Non-traditional data-driven approaches to epidemiology

Vasileios Lampos

Department of Computer Science
University College London

www lampos.net

@lampos

Engineering and Physical Sciences Research Council

Medical Research Council

i-sense

Research team, collaborations & funding

- UCL Computer Science, UCL AI Centre
- UCL Health Informatics, Epidemiology, Institute for Global Health, Division of Medicine
- Imperial College London, Harvard, Bar-Ilan universities
- Public Health England, World Health Organisation
- Microsoft Research, Google Health
- i-sense (i-sense.org.uk), VirusWatch (ucl-virus-watch.net) projects
- Funding: EPSRC, MRC, Google (> £19 million since 2014)

Engineering and Physical Sciences Research Council

Medical Research Council

Public Health England

Traditional epidemiology

- Traditional ≈ *conventional*, *established*
- Data streams based on interactions with health services
- Methods: statistics, mechanistic models, rarely machine learning
- Challenges
  - Biases in the cohorts (*sampling bias*)
  - Reporting latency
  - Non-established health systems
  - A pandemic!

Alternative data streams for epidemiology

Microsoft Bing

Google

- Web search activity, social media
- Different *data*, different *methods*?
- Complementary to conventional approaches
  - Larger cohorts, broader/different demographic and geographic coverage
  - Reduced latency
  - Lower cost
  - Not particularly affected by closure days and *pandemics*
  - Applicable in locations where health surveillance is less established

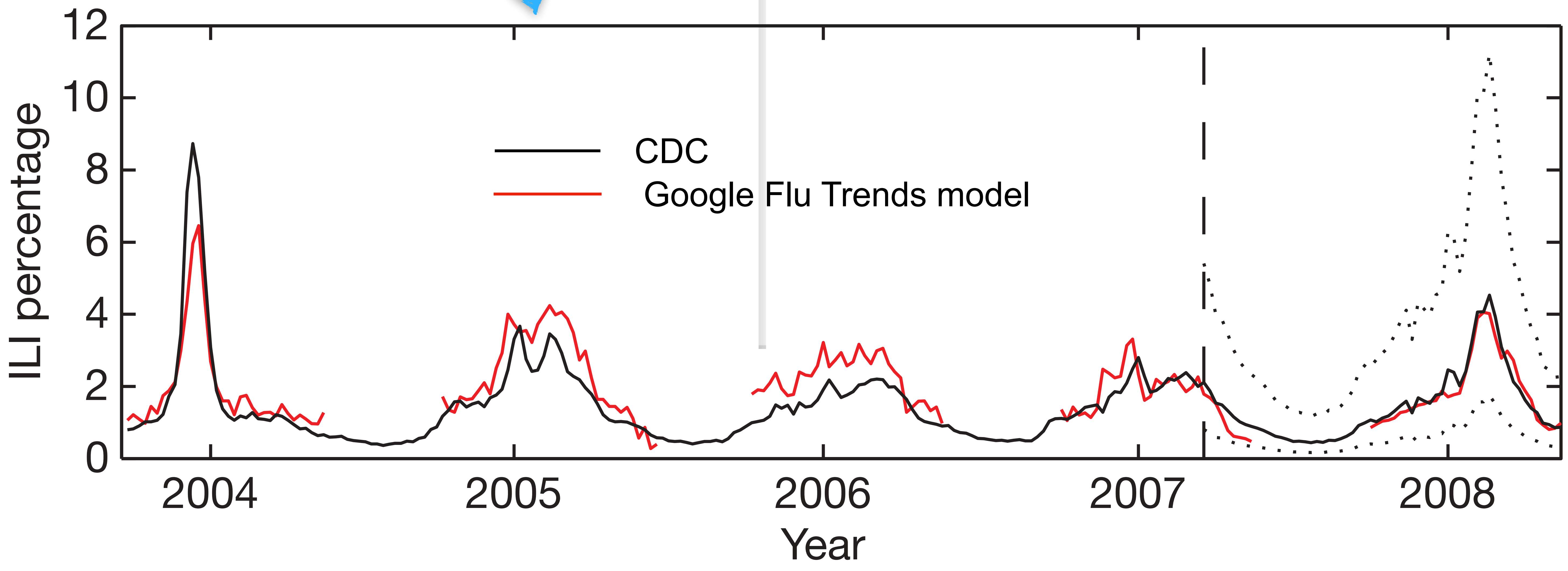
Wagner, Lampos, Cox, Pebody. *Sci. Rep.* (2018), doi:10.1038/s41598-018-32029-6

Mapping web search activity to disease rate estimates

Google

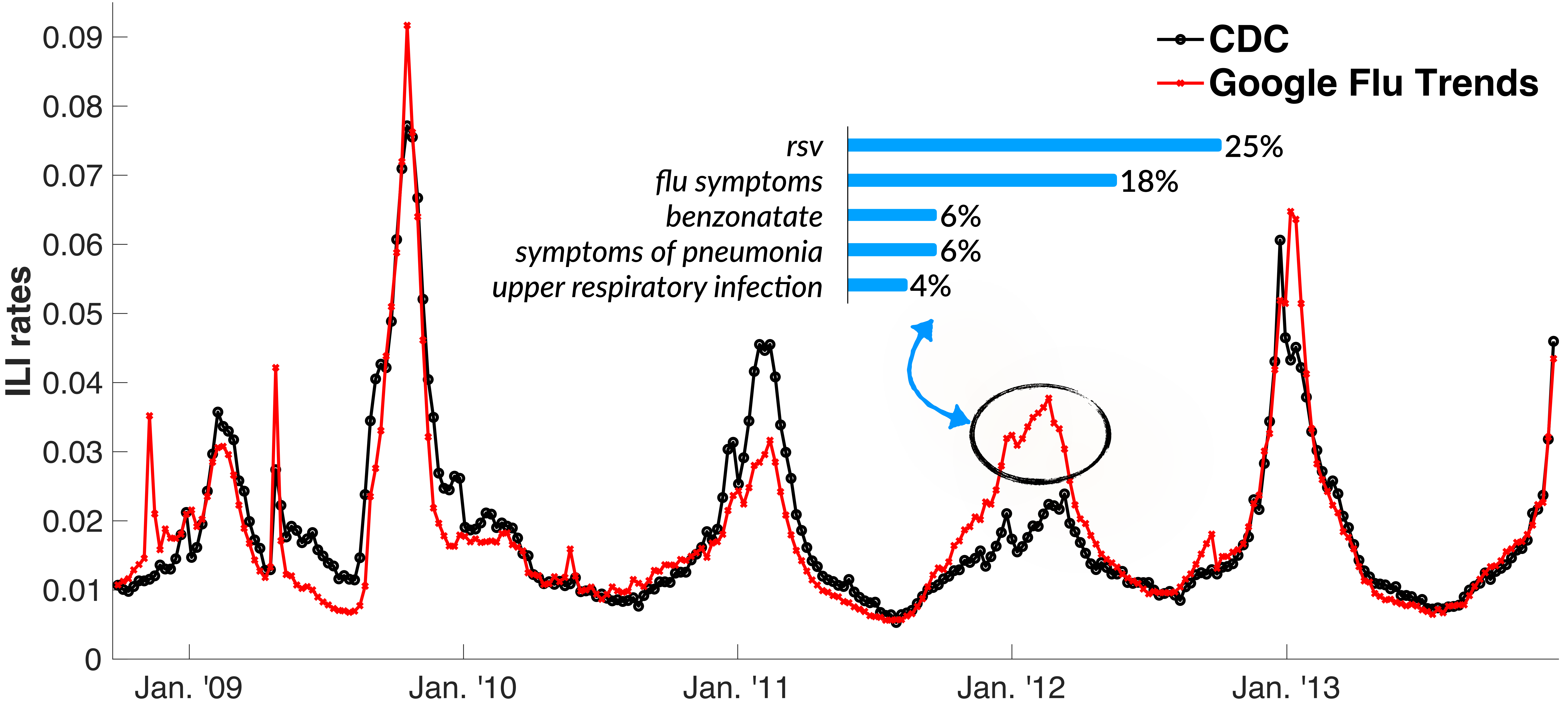
flu treatment

flu treatment
flu treatment **kids**
flu treatment **otc**
flu treatment **natural**
flu treatment **medication**
flu treatment **toddler**



Eysenbach, *AMIA* (2006); Polgreen et al. *Clin. Infect. Dis.* (2008); Ginsberg et al. *Nature* (2009)

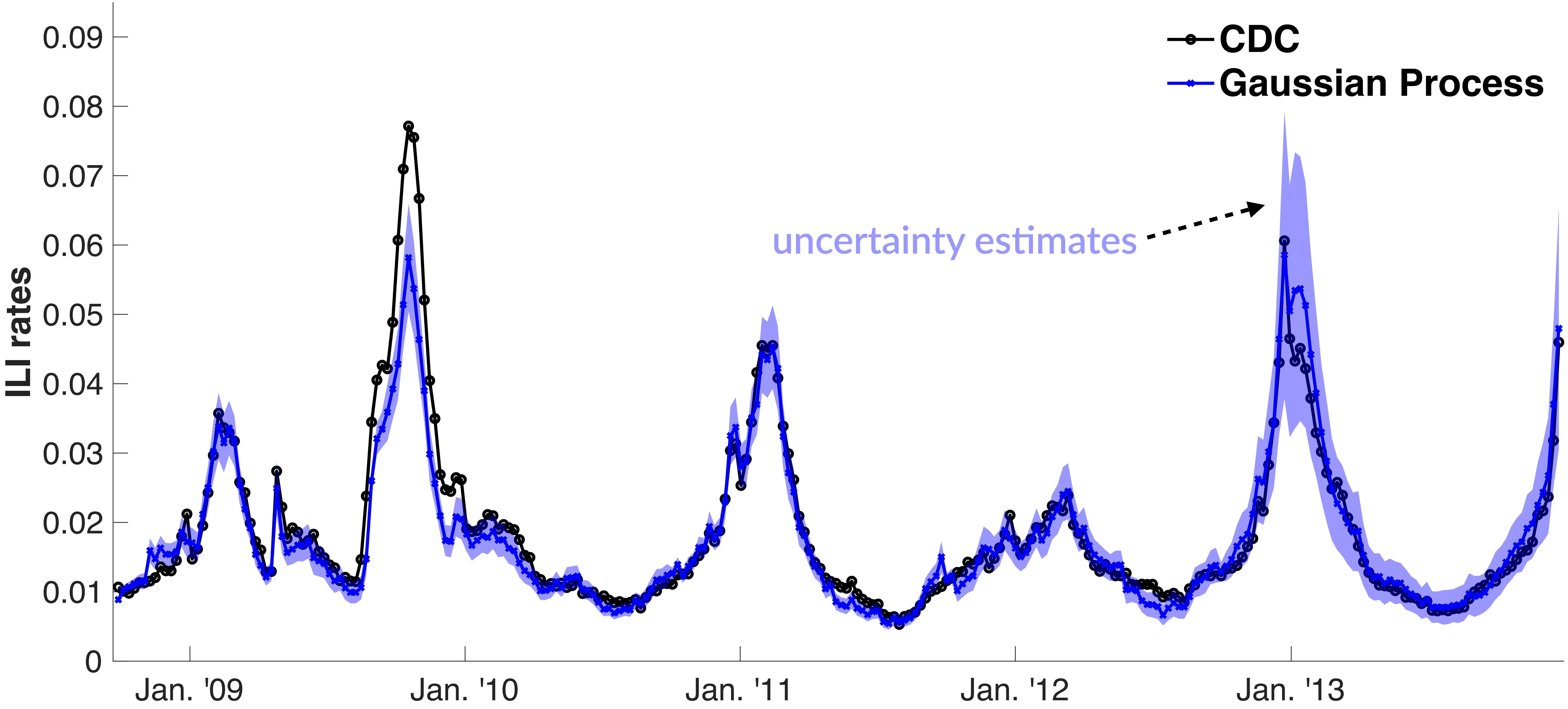
What went wrong with Google Flu Trends?



Influenza-like illness (**ILI**) rate estimates in the US during the 2011/12 flu season were greatly affected by web searches that were not related to flu.

Lampos, Miller, Crossan, Stefansen. *Sci. Rep.* (2015), doi:10.1038/srep12760

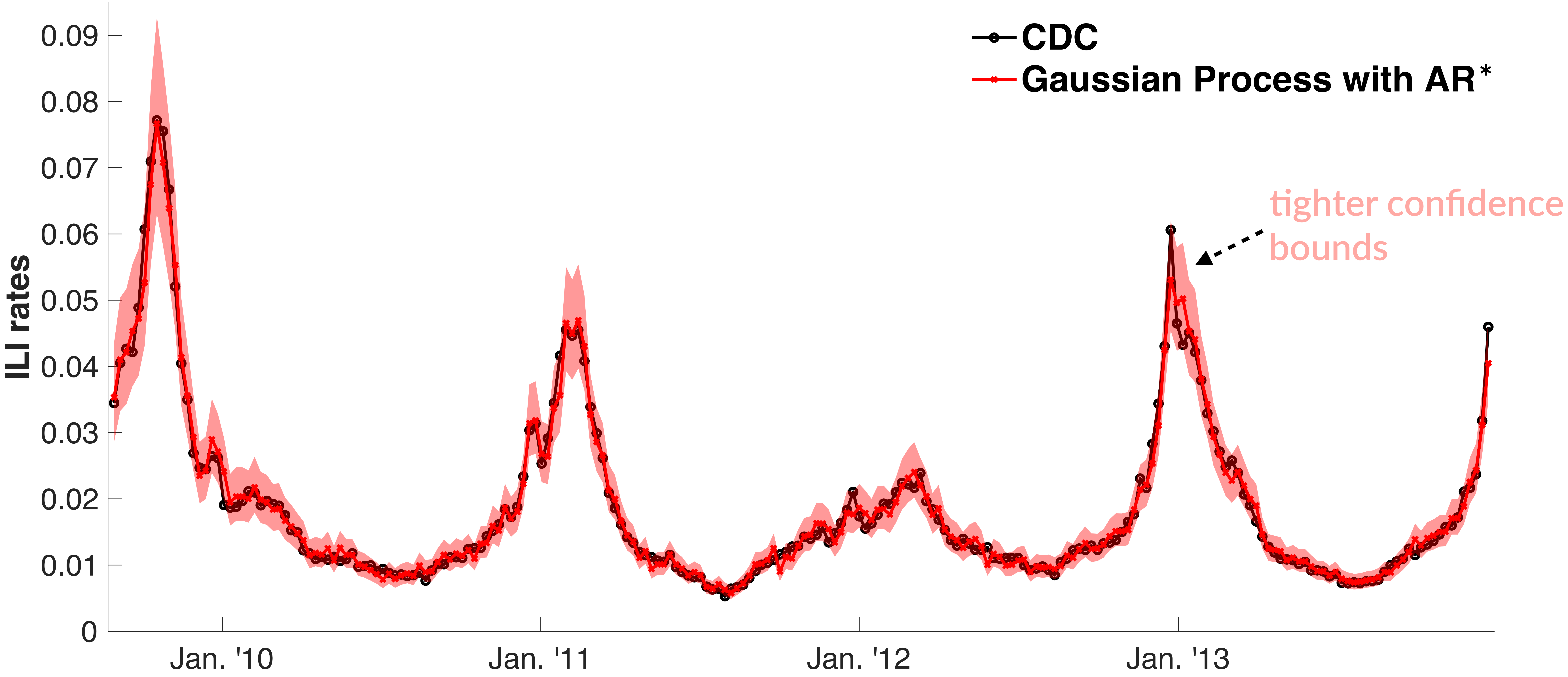
Estimating flu rates using web search activity (*US*)



- **Model:** *Gaussian Process* covariance functions on clusters (*temporal topics*) of search queries
- **42%** mean absolute error reduction compared to Google Flu Trends

Lampos, Miller, Crossan, Stefansen. *Sci. Rep.* (2015), doi:10.1038/srep12760

Estimating flu rates using web search activity (*US*)

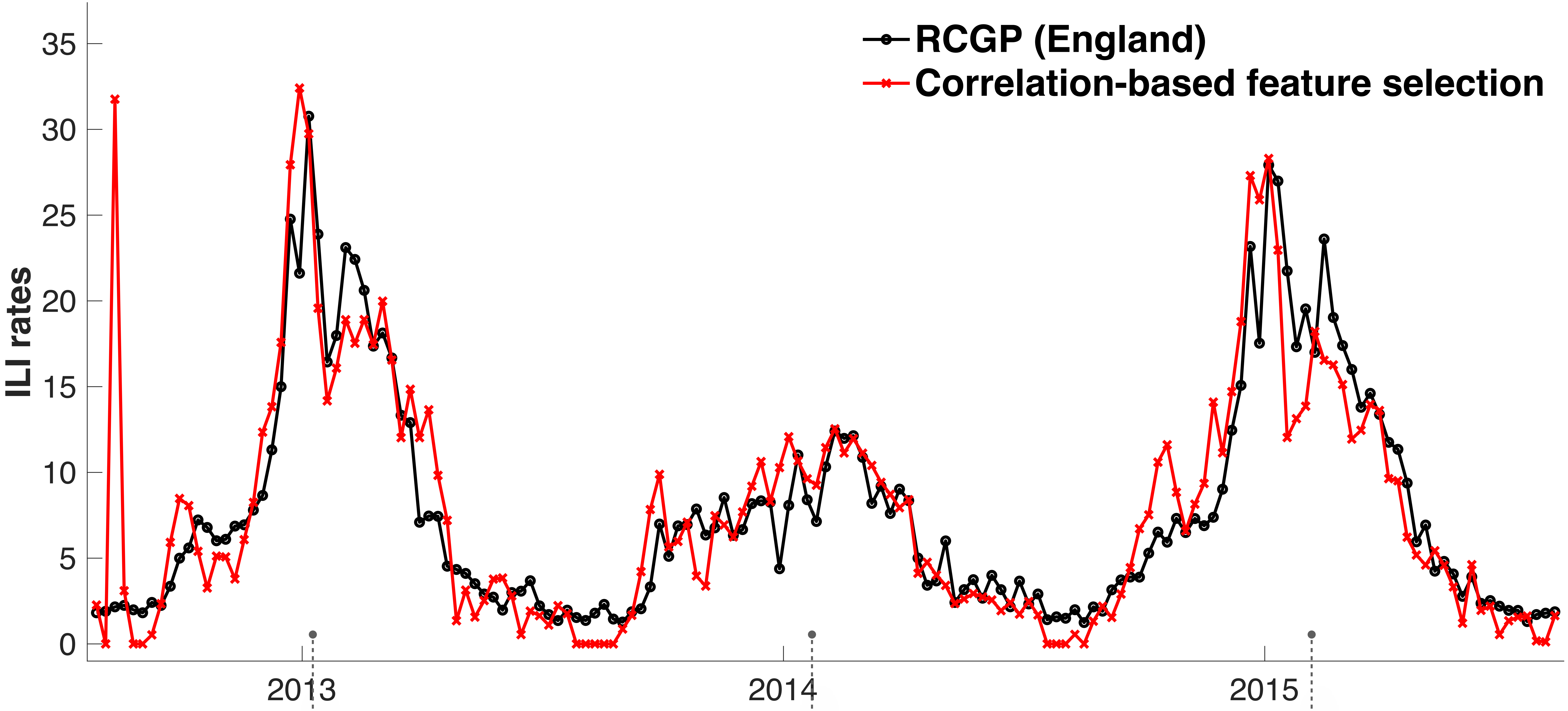


- **Model:** *Gaussian Process* flu rate estimates in ARMAX — 1 week lag for the CDC rates
- 27% mean absolute error reduction compared to using Google Flu Trends estimates in ARMAX

* *AR reinforces systemic biases (not always desirable)*

Lampos, Miller, Crossan, Stefansen. *Sci. Rep.* (2015), doi:10.1038/srep12760

Feature (*search query terms*) selection

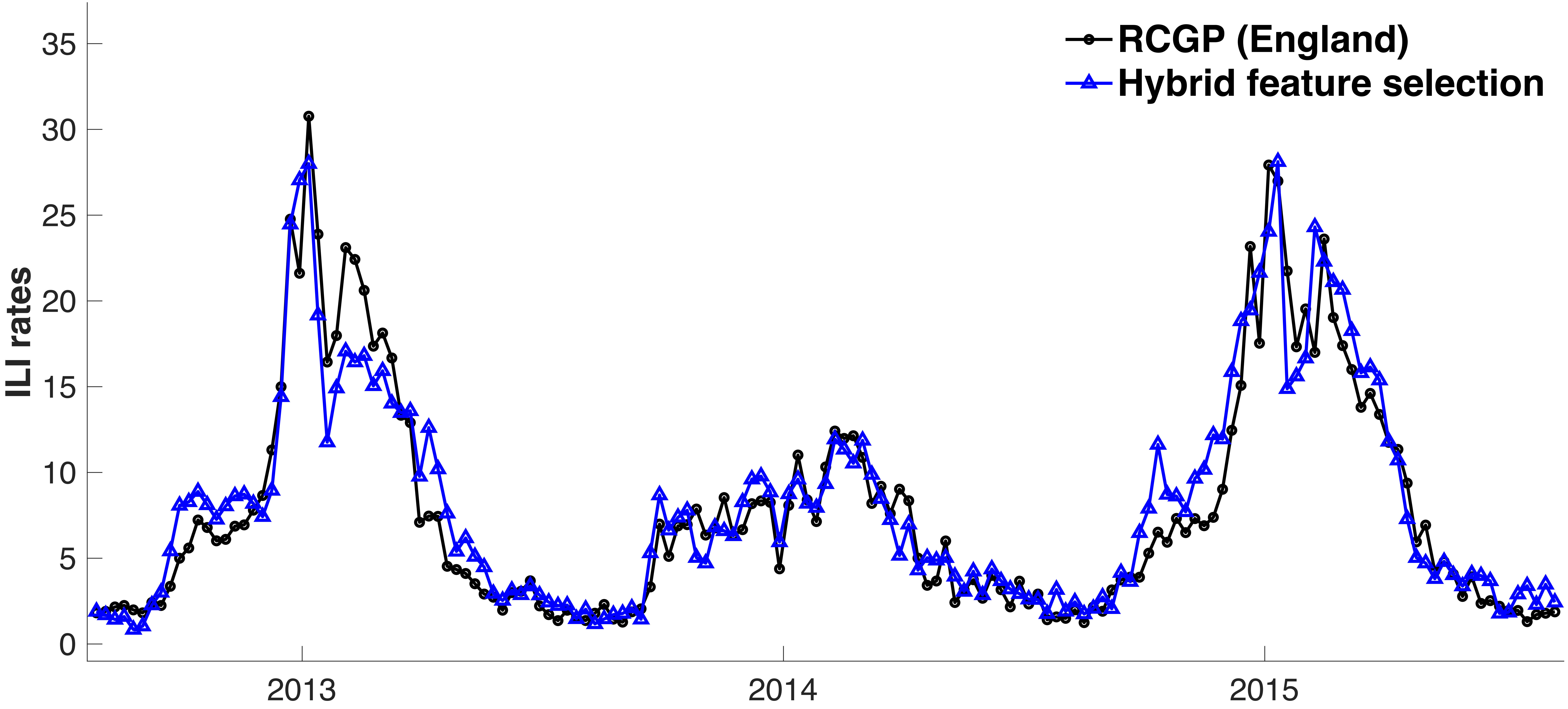


spurious search queries

'Prof. *Surname*', 'heal the world', 'heating oil', 'The White Company'	'Broken Sword', 'Snow Rock', 'gas homecare', 'North Face', 'low cost flights'	'Florence and the Machine Lungs', 'pleurisy symptoms', 'Boots sale', '*Name Surname*'

Lampos, Zou, Cox. *WWW* (2017), doi:10.1145/3038912.3052622

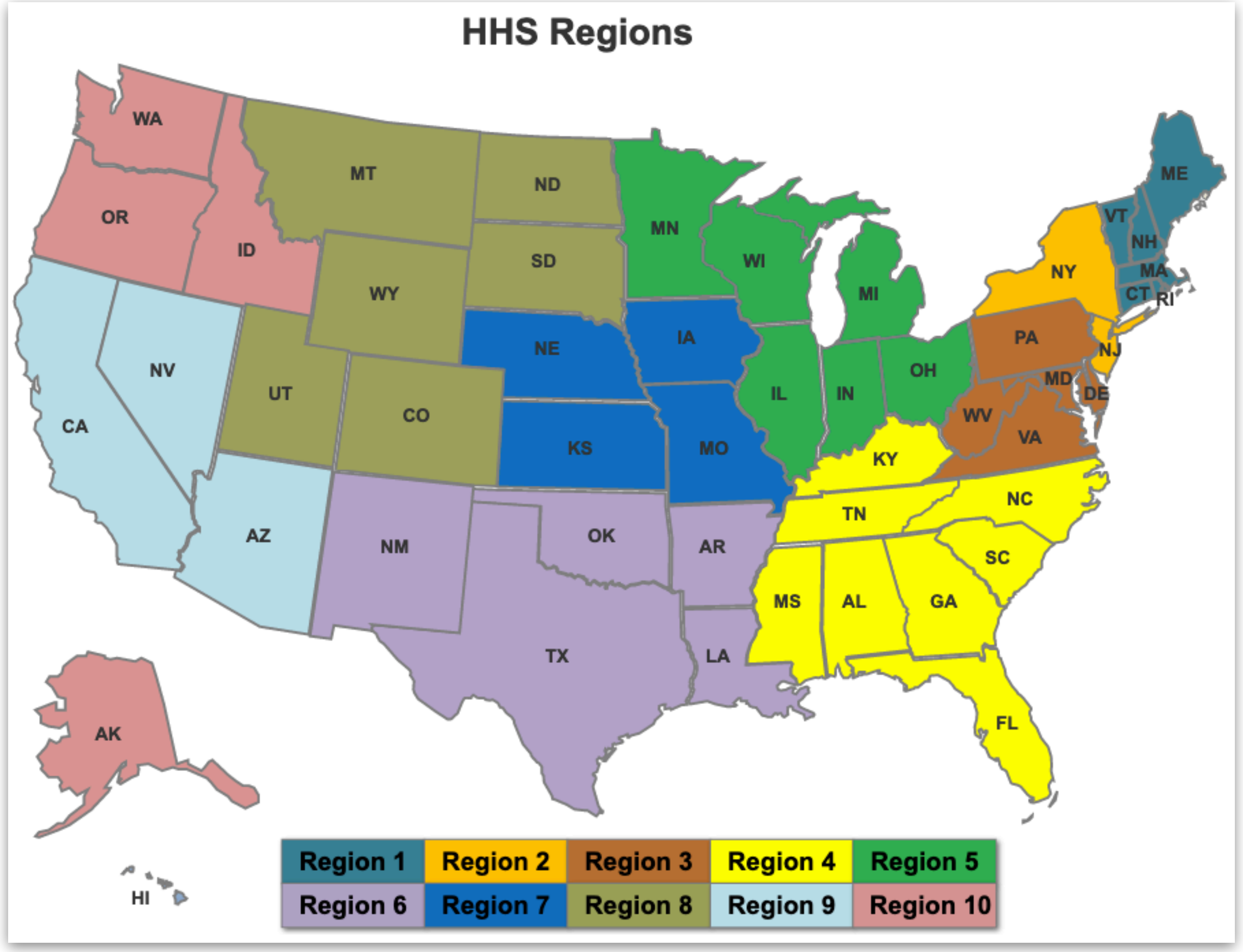
Feature (*search query terms*) selection

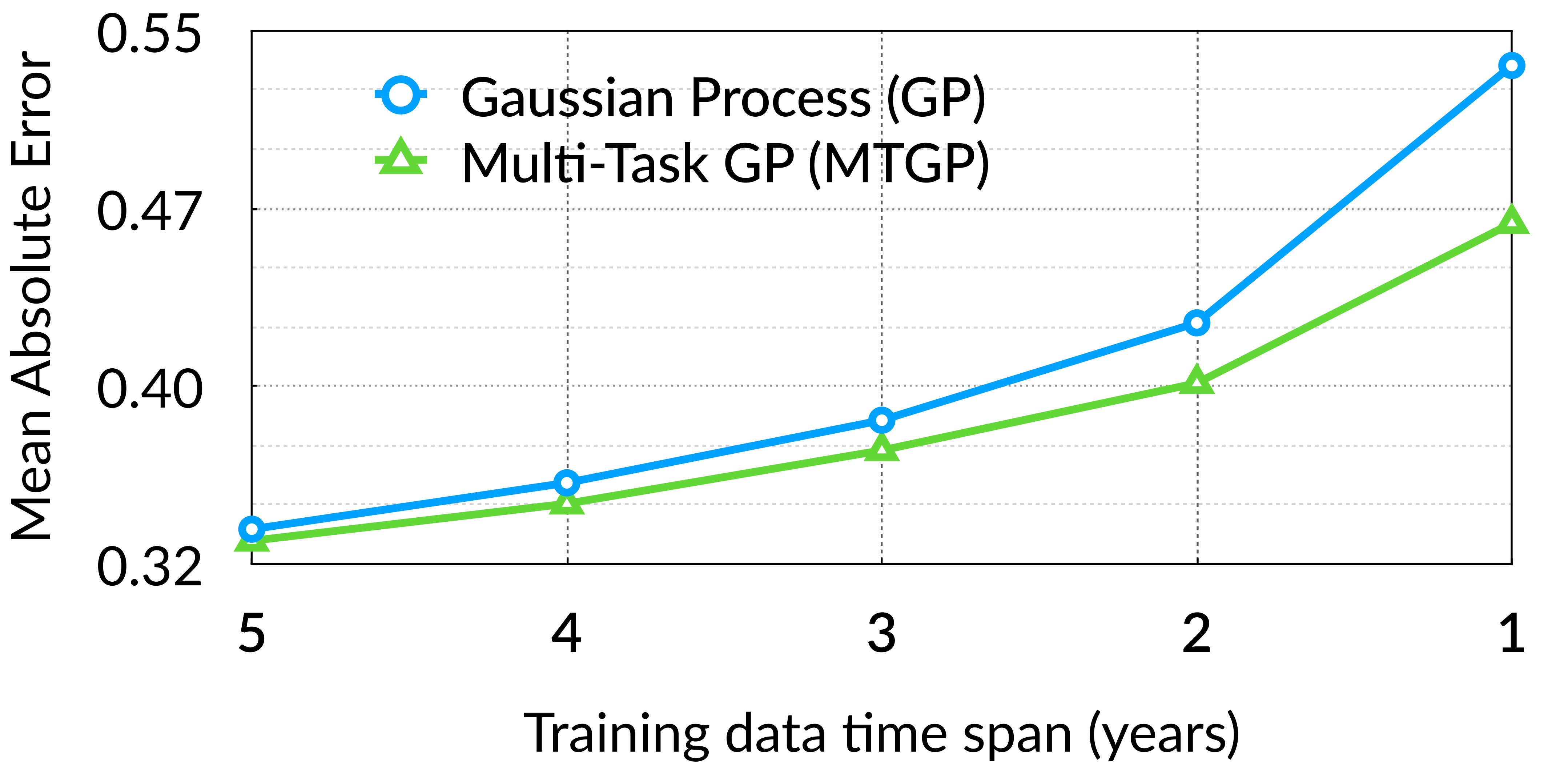


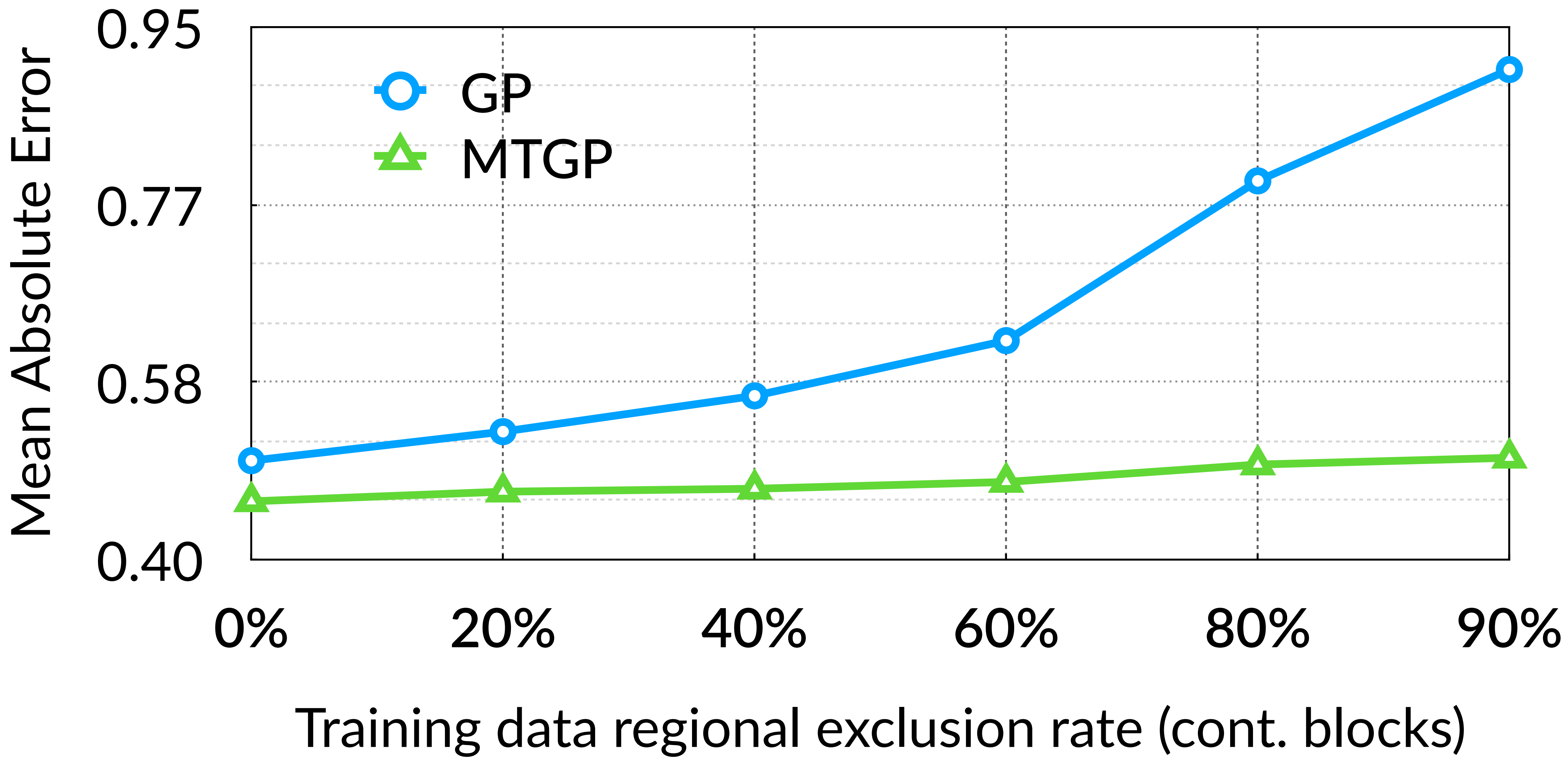
- Accuracy improved by **12.3%** (*mean absolute error*)
- Bivariate correlation of **.91**

Lampos, Zou, Cox. *WWW* (2017), doi:10.1145/3038912.3052622

Multi-task learning for robust *subregional* flu models (*US*)





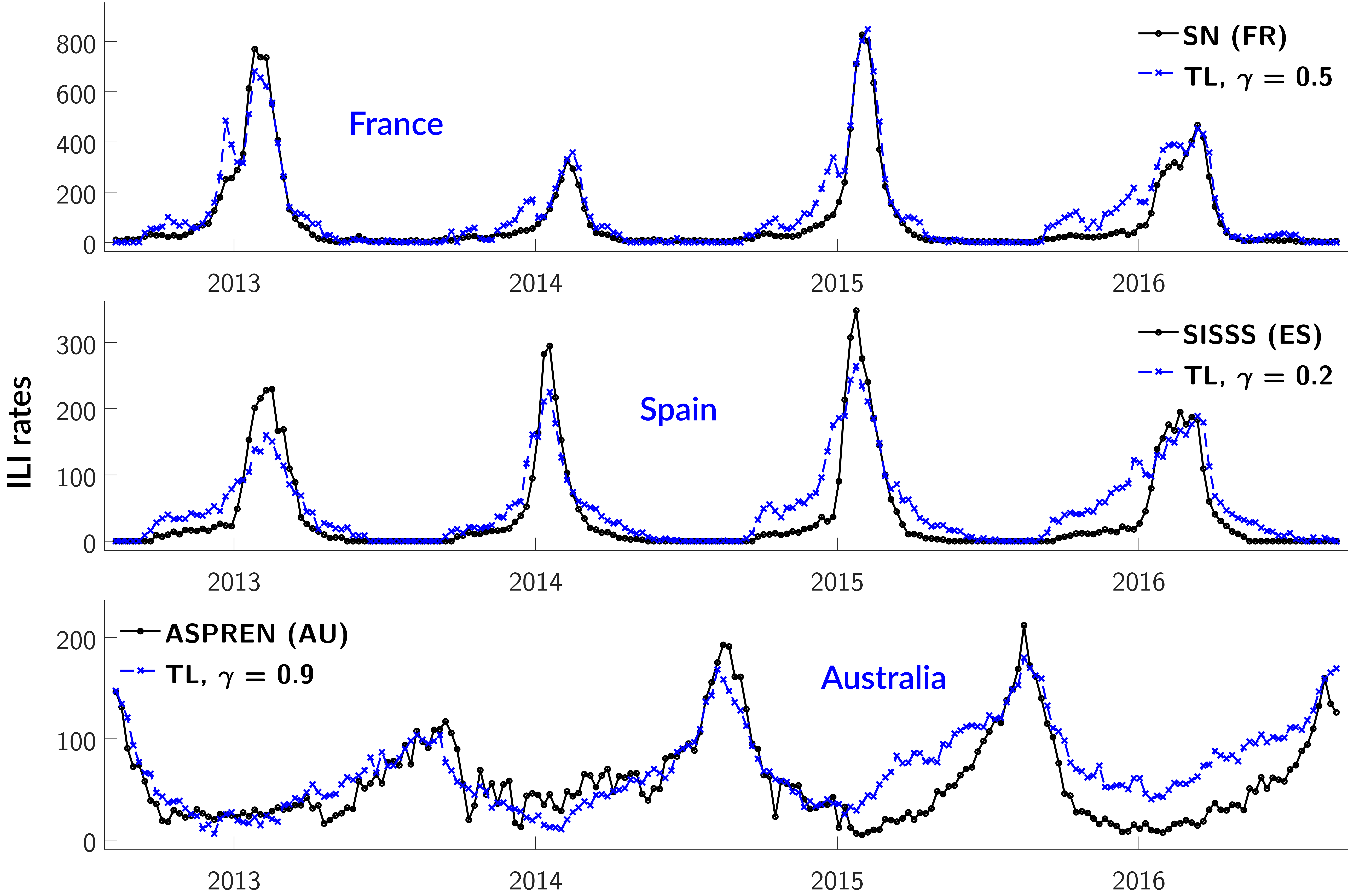


Multi-task learning models estimating flu rates in US regions:

- yield stat. sig. benefits when historical training data is limited (< 3 years)
- are mildly affected by data loss

Zou, Lampos, Cox. *WWW* (2018), doi:10.1145/3178876.3186050

Transferring flu models based on web search activity

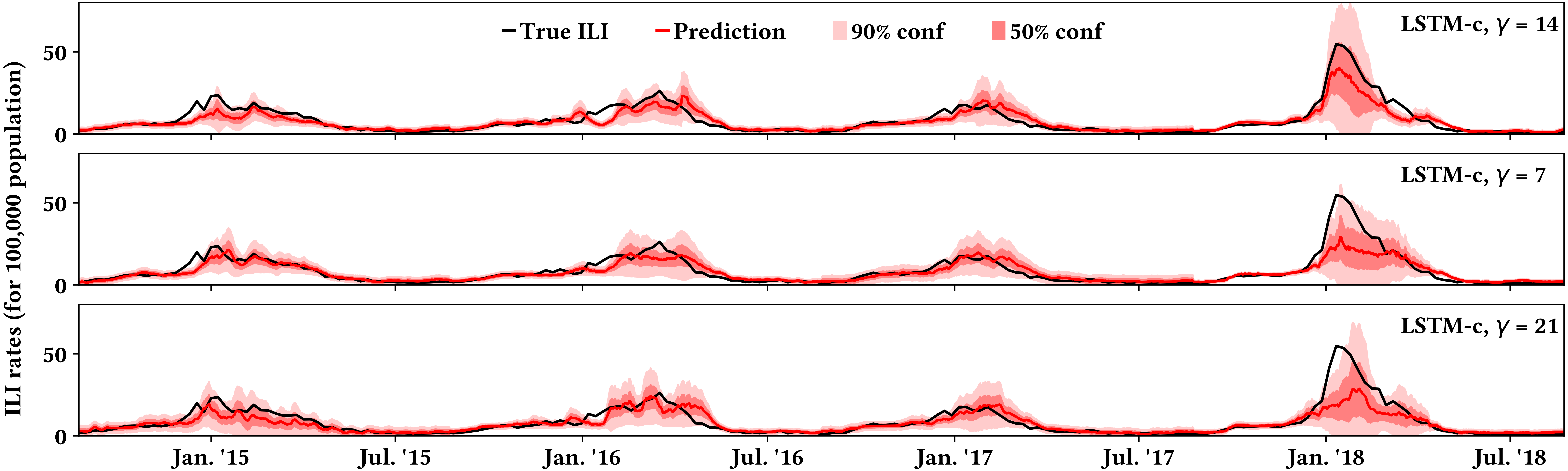


Transfer learning

- supervised flu model using data from the US
- transfer it to target countries (*no historical flu rates*, *no calibration*)
- γ controls the balance between the temporal and semantic similarity of source (US) and target searches

Zou, Lampos, Cox. *WWW* (2019), doi:10.1145/3308558.3313477

Forecasting flu rates using web search activity (*England*)

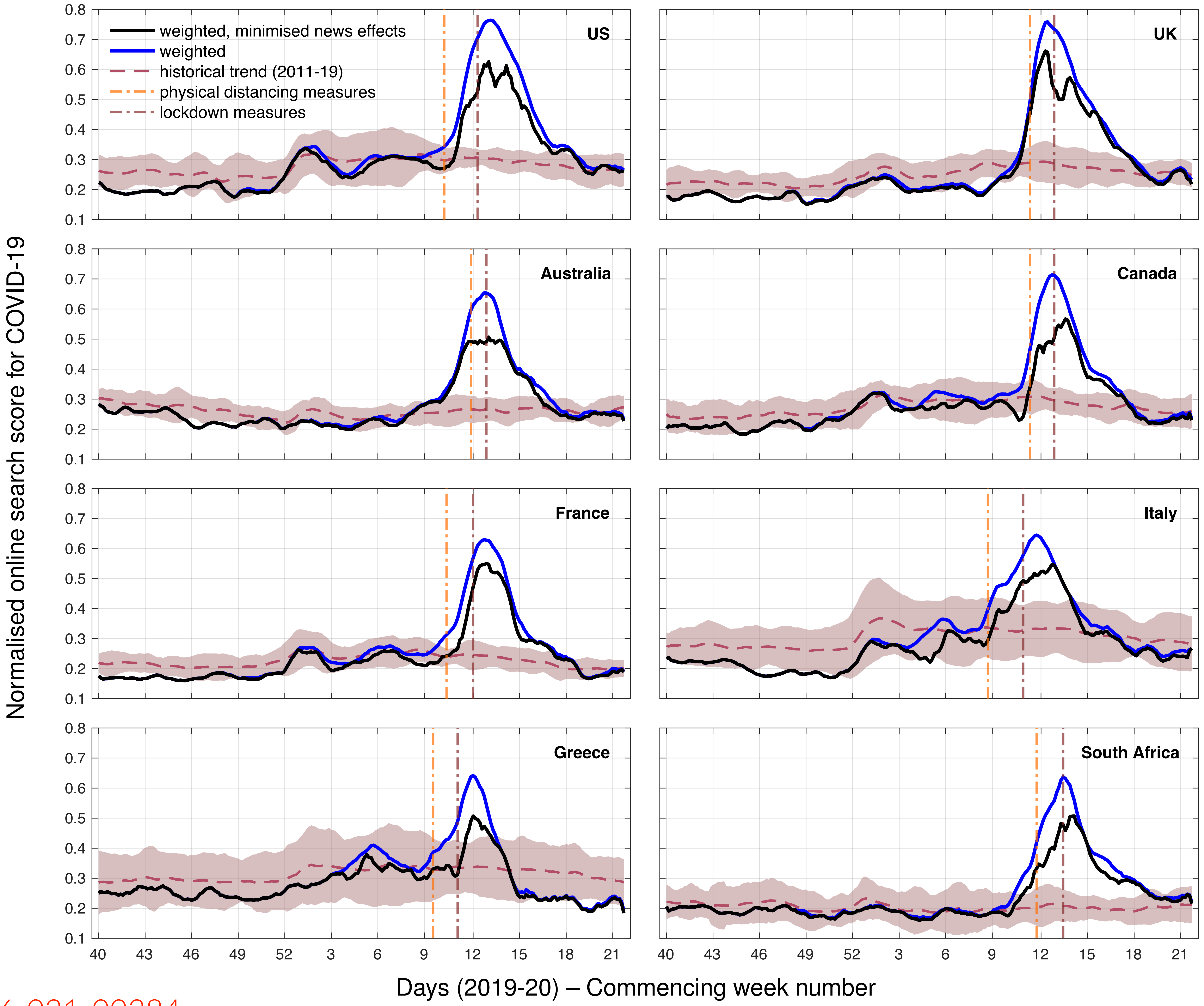


- **Bayesian Neural Networks** can provide forecasts (γ days ahead) with uncertainty without significant accuracy loss
- Combine model (*epistemic*) and data (*aleatoric*) uncertainties
- Web search activity data is key for improving forecasting accuracy

Morris, Hayes, Cox, Lampos. *arXiv preprint* (2021), arXiv:2105.12433

Modelling COVID-19 using web search activity

- Unsupervised learning
- 8 countries — national signals
- Attempt to reduce news media effects
- Scores reduced by 16.4% on average during peak moments



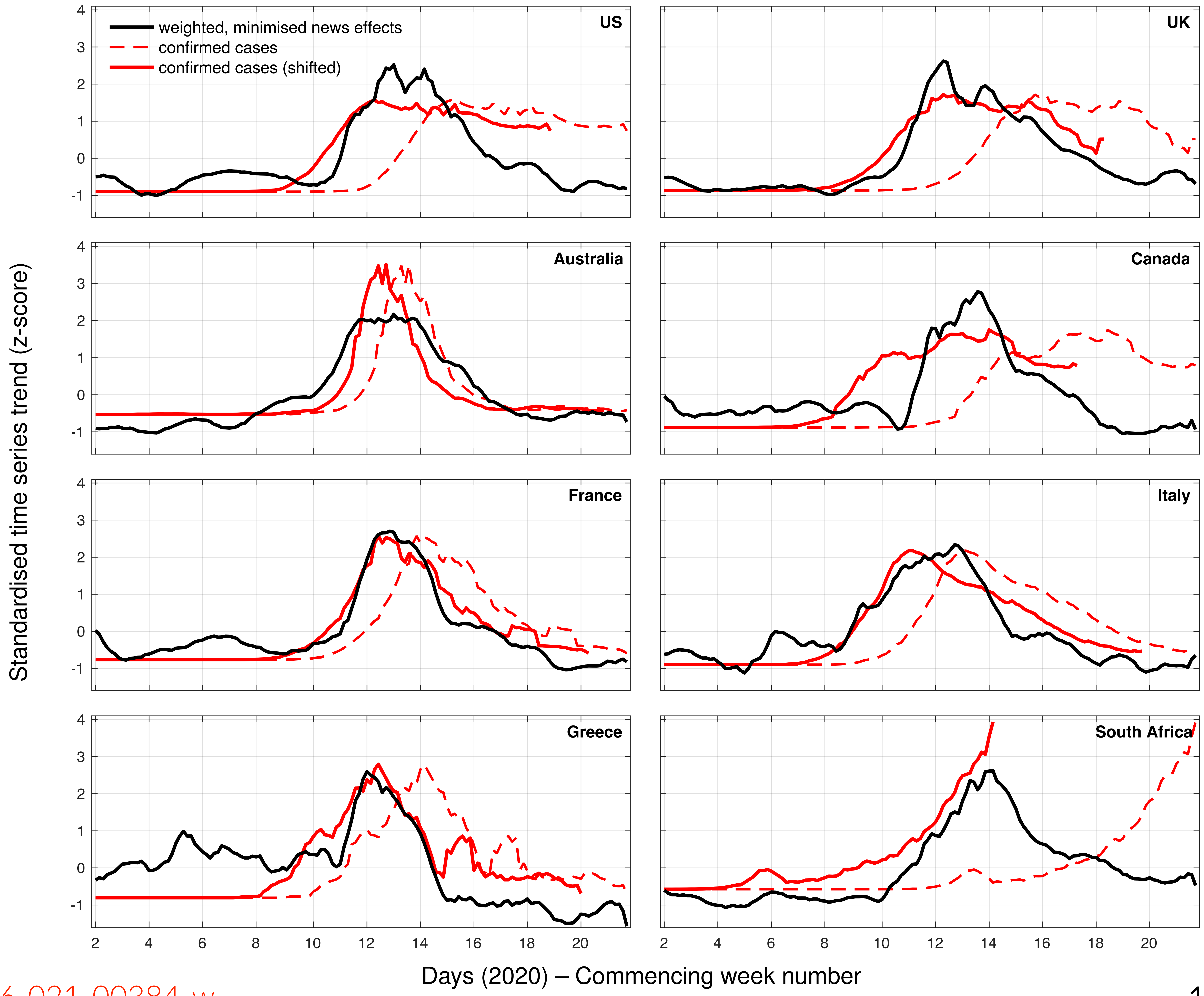
Lampos et al. *npj Digit. Med.* (2021), doi:10.1038/s41746-021-00384-w

Modelling COVID-19 using web search activity

Comparison with confirmed COVID-19 cases

- Average **early-warning**
  16.7 days, CI: 10.2–23.2 days
- Average bivariate correlation
  r = .83, CI: .74–.92

<u>Note</u>: *South Africa is excluded from this analysis*



Lampos et al. *npj Digit. Med.* (2021), doi:10.1038/s41746-021-00384-w

Modelling COVID-19 using web search activity

Comparison with deaths of people diagnosed with COVID-19

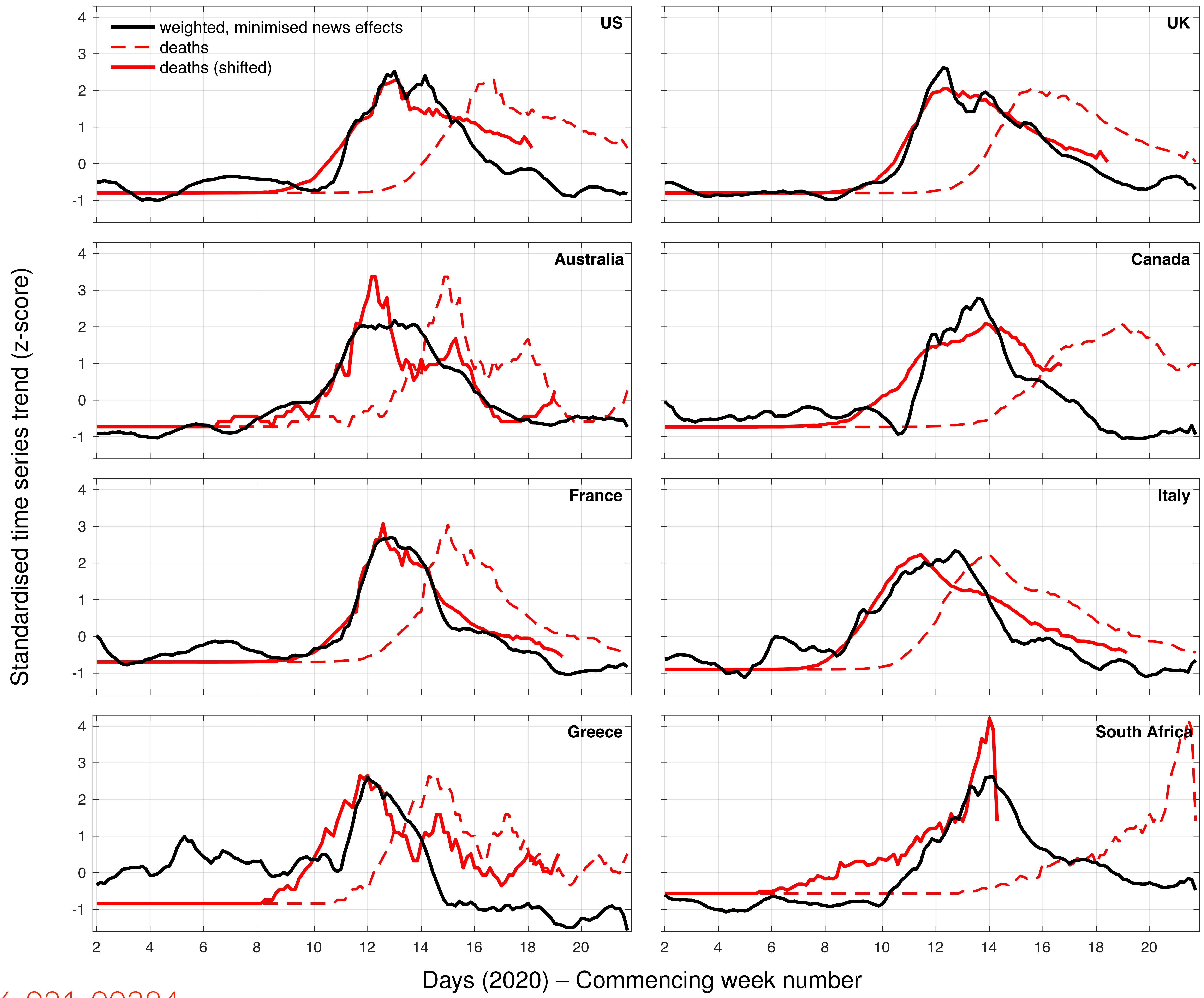
- Average **early-warning**

  22.1 days, CI: 17.4–26.9 days

- Average bivariate correlation

  r = .85, CI: .70–.99

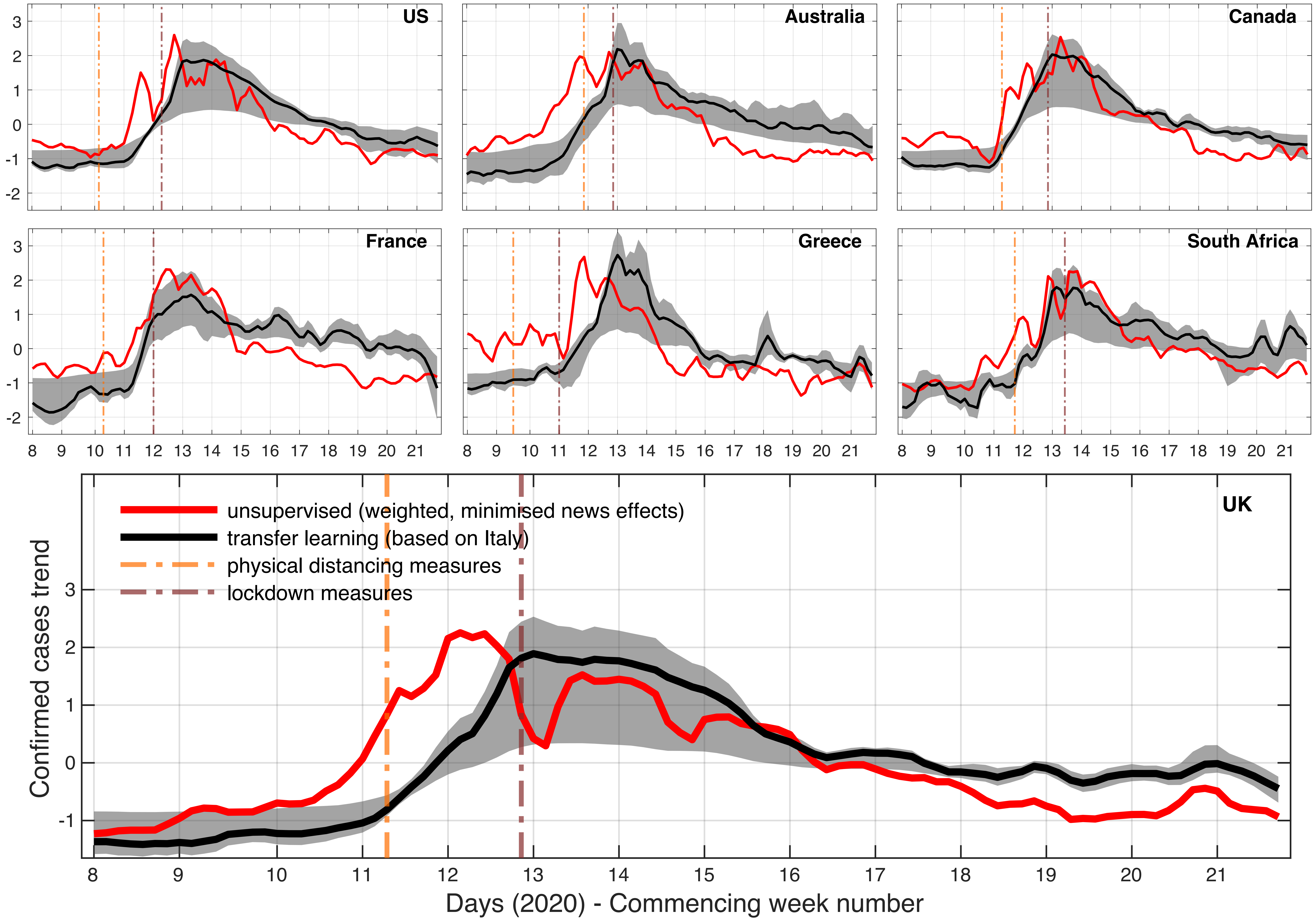
<u>Note</u>: *South Africa is excluded from this analysis*



Lampos et al. *npj Digit. Med.* (2021), doi:10.1038/s41746-021-00384-w

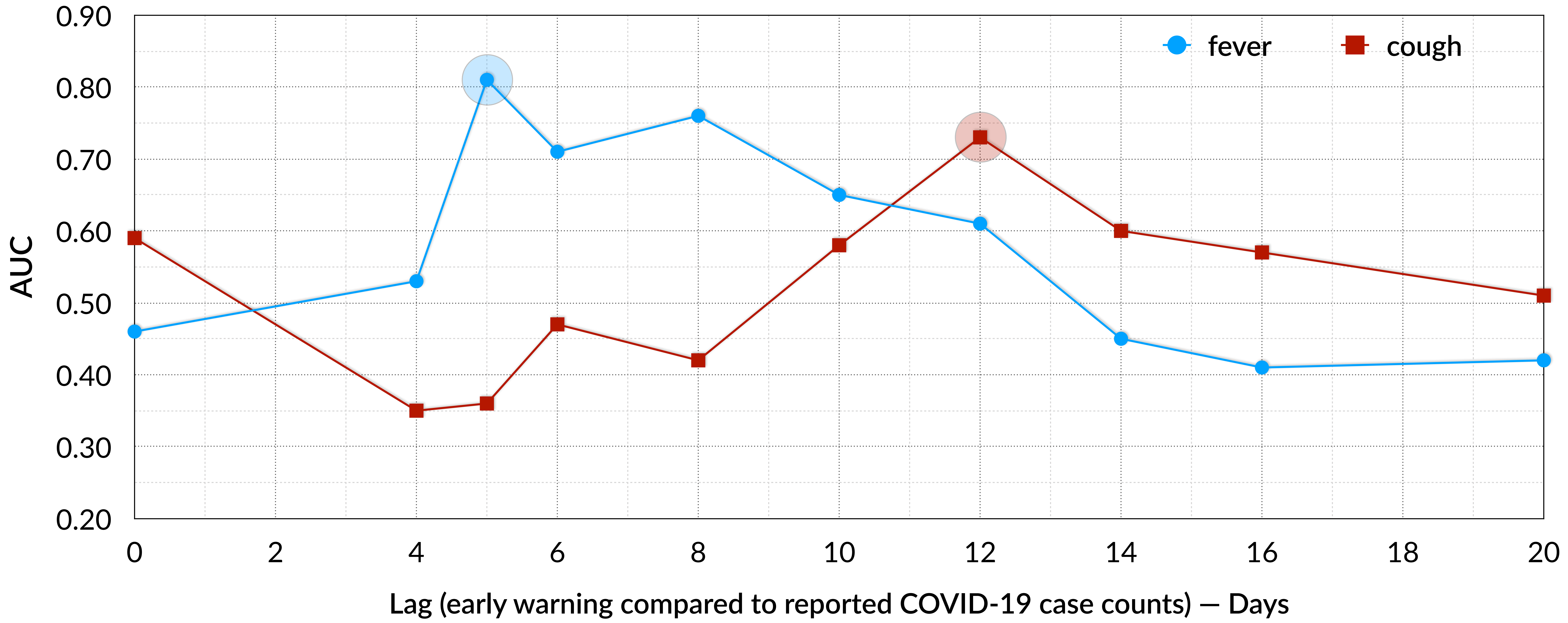
Modelling COVID-19 using web search activity

- Transfer a model trained on data from Italy (*confirmed cases, web search activity*)
- Unsupervised learning vs. transfer learning
  - 5 days earlier warning for the unsupervised models
  - curves are similar when aligned temporally



Lampos et al. *npj Digit. Med.* (2021), doi:10.1038/s41746-021-00384-w

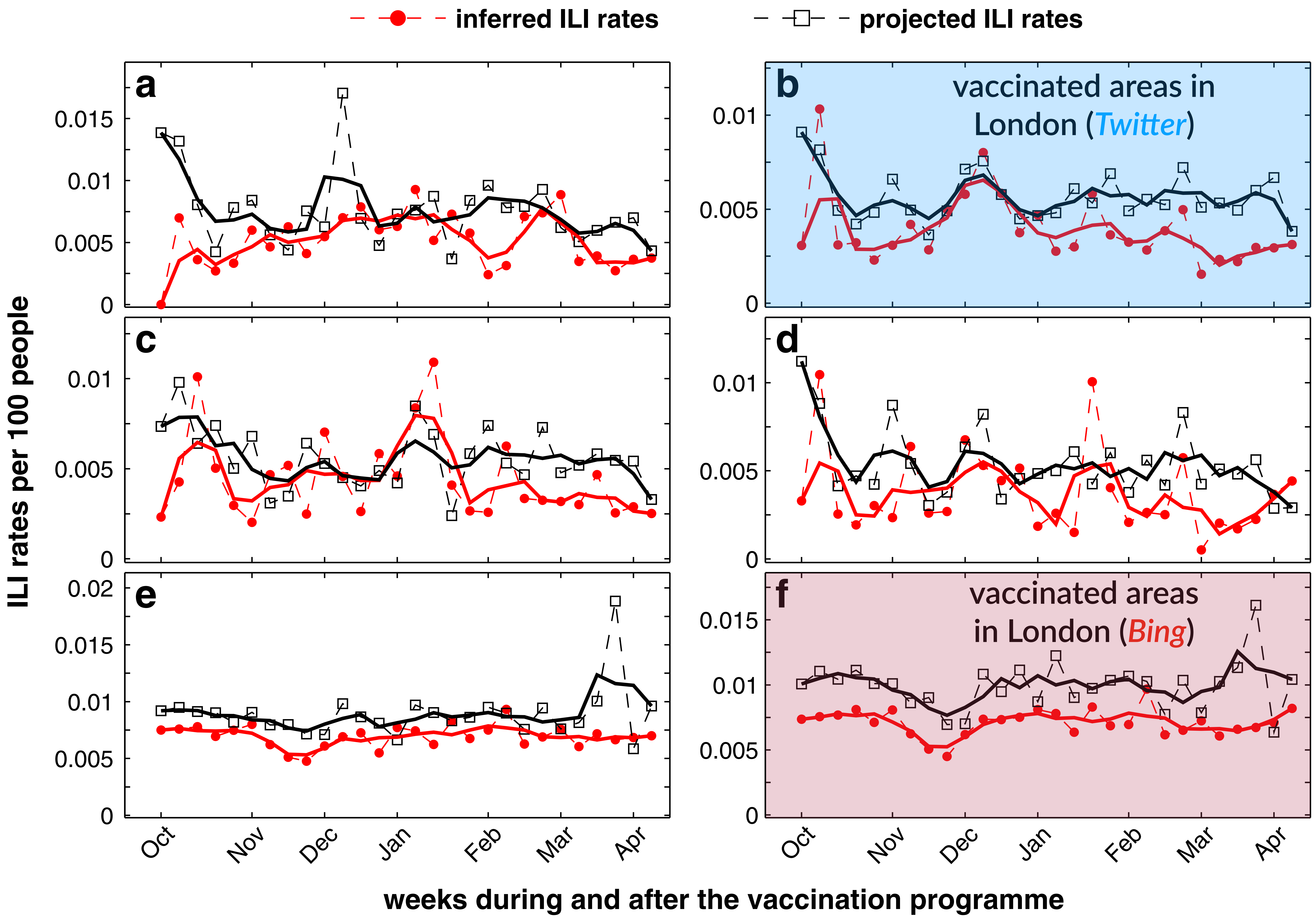
Using web search anomalies to predict COVID-19 outbreaks (*England*)



- Identify anomalies in web searches about COVID-19-related symptoms in local authorities in England (*difference-in-difference* methodology)
- Predict local outbreaks with a substantial early-warning — Caveat: *hard to assess accuracy*!

Yom-Tov, Lampos, Inns, Cox, Edelstein. *arXiv preprint* (2020) — *updated manuscript under review*, arXiv:2007.11821

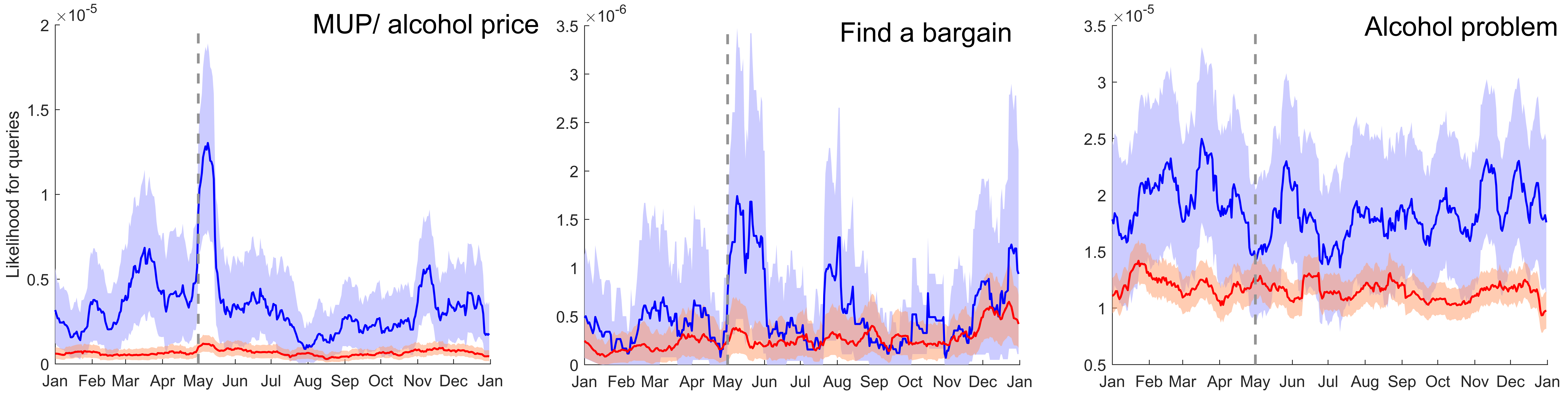
What is the impact of a vaccination campaign?



- Flu vaccination campaign by NHS/PHE (*schools*)
- Launched in a few areas — hard to assess the impact
- What would the flu rates be had the vaccination not taken place?
- **Twitter**: 32.8% reduction
- **Bing**: 21.1% reduction

Lampos, Yom-Tov, Pebody, Cox. *Data Min. Knowl. Disc.* (2015), doi:10.1007/s10618-015-0427-9
Wagner, Lampos, Yom-Tov, Pebody, Cox. *J. Med. Internet Res.* (2017), doi:10.2196/jmir.8184

Minimum Unit Pricing (MUP) of alcohol in Scotland



Search trends (Bing search engine) for alcohol-related topics in England (red) and Scotland (blue) during 2018.

- Alcohol MUP is a public health intervention aimed at reducing alcohol-related ill health in Scotland
- Search trends reflect on the policy introduction (May 1, 2018)
- Attempts to buy cheaper alcohol, circumvent the policy, no observable impact (*at the time of the analysis*)

Leon et al. *Addiction* (2021), doi:10.1111/add.15388

Estimation of secondary attack rates (SAR) from social media activity

Bill Lampos @lampos · Mar 30
My father has had a #COVID19 PCR test a couple days ago and it came back positive. Oh no!

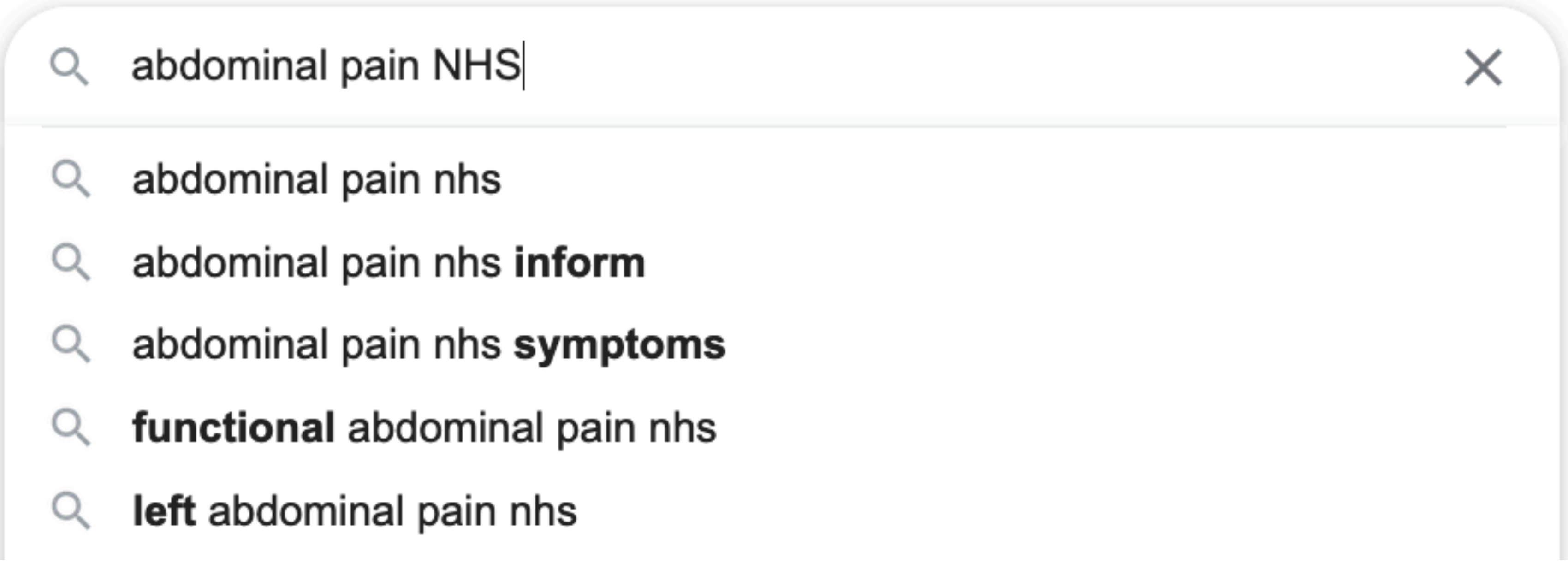
Bill Lampos @lampos · Apr 8
While my father seems to be OK now, I am starting to feel really unwell and my lunch did not taste right. #COVID19

- SAR: *probability of infection following contact with an infectious person*
- Original model applied to estimate familial (household) SAR (fSAR) for influenza in the UK
- Ongoing work to estimate fSAR for COVID-19

Yom-Tov, Cox, Lampos, Hayward. *Influenza Other Respir. Viruses* (2015), doi:10.1111/irv.12321
and an ongoing project led by Tomasz Czernuszenko

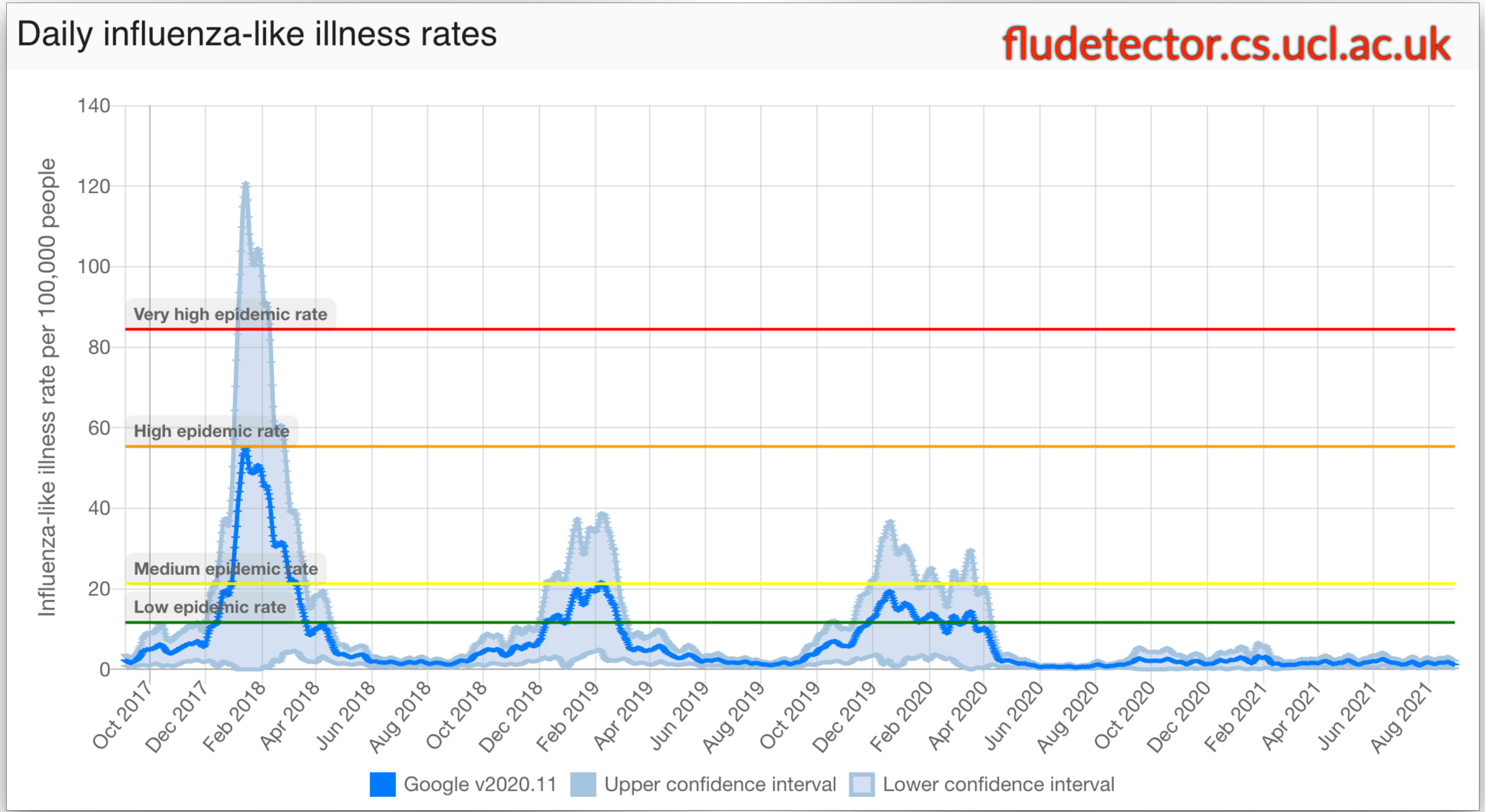
Gynaecological cancer risk prediction through web search activity

- Late diagnosis attributed to vague clinical presentation
- 10-year survival rate for ovarian cancer
  Stage I: 75%, Stage III: 21%, Stage IV: 5%
- Web search activity
  - early-warning to visit a specialist
  - investigate symptom patterns in larger cohorts
- Collaboration with Imperial College London NHS Trust
- Ethics approval, recruiting patients since late 2020
- Data: medical history, web search history



This is an ongoing project that currently involves S. Saso, J. Barcroft, D. Guzman, E. Yom-Tov, I. J. Cox, *and* myself

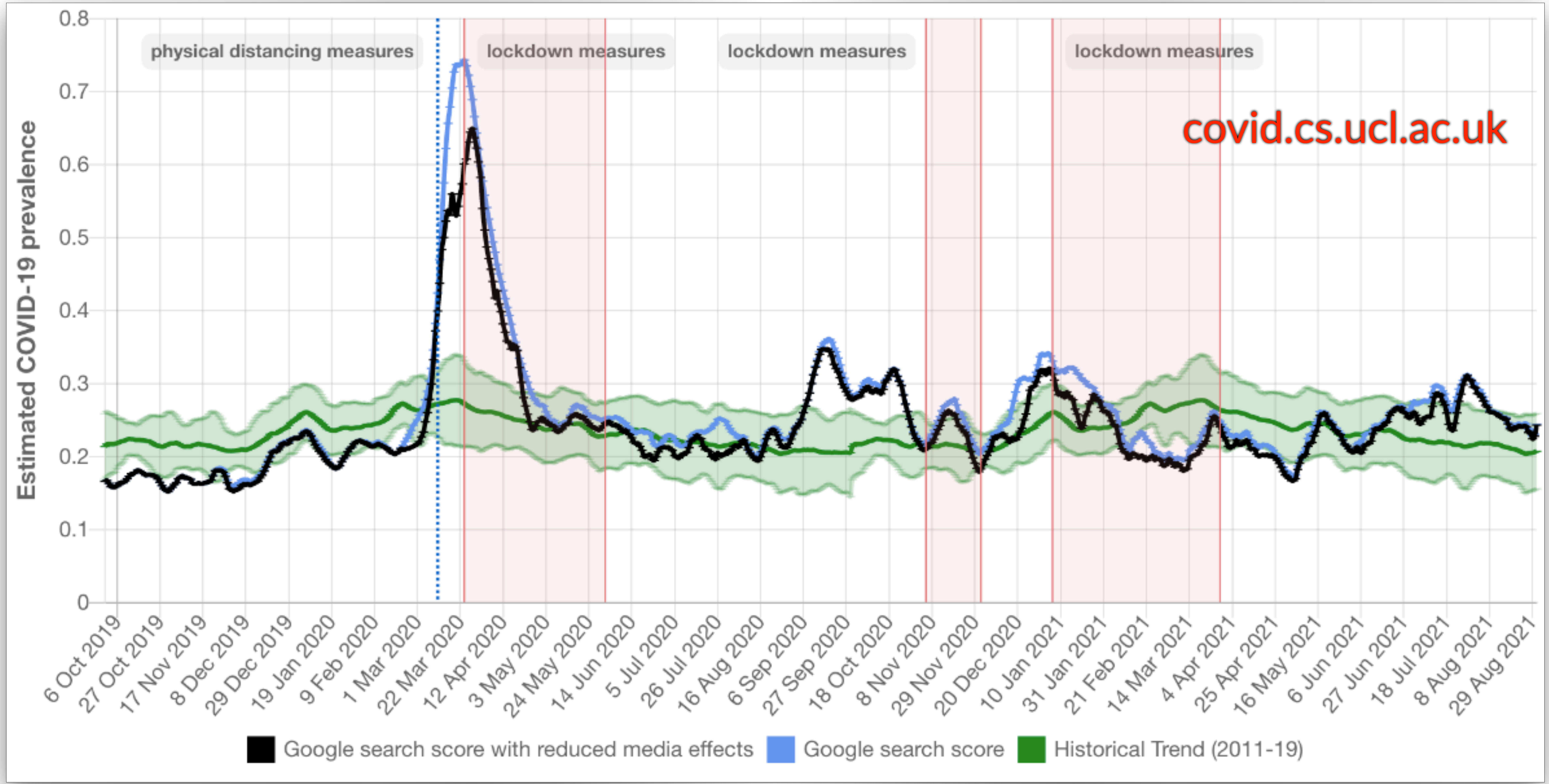
Real-world impact — Flu rate estimates (*England*)



Public Health England

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports

Real-world impact — COVID-19 prevalence estimates (*England*)



Public Health England

gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2021-to-2022-season

The future

- ✓ Integrate
- ✓ Evolve
- ✓ Expand
- ✓ Educate
- ✓ Regulate

Acknowledgements

Thank you!

Key collaborators
— Ingemar J. Cox (*UCL*)
— Elad Yom-Tov (*Microsoft Research*)
— Richard Pebody (*WHO*, previously *PHE*)

Collaborators (*in aforementioned research*)
Bin Zou, David Guzman, Michael Morris, Tomasz Czernuszenko, Aarzoo Dhiman, Michael Edelstein, Maimuna Majumder, Rachel McKendry, Srdjan Saso, Jennifer Barcroft, Moritz Wagner, Andrew Miller, David Leon, Anne Johnson, Evgeniy Gabrilovich, Andrew Hayward, Molebogeng X. Rangaka, Yohhei Hamada

Data providers
Google, Microsoft, Twitter, Royal College of General Practitioners, Public Health England, Centers for Disease Control and Prevention, NHS patients

Projects / Funding
i-sense (EPSRC), Virus Watch (MRC), Google

?

Non-traditional data-driven approaches to epidemiology

Vasileios Lampos

Department of Computer Science
University College London

www lampos.net

@lampos

Engineering and Physical Sciences Research Council

Medical Research Council

i-sense