Modelling influenza-like illness using online search

Vasileios Lampos

Computer Science, UCL

Engineering and Physical Sciences Research Council

@lampos

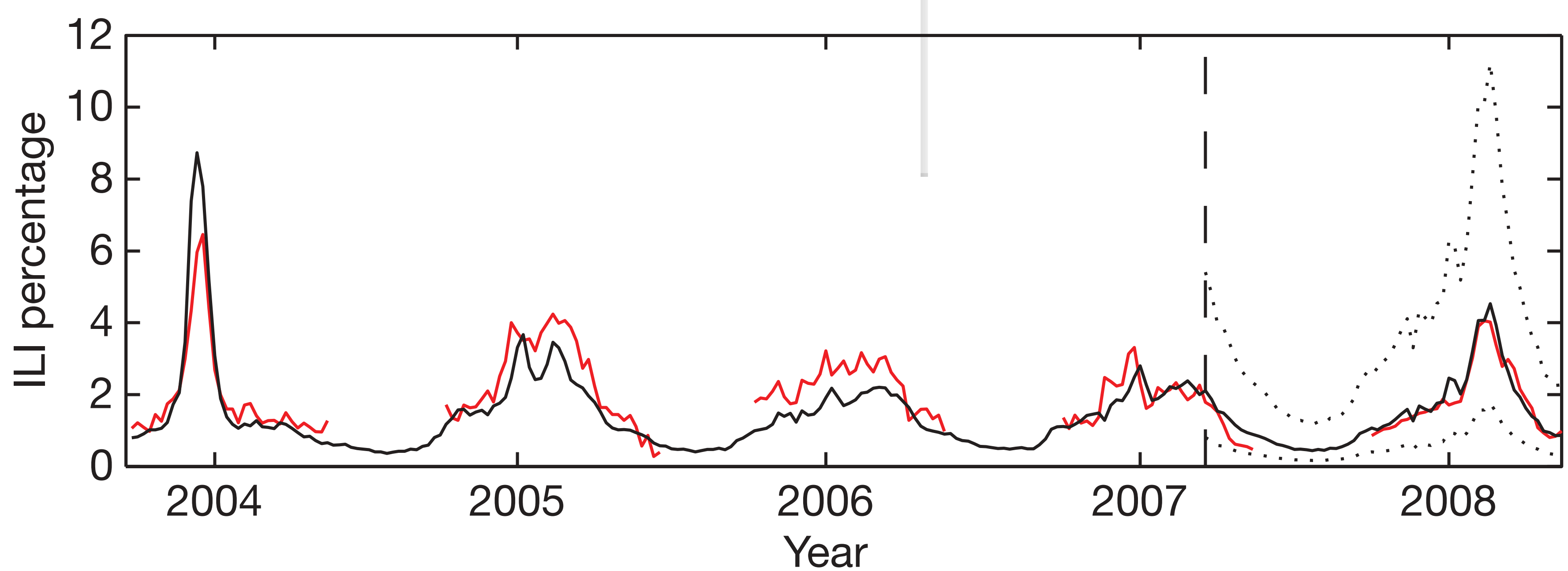
lampos.net

Mapping online search to flu estimates

Google

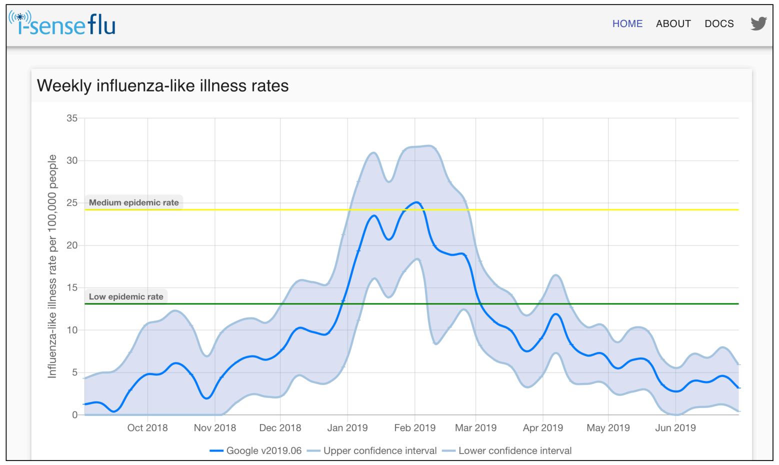
flu treatment

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



Why estimate flu rates from online search?

- *Complement* traditional syndromic surveillance
  - timeliness
  - broader demographic coverage, larger cohort
  - broader geographic coverage
  - not affected by closure days
  - lower cost
- Applicable to locations that *lack* an established healthcare system

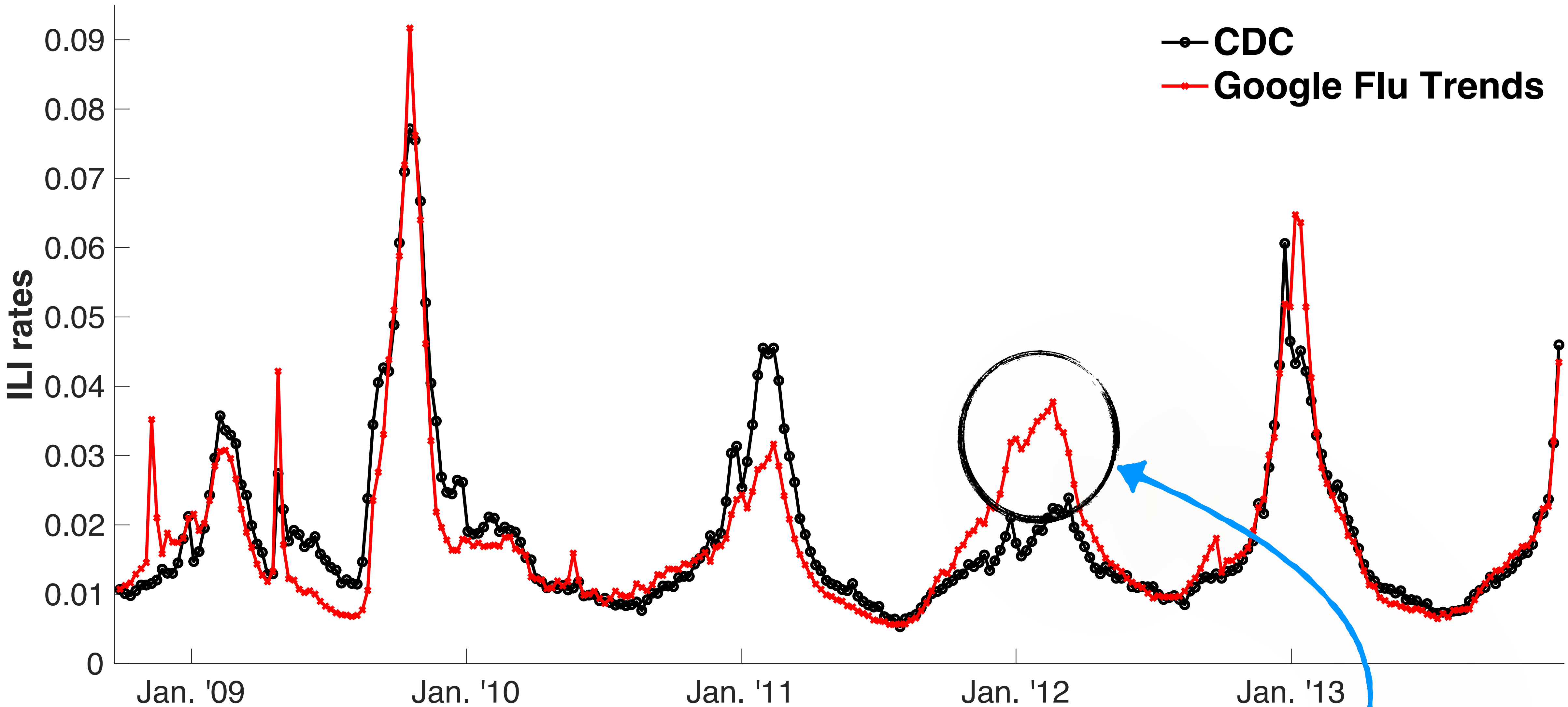


Google Flu Trends — *discontinued*

— popularising an ***established*** idea

Ginsberg *et al.* (2009); Eysenbach (2006); Polgreen *et al.* (2008)

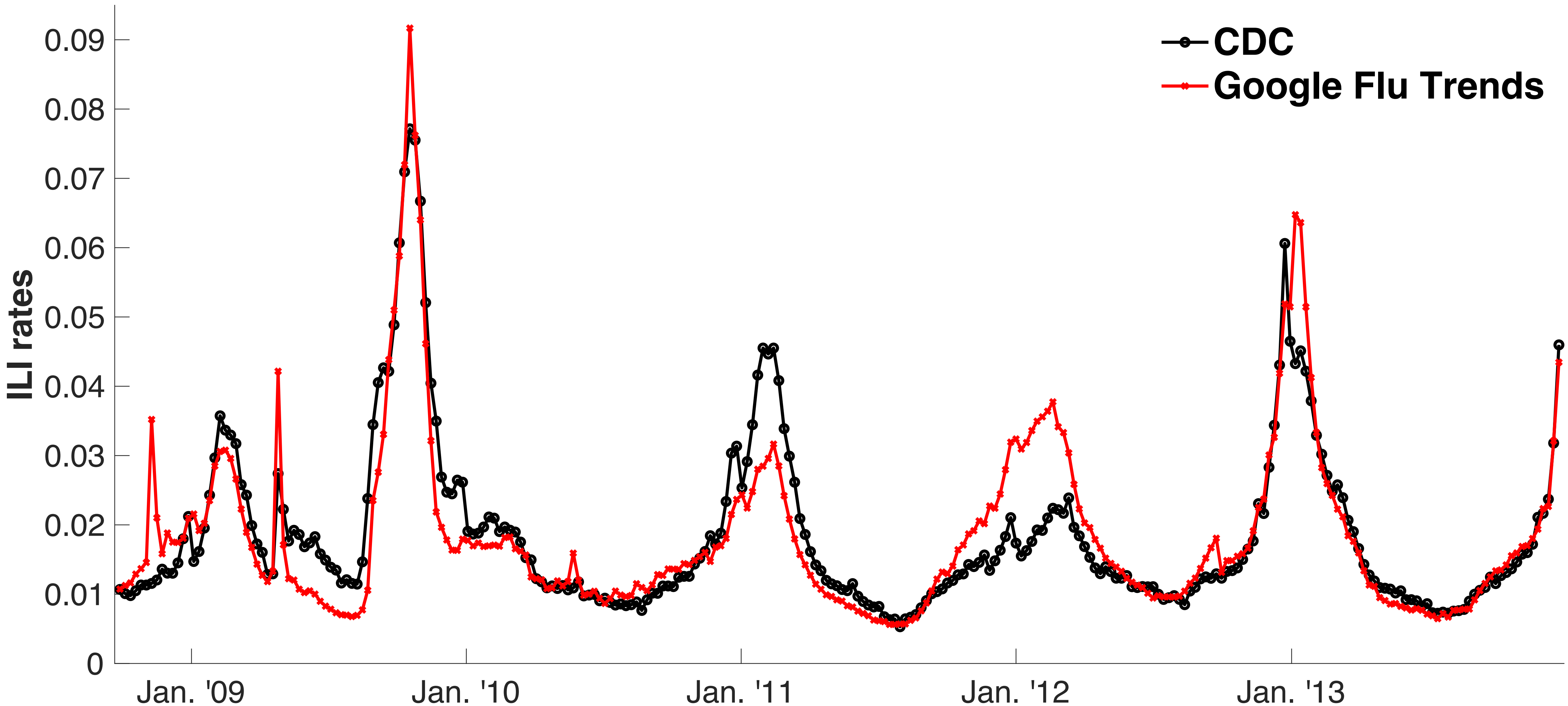
Google Flu Trends — *why did it fail?*



ILI rate $= \beta_0 + \beta_1 \times Q$,
where Q is the average query frequency

rsv **— 25%**
flu symptoms — 18%
benzonatate — 6%
symptoms of pneumonia — 6%
upper respiratory infection — 4%

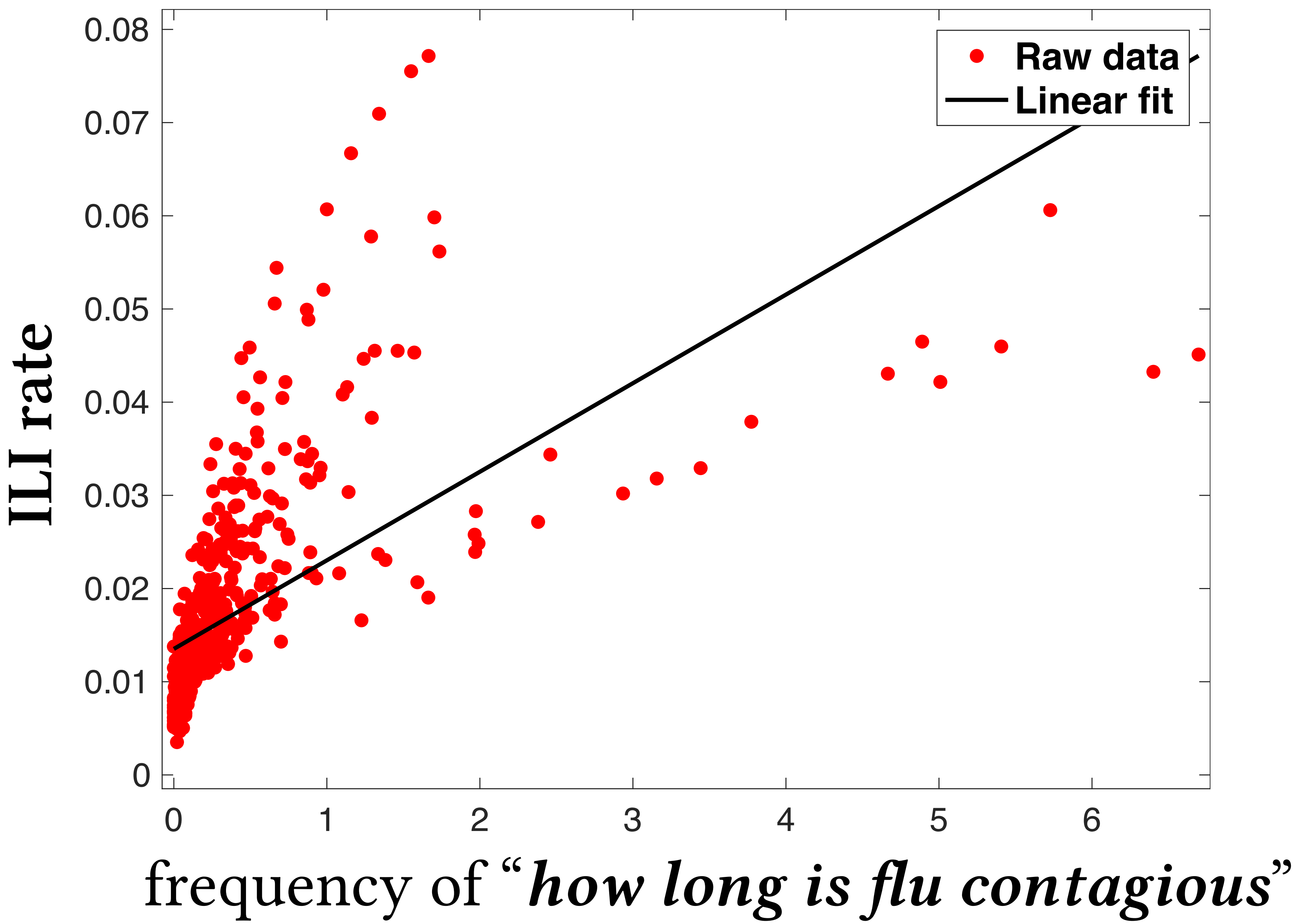
Google Flu Trends — *why did it fail?*



- **non-ideal query selection, model simplicity**
- **inappropriate evaluation** (*less than 1 flu season!*)

Multivariate, nonlinear, generative models

- Treat single search queries as **distinct variables**
- Model **nonlinearities**



Multivariate, nonlinear, generative models

- Treat single search queries as **distinct variables**
- Model **nonlinearities**
- Model groups of queries that share common temporal patterns

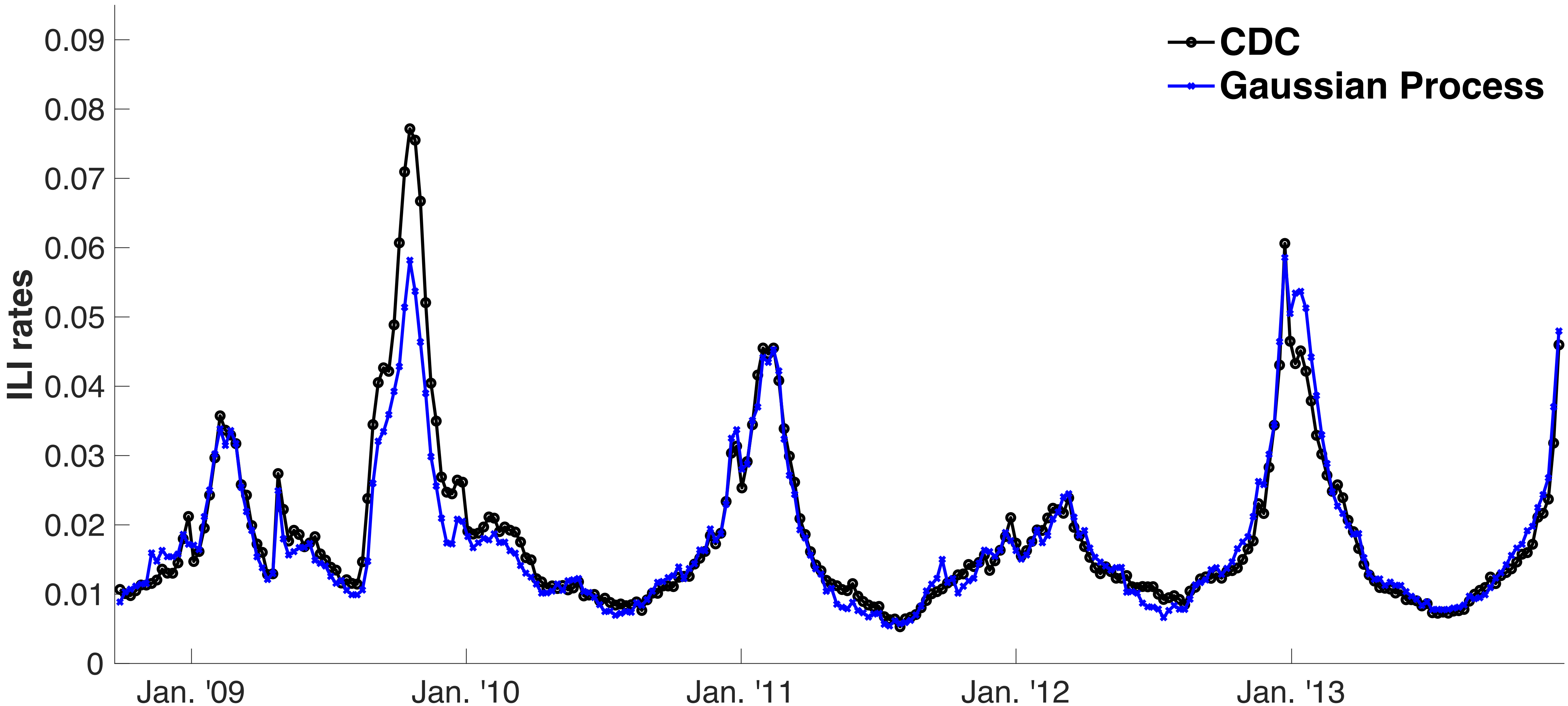
Gaussian Processes (GPs)

— distribution over functions that can explain the data

— allow some room for model interpretability

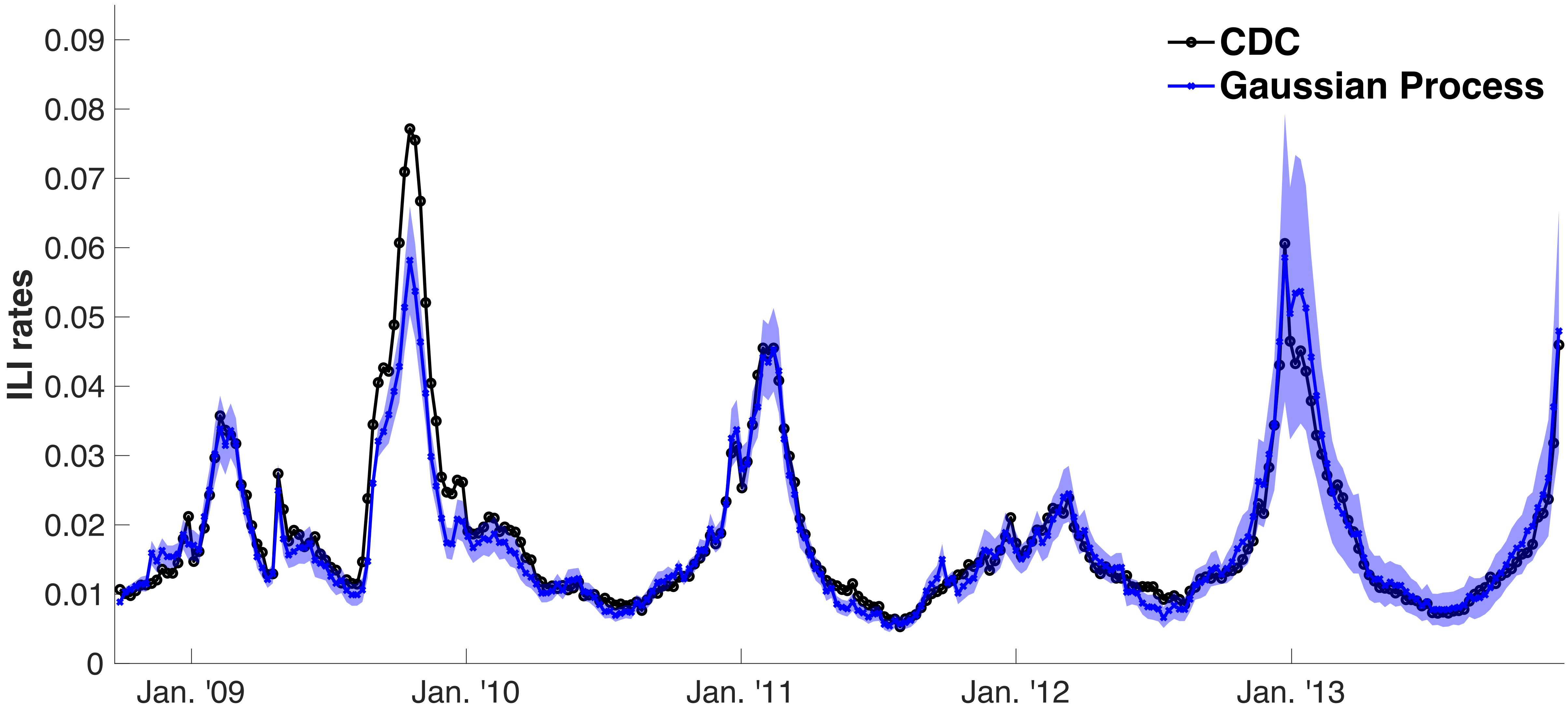
— can model uncertainty

Correcting the deficiencies of Google Flu Trends

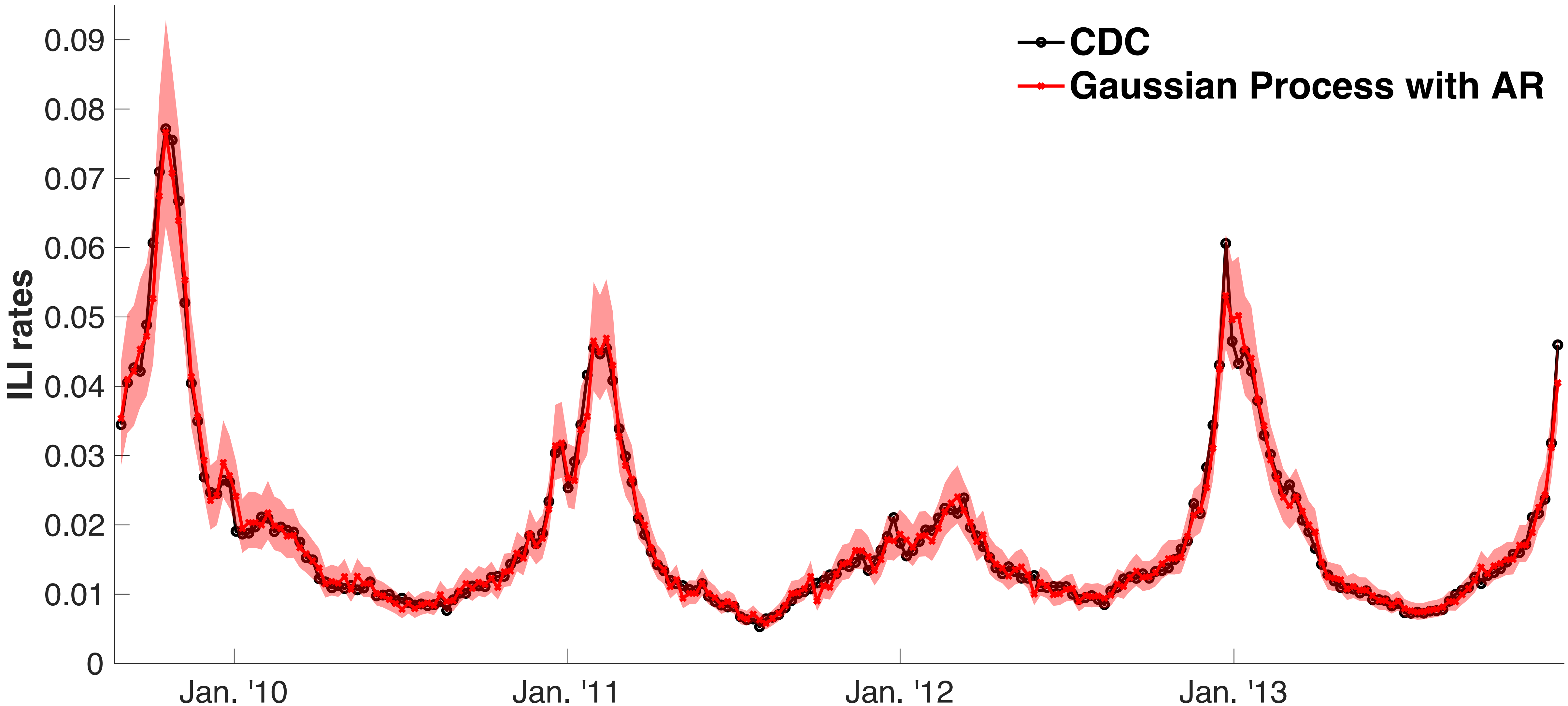


- **42%** mean absolute error reduction compared to Google Flu Trends
- **.95** Pearson correlation (*previously* .89) with CDC

Modelling uncertainty



Combining GPs with autoregression (AR)



- **1 week** delay in incorporating historical CDC estimates
- **27%** mean absolute error reduction over GFT with AR
- **52%** mean absolute error reduction over GP *without* AR
- **.99** Pearson correlation with CDC

Query selection based on *meaning*

- Select search queries based on their **semantic similarity** to the topic of flu

- Make this possible by using **word embeddings**, *i.e.* word representations in a common vector space — learn them using a corpus of 215 million tweets

Query selection based on *meaning*

- Select search queries based on their semantic similarity to the topic of flu
- Make this possible by using word embeddings, *i.e.* word representations in a common vector space
  - learn them using a corpus of 215 million tweets

Analogy: A (**is to**) $\longrightarrow$ B what X (**is to**) $\longrightarrow$?

Rome $\longrightarrow$ *Italy*	*London* $\longrightarrow$ [**UK**, Denmark, Sweden]
go $\longrightarrow$ *went*	*do* $\longrightarrow$ [**did**, doing, happened]
Messi $\longrightarrow$ *football*	*Lebron* $\longrightarrow$ [**basketball**, bball, NBA]
Elvis $\longrightarrow$ *Presley*	*Aretha* $\longrightarrow$ [**Franklin**, Ruffin, Vandross]
UK $\longrightarrow$ *Brexit*	*Greece* $\longrightarrow$ [**Grexit**, Syriza, Tsipras]
UK $\longrightarrow$ *Farage*	*USA* $\longrightarrow$ [**Trump**, Farrage, Putin]

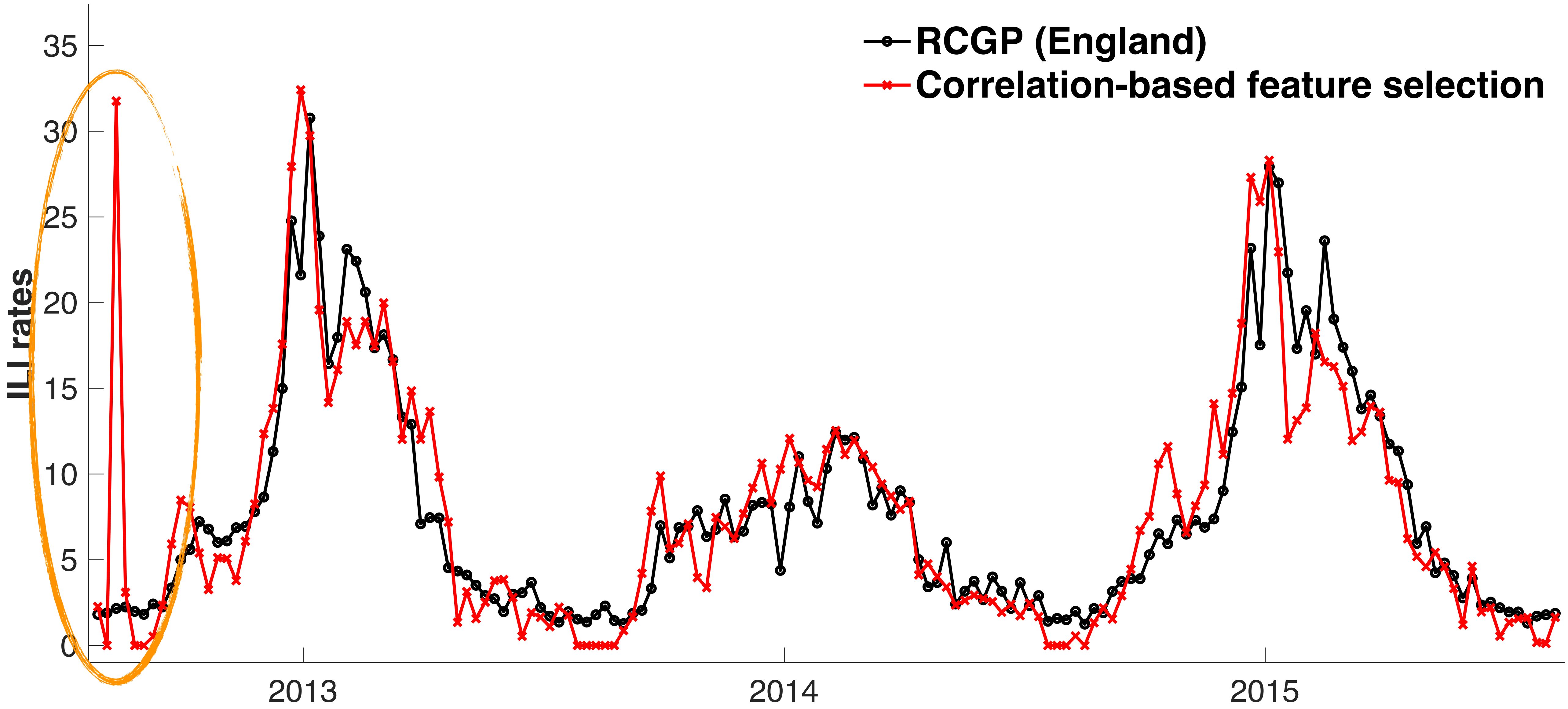
Query selection based on *meaning*

- Select search queries based on their **semantic similarity** to the topic of flu

- Make this possible by using **word embeddings**, *i.e.* word representations in a common vector space — learn them using a corpus of 215 million tweets

- **Combine** temporal correlation with semantic similarity (***hybrid similarity***) for optimal feature selection

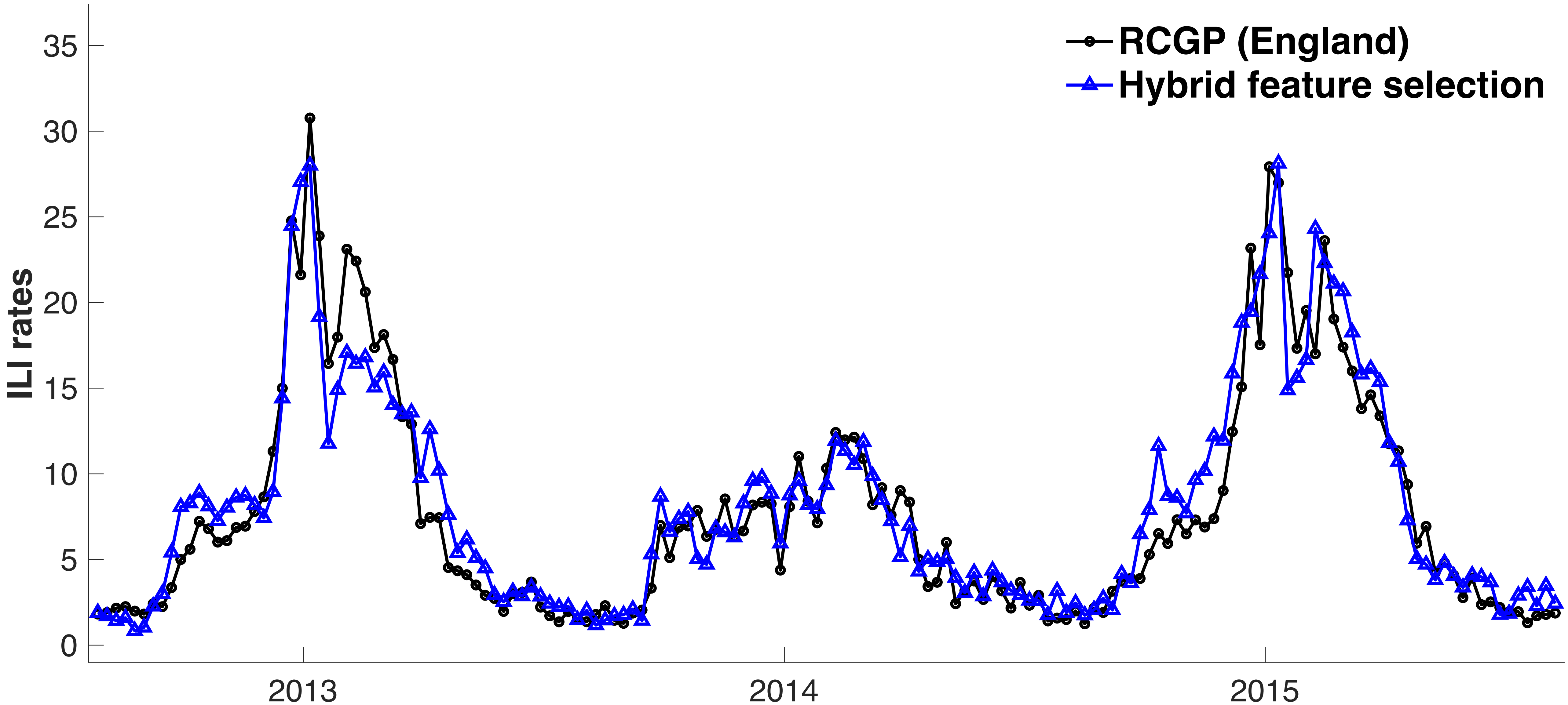
Query selection based on *meaning* — Results



Examples of spurious selected queries

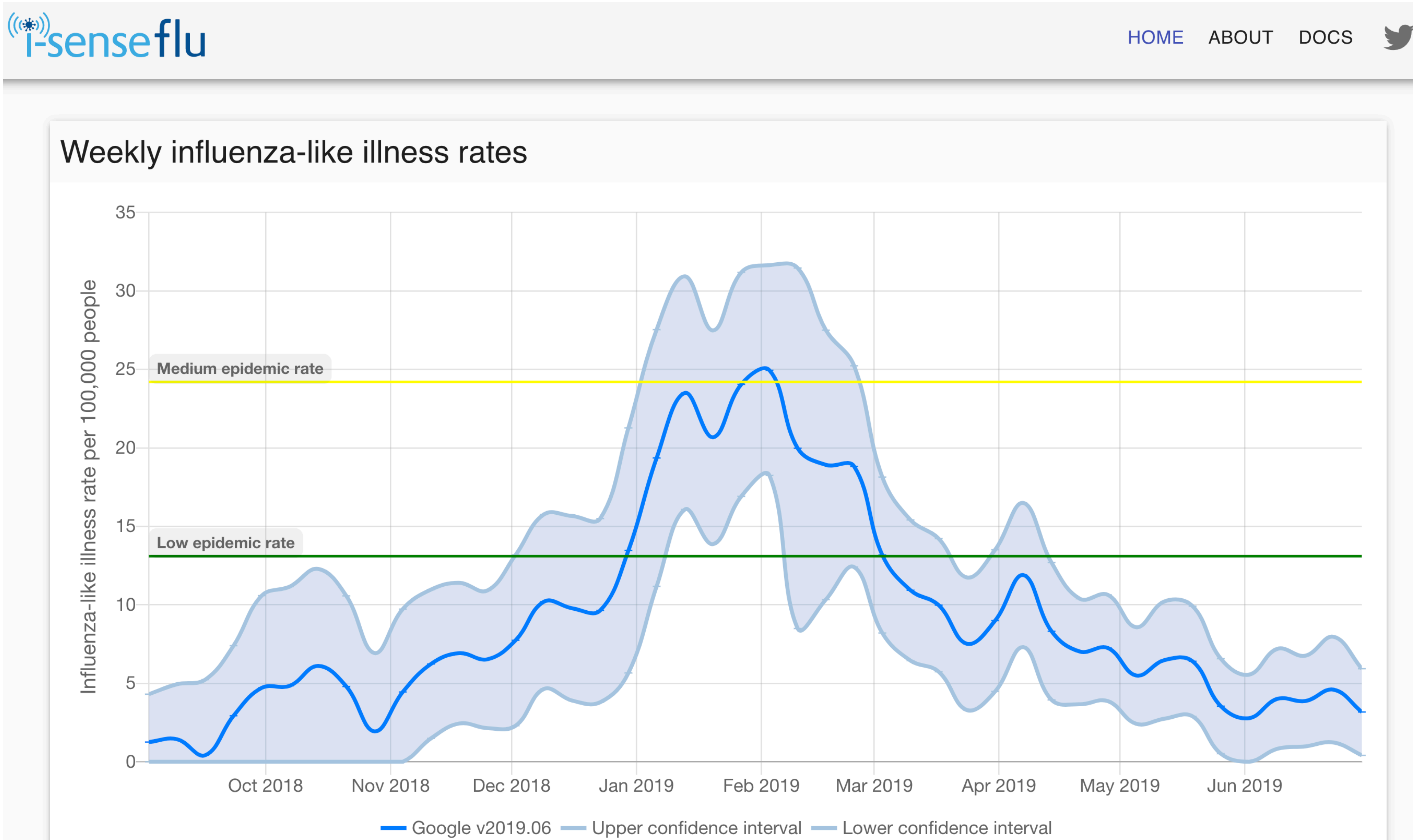
prof. *surname* (70%)
name surname (27%)
heating oil (21%)
name surname recipes (21%)
blood game (12.3%)
swine flu vaccine side effects (7.2%)

Query selection based on *meaning* — Results



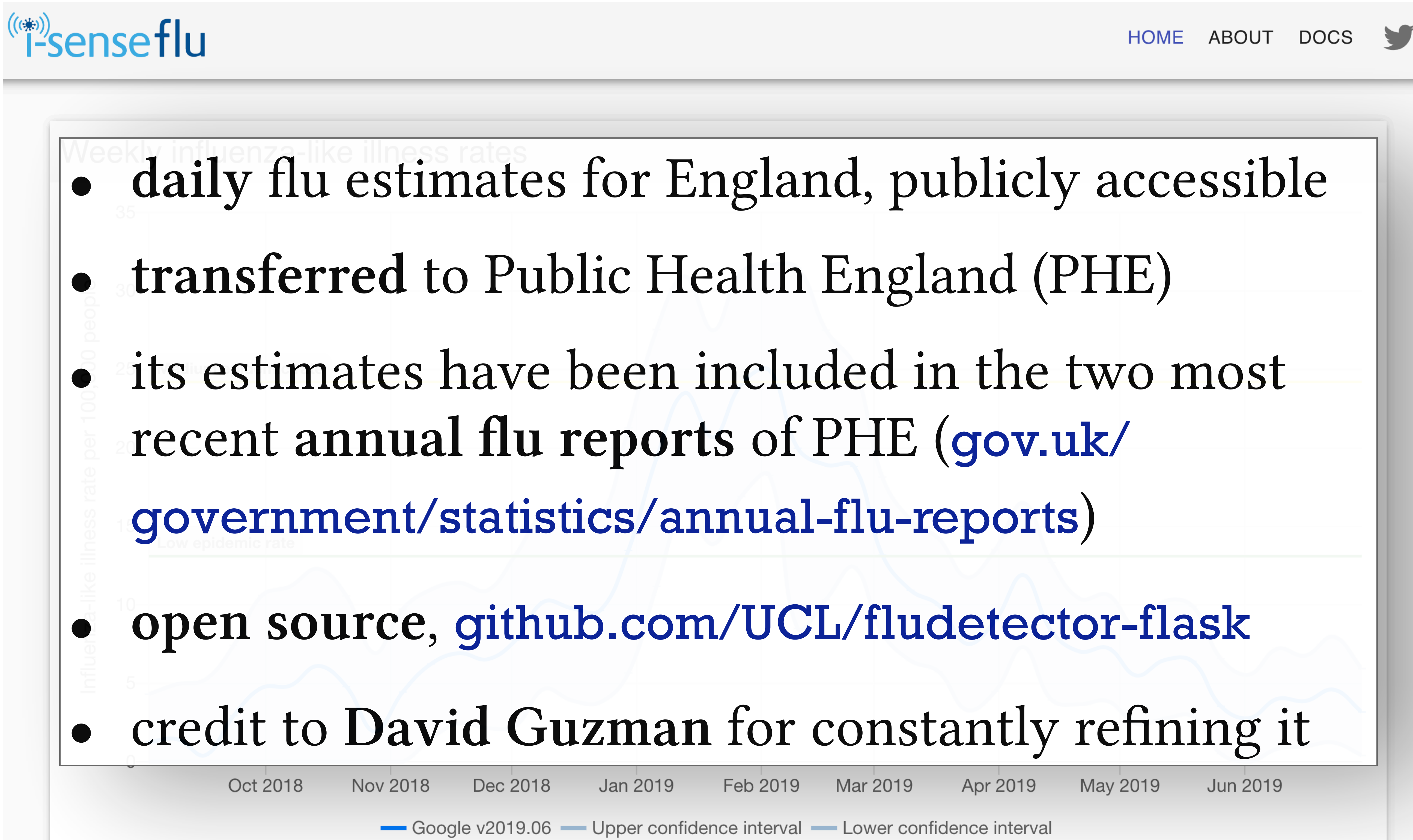
- **12.3%** performance improvement
- .913 Pearson correlation with RCGP ILI rates

i-sense flu (*Flu Detector*)



🔗 **fludetector.cs.ucl.ac.uk** @isenseflu

i-sense flu (*Flu Detector*)

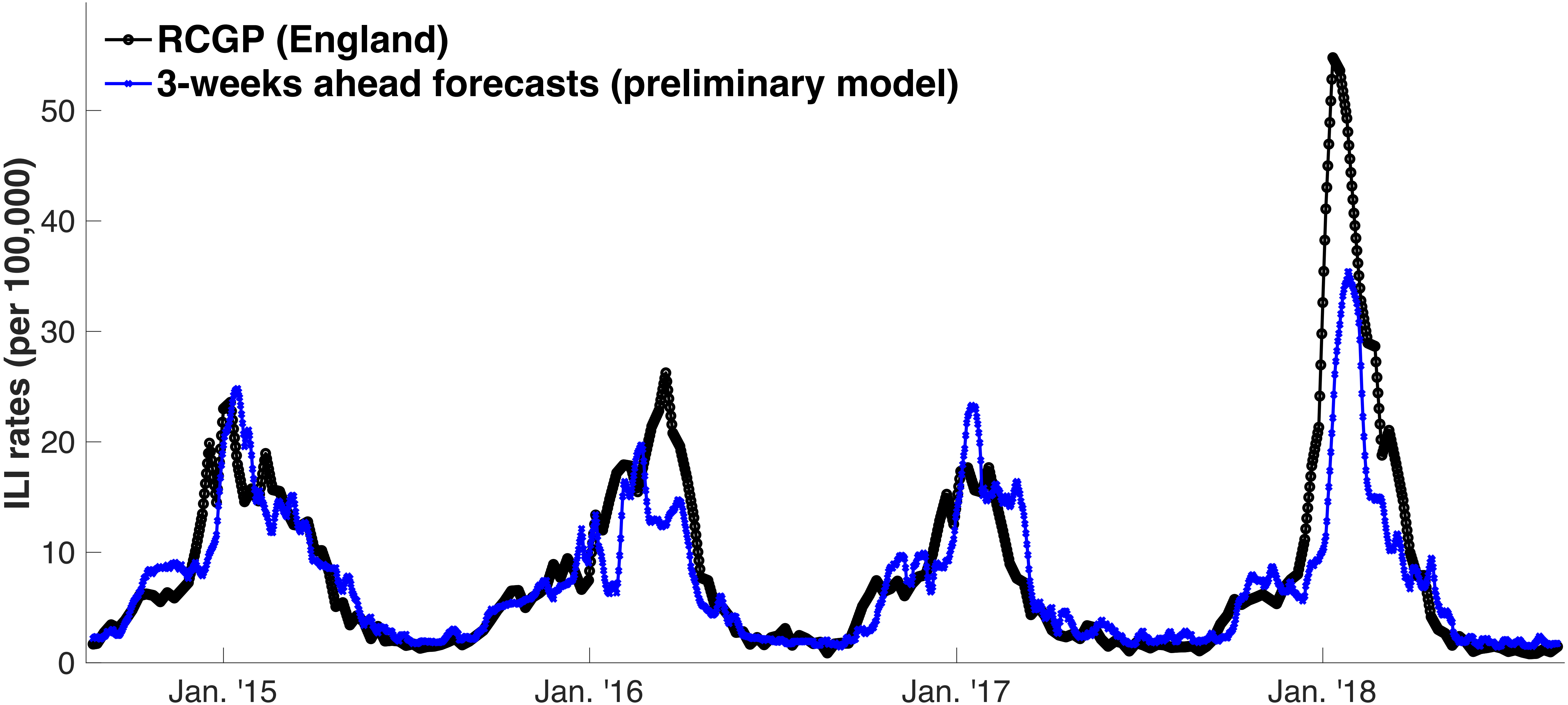


- **daily** flu estimates for England, publicly accessible
- **transferred** to Public Health England (PHE)
- its estimates have been included in the two most recent **annual flu reports** of PHE (gov.uk/government/statistics/annual-flu-reports)
- **open source**, github.com/UCL/fludetector-flask
- credit to **David Guzman** for constantly refining it

🔗 fludetector.cs.ucl.ac.uk

@isenseflu

Forecasting flu rates — Ongoing work

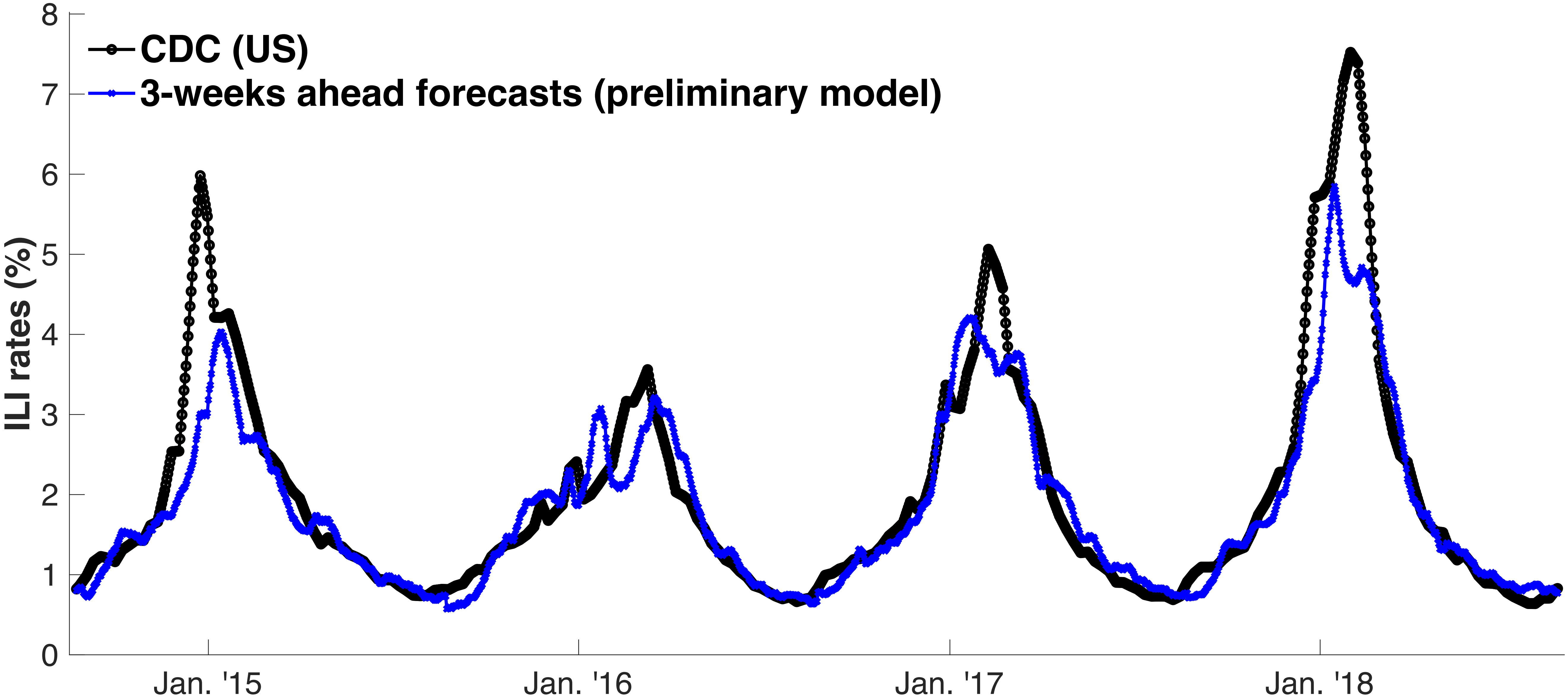


mean absolute error = **2.56** (cases per 100,000)

r = **.901**

led by **Simon Moura**

Forecasting flu rates (US) — Ongoing work



mean absolute error = **0.33%**

r = **.927**

led by **Simon Moura**

Multi-task learning for flu

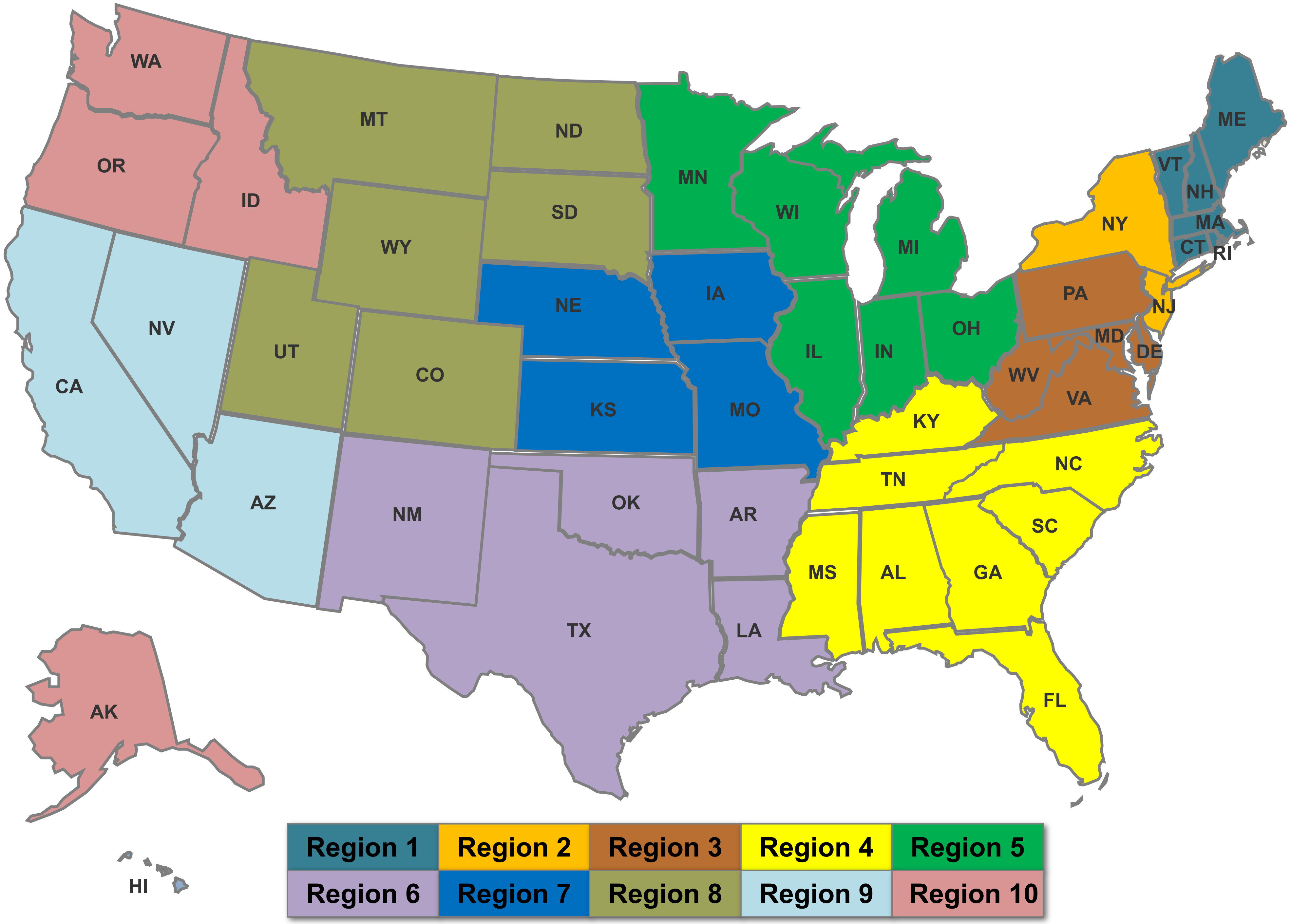
Multi-task learning (MTL) vs. single-task learning (STL)

- learns models ***jointly*** instead of independently
- for *related tasks* is performing better than STL solutions
- provides good performance with *fewer training samples*

Flu models with MTL

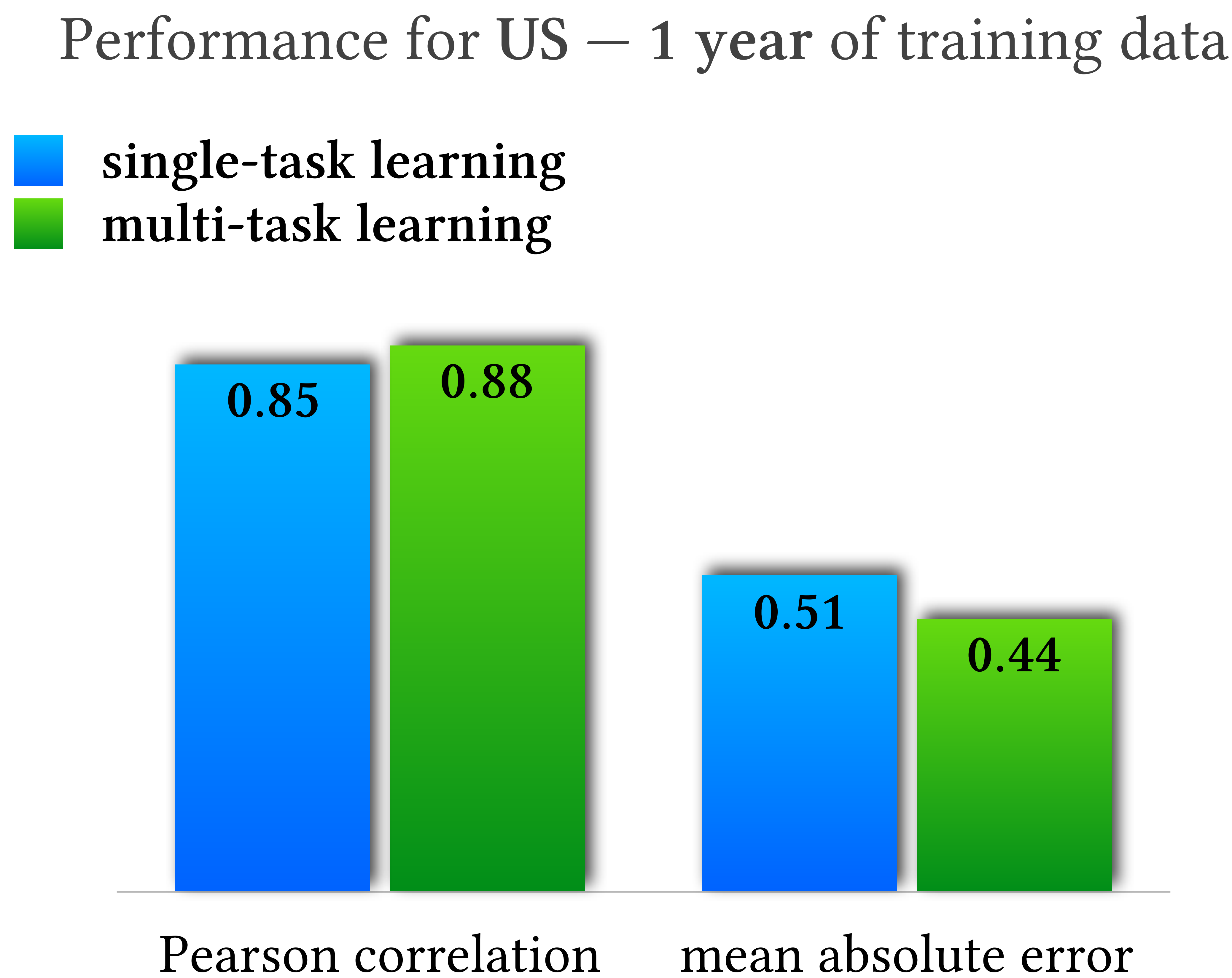
- limit performance loss under **sporadic training data**
- improve **accuracy**
  - of **regional** models within a country
  - across **different countries**

Modelling flu across US regions with MTL



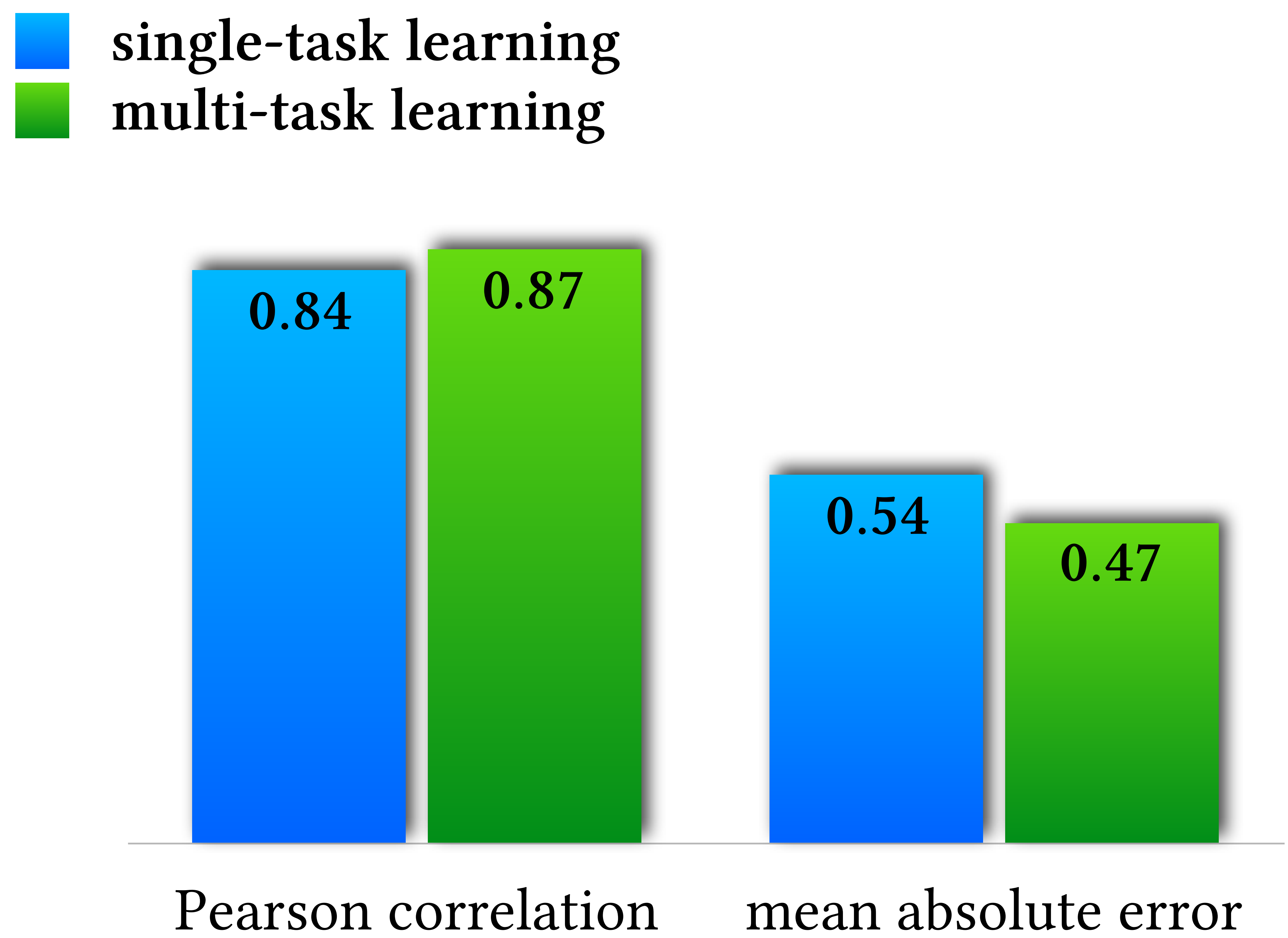
Train 10 US regional models for flu jointly

MTL across US and US regions



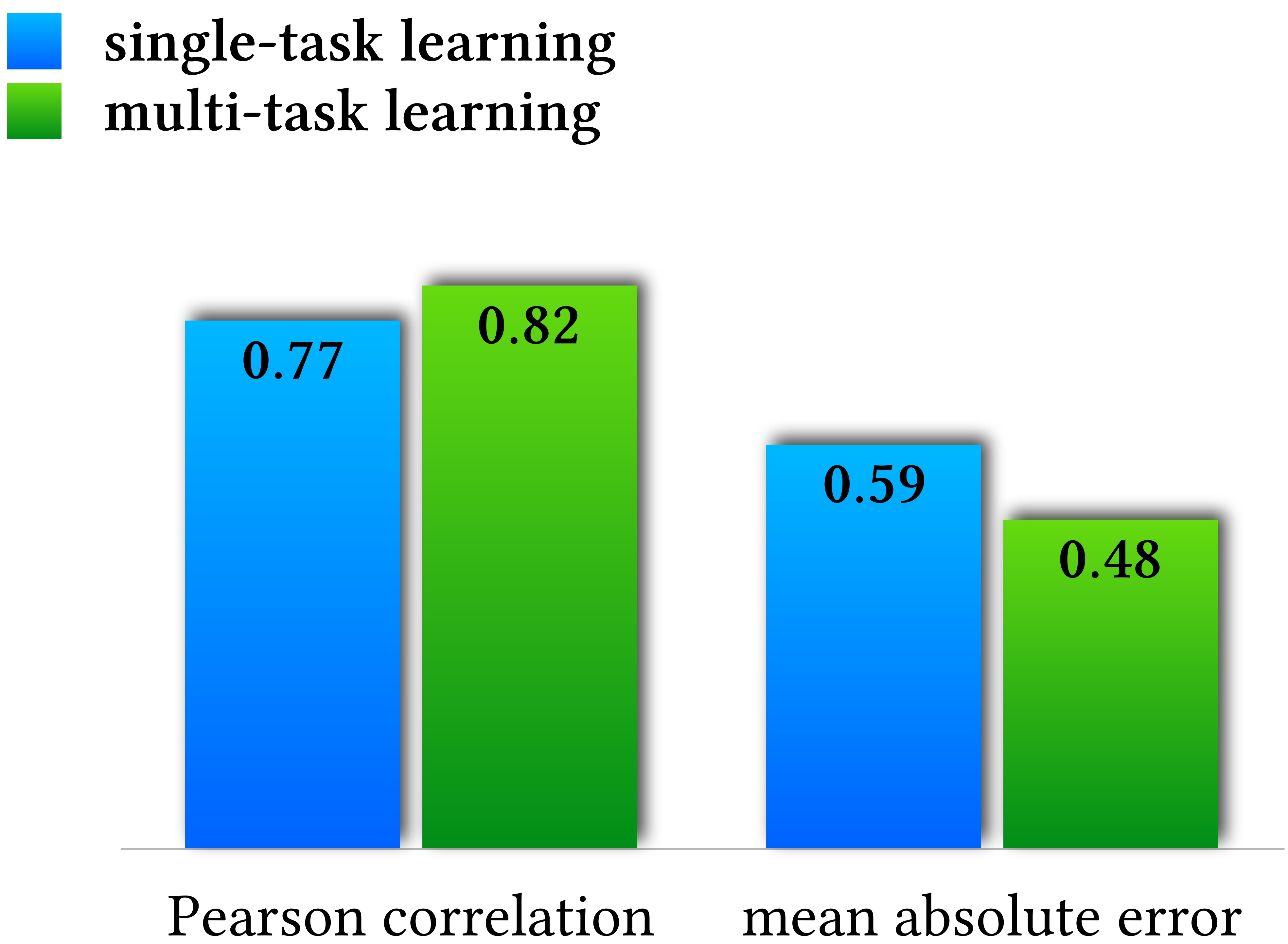
MTL across US and US regions

Performance for US regions — 1 year of training data



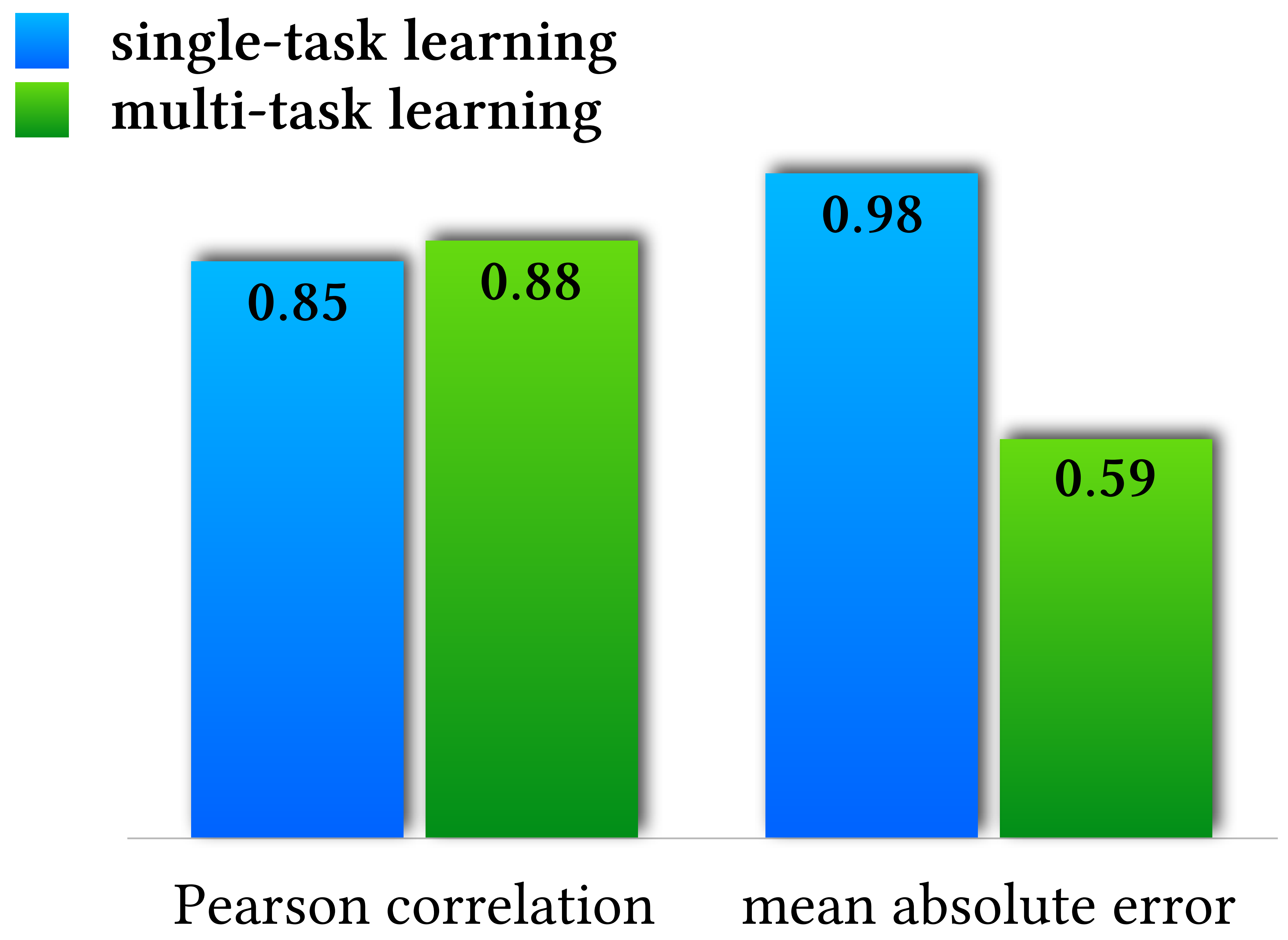
MTL across US and US regions

Performance for US regions — 1 year of training data
50% of the data lost



MTL across US and England

Performance for England — 1 year of training data



Why estimate flu rates from online search?

- ***Complement*** traditional syndromic surveillance
  - timeliness
  - broader demographic coverage, larger cohort
  - broader geographic coverage
  - not affected by closure days
  - lower cost

- Applicable to locations that ***lack*** an established healthcare system
  - **oxymoron**: *healthcare data is required for training the models!*

Transfer learning for flu modelling

Main task

- train a model for a *source* location where historical syndromic surveillance data **is available**
- transfer it to a *target* location where syndromic surveillance data **is not available** or, *in our experiments*, ignored

Transfer learning steps

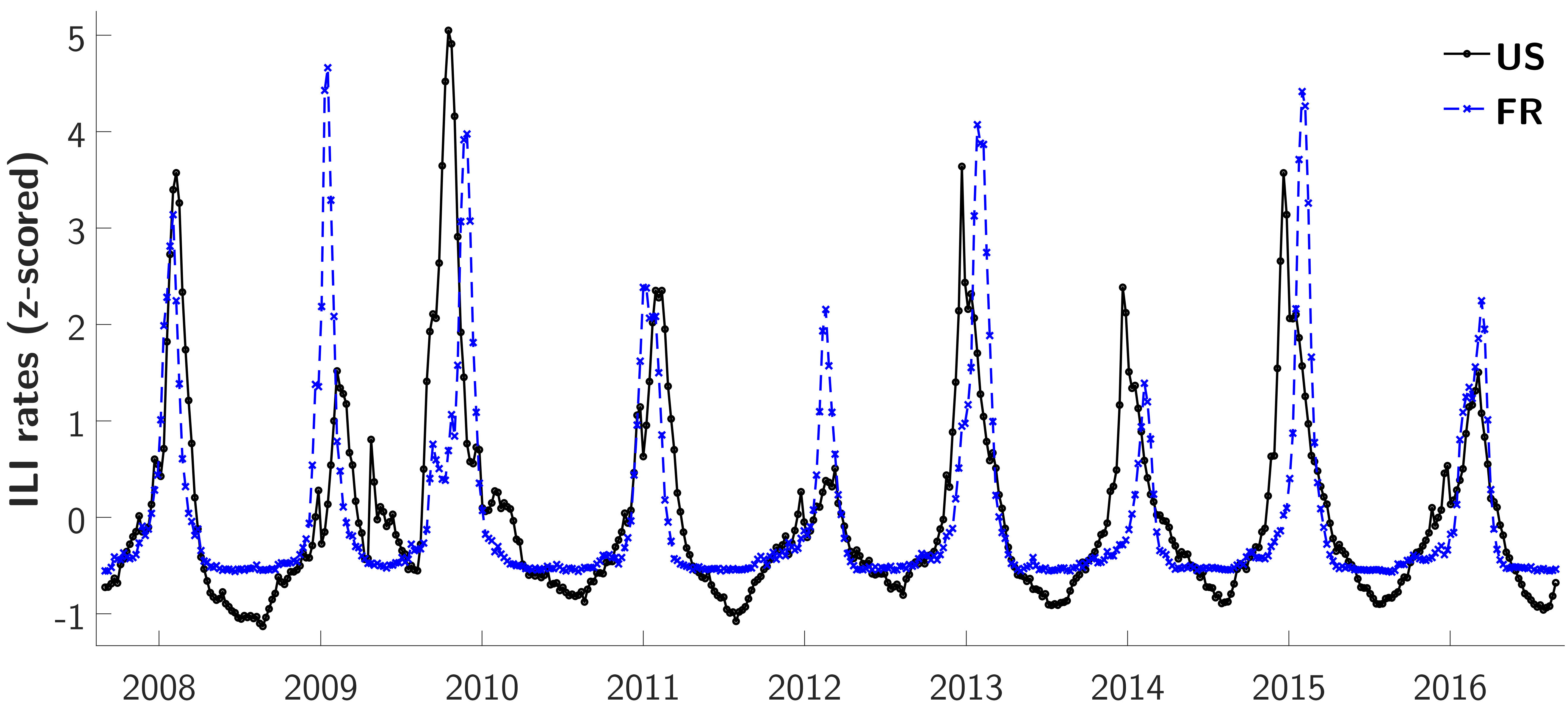
1. Learn a regression model for a source location
2. Map search queries from the source to the target domain
3. Transfer the source regression weights to the target domain

Mapping source to target queries

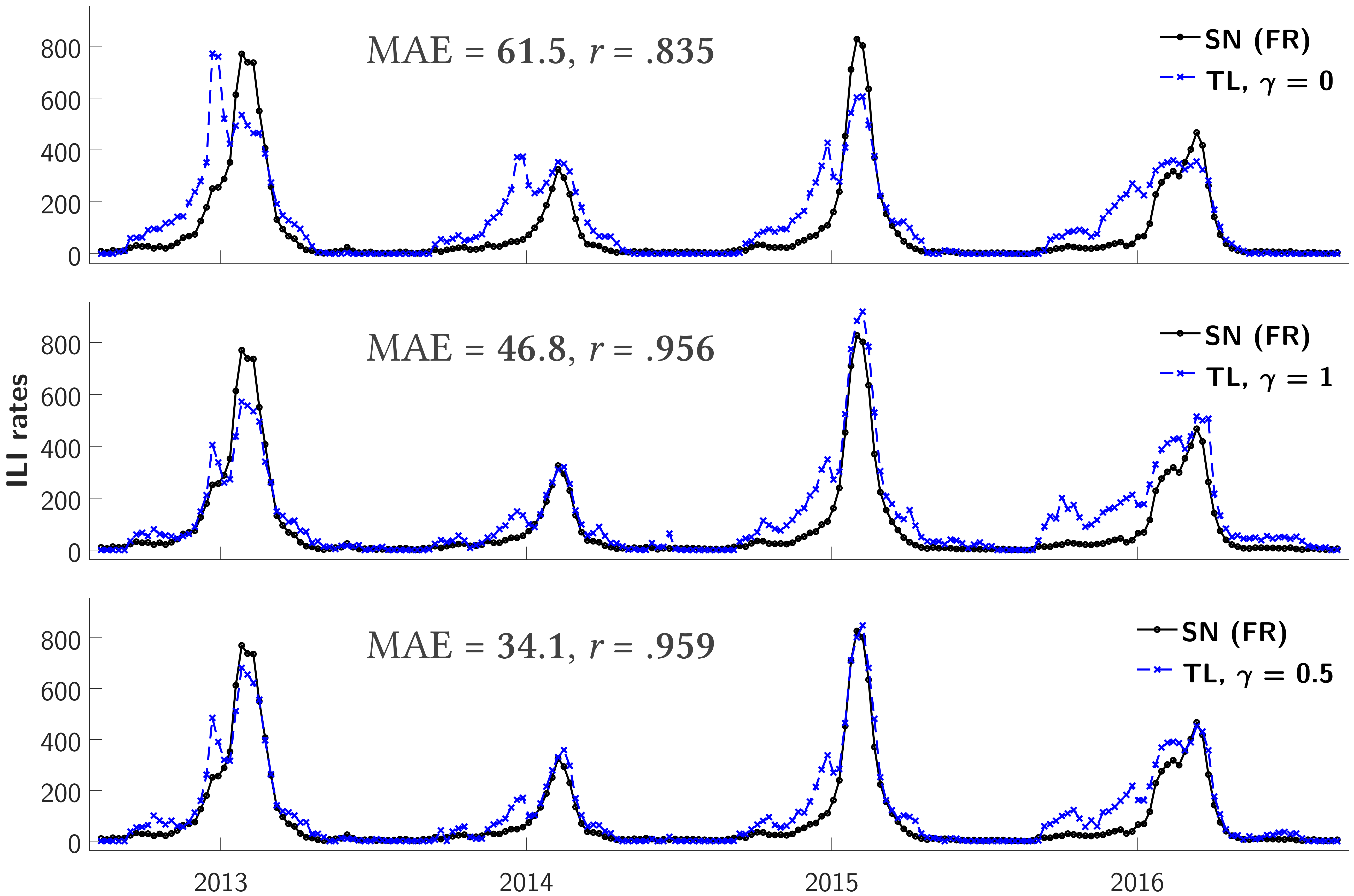
- Direct translation *does not work*
- Two similarity components
  - **Semantic similarity** (*meaning*) using cross-lingual word embedding representations (Θ_s)
  - **Temporal similarity** based on their frequency time series (Θ_c)
- **Joint similarity**: $\Theta = \gamma\Theta_s + (1-\gamma)\Theta_c, \quad \gamma \in [0,1]$

Source: US, Target: France

How similar are their flu rates?

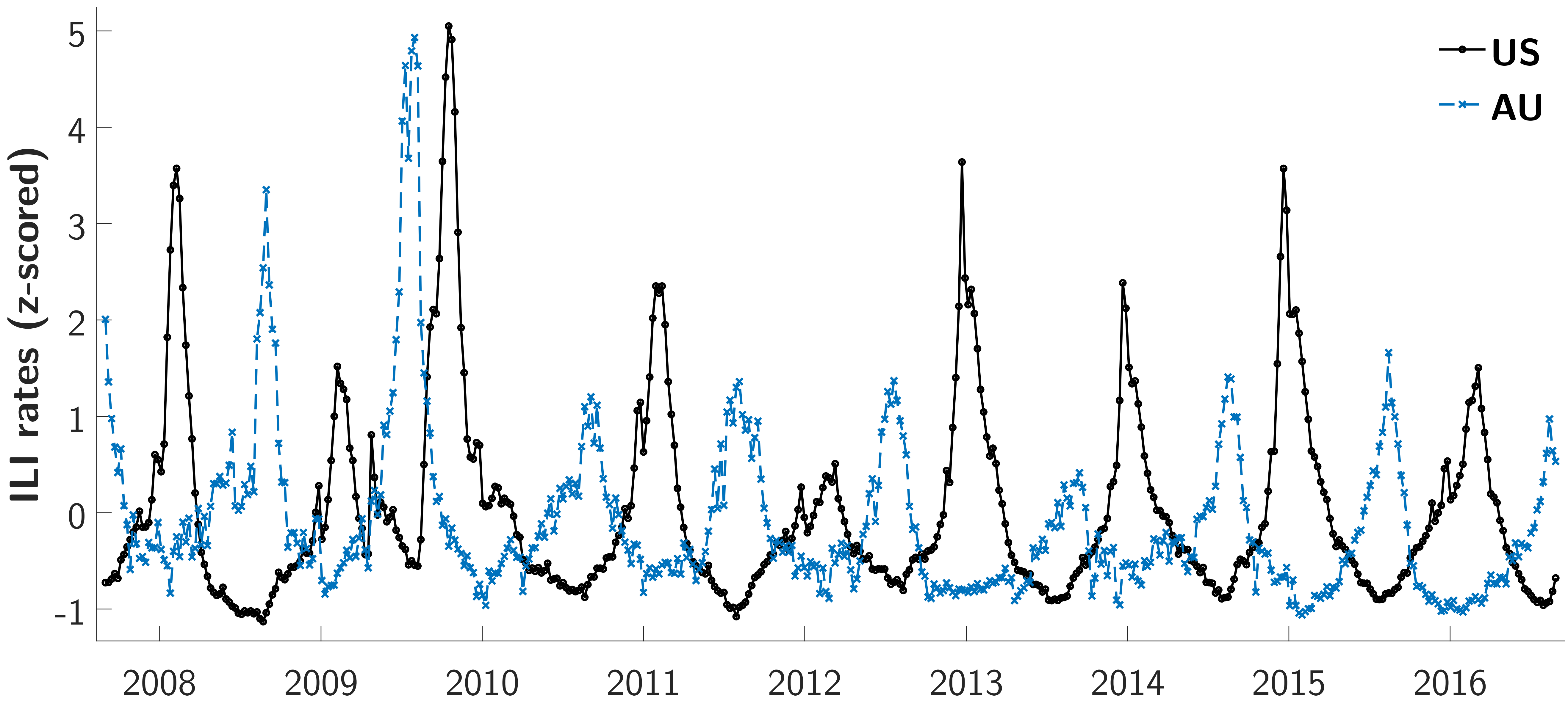


Source: US, Target: France

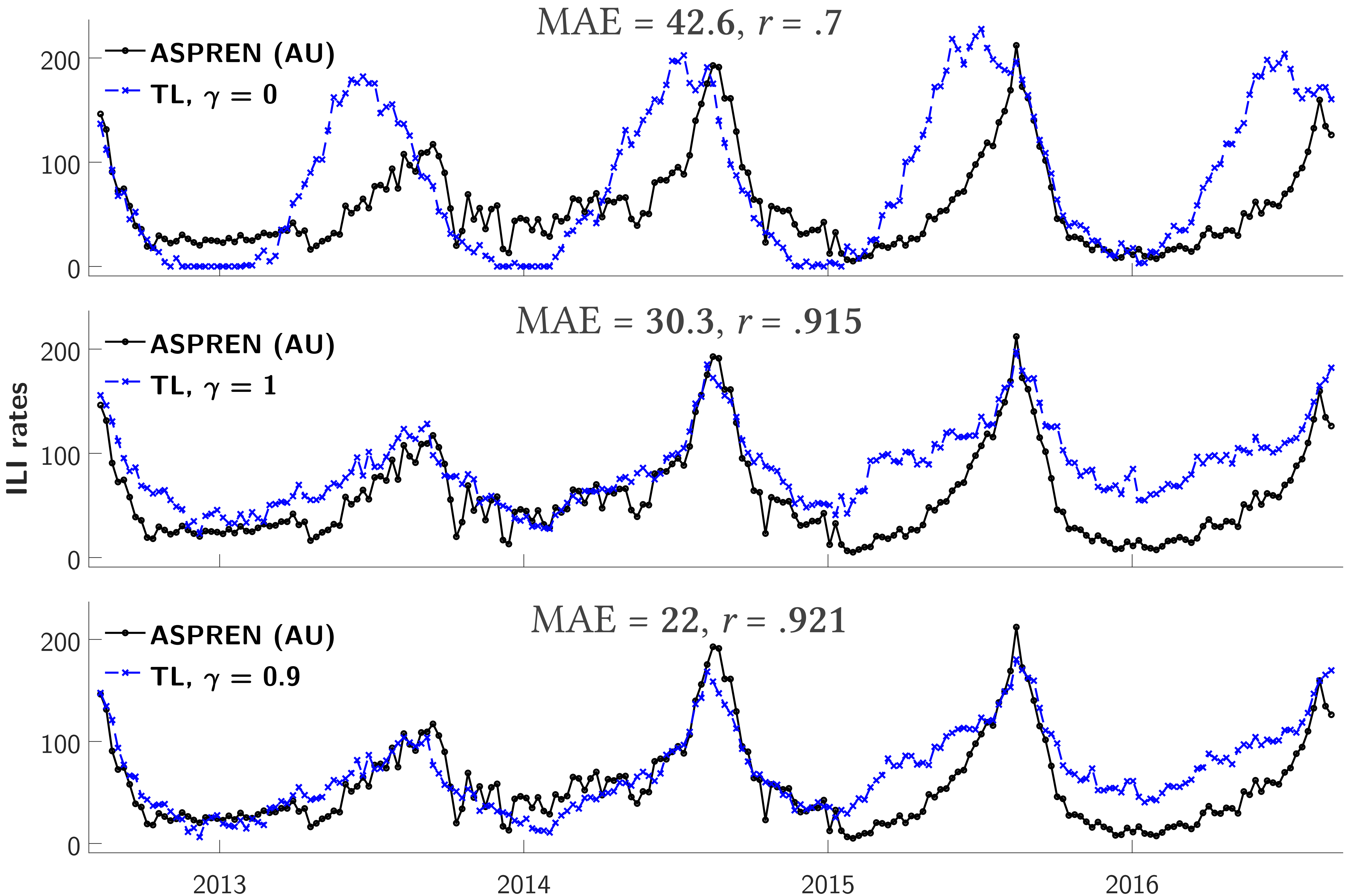


Source: US, Target: Australia

How similar are their flu rates?



Source: US, Target: Australia



Conclusions

We have shown that we can

- estimate flu rates from online search
  - right **modelling** approach
  - right **query selection** approach
- utilise **multi-task learning** to improve models
- **transfer models** when healthcare data is not available

Future work within i-sense includes

- **forecasting** flu rates
- **translation** of our research to public health solutions

Acknowledgements

Collaborators: Ingemar J. Cox, Elad Yom-Tov, Richard Pebody, Bin Zou, Andrew Miller, Moritz Wagner, Simon Moura

Organisations: Microsoft Research, Google, Royal College of General Practitioners (RCGP), Public Health England (PHE)

Funding: EPSRC IRC "i-sense"

EPSRC Engineering and Physical Sciences Research Council

i-sense

@lampos

lampos.net

References

Key papers from our group

- Lampos, Miller, Crossan and Stefansen (2015). *Advances in Nowcasting Influenza-like Illness Rates using Search Query Logs*. Scientific Reports, 5(12760).
- Lampos, Zou and Cox (2017). *Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance*. Proc. of the 26th International Conference on World Wide Web, pp. 695–704.
- Zou, Lampos and Cox (2018). *Multi-Task Learning Improves Disease Models from Web Search*. Proc. of the 2018 World Wide Web Conference, pp. 87–96.
- Zou, Lampos and Cox (2019). *Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data*. Proc. of the 2019 World Wide Web Conference, pp. 2505–2516.

Other papers mentioned in the slides

- Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009). *Detecting Influenza Epidemics using Search Engine Query Data*. Nature, 457(7232):1012–1014.
- Eysenbach (2006). *Infodemiology: tracking flu-related searches on the web for syndromic surveillance*. Proc. of AMIA Annual Symposium, pp. 244–248.
- Polgreen, Chen, Pennock, Nelson and Weinstein (2008). *Using Internet Searches for Influenza Surveillance*. Clinical Infectious Diseases, 47(11):1443–1448.