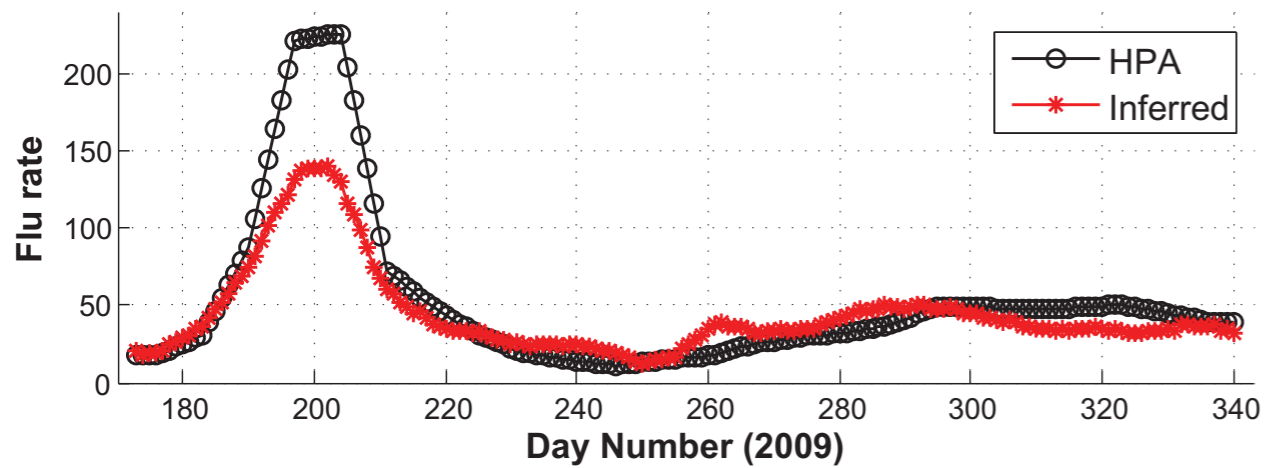
Fit looks better than in the previous modelling attempt



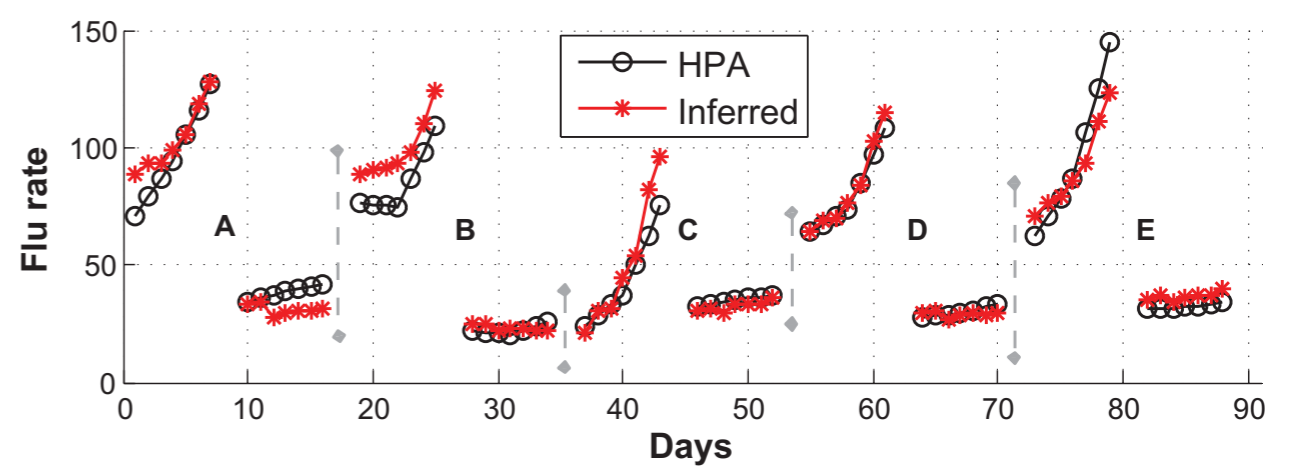
(a) Region C - Correlation: 93.49% (p-value: 1.39e-76)



(b) Region D - Correlation: 96.77% (p-value: 2.98e-101)



(c) Region E - Correlation: 97.55% (p-value: 3.85e-111)



(d) Inference on the aggregated data set for weeks 28 and 41 - Correlation: 97.13% (p-value: 3.96e-44)

Twitter 'Flu Trends': *the results* (3)

keywords that have been selected (non zero weights)
by lasso regression

lung	unwel	temperatur	like	headach	season
unusu	chronic	child	dai	appetit	stai
symptom	spread	diarrhoea	start	muscl	weaken
immun	feel	liver	plenti	antivir	follow
sore	peopl	nation	small	pandem	pregnant
thermomet	bed	loss	heart	mention	condit
high	group	tired	import	risk	carefulli
work	short	stage	page	diseas	recognis
servic	wors	case	similar	term	home
increas	exist	ill	sens	counter	better
cough	vomit	earli	neurolog	catch	onlin
fever	concern	check	drink	long	far
consid	ach	breath	flu	member	kidnei
mild	number	sick	throat	famili	water
read	includ	swine	confirm	need	nose
medic	phone	cancer	disord	unsur	suddenli
runni					

6. *Further regression basics: Elastic net and Gaussian Processes in a nutshell*

Regularised regression: the *elastic net*

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

$$\operatorname{argmin}_{\mathbf{w}_*} \left\{ \underbrace{\|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2}_{\text{OLS}} + \underbrace{\lambda_1 \|\mathbf{w}\|_{\ell_2}^2}_{\text{RR reg.}} + \underbrace{\lambda_2 \|\mathbf{w}\|_{\ell_1}}_{\text{Lasso reg.}} \right\}$$

elastic net **combines** L2-norm (ridge)
and L1-norm (lasso) regularisation

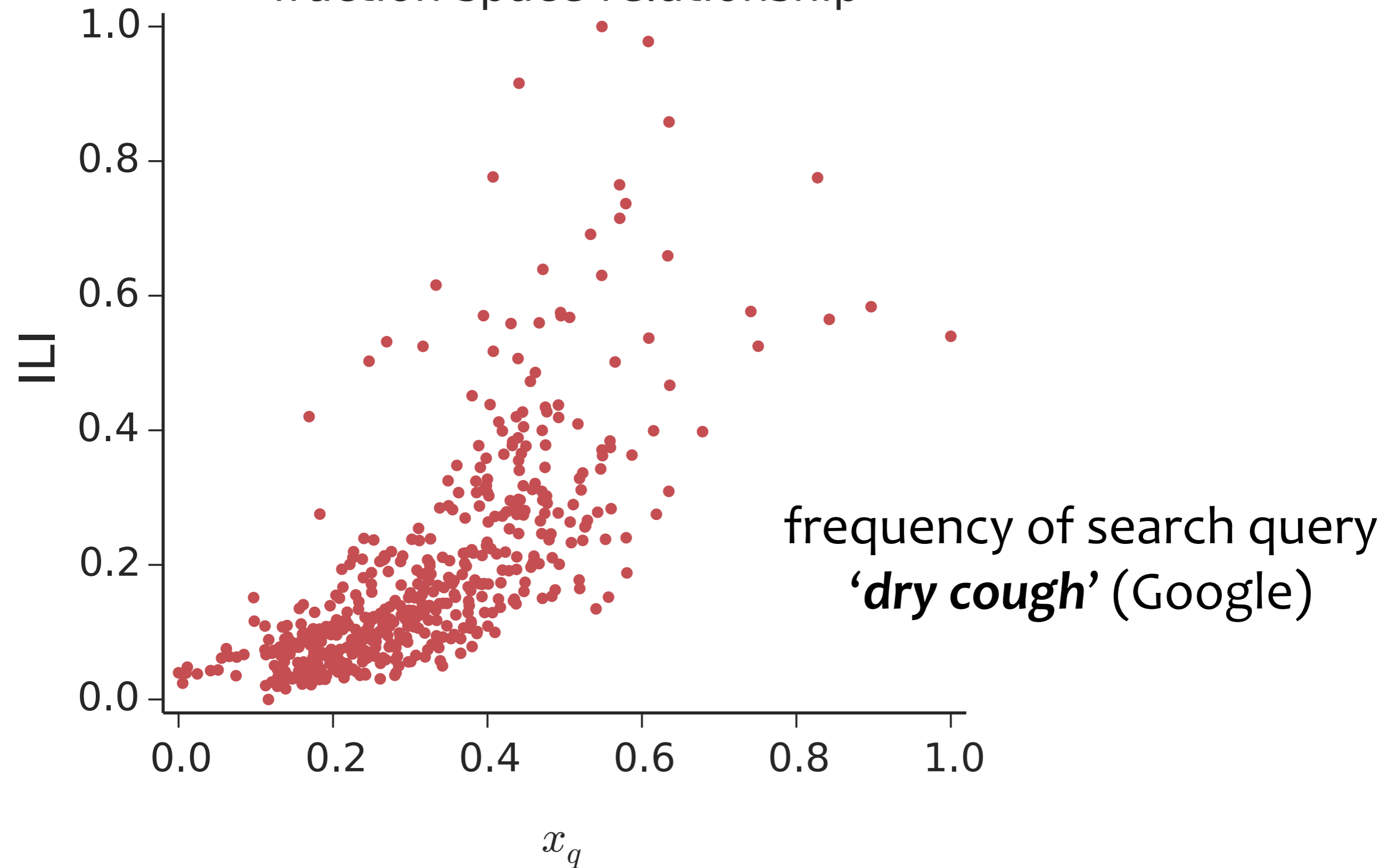
Pros and Cons of elastic net

$$\operatorname{argmin}_{\mathbf{w}_*} \left\{ \underbrace{\|\mathbf{X}_* \mathbf{w}_* - \mathbf{y}\|_{\ell_2}^2}_{\text{OLS}} + \underbrace{\lambda_1 \|\mathbf{w}\|_{\ell_2}^2}_{\text{RR reg.}} + \underbrace{\lambda_2 \|\mathbf{w}\|_{\ell_1}}_{\text{Lasso reg.}} \right\}$$

- + ‘compromise’ between ridge regression (**handles collinear predictors**) and lasso (**favours sparsity**)
- + entire regularisation path can be explored by modifying **LARS** algorithm
- + if $m > n$, # of selected variables is **not limited to n**
- it may select redundant variables
- has two regularisation parameters to validate (although there are ways to mitigate this, e.g. by setting $\lambda_1 = \alpha \lambda_2$)

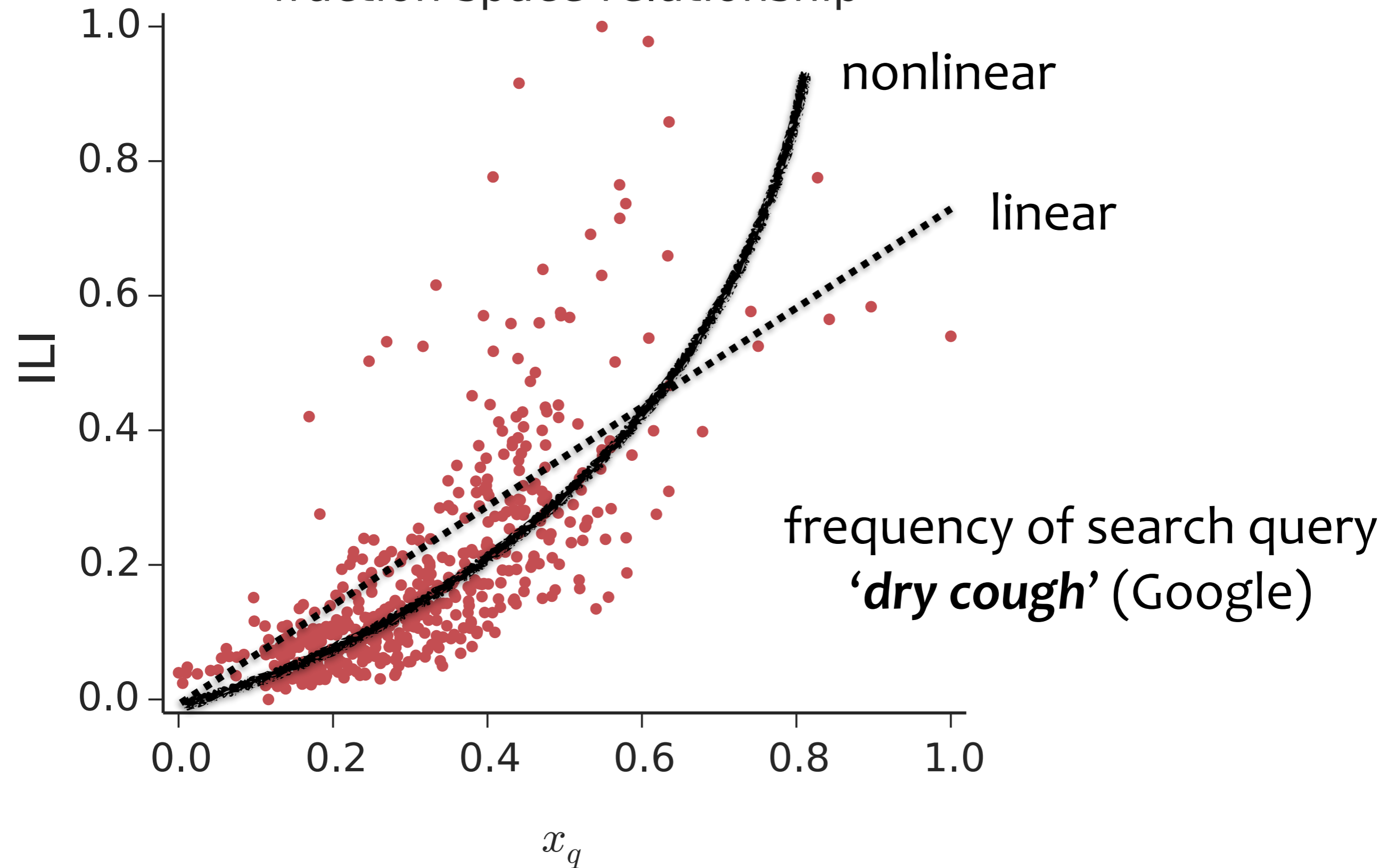
Nonlinearities in the data (1)

fraction space relationship



Nonlinearities in the data (2)

fraction space relationship



Gaussian Processes (GPs)

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

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

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

mean function
drawn on inputs

covariance function (or kernel)
drawn on pairs of inputs

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

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Gaussian Processes (GPs)

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

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

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

mean function
drawn on inputs

covariance function (or kernel)
drawn on pairs of inputs

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

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Common covariance functions (kernels)

Kernel name:

Squared-exp (SE)

Periodic (Per)

Linear (Lin)

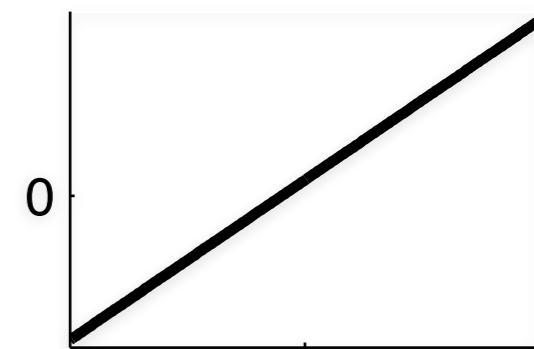
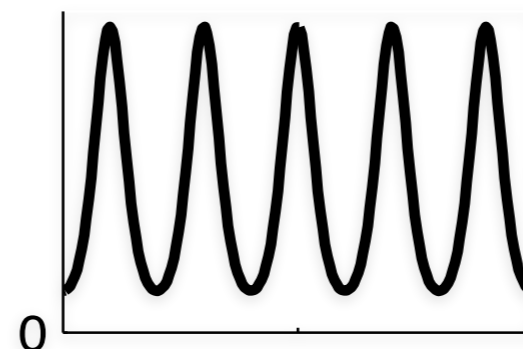
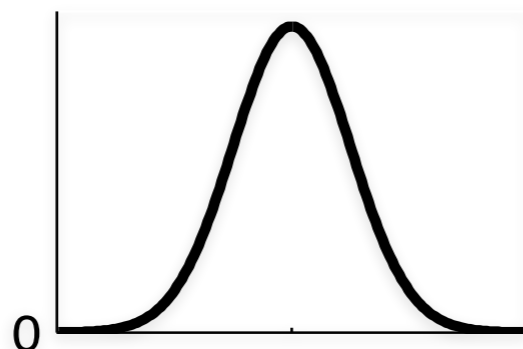
$$k(x, x') =$$

$$\sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

$$\sigma_f^2 \exp\left(-\frac{2}{\ell^2} \sin^2\left(\pi \frac{x-x'}{p}\right)\right)$$

$$\sigma_f^2 (x-c)(x'-c)$$

Plot of $k(x, x')$:

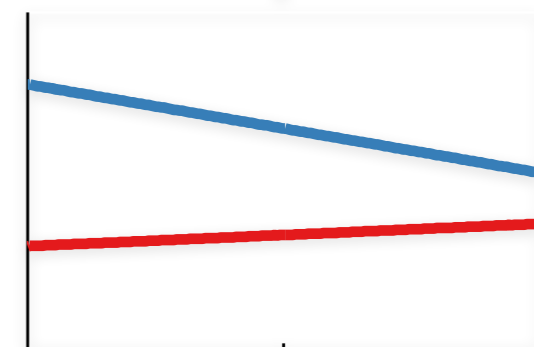
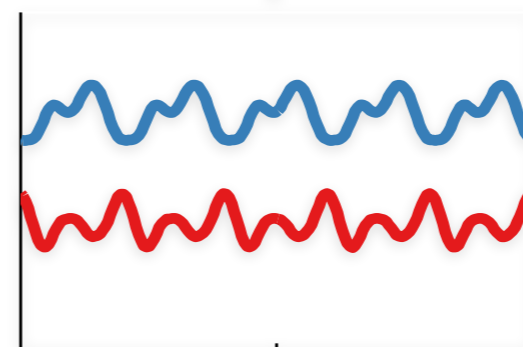
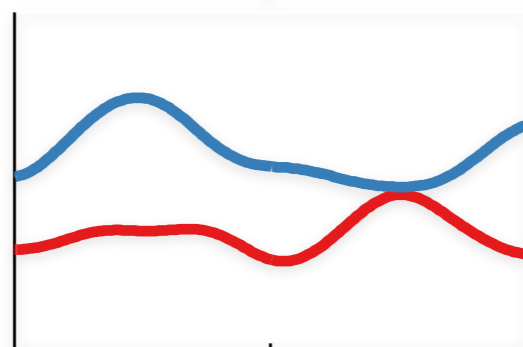


$x - x'$

$x - x'$

x (with $x' = 1$)

Functions $f(x)$
sampled from
GP prior:



x

x

x

Type of structure:

local variation

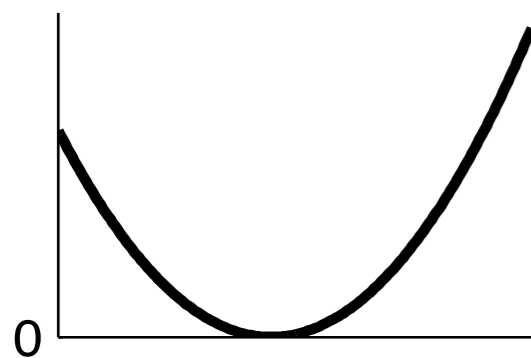
repeating structure

linear functions

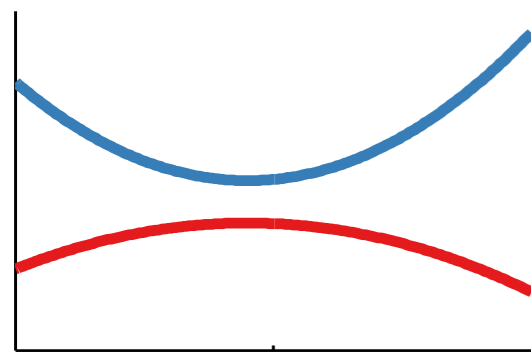
Combining kernels in a GP

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

Lin \times Lin

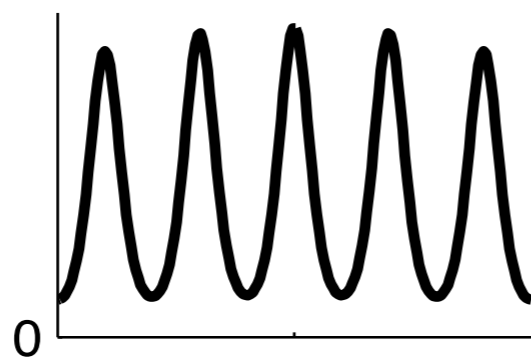


x (with $x' = 1$)

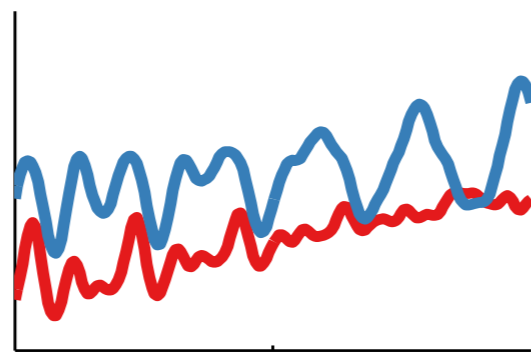


quadratic functions

SE \times Per

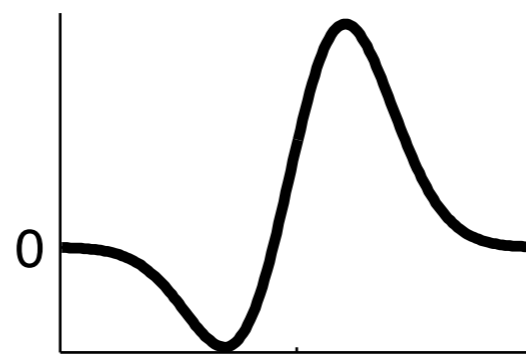


$x - x'$

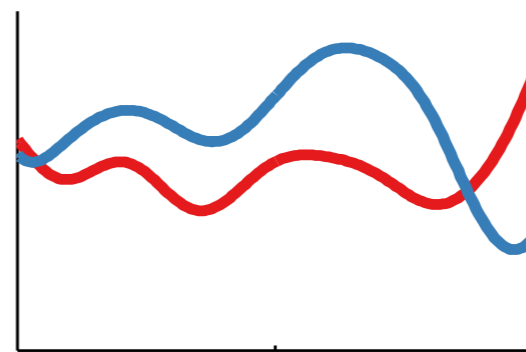


locally periodic

Lin \times SE

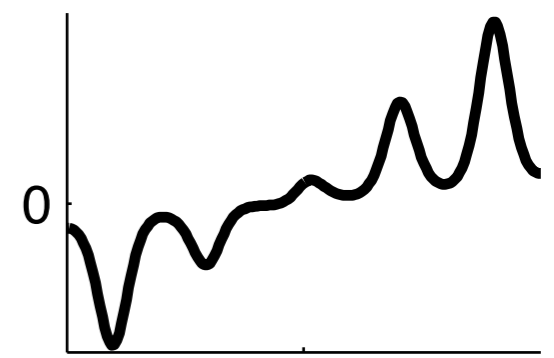


x (with $x' = 1$)

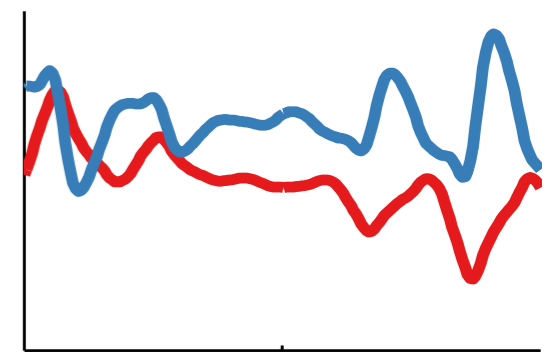


increasing variation

Lin \times Per



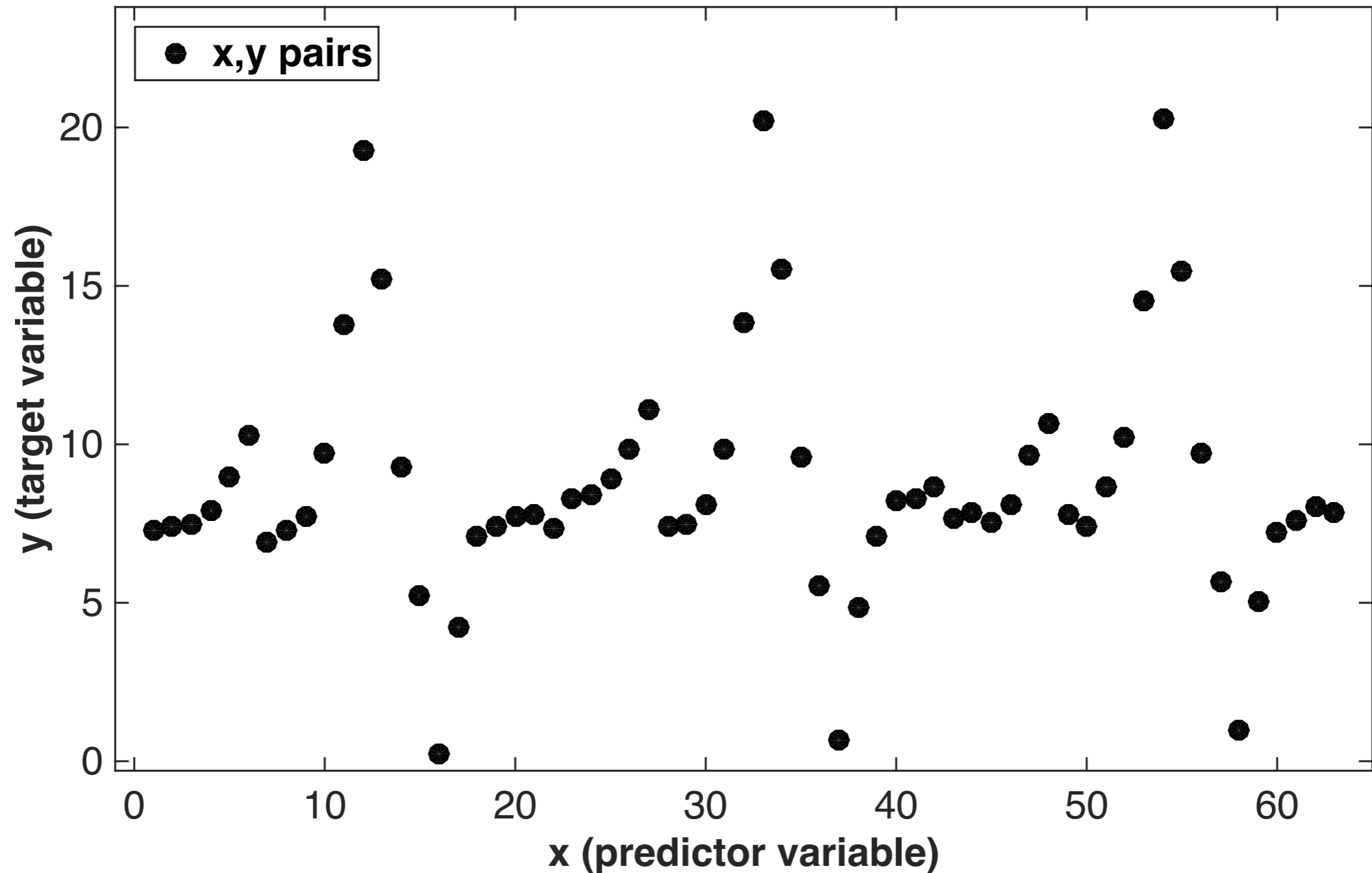
x (with $x' = 1$)



growing amplitude

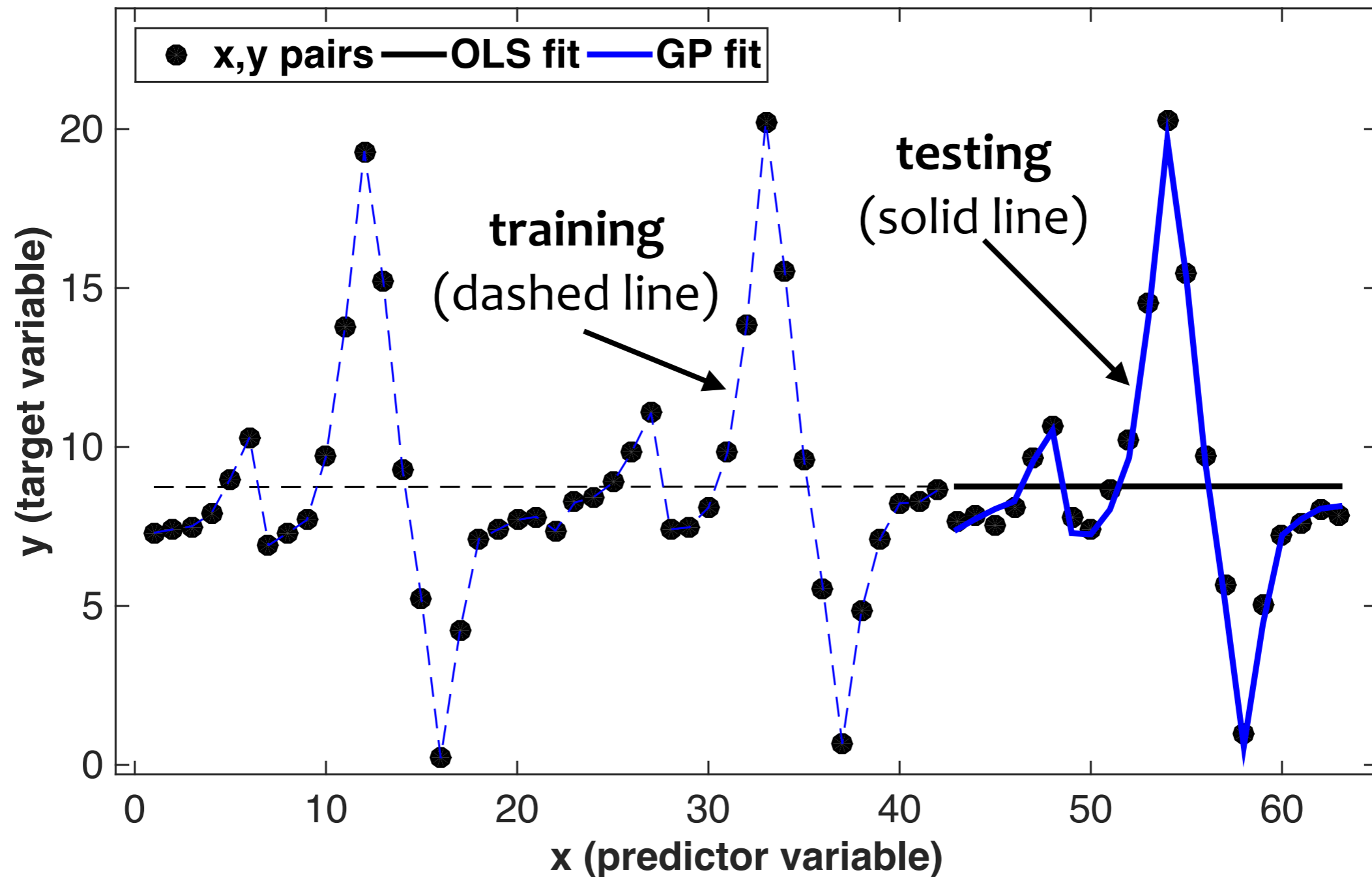
GPs for regression: An example (1)

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



GPs for regression: An example (2)

Addition of 2 GP kernels:
periodic + squared exponential



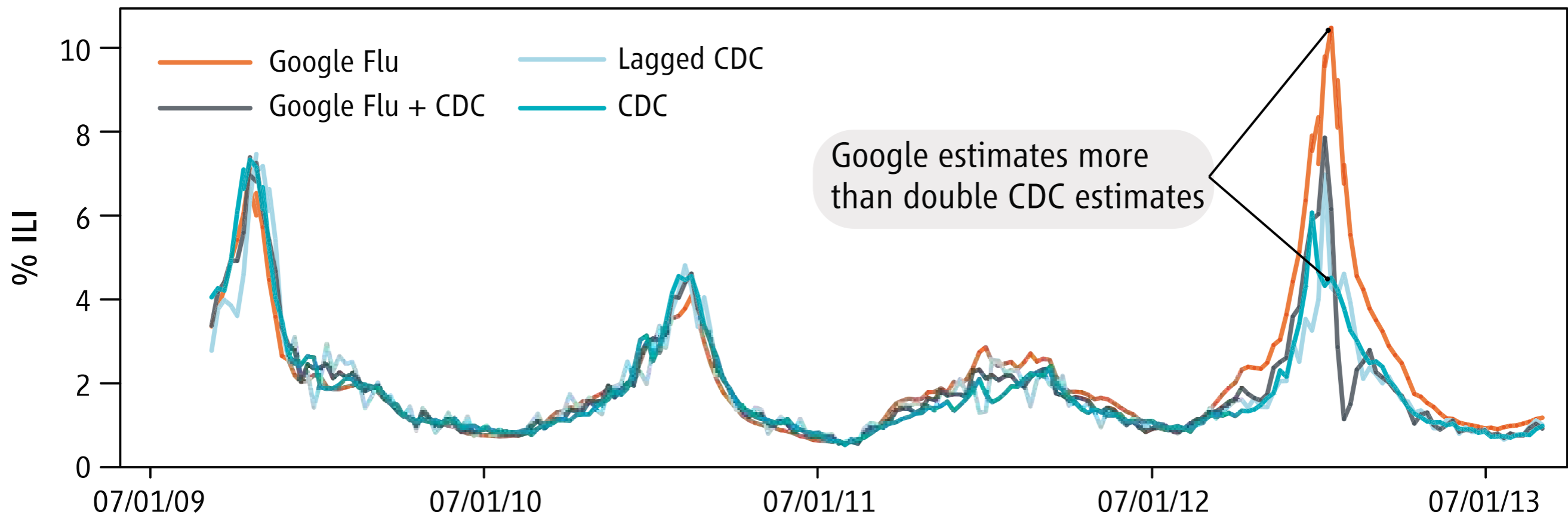
More information about GPs

- + 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://videlectures.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/>

7. *Improving the Google Flu Trends modelling approach*

(Lampos, Miller, Crossan and Stefansen, 2015)

Failures of the previous modelling



The estimates of the online Google Flu Trends tool were approx. two times larger than the ones from CDC

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 will assume that they are generally a good representation of the real signal)

Google Flu Trends revised: *the data* (1)

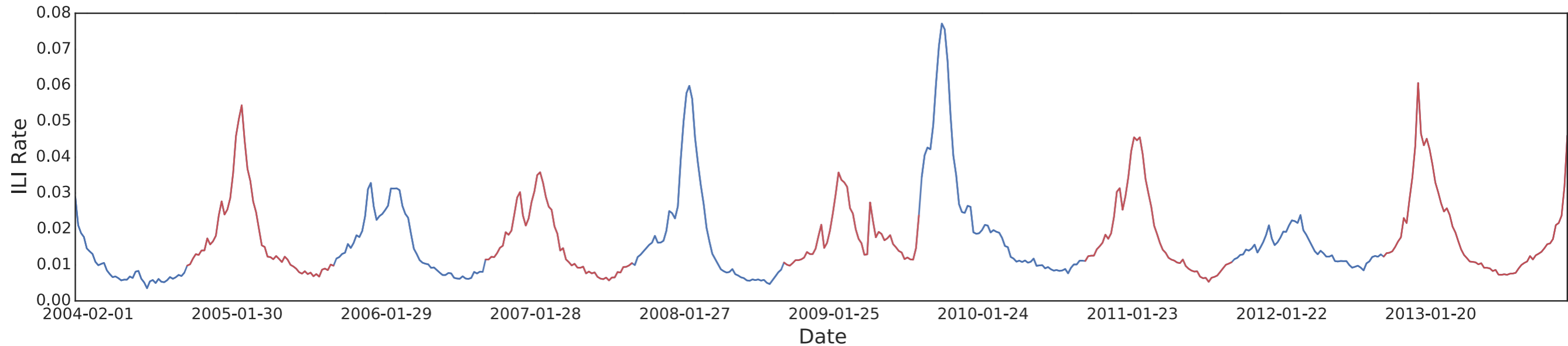
Google search query logs

- > 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

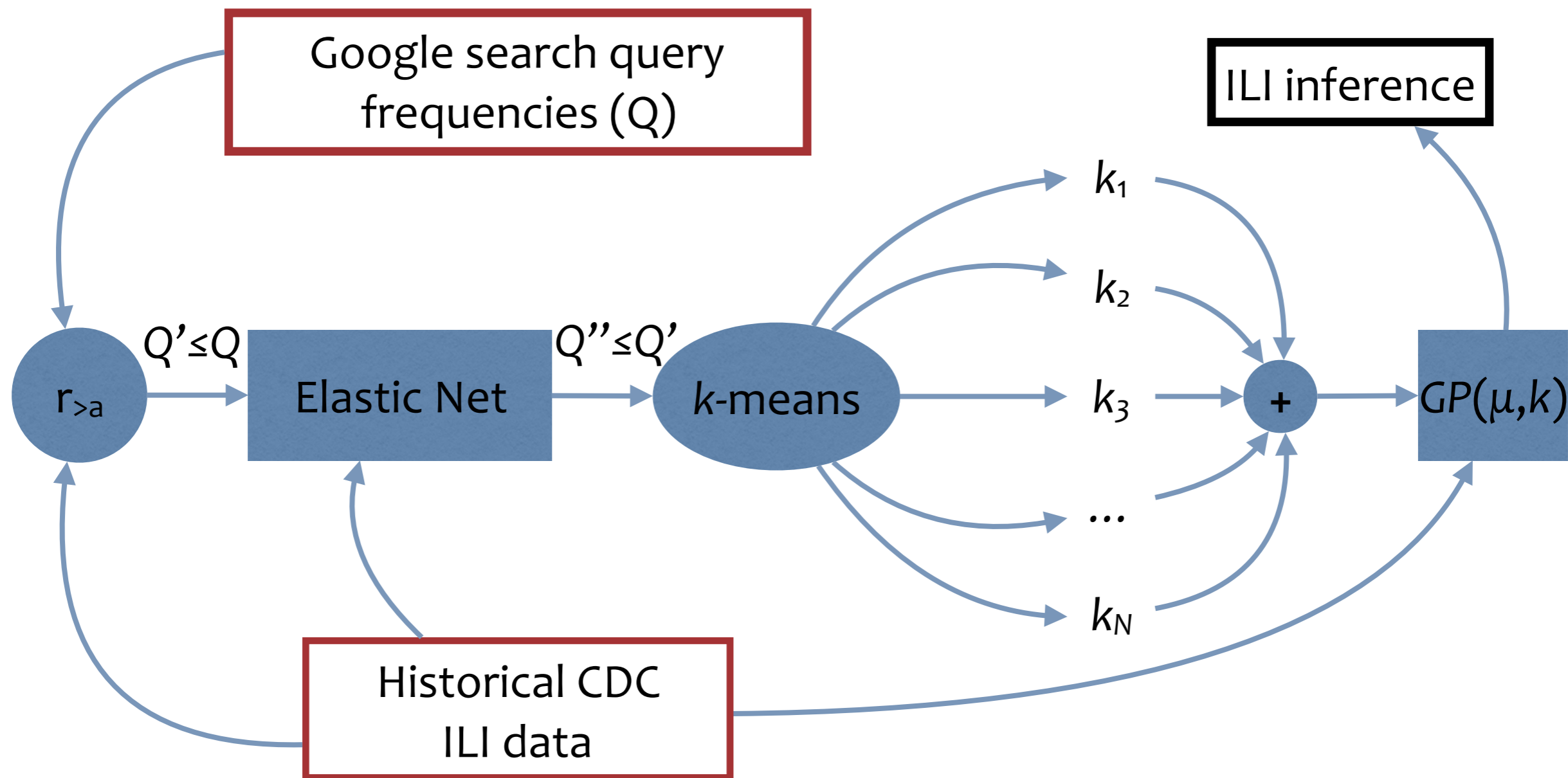
Google Flu Trends revised: *the data* (2)

Corresponding ILI rates from the CDC



different colouring per flu season

Google Flu Trends revised: *the methods* (1)



Google Flu Trends revised: *the 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

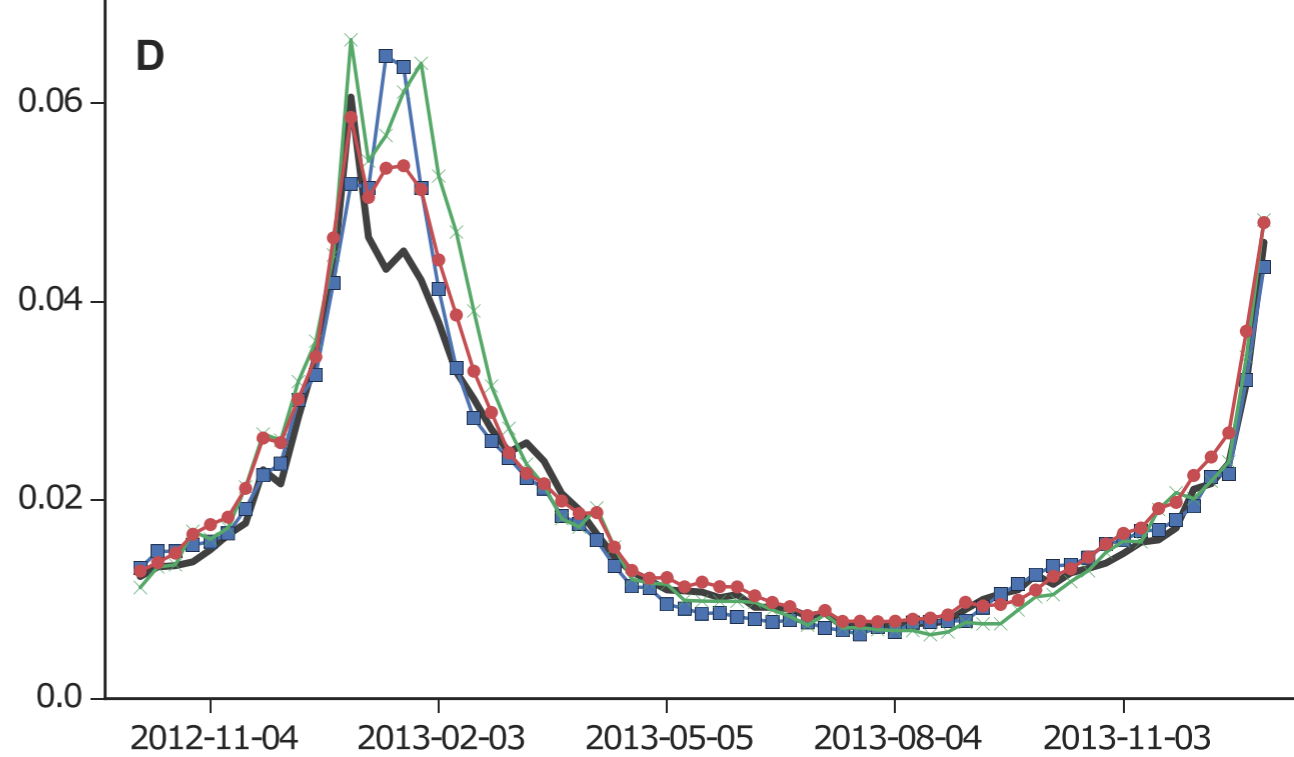
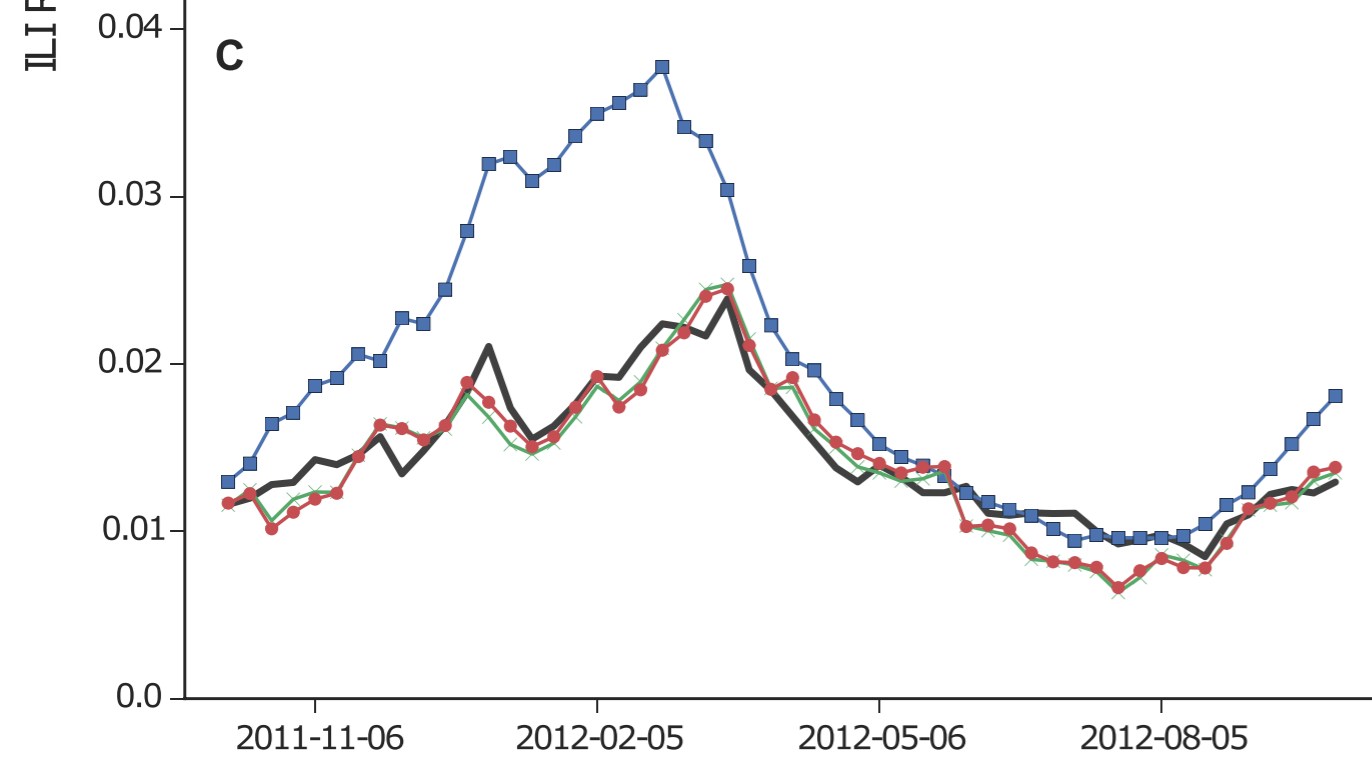
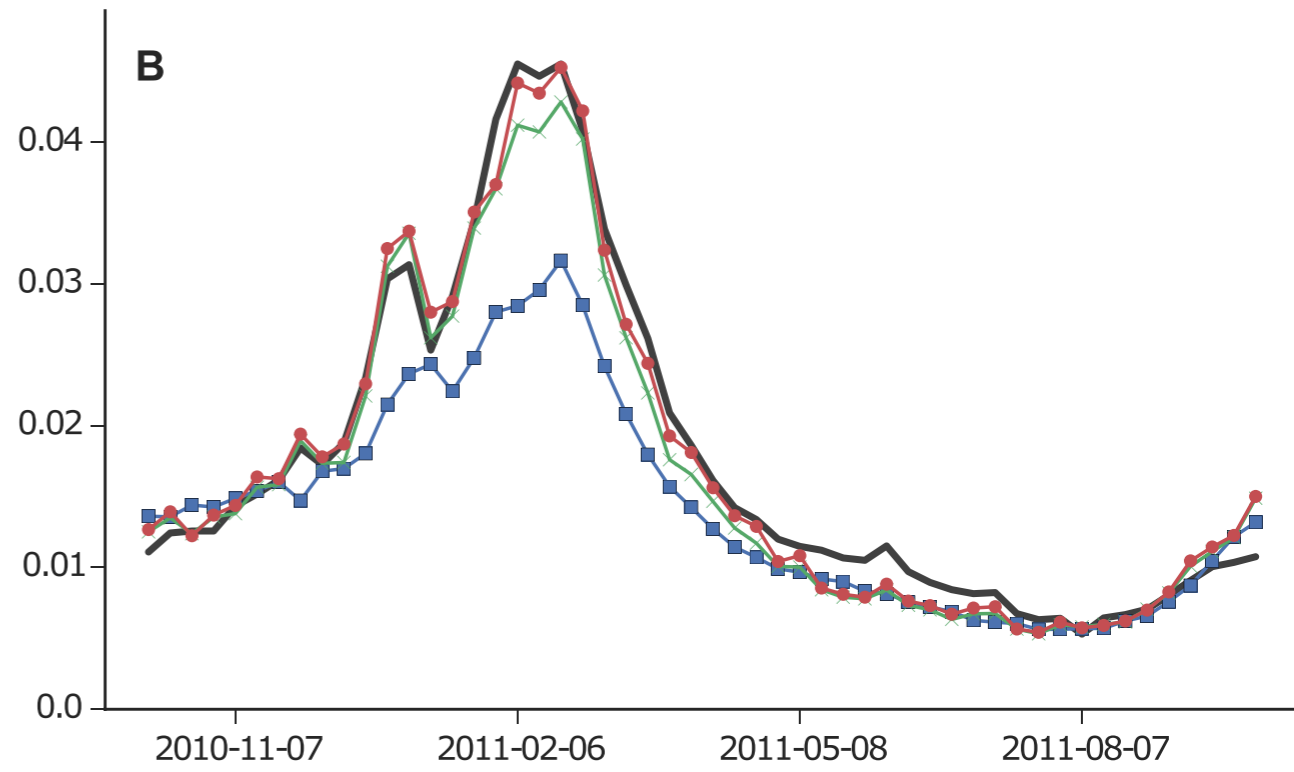
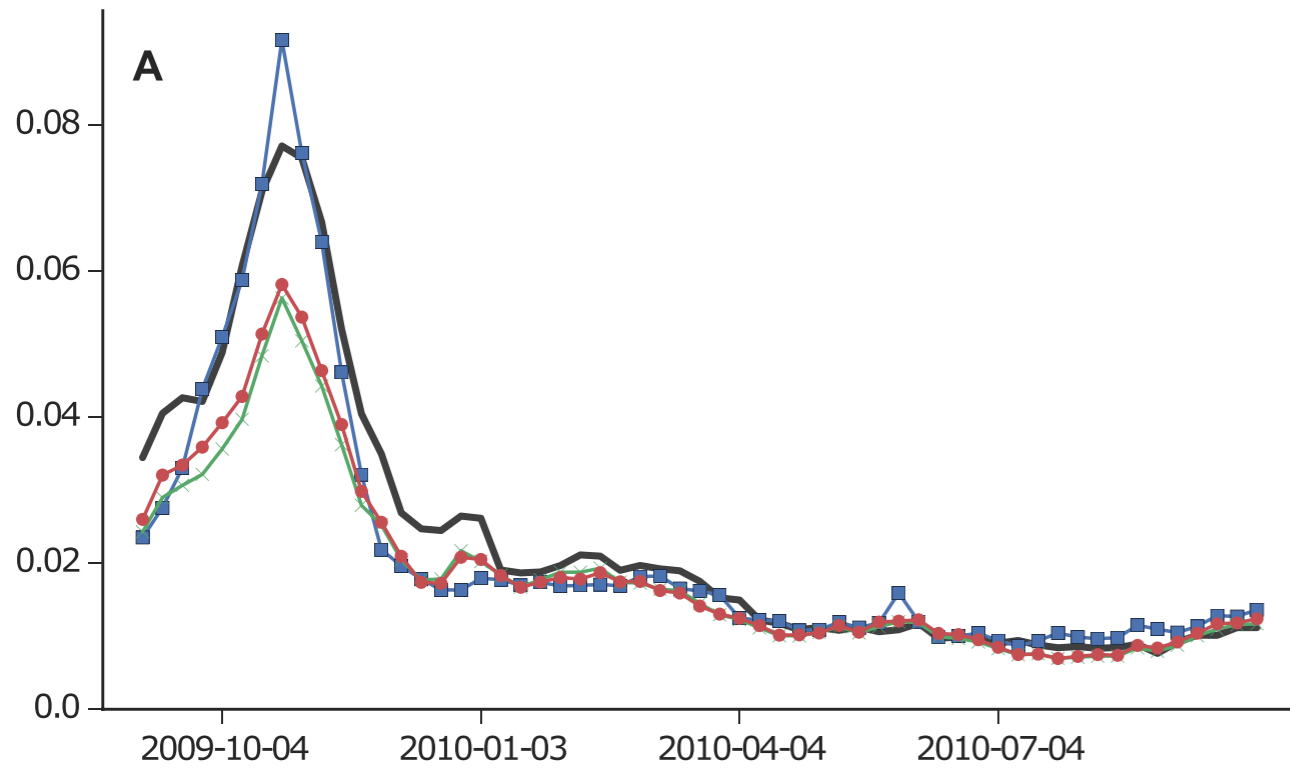
Google Flu Trends revised: *the methods* (3)

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

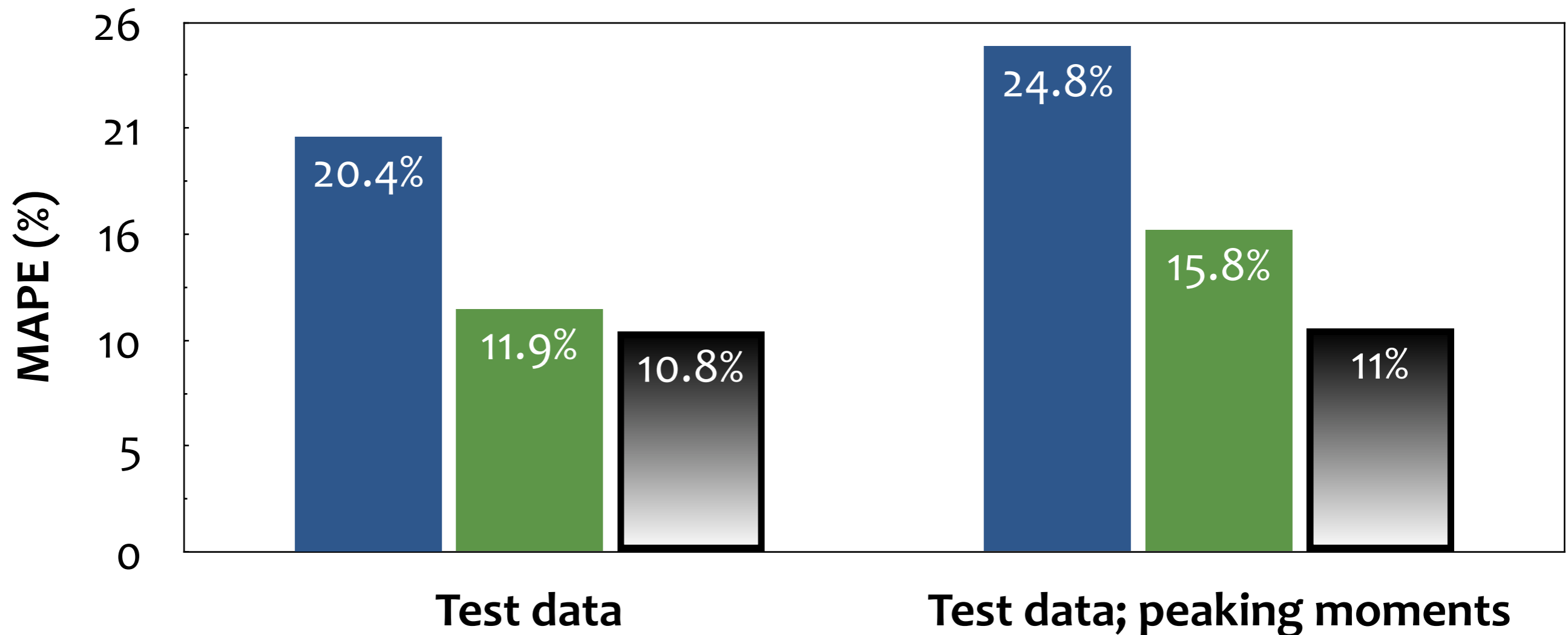
- + **protect a model from radical changes** in the frequency of single queries that are not representative of a cluster
- + model the **contribution of various thematic concepts** (captured by different clusters) to the final prediction
- + learning a sum of lower-dimensional functions: significantly 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: *the results* (1)

— CDC —■— GFT —×— Elastic Net —●— GP



Google Flu Trends revised: *the results* (2)



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

Google Flu Trends revised: *the 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: *the results* (4)

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

elastic net

‘ear thermometer’ — 3%

‘musinex’ — 2%

‘how to break a fever’ — 2%

‘flu like symptoms’ — 2%

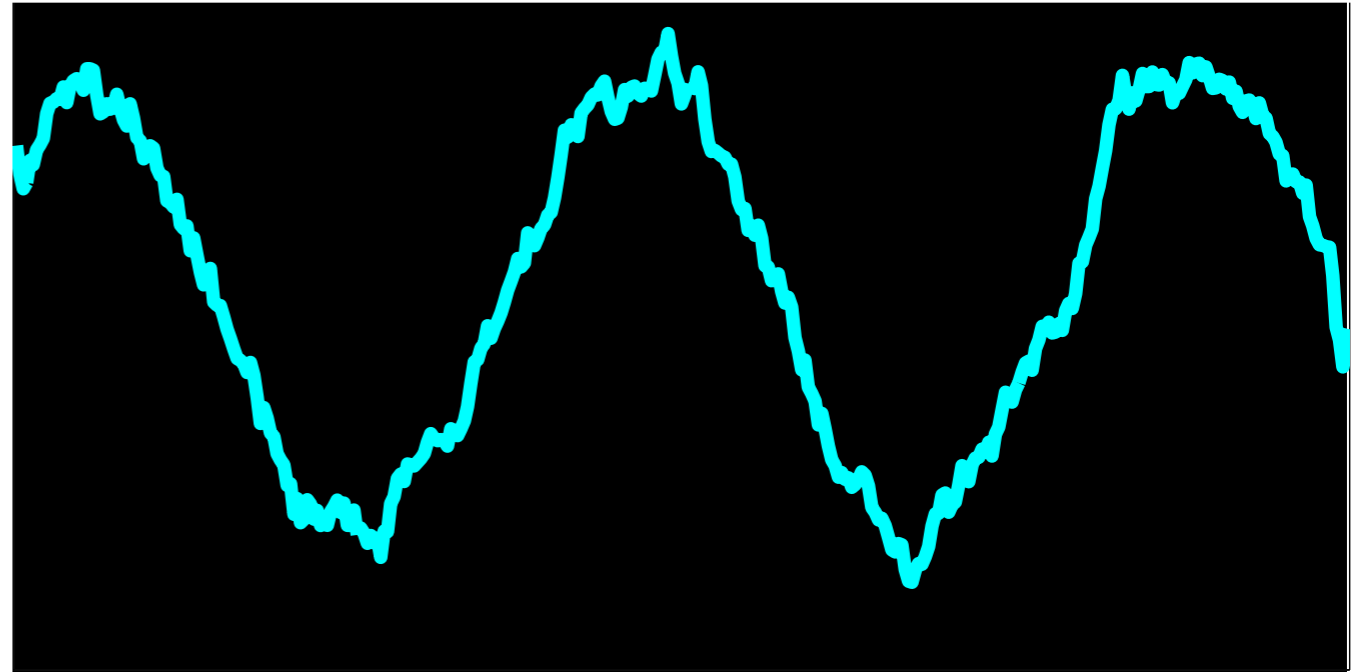
‘fever reducer’ — 2%

8. *Assessing the impact of a health intervention using Internet data*

(Lampos, Yom-Tov, Pebody and Cox, 2015)

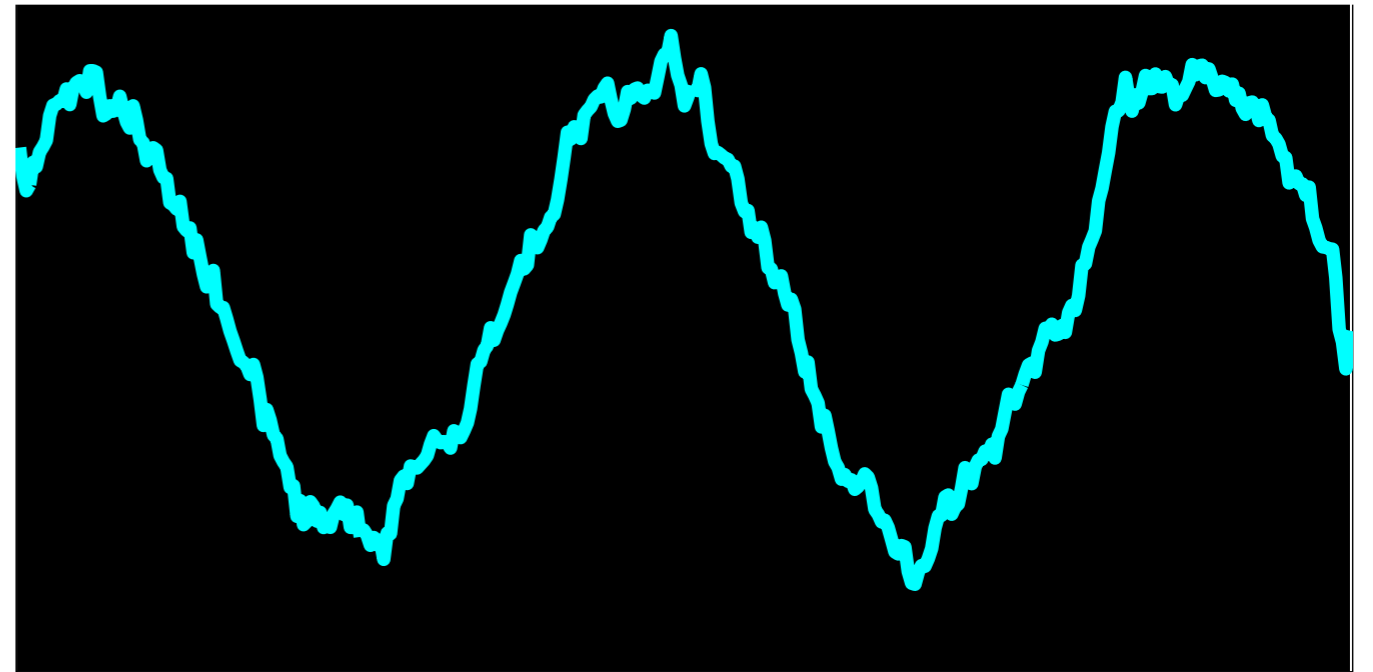
Intervention impact: *the idea*

disease rates in the population



Intervention impact: *the idea*

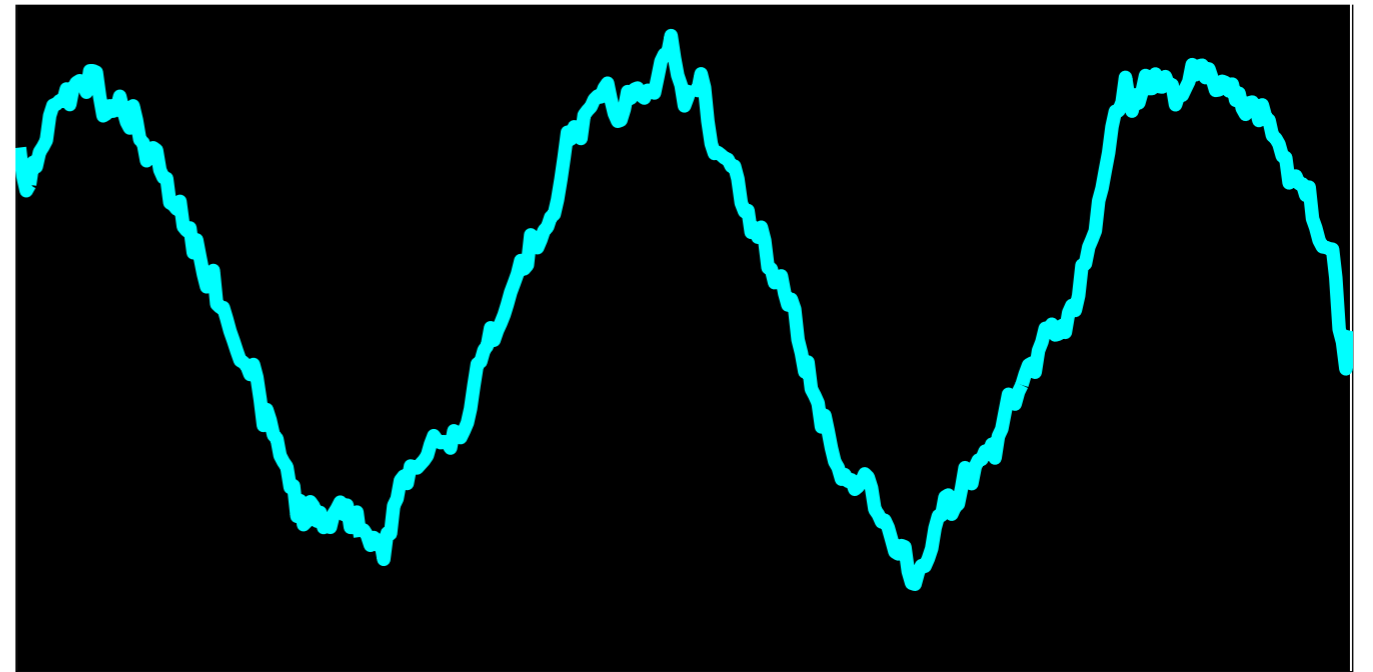
disease rates in the population



Health intervention
e.g. a vaccination campaign

Intervention impact: *the idea*

disease rates in the population



impact ?

Health intervention
e.g. a vaccination campaign

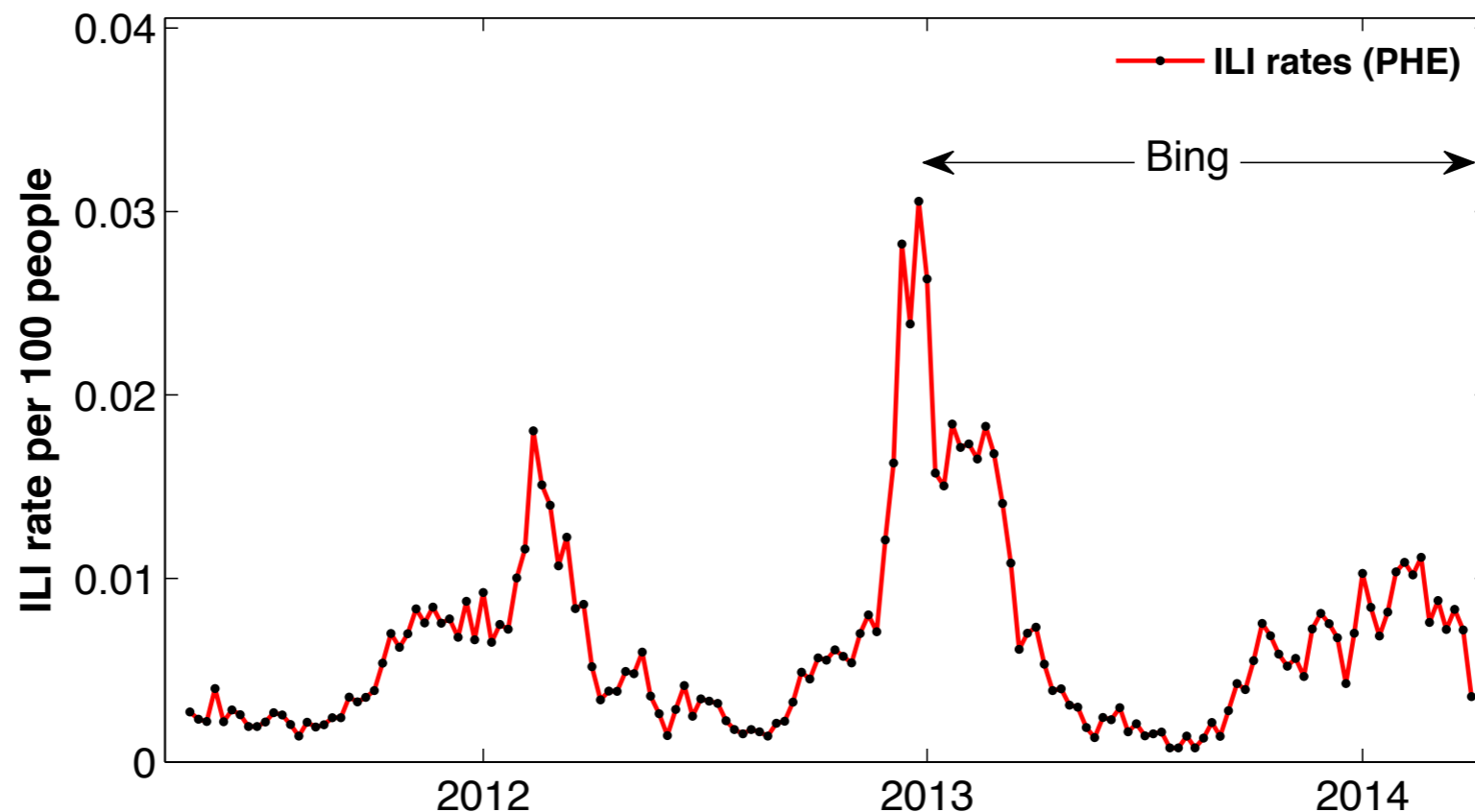
Intervention impact: *the data*

308 million **tweets** exactly geolocated in England
2 May 2011 to 13 Apr. 2014 (154 weeks)

Query frequencies from **Bing**, geolocated in England
31 Dec. 2012 to 13 Apr. 2014 (67 weeks)

generally larger numbers from the Twitter data

ILI rates for England obtained from Public Health England



Intervention impact: *the methods* (1)

Feature extraction was performed as follows:

- + Start with a manually crafted seed list of **36 textual markers**, e.g. *flu, headache, doctor, cough*
- + Extract frequent co-occurring n-grams from a corpus of 30 million UK tweets (February & March, 2014) after removing stop-words
- + Set of markers expanded to **205 n-grams** ($n \leq 4$)
e.g. *#flu, #cough, annoying cough, worst sore throat*
- + Relatively small set of features motivated by previous work

Intervention impact: *the methods* (2)

the produced
n-grams (*features*)

bolded *n*-grams
denote the seed
terms

1-grams: #chills, #cough, #disease, #dizzy, #doctor, #fatigue, #fever, #flu, #gp, #headache, #illness, #infected, #infection, #medicine, #nausea, #shiver, #shivering, #sneeze, #unwell, #vomit, **chills, cough, coughed, coughing, diarrhoea, disease, dizzy, doctor, fatigue, fatigued, fever, flu, gp,** hay-fever, **headache, illness, infected, infection, influenza,** man-flu, **medicine, nausea, shiver, shivering, sneeze, sneezed, sneezing, thermometer, tonsil, tonsils, unwell, vomit, vomited, vomiting**

2-grams: annoying cough, awful headache, bad cough, bad headache, banging headache, bed flu, bed headache, biggest headache, **blocked nose, body ache, body aches,** chest infection, chesty cough, cold cough, cold flu, constant headache, cough cough, cough fuck, cough medicine, cough sneeze, cough sore, cough syrup, cough worse, coughing blood, coughing guts, coughing lungs, coughing sneezing, day doctor, day headache, disease nation, doctor cure, doctor experience, doctor today, doctor told, dying flu, ear infection, eye infection, feel dizzy, feel sick, feel unwell, feeling dizzy, feeling sick, feeling unwell, fever pitch, flu feel, flu jab, flu tablets, fucking headache, gonna vomit, good doctor, hate flu, hate unwell, hay fever, headache coming, headache days, headache feel, headache feeling, headache fuck, headache good, headache hell, headache hours, headache morning, headache night, headache sleep, headache sore, headache time, headache today, headache work, headache worse, heart disease, horrible disease, horrible headache, infected restless, kidney infection, killer headache, love doctor, love sneezing, major headache, man flu, massive headache, mental illness, **muscles ache,** new doctor, night coughing, night fever, people cough, pounding headache, rare disease, rid headache, runny nose, shiver spine, sick dizzy, sick headache, sleep coughing, sneeze sneeze, sneezing fit, sore throat, splitting headache, start fever, stomach ache, stuffy nose, stupid cough, swine flu, taste medicine, terminal illness, throat cough, throat headache, throat infection, tickly cough, tired headache, viral infection, waiting doctor, waking headache, wanna vomit, watch doctor, watching doctor, wine headache, woke headache, worst cough, worst headache

3-grams: blocked nose sore, cold flu tablets, cold sore throat, cough cough cough, day feel sick, eat feel sick, feel sick eating, feel sick feel, feel sick stomach, feel sore throat, food feel sick, hate feeling sick, headache feel sick, headache sore throat, hungry feel sick, literally feel sick, nose sore throat, risk heart disease, sleep feel sick, sore throat blocked, sore throat coming, sore throat cough, throat blocked nose, tired feel sick, today feel sick, woke sore throat, worlds worst headache, worst sore throat, worst stomach ache

4-grams: blocked nose sore throat, cough cough cough cough

Intervention impact: *the methods* (3)

First, we come up with an ILI model using (*and comparing*):

1. Ridge regression
2. Elastic Net
3. **A GP model**

using a kernel per n -gram category

Main kernel function

$$k(\mathbf{x}, \mathbf{x}') = \left(\sum_{n=1}^C k_{\text{RQ}}(\mathbf{g}_n, \mathbf{g}'_n) \right) + k_{\text{N}}(\mathbf{x}, \mathbf{x}')$$

Rational Quadratic kernel
(infinite sum of squared
exponential kernels)

$$k_{\text{RQ}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\alpha\ell^2} \right)^{-\alpha}$$

Intervention impact: *the methods* (4)

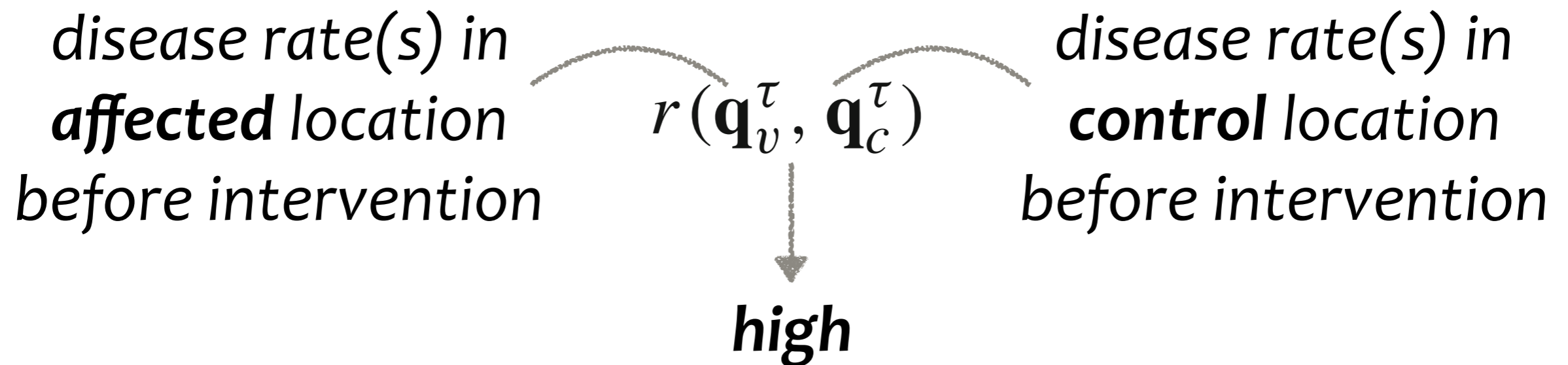
1. Disease intervention launched (to a set of areas)
2. Define a distinct set of control areas
3. Estimate disease rates in all areas
4. Identify pairs of areas with strong historical correlation in their disease rates
5. Use this relationship during and slightly after the intervention to infer diseases rates in the affected areas had the intervention not taken place

Intervention impact: *the methods* (5)

$\tau = \{t_1, \dots, t_N\}$ time interval(s) before the intervention

v location(s) where the intervention took place

c control location(s)



$$f(w, \beta) : \mathbb{R} \rightarrow \mathbb{R} \text{ such that } \operatorname{argmin}_{w, \beta} \sum_{i=1}^N (q_c^{t_i} w + \beta - q_v^{t_i})^2$$

Intervention impact: *the methods* (6)

$$f(w, \beta) : \mathbb{R} \rightarrow \mathbb{R} \text{ such that } \operatorname{argmin}_{w, \beta} \sum_{i=1}^N (q_c^{t_i} w + \beta - q_v^{t_i})^2$$

estimate projected rate(s) in affected location during/after intervention $\longrightarrow \mathbf{q}_v^* = \mathbf{q}_c^* w + \mathbf{b}$

$\mathbf{q}_v \longrightarrow$ disease rate(s) in affected location during/after intervention

absolute difference

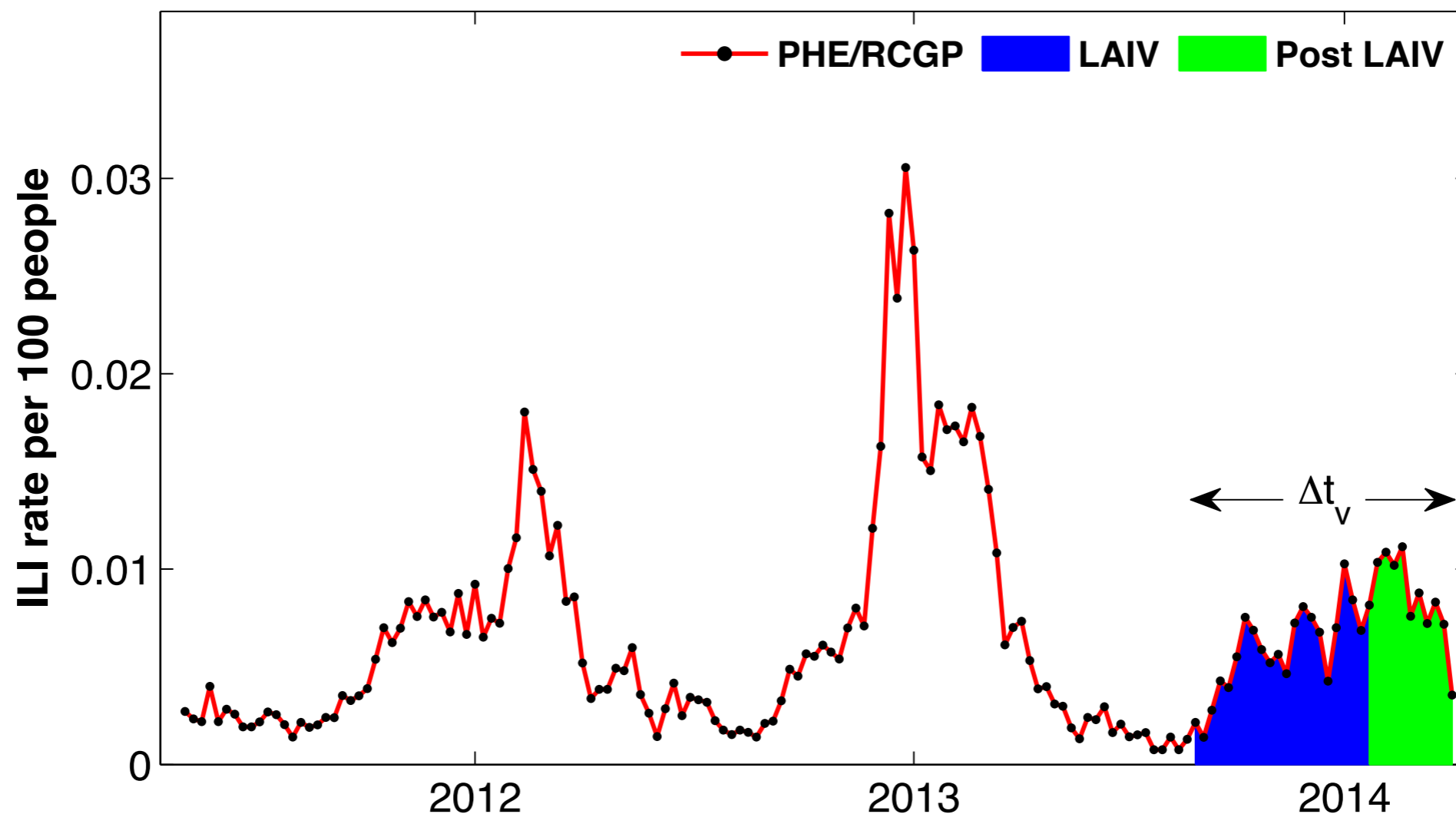
$$\delta_v = \bar{\mathbf{q}}_v - \bar{\mathbf{q}}_v^*$$

relative difference (**impact**)

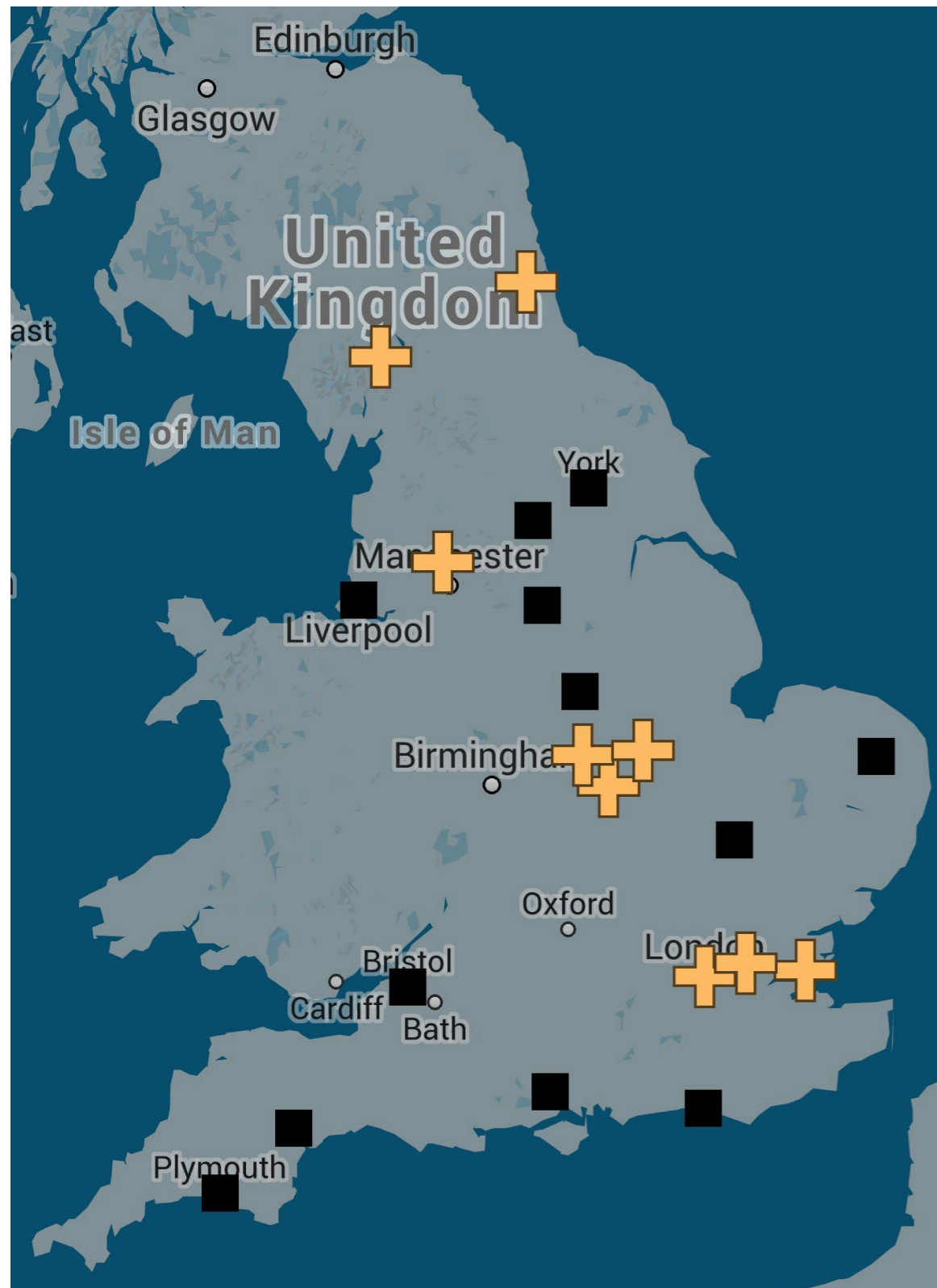
$$\theta_v = \frac{\bar{\mathbf{q}}_v - \bar{\mathbf{q}}_v^*}{\bar{\mathbf{q}}_v^*}$$

Intervention impact: *results* (1)

- > Vaccination programme for children (4 to 11 years) in pilot areas of England during the 2013/14 flu season
- > Vaccination period (**blue**): Sept. 2013 to Jan. 2014
- > Post-vaccination period (**green**): Feb. to April 2014



Intervention impact: *results* (2)



Vaccinated areas

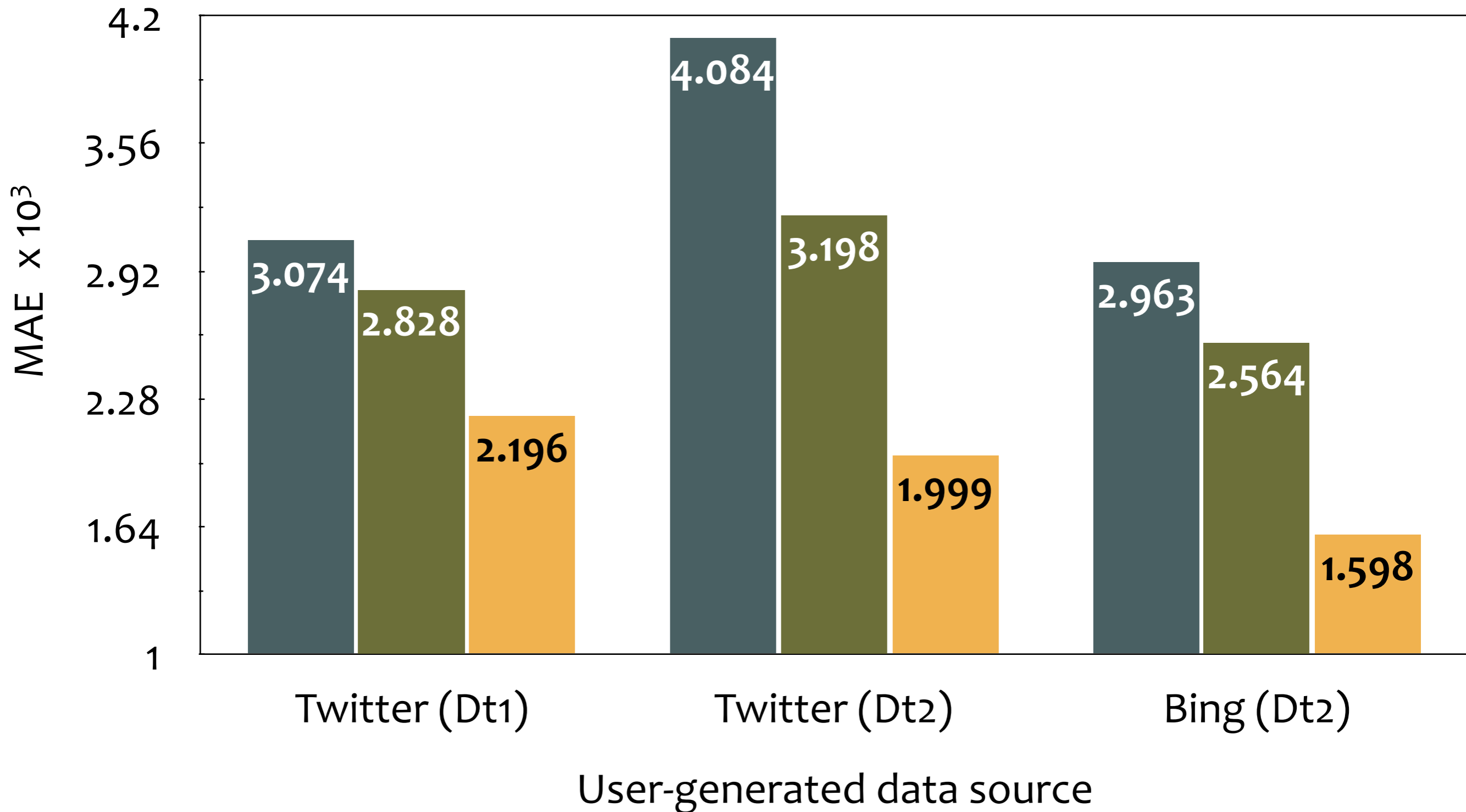
Bury • Cumbria • Gateshead
Leicester • East Leicestershire
Rutland • South-East Essex
Havering (London)
Newham (London)

Control areas

Brighton • Bristol • Cambridge
Exeter • Leeds • Liverpool
Norwich • Nottingham • Plymouth
Sheffield • Southampton • York

Intervention impact: *results* (3)

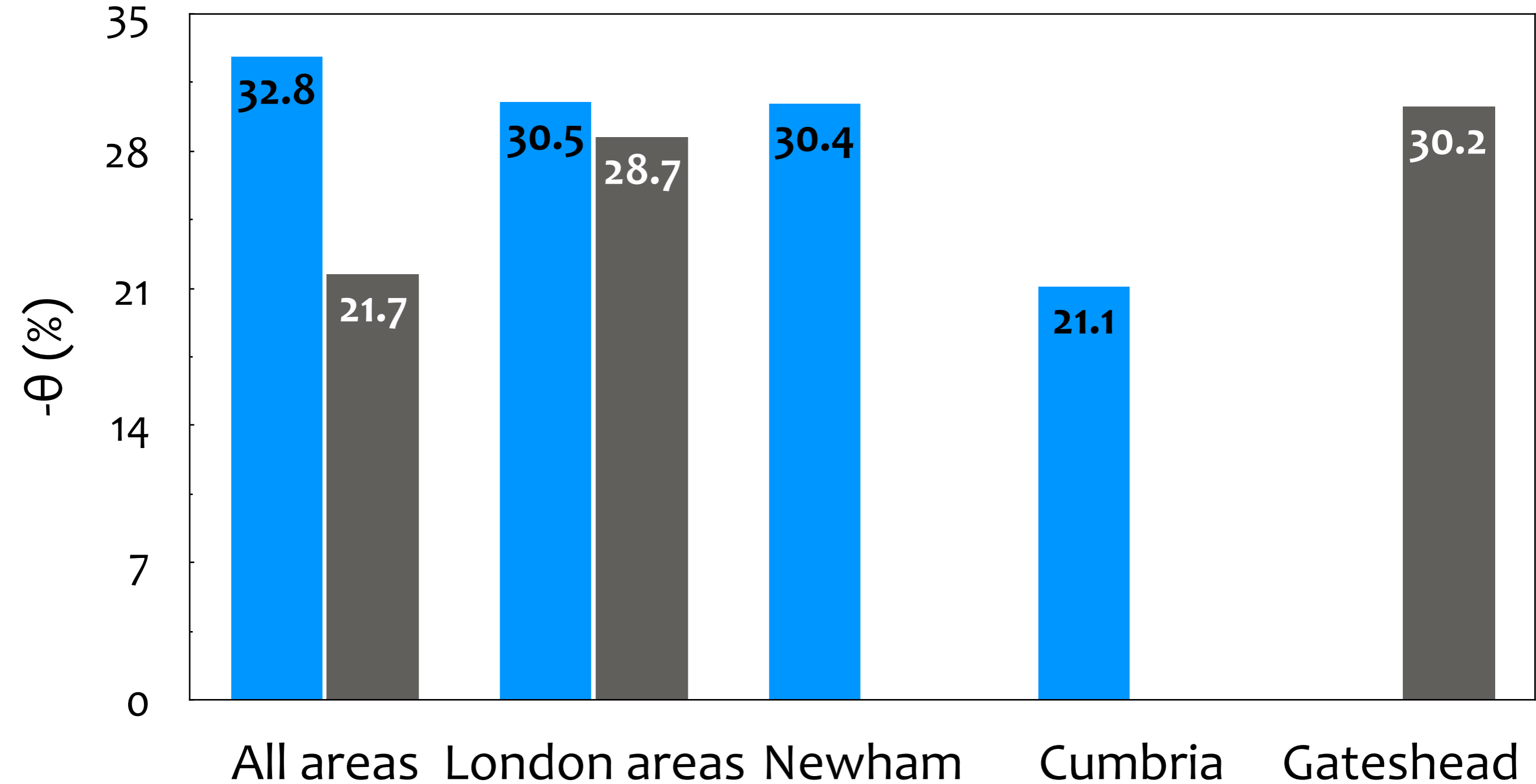
■ Ridge Regression ■ Elastic Net ■ Gaussian Process



Intervention impact: *results* (4)

Twitter

Bing



9. *Recap and concluding remarks*

What has been presented today

- a. Essentials on public health surveillance
- b. Basics on linear, regularised (ridge, lasso, elastic net) and nonlinear regression using Gaussian Processes
- c. Original Google Flu Trends model and why it failed, plus an improved approach
- d. From Twitter to flu using regularised regression
- e. A framework for assessing the impact of a health intervention using social media and search query data

A lot of things not presented today

- + There is a growing research interest on digital disease surveillance (*lots of interesting research projects and papers*); we just **scratched the surface** today!
- + Digging further into **methodological details**
 - > Machine Learning / Statistical aspects
 - > Natural Language Processing / Information Retrieval
 - > Epidemiology
- + **Negative results**

Things to take away from this lecture

- + **User-generated data** can be used to **assist** traditional health surveillance methods
- + **Useful:** (a) more information — better decisions, and (b) under-developed parts of the world may benefit
- + Techniques may not always be straightforward (or simplistic); they require **rigorous evaluation** (*although not always possible!*)
- + Key elements in this procedure are (a) the **better understanding of natural language**, and (b) the **statistical machine learning methods** that will capture and translate this *understanding* to *correct* estimates

Research opportunities

- + In our research group at UCL, we focus on user-generated content analysis
- + Themes of interest are not only health-based, e.g. applications for **inferring characteristics of social media users**, use of social media in other predictive tasks such as modelling **voting intention** etc.
- + If you are interested in this or similar research ideas and want to do a Ph.D., get in touch, funding may be available (email: v.lampos@ucl.ac.uk)

Acknowledgments

All people I collaborated with in research mentioned today

Ingemar J. Cox (*University College London & University of Copenhagen*), **Nello Cristianini** (*University of Bristol*), **Steve Crossan** (*Google*), **Andrew C. Miller** (*Harvard*), **Richard Pebody** (*Public Health England*), **Christian Stefansen** (*Google*), and **Elad Yom-Tov** (*Microsoft Research*)

Currently funded by



Thank you.

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

References

David Duvenaud. Automatic Model Construction with Gaussian Processes (Ph.D. Thesis, University of Cambridge, 2014)

Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance (AMIA Annual Symposium, 2006)

Ginsberg et al. Detecting influenza epidemics using search engine query data (Nature, 2009)

Hastie, Tibshirani & Friedman. The Elements of Statistical Learning (Springer, 2009)

Lamos & Cristianini. Tracking the flu pandemic by monitoring the Social Web (CIP, 2010)

Lamos, Miller, Crossan & Stefansen. Advances in nowcasting influenza-like illness rates using search query logs (Nature Scientific Reports, 2015)

Lamos, Yom-Tov, Pebody & Cox. Assessing the impact of health intervention via user-generated Internet content (Data Mining and Knowledge Discovery, 2015)

Lazer, Kennedy, King and Vespignani. The Parable of Google Flu: Traps in Big Data Analysis (Science, 2014)

Rasmussen and Williams. Gaussian Processes for Machine Learning (MIT Press, 2006)