EPSRC IRC in Early Warning Sensing Systems for Infectious Diseases

Assessing the impact of a health intervention via user-generated Internet data

Data Mining and Knowledge Discovery 29(5), pp. 1434–1457, 2015

Vasileios Lampos, Elad Yom-Tov,
Richard Pebody and Ingemar J. Cox

ECML PKDD 2015, Porto, Portugal

- **Background and motivation**
- Nowcasting disease rates from online text
- Estimating the impact of a health intervention
- Case study: influenza vaccination impact
- Conclusions & future work

1%

Online, user-generated data

+ Social media, blogs, search engine query logs
+ Proxy of real-world (*online+offline*) behaviour
+ Complementary information sensors to more 'traditional' crowdsourcing efforts
+ Can answer questions difficult to resolve otherwise
+ Strong predictive power

Online, user-generated data — Applications

- **Politics**
  - *voting intention* (*Lampos, Preotiuc-Pietro & Cohn, 2013*)
  - *result of an election* (*Tumasjan et al., 2010*)
- **Finance**
  - *financial indices* (*Bollen, Mao & Zeng, 2011*)
  - *tourism patterns* (*Choi & Varian, 2012*)
- **User profiling**
  - *age* (*Rao et al., 2010*)
  - *gender* (*Burger et al., 2011*)
  - *occupation* (*Preotiuc-Pietro, Lampos & Aletras, 2015*)

Online, user-generated data for health

Traditional disease surveillance

- does not cover the entire population
- not present everywhere (cities / countries)
- not always timely

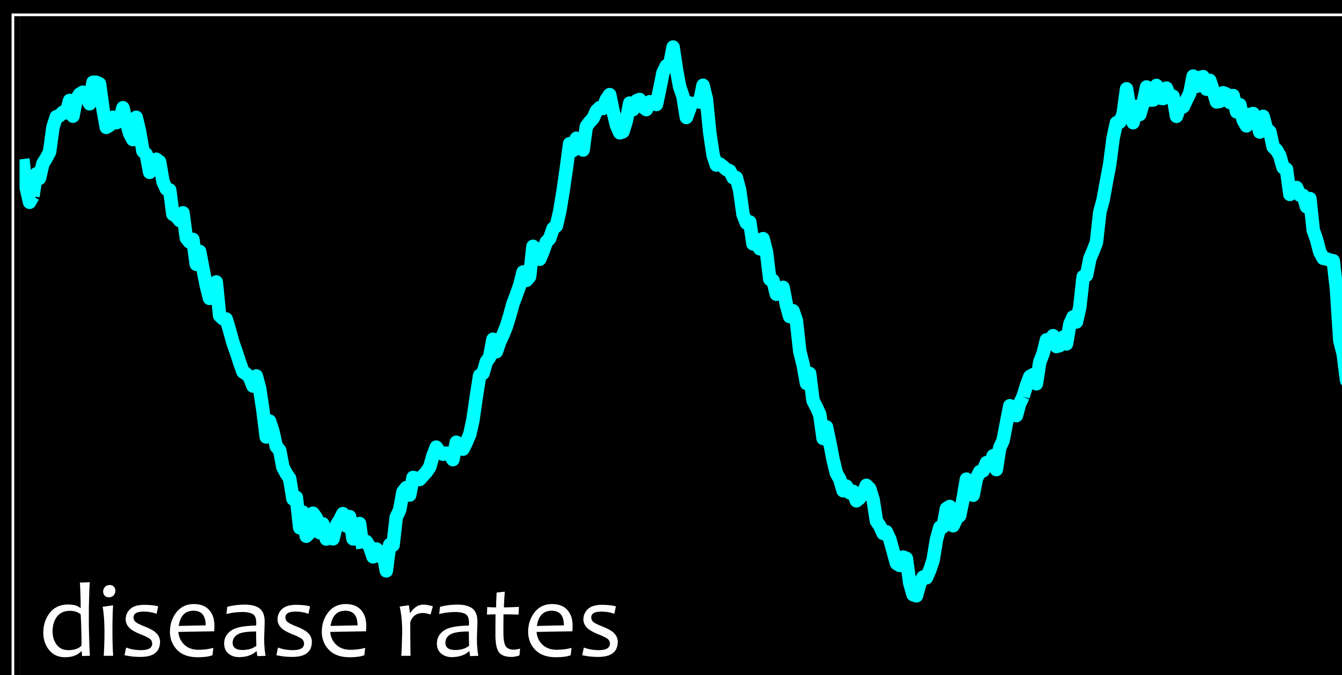
Digital disease surveillance

+ different or better population coverage
+ better geographical granularity
+ useful in underdeveloped parts of the world
+ almost instant

- *noisy, unstructured information*

e.g. (*Lampos & Cristianini, 2010 & 2012*), (*Lamb, Paul & Dredze, 2013*), (*Lampos et al., 2015*)

What this work is all about

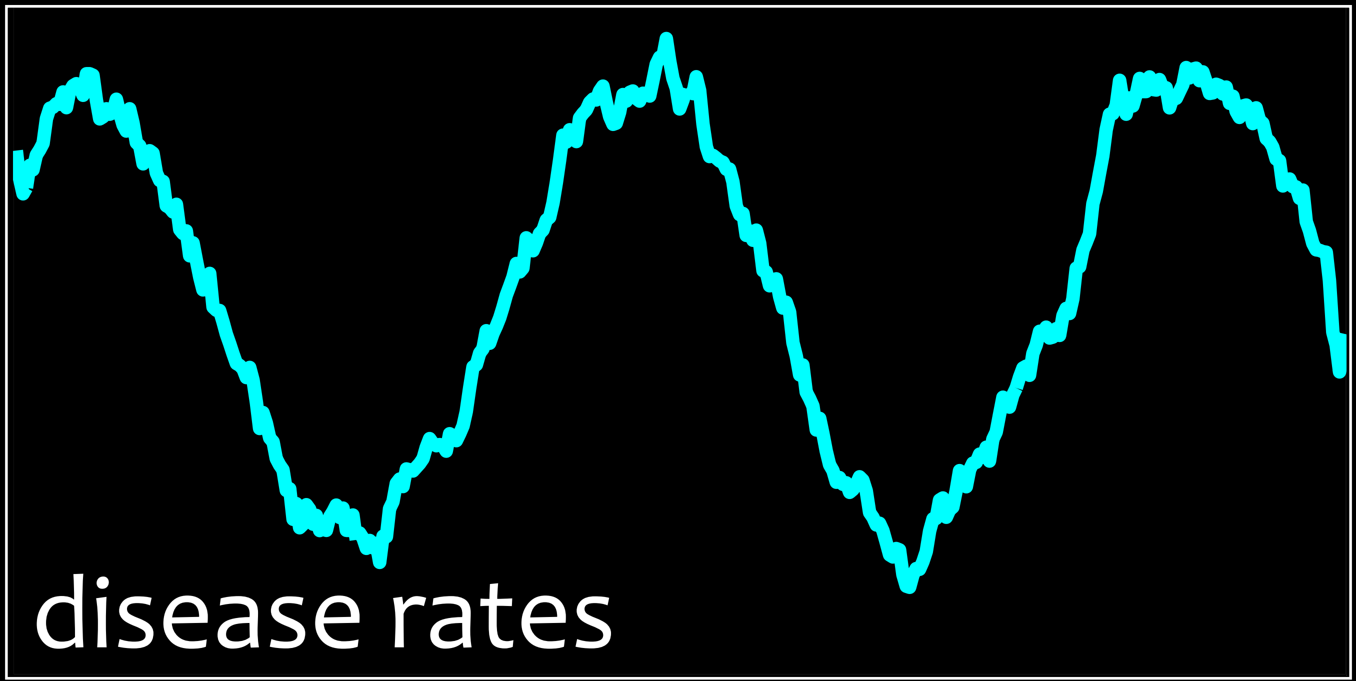


impact ?

(Pebody & Cox, 2015

Health intervention

What this work is all about



impact ?

(*Lampos, Yom-Tov, Pebody & Cox, 2015*)

Health intervention

- ✓ Background and motivation
- ◉ **Estimating disease rates from online text**
- ◉ Estimating the impact of a health intervention
- ◉ Case study: influenza vaccination impact
- ◉ Conclusions & future work

15%

Assessing the impact of a health intervention via online content

Estimating disease rates from online text

time intervals	N
n-grams	M
frequency of n-grams during the time intervals	$\mathbf{X} \in \mathbb{R}^{N \times M}$
disease rates during the time intervals	$\mathbf{y} \in \mathbb{R}^{N}$

Ridge regression

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{N} (\mathbf{x}_i \mathbf{w} + \beta - y_i)^2 + \kappa \sum_{j=1}^{M} w_j^2 \right)$$

(*Hoerl & Kennard, 1970*)

Elastic net

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{N} (\mathbf{x}_i \mathbf{w} + \beta - y_i)^2 + \lambda_1 \sum_{j=1}^{M} |w_j| + \lambda_2 \sum_{j=1}^{M} w_j^2 \right)$$

(*Zou & Hastie, 2005*)

Estimating disease rates from online text

Gaussian Process $f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}) = 0, k(\mathbf{x}, \mathbf{x}')\right)$

(*Rasmussen & Williams, 2006*)

Rational Quadratic covariance function (kernel)

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

infinite sum of squared exponential (RBF) kernels

One kernel per n-gram category

varied usage patterns, increasing semantic value

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

see also (

Estimating disease rates from online text

Gaussian Process $f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}) = 0, k(\mathbf{x}, \mathbf{x}')\right)$

(*Rasmussen & Williams, 2006*)

Rational Quadratic covariance function (kernel)

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

infinite sum of squared exponential (RBF) kernels

One kernel per n-gram category

varied usage patterns, increasing semantic value

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

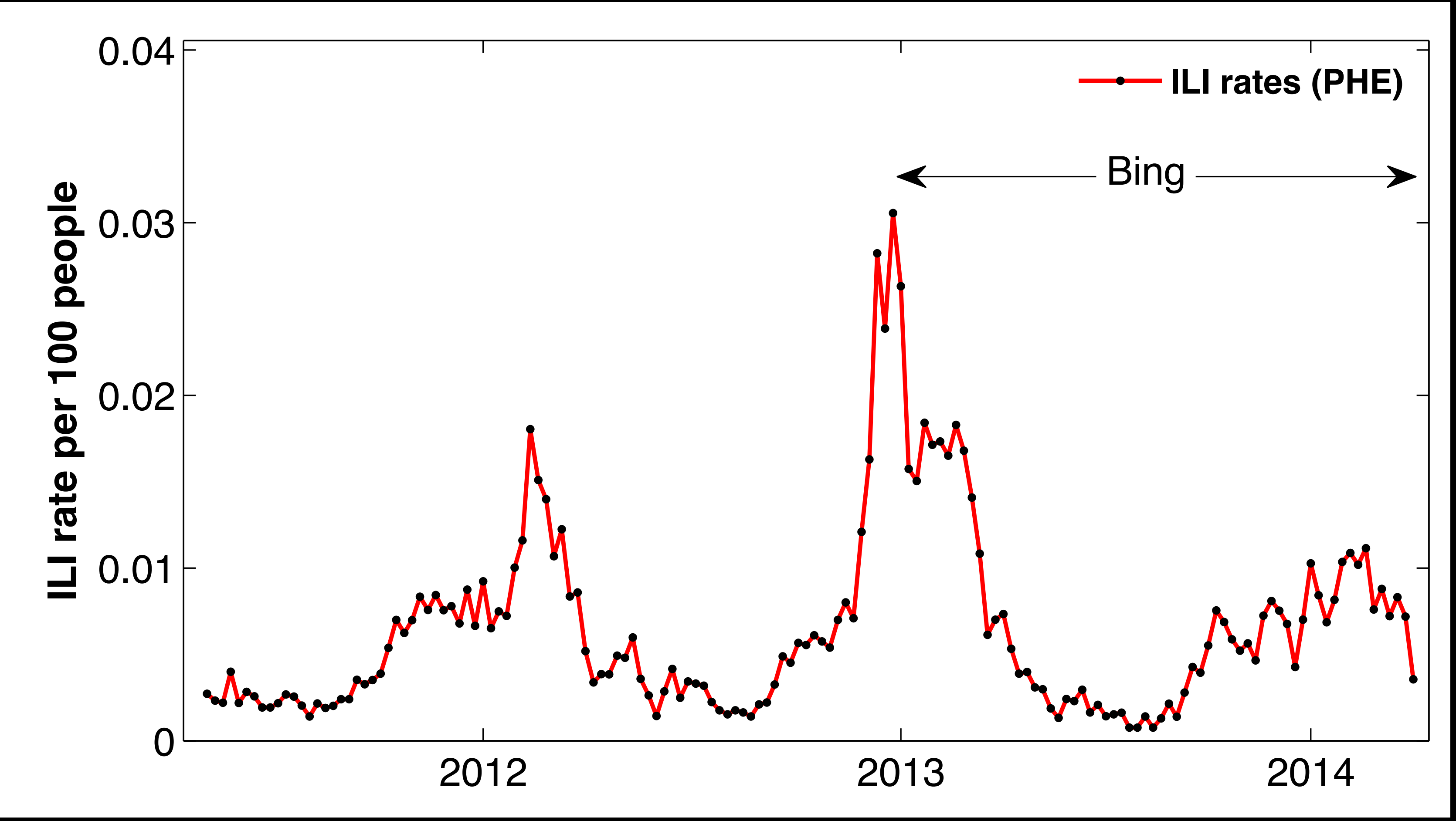
see also (*Lampos et al., 2015*)

Estimating influenza-like illness (ILI) rates — Data

User-generated data, geolocated in England

- Twitter: May 2011 to April 2014 (*308 million tweets*)
- Bing: end of December 2012 to April 2014

ILI rates from Public Health England (PHE)



Estimating ILI rates — Feature extraction

- Start with a manually crafted 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 (*Culotta, 2013*)

Estimating ILI rates — Experimental setup

Two time intervals based on the different temporal coverage of Twitter and Bing data

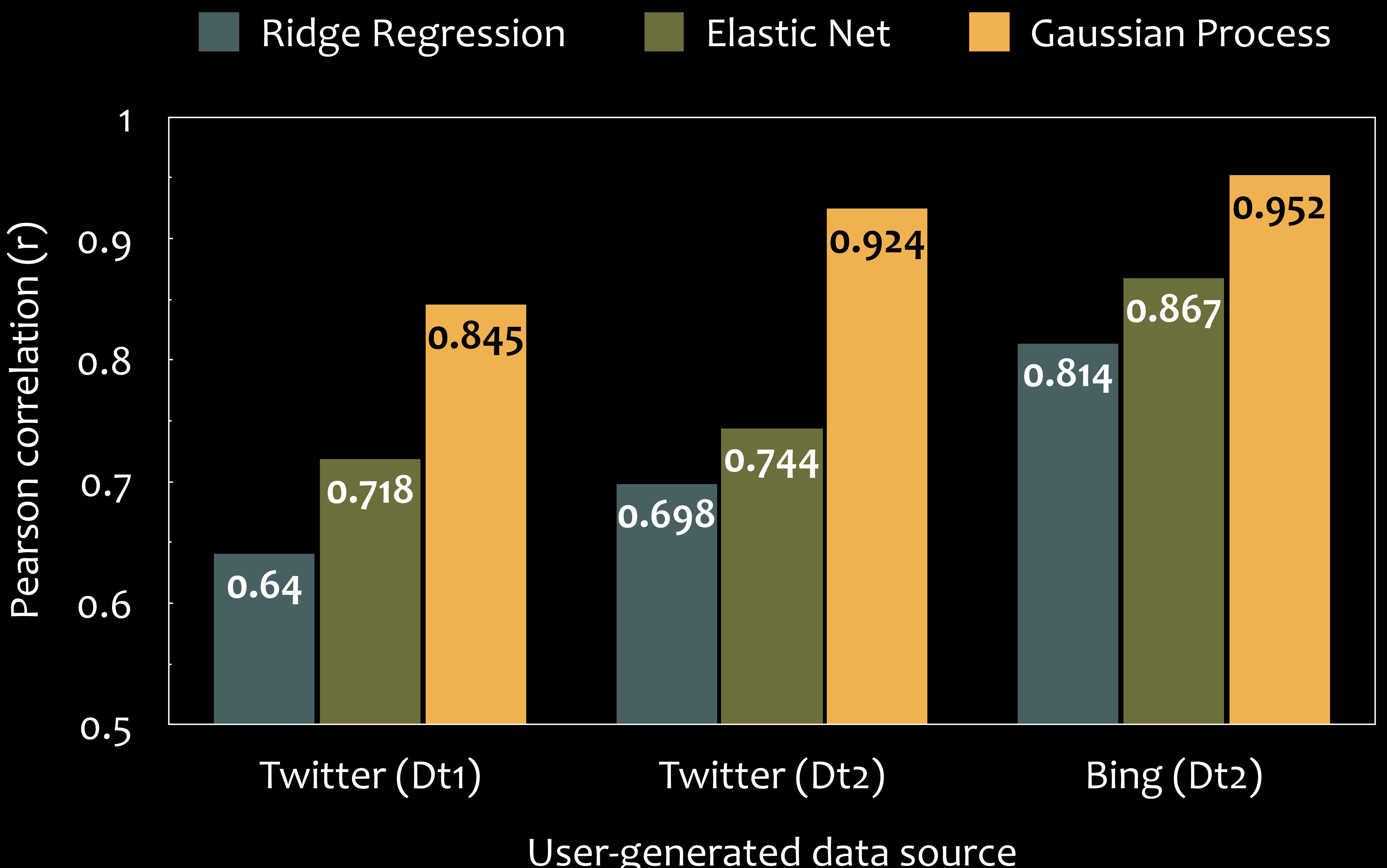
- **Dt1:** 154 weeks (May 2011 to April 2014)
- **Dt2**: 67 weeks (December 2012 to April 2014)

Stratified **10-fold cross validation**

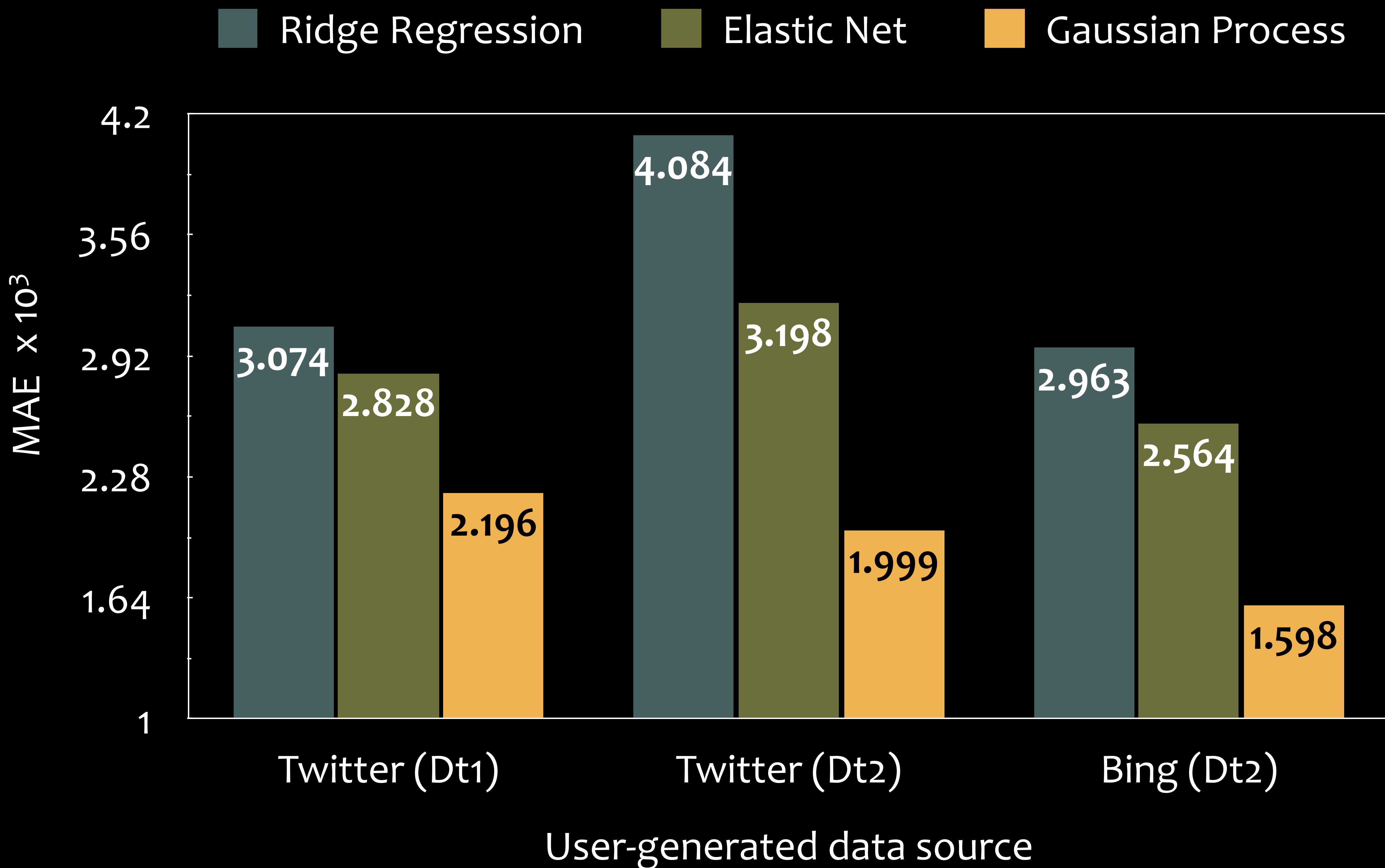
Error metrics

- Pearson correlation (**r**)
- Mean Absolute Error (**MAE**)

Estimating ILI rates — Performance



Estimating ILI rates — Performance



- ✓ Background and motivation
- ✓ Estimating disease rates from online text
- **Estimating the impact of a health intervention**
- Case study: influenza vaccination impact
- Conclusions & future work

41%

Assessing the impact of a health intervention via online content

Estimating the impact of a health intervention

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

Estimating the impact of a health intervention

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

v location(s) where the intervention took place

c control location(s)

*disease rate(s) in **affected** location before intervention*

$$r(\mathbf{q}_v^{\tau}, \mathbf{q}_c^{\tau})$$

*disease rate(s) in **control** location before intervention*

high

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

Estimating the impact of a health intervention

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

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

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

absolute difference

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

relative difference (**impact**)

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

(*Lambert & Pregibon, 2008*

Estimating the impact of a health intervention

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

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

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

absolute difference

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

relative difference (**impact**)

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

(*Lambert & Pregibon, 2008*)

✓ Background and motivation

✓ Estimating disease rates from online text

✓ Estimating the impact of a health intervention

◉ **Case study: influenza vaccination impact**

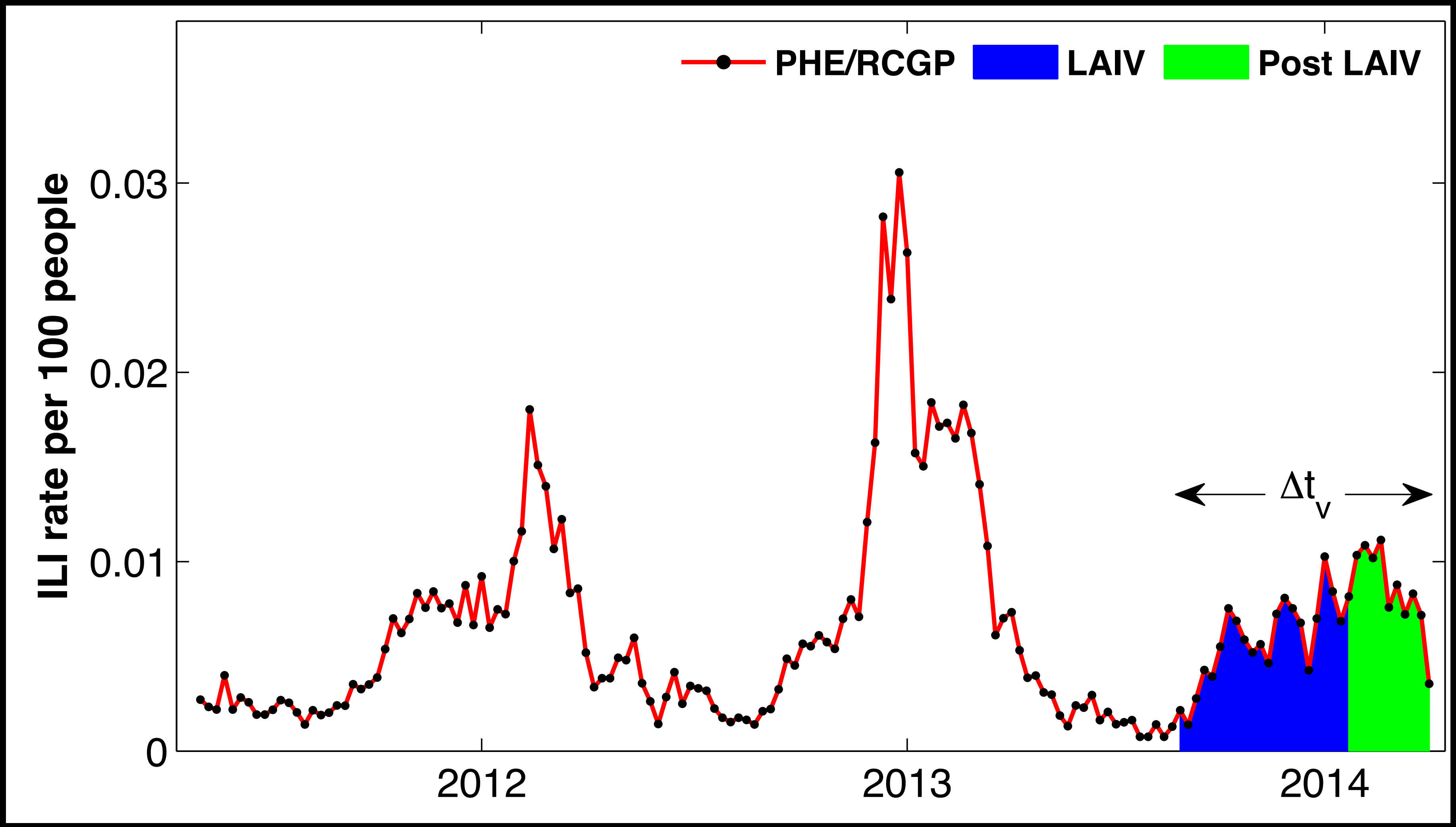
◉ Conclusions & future work

52%

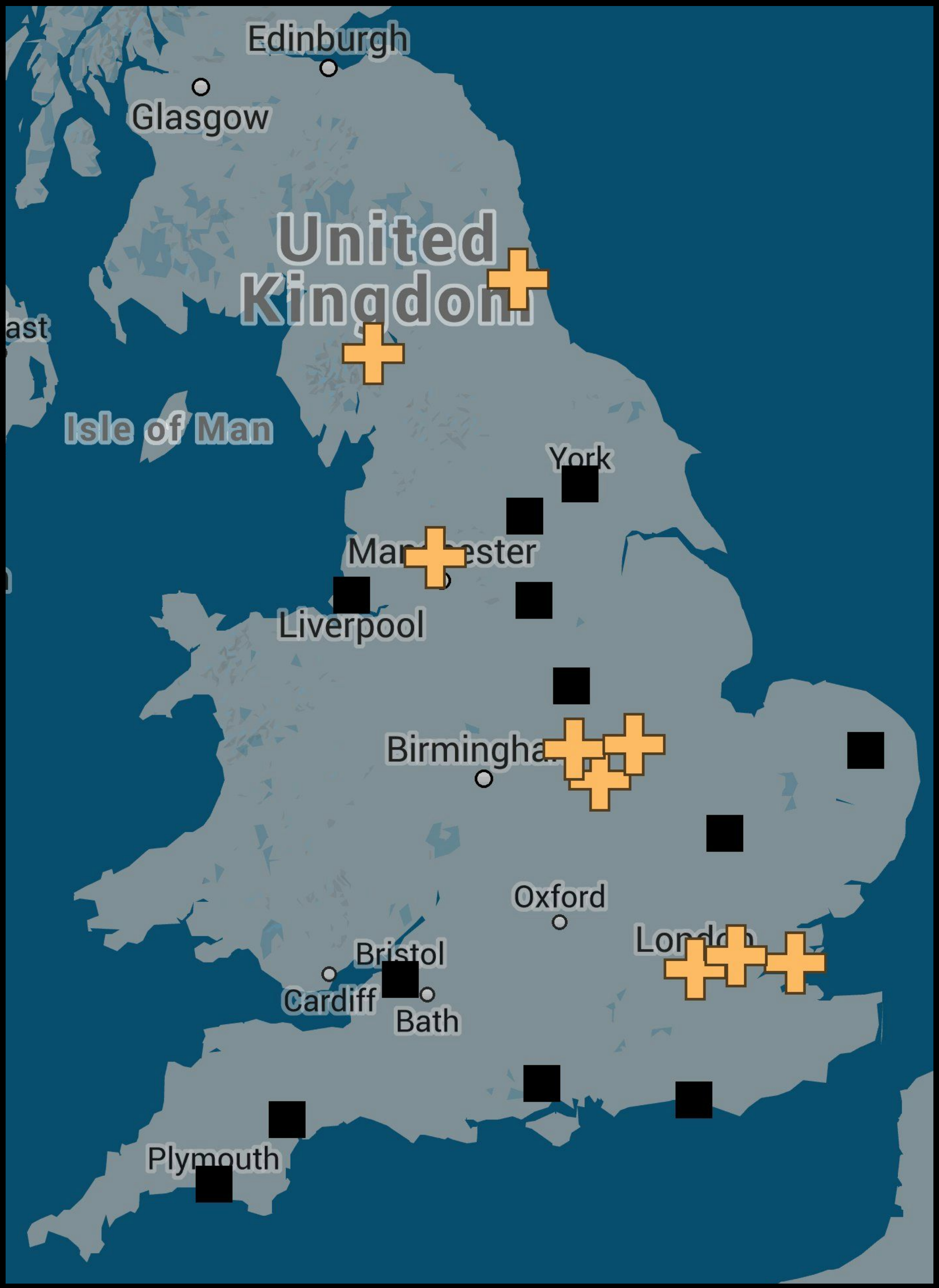
Assessing the impact of a health intervention via online content

Live Attenuated Influenza Vaccine (LAIV) campaign

- LAIV 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



Target (vaccinated) & control areas



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

Applying the impact estimation framework

Target vs. control areas

- Use previous flu season only to establish relationships
- Find the best correlated areas or **supersets** of them

Confidence intervals

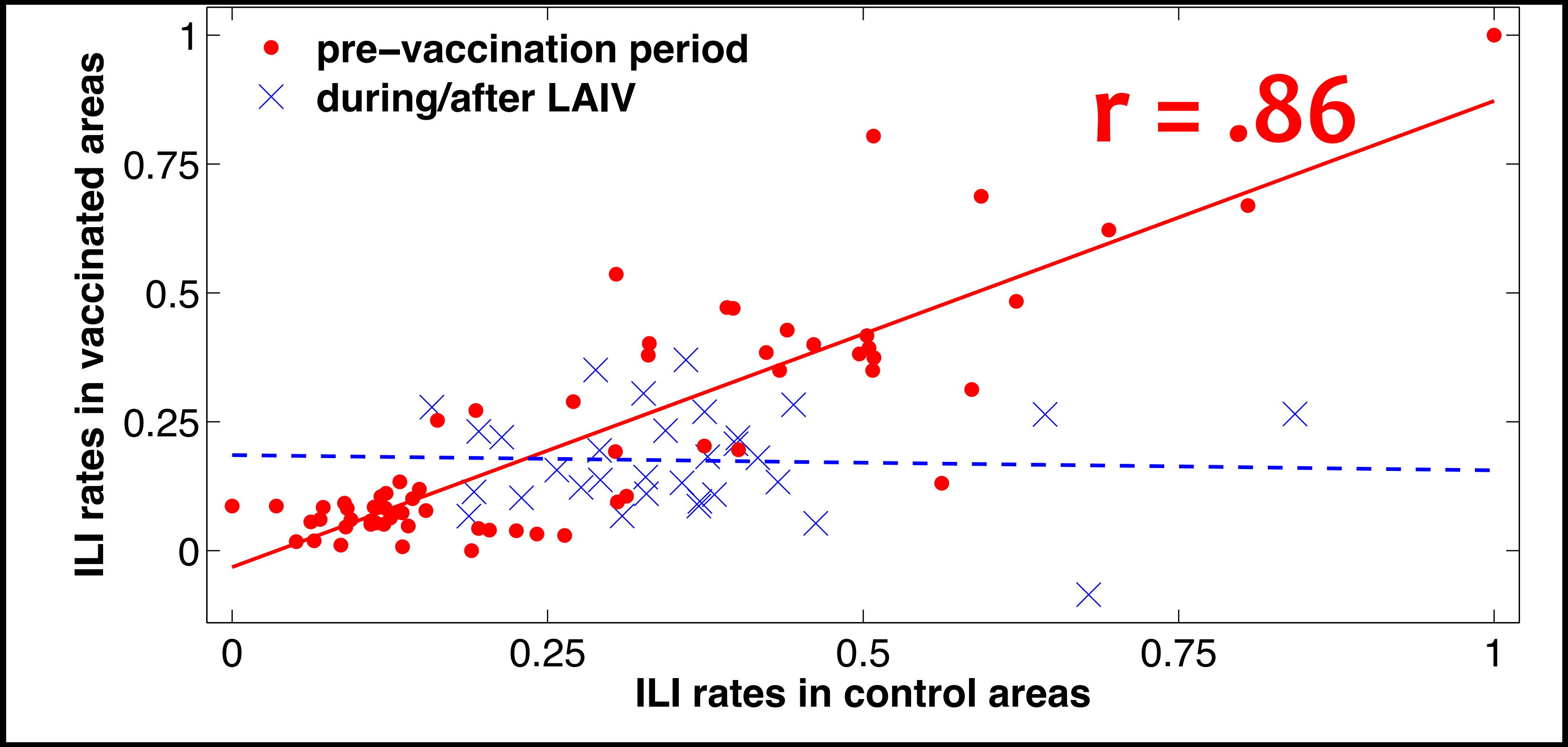
- Bootstrap sampling of the regression residuals (*mapping function of control to vaccinated areas*)
- Bootstrap sampling of data prior to the application of the bootstrapped regressor
- 10^5 bootstraps; use the .025 and .975 quantiles

Statistical significance assessment

- Impact estimate (abs.) > 2σ of the bootstrap estimates

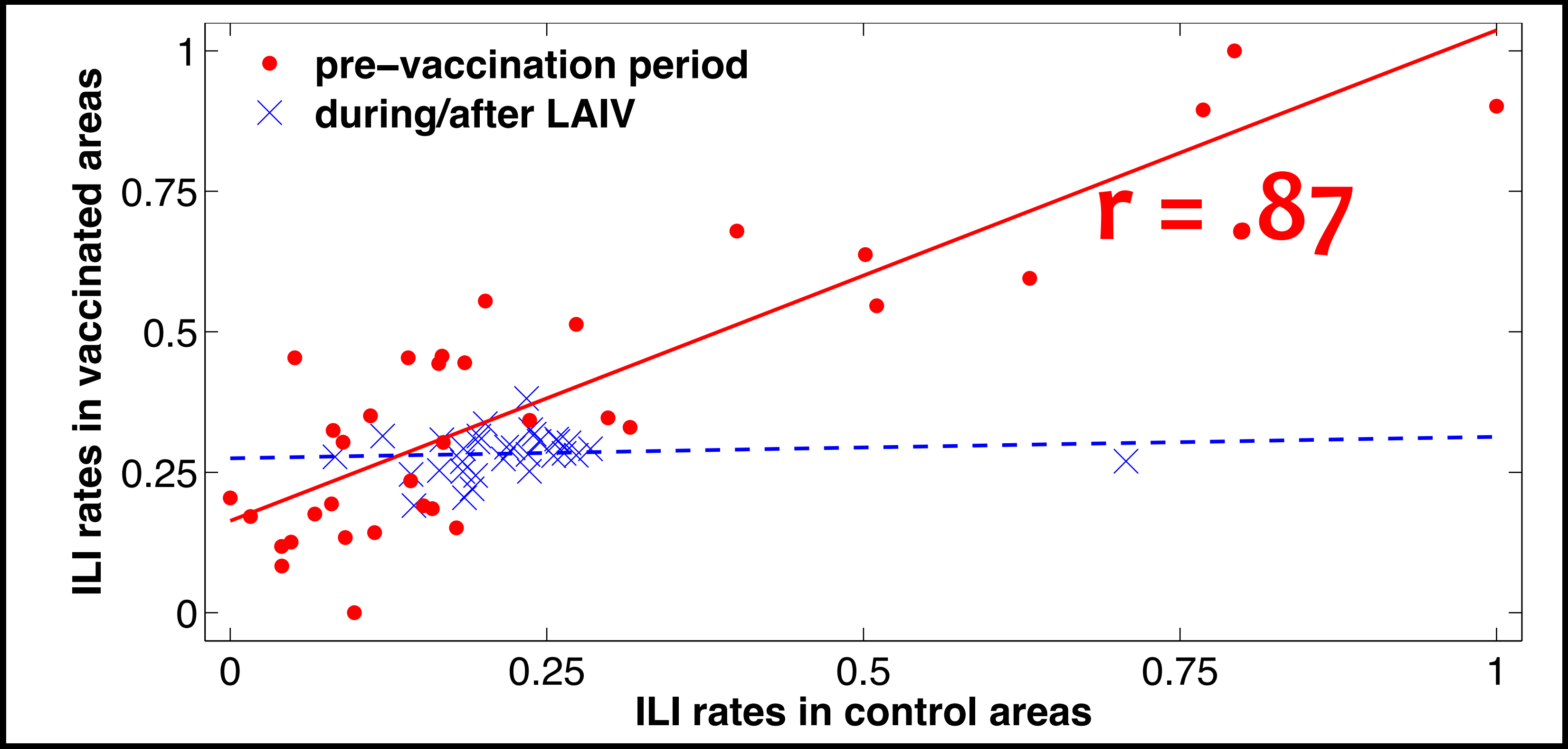
Relationship between vaccinated & control areas

Twitter — All areas



axes normalised from 0 to 1

Bing — All areas

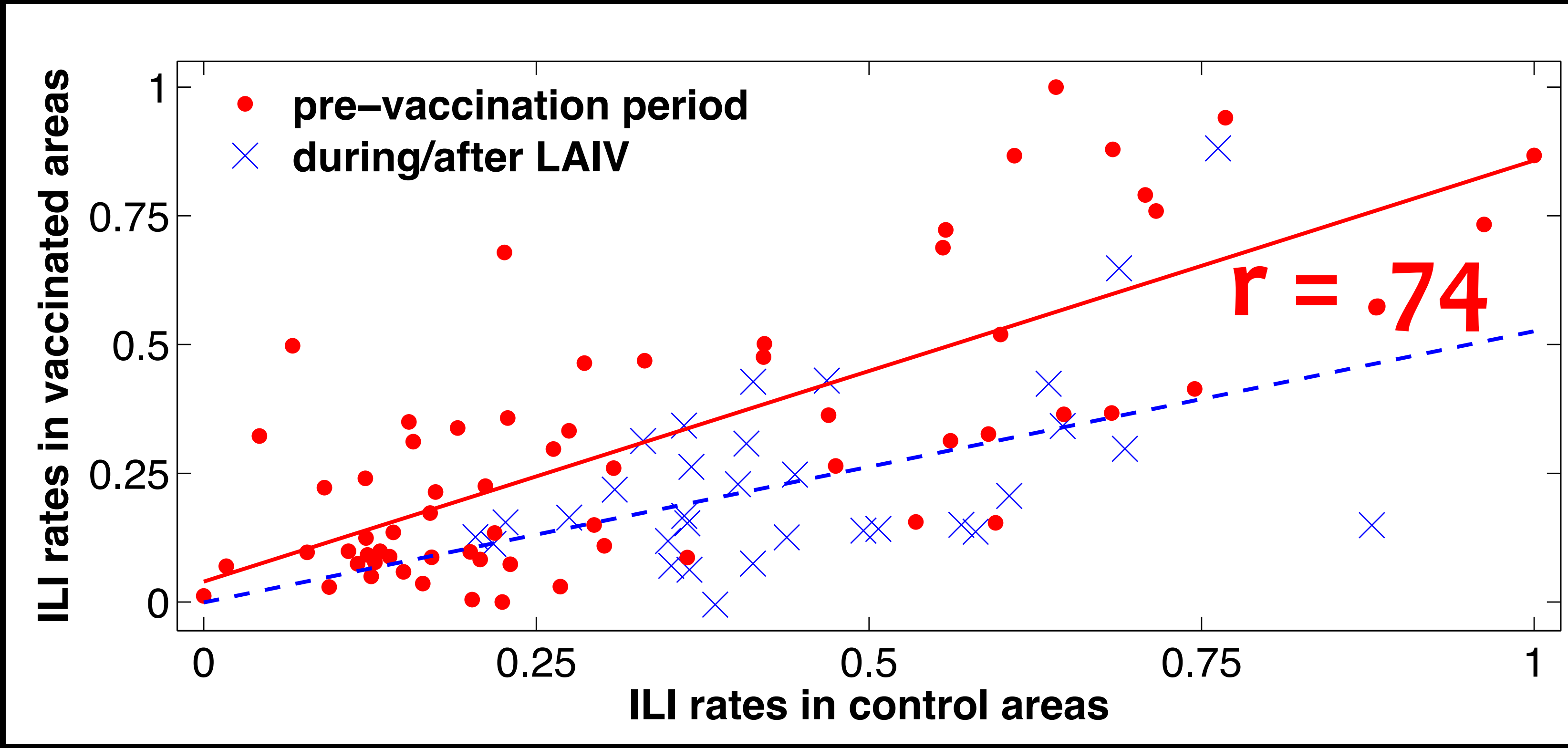


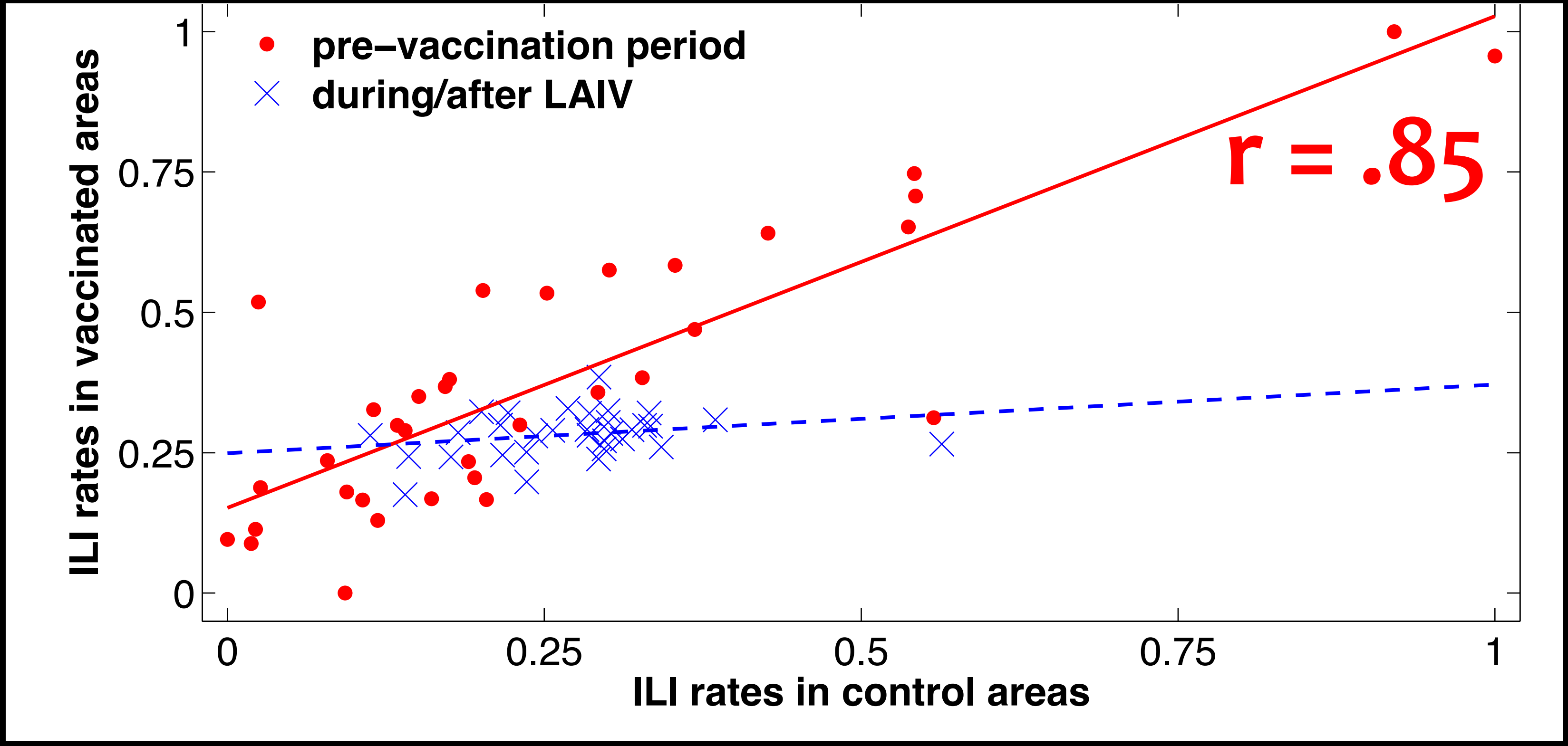
Relationship between vaccinated & control areas

Twitter — London areas

axes normalised from 0 to 1

Bing — London areas





Impact estimation results (*strongly correlated controls*)

Source	Target	r	$\delta \times 10^3$	θ (%)
Twitter	All areas	.861	**-2.5** (-4.1, -1.0)	**-32.8** (-47.4, -15.6)
Bing	All areas	.866	**-1.9** (-3.2, -0.7)	**-21.7** (-32.1, -9.10)
Twitter	London areas	.738	**-1.7** (-2.5, -0.9)	**-30.5** (-41.8, -17.5)
Bing	London areas	.848	**-2.8** (-4.1, -1.6)	**-28.4** (-36.7, -17.9)

Impact estimation results (*strongly correlated controls*)

Source	Target	r	$\delta \times 10^3$	θ (%)
Twitter	All areas	.861	**-2.5** (-4.1, -1.0)	**-32.8** (-47.4, -15.6)
Bing	All areas	.866	**-1.9** (-3.2, -0.7)	**-21.7** (-32.1, -9.10)
Twitter	London areas	.738	**-1.7** (-2.5, -0.9)	**-30.5** (-41.8, -17.5)
Bing	London areas	.848	**-2.8** (-4.1, -1.6)	**-28.4** (-36.7, -17.9)

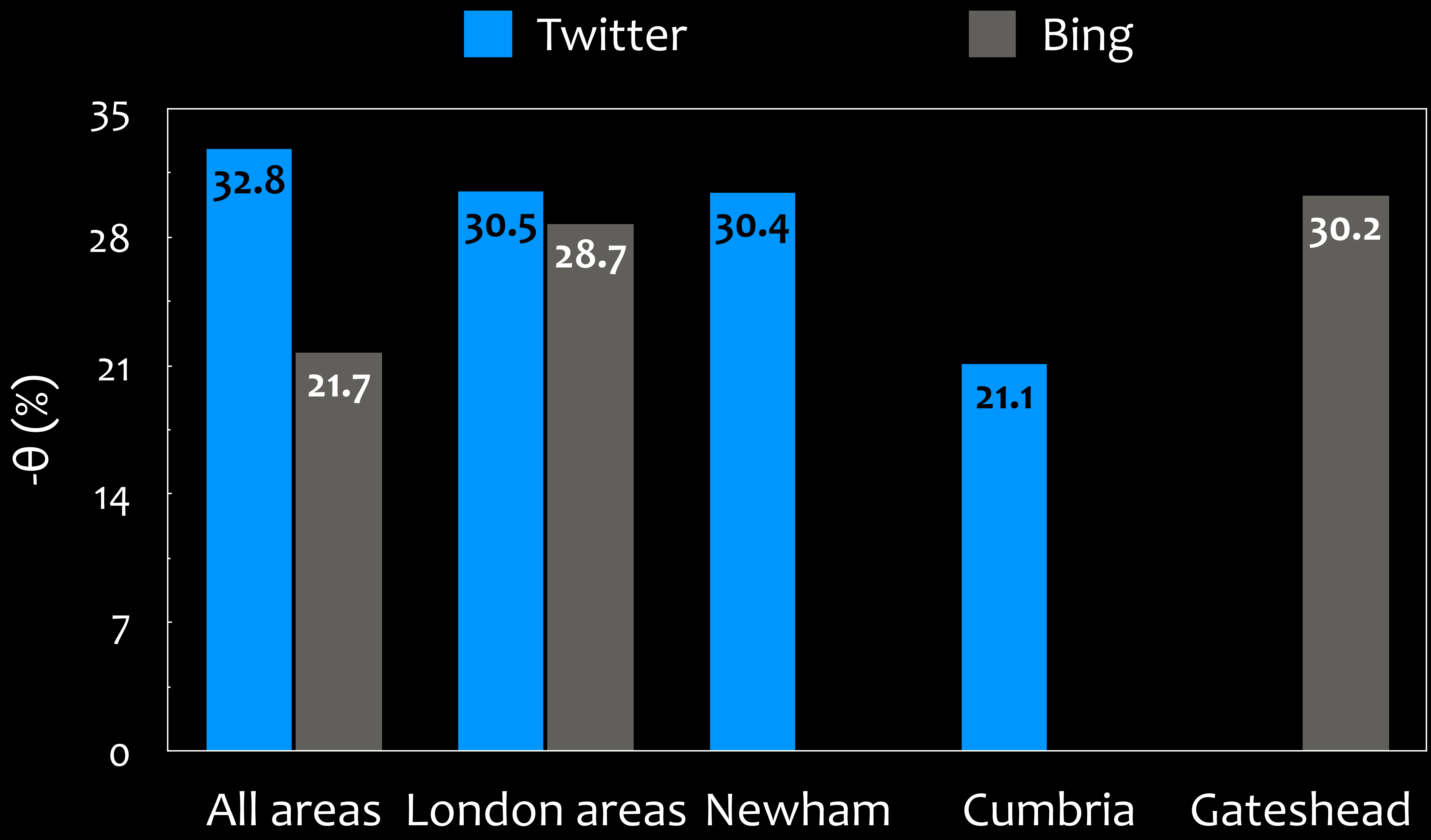
Impact estimation results (*strongly correlated controls*)

Source	Target	r	$\delta \times 10^3$	θ (%)
Twitter	All areas	.861	**-2.5** (-4.1, -1.0)	**-32.8** (-47.4, -15.6)
Bing	All areas	.866	**-1.9** (-3.2, -0.7)	**-21.7** (-32.1, -9.10)
Twitter	London areas	.738	**-1.7** (-2.5, -0.9)	**-30.5** (-41.8, -17.5)
Bing	London areas	.848	**-2.8** (-4.1, -1.6)	**-28.4** (-36.7, -17.9)

Impact estimation results (*strongly correlated controls*)

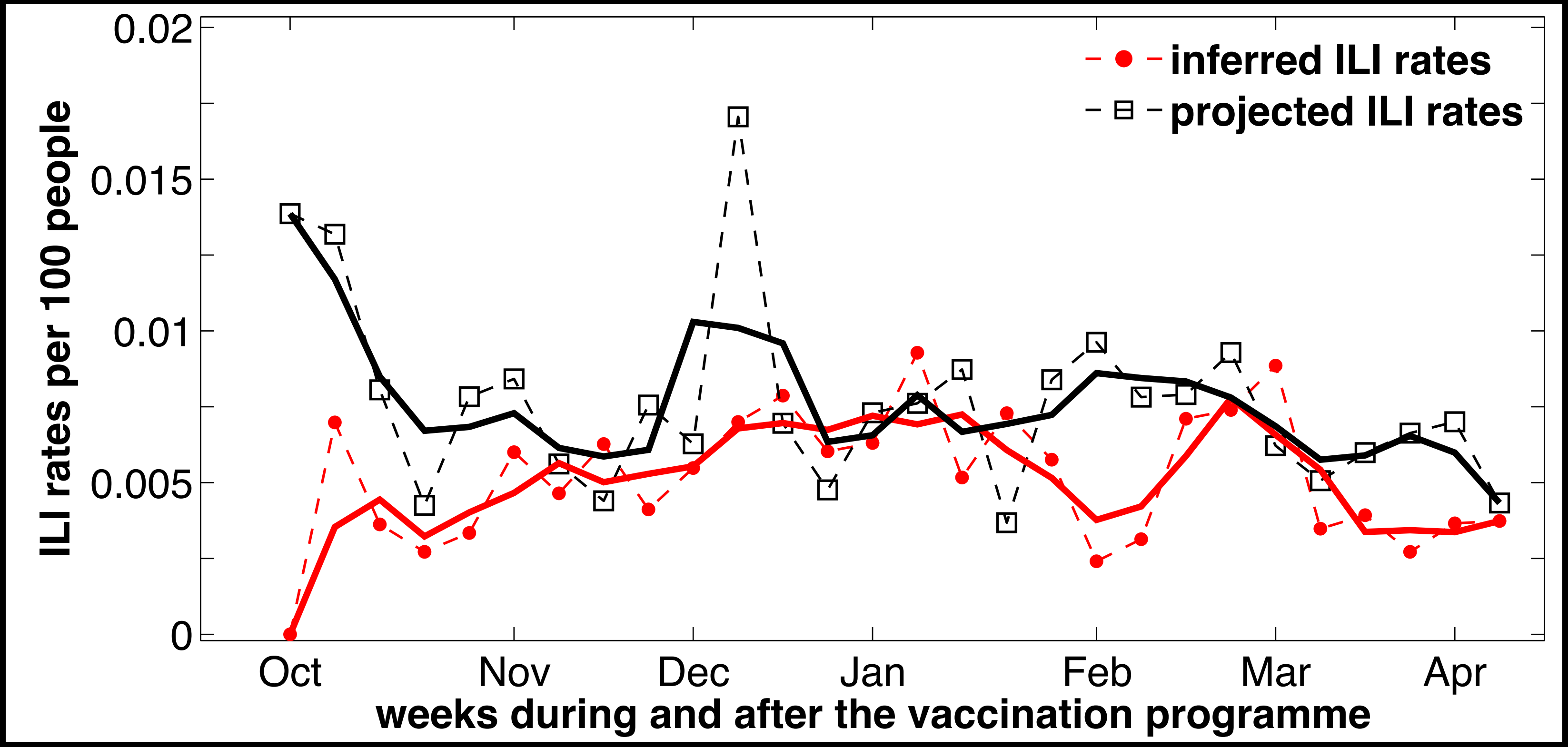
Source	Target	r	δ x 10^3	θ (%)
Twitter	All areas	.861	**-2.5** (-4.1, -1.0)	**-32.8** (-47.4, -15.6)
Bing	All areas	.866	**-1.9** (-3.2, -0.7)	**-21.7** (-32.1, -9.10)
Twitter	London areas	.738	**-1.7** (-2.5, -0.9)	**-30.5** (-41.8, -17.5)
Bing	London areas	.848	**-2.8** (-4.1, -1.6)	**-28.4** (-36.7, -17.9)

Impact estimation results (*stat. sig.*)

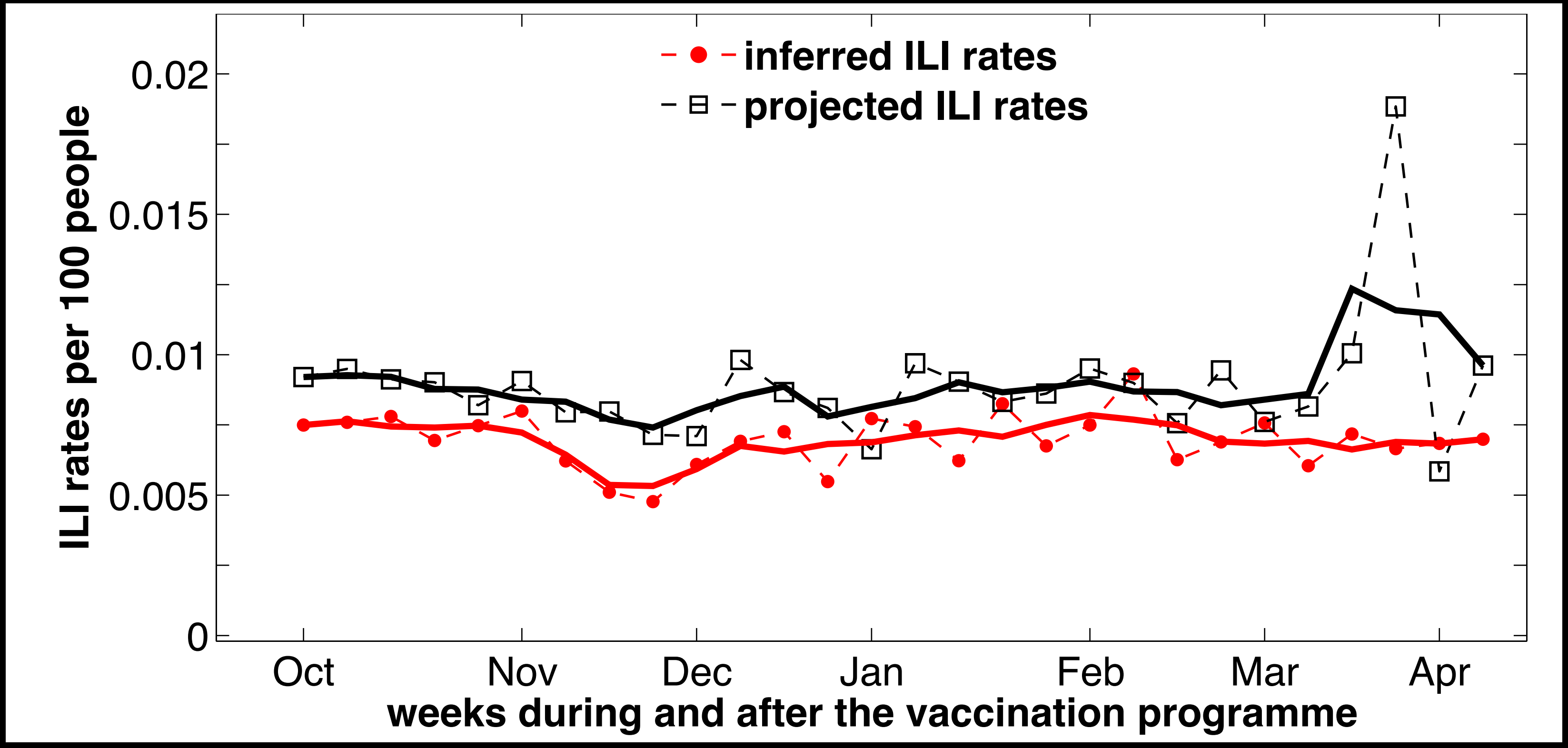


Projected vs. inferred ILI rates in vaccinated locations

Twitter — All areas

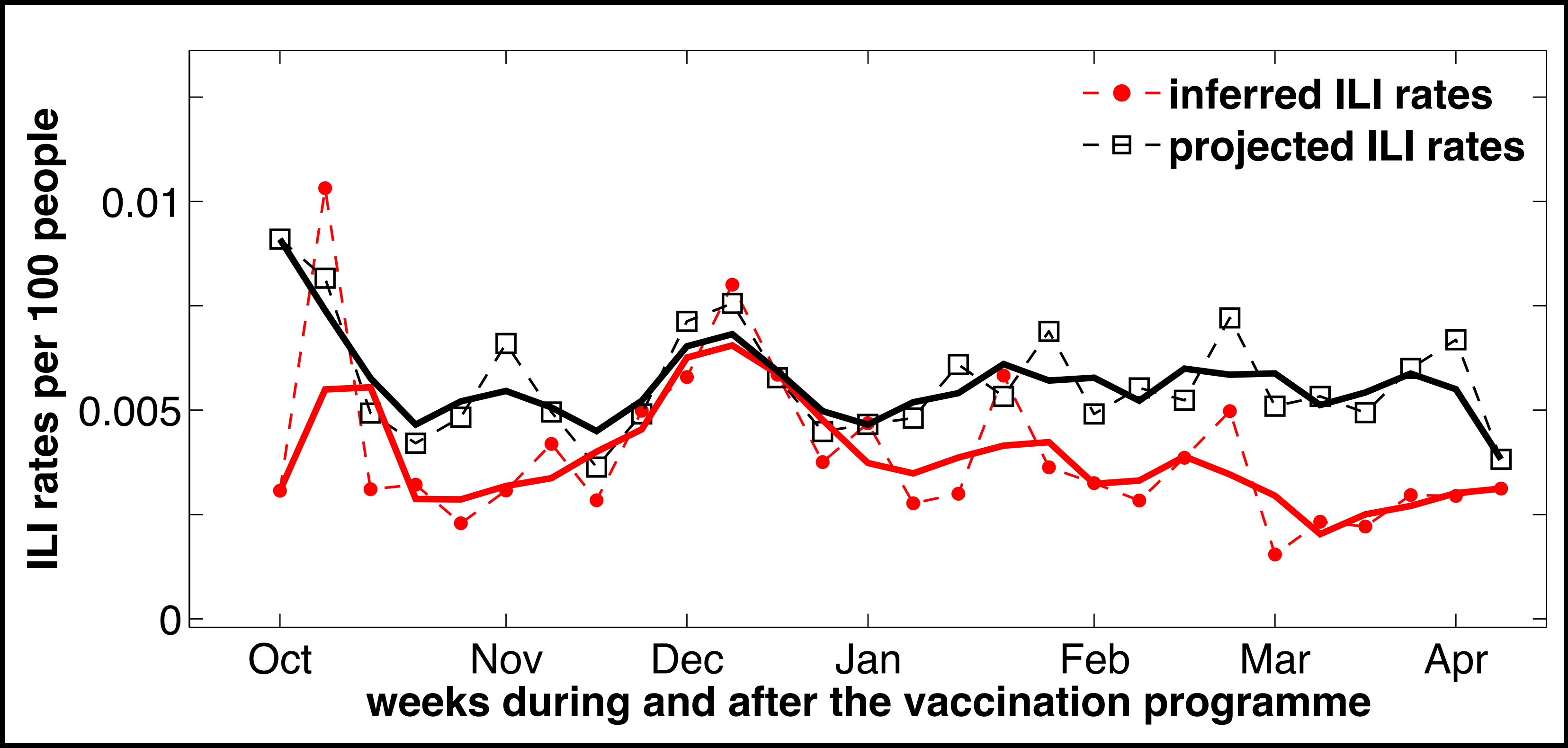


Bing — All areas

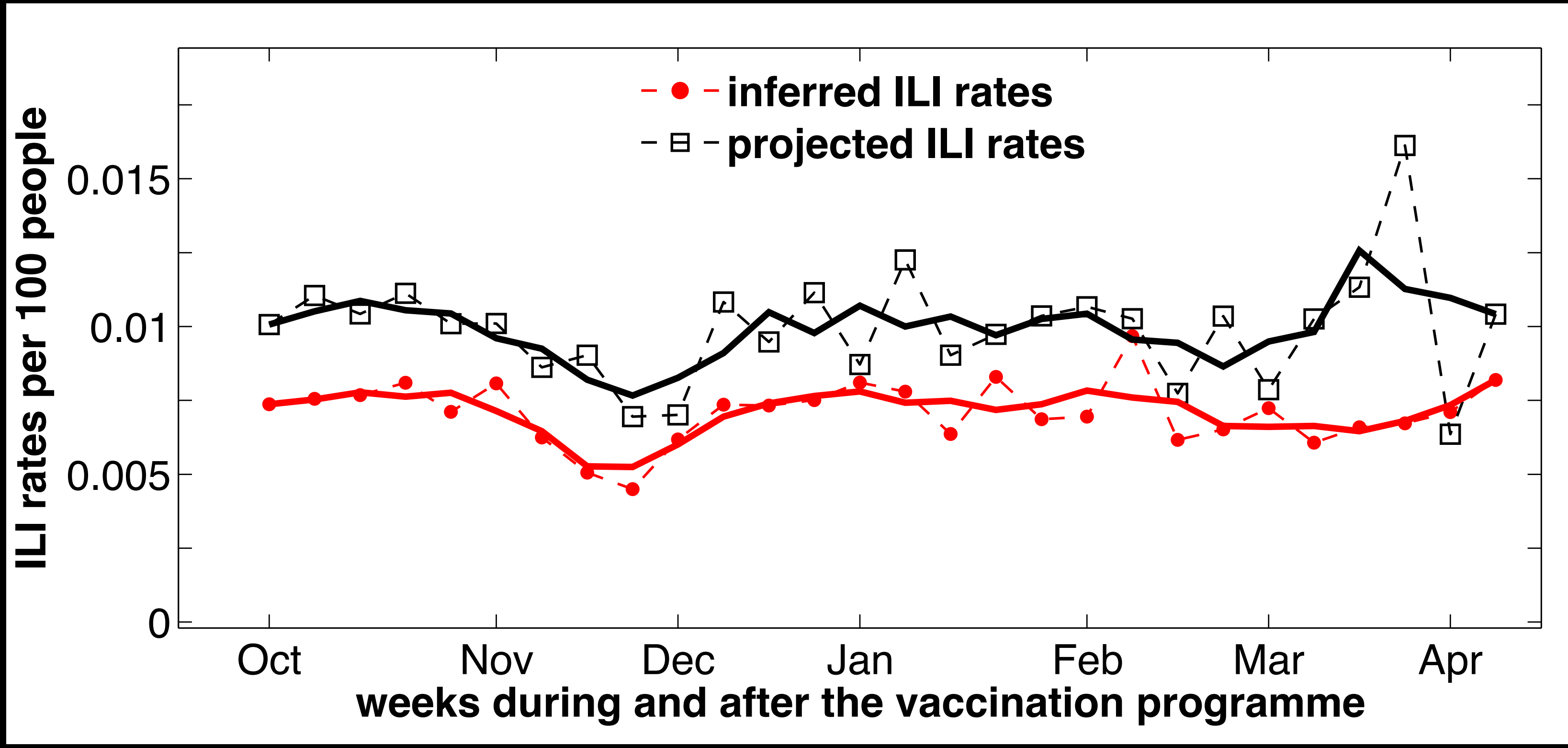


Projected vs. inferred ILI rates in vaccinated locations

Twitter — London areas



Bing — London areas



Sensitivity of impact estimates to variable controls

- Repeat the impact estimation for the N controls (up to a 100) with $r \geq 95\%$ of the best r —> $\mu(\delta)$ and $\mu(\theta)$ (%)
- Measure % of difference, $\boldsymbol{\Delta(\theta)}$, between θ and $\mu(\theta)$

Source	Target	N	$\mu(r)$	$\mu(\delta) \times 10^3$	$\mu(\theta)$ (%)	$\Delta\theta$ (%)
Twitter	All areas	100	0.84	-2.5 (0.2)	-32.7 (2.1)	0.10
Bing	All areas	46	0.85	-1.4 (0.4)	-16.4 (3.6)	24.4
Twitter	London areas	79	0.70	-1.5 (0.1)	-27.9 (2.0)	8.32
Bing	London areas	100	0.84	-1.4 (0.2)	-16.9 (1.8)	40.4

Sensitivity of impact estimates to variable controls

- Repeat the impact estimation for the N controls (up to a 100) with r ≥ 95% of the best r —> $\mu(\delta)$ and $\mu(\theta)$ (%)
- Measure % of difference, $\boldsymbol{\Delta(\theta)}$, between θ and $\mu(\theta)$

Source	Target	N	$\mu(r)$	$\mu(\delta) \times 10^3$	$\mu(\theta)$ (%)	$\Delta\theta$ (%)
Twitter	**All areas**	**100**	**0.84**	**-2.5 (0.2)**	**-32.7 (2.1)**	**0.10**
Bing	All areas	46	0.85	-1.4 (0.4)	-16.4 (3.6)	24.4
Twitter	London areas	79	0.70	-1.5 (0.1)	-27.9 (2.0)	8.32
Bing	London areas	100	0.84	-1.4 (0.2)	-16.9 (1.8)	40.4

Sensitivity of impact estimates to variable controls

- Repeat the impact estimation for the N controls (up to a 100) with $r \geq 95\%$ of the best r —> $\mu(\delta)$ and $\mu(\theta)$ (%)
- Measure % of difference, **$\Delta(\theta)$**, between θ and $\mu(\theta)$

Source	Target	N	$\mu(r)$	$\mu(\delta) \times 10^3$	$\mu(\theta)$ (%)	$\Delta\theta$ (%)
Twitter	All areas	100	0.84	-2.5 (0.2)	-32.7 (2.1)	0.10
Bing	**All areas**	**46**	**0.85**	**-1.4 (0.4)**	**-16.4 (3.6)**	**24.4**
Twitter	London areas	79	0.70	-1.5 (0.1)	-27.9 (2.0)	8.32
Bing	London areas	100	0.84	-1.4 (0.2)	-16.9 (1.8)	40.4

- ✓ Background and motivation
- ✓ Estimating disease rates from online text
- ✓ Estimating the impact of a health intervention
- ✓ Case study: influenza vaccination impact
- **Conclusions & future work**

89%

Assessing the impact of a health intervention via online content

Conclusions & points for discussion

- Framework for estimating the impact of a health intervention based on online content
- Access to different & larger parts of the population

Evaluation is hard, however:

- PHE's impact estimates: -66% based on sentinel surveillance, -24% laboratory confirmed (*Pebody et al., 2014*)
- Correlation between actual vaccination uptake and our study's estimated impacts

Why are Bing and Twitter estimations different?

- Different user demographics (?) — *this can be useful*
- Different temporal resolution

Potential future work directions

- Improve **supervised learning** models
  - better natural language processing / machine learning modelling
  - combination of different data sources
- Work on **unsupervised techniques**
  - inferring / understanding the demographics of the online medium will be essential
- **More rigorous evaluation**

Collaborators, acknowledgements & material

Elad Yom-Tov, Microsoft Research
Richard Pebody, Public Health England
Ingemar J. Cox, UCL & University of Copenhagen

Jens Geyti, UCL (Software Engineer)
Simon de Lusignan, University of Surrey & RCGP

i-sense
EPSRC IRC in Early Warning Sensing Systems for Infectious Diseases

i-sense.org.uk

Paper: ow.ly/RN9J2

Slides: ow.ly/RN7MZ

References

Bollen, Mao & Zeng. Twitter mood predicts the stock market. J Comp Science, 2011.

Burger, Henderson, Kim & Zarrella. Discriminating Gender on Twitter. EMNLP, 2011.

Choi & Varian. Predicting the Present with Google Trends. Economic Record, 2012.

Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Lang Resour Eval, 2013.

Hoerl & Kennard. Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 1970.

Lamb, Paul & Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. NAACL, 2013.

Lambert & Pregibon. Online effects of offline ads. Data Mining & Audience Intelligence for Advertising, 2008.

Lampos & Cristianini. Tracking the flu pandemic by monitoring the Social Web. CIP, 2010.

Lampos & Cristianini. Nowcasting Events from the Social Web with Statistical Learning. ACM TIST, 2012.

Lampos, Miller, Crossan & Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep, 2015.

Lampos, Yom-Tov, Pebody & Cox. Assessing the impact of a health intervention via user-generated Internet content. DMKD, 2015.

Pebody et al. Uptake and impact of a new live attenuated influenza vaccine programme in England: early results of a pilot in primary school-age children, 2013/14 influenza season. Eurosurveillance, 2014.

Preotiuc-Pietro, Lampos & Aletras. An analysis of the user occupational class through Twitter content. ACL, 2015.

Rao, Yarowsky, Shreevats & Gupta. Classifying Latent User Attributes in Twitter. SMUC, 2010.

Rasmussen & Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.

Tumasjan, Sprenger, Sandner & Welpe. Predicting Elections with Twitter: What 140 characters Reveal about Political Sentiment. ICWSM, 2010.

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