

# Early online behavior during pandemics: why it matters for now-casting

Sara Mesquita<sup>1,2</sup>, Lília Perfeito<sup>1</sup>, João Loureiro<sup>1</sup>, Cláudio Haupt-Vieira<sup>2</sup>, and Joana Gonçalves-Sá<sup>1,3</sup>

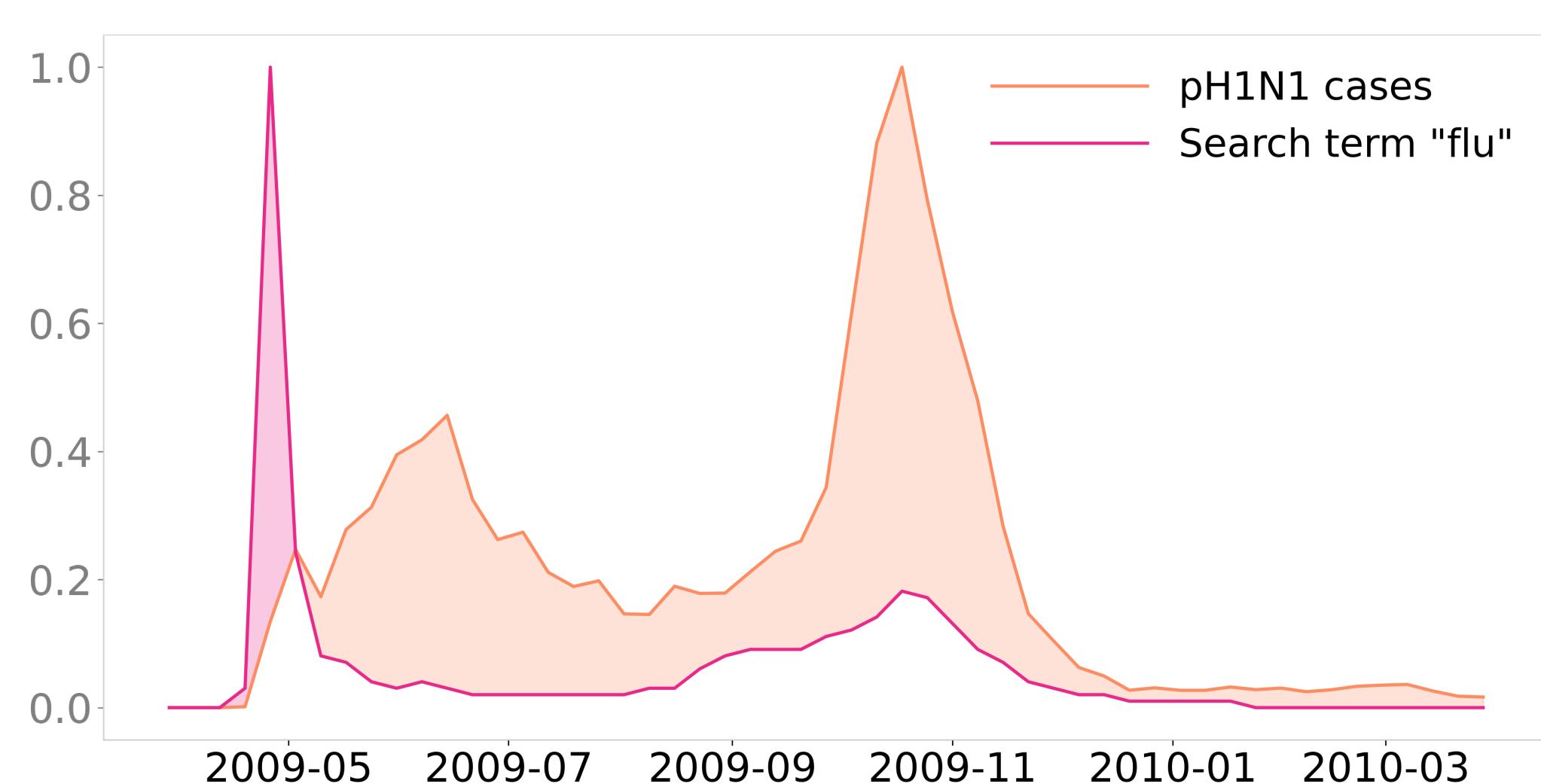
<sup>1</sup>LIP, Lisbon, Portugal, <sup>2</sup>Nova Medical School, Lisbon, Portugal, <sup>3</sup>Nova School of Business and Economics, Carcavelos, Portugal  
 smesquita@lip.pt



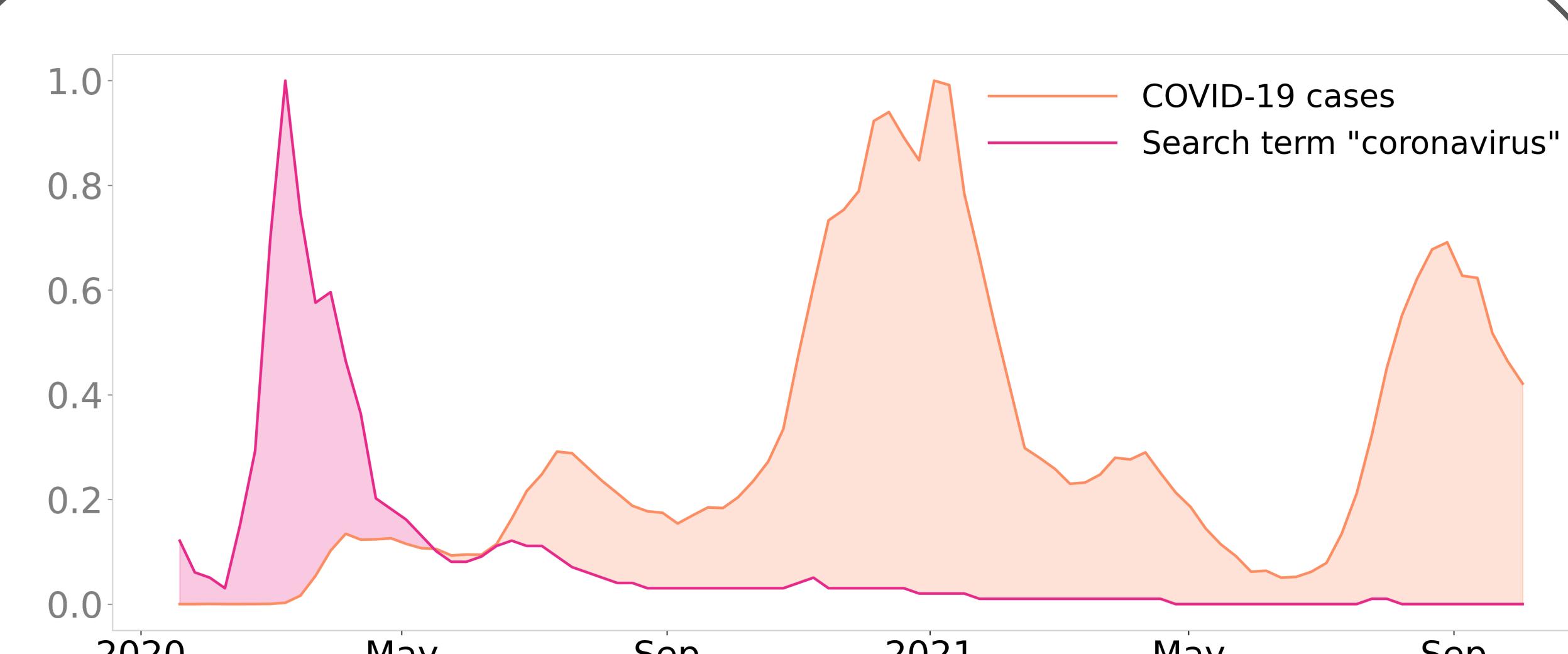
@MesquitaSra

Online behavior has been used as a tool for close to real-time study of different health-related behaviors, including identifying disease outbreaks. However, its validity in predicting infections, has been questioned by many, particularly during extraordinary events, such as pandemics. Here we hypothesized that (1) the earlier period of the pandemics could be used to disentangle between searches driven by media attention from searches driven by actual disease. The rationale is that by having periods in which media attention and cases are decoupled, it should be possible to identify which search-terms are more sensitive to media hype. This assumes that (2) models can be improved not necessarily by blindly increasing the size of the training dataset, but by integrating prior information.

## 2009 H1N1 Pandemic

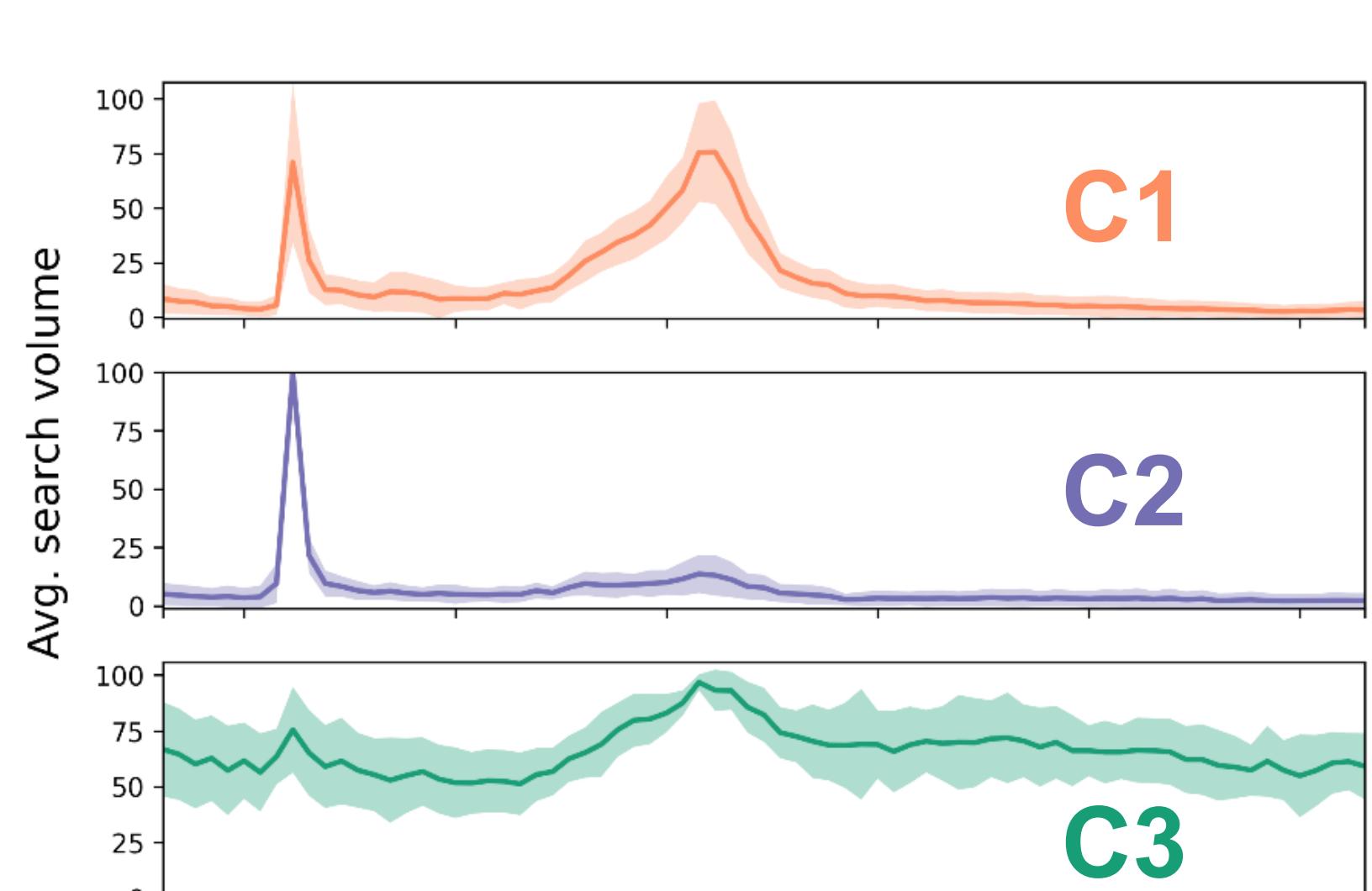
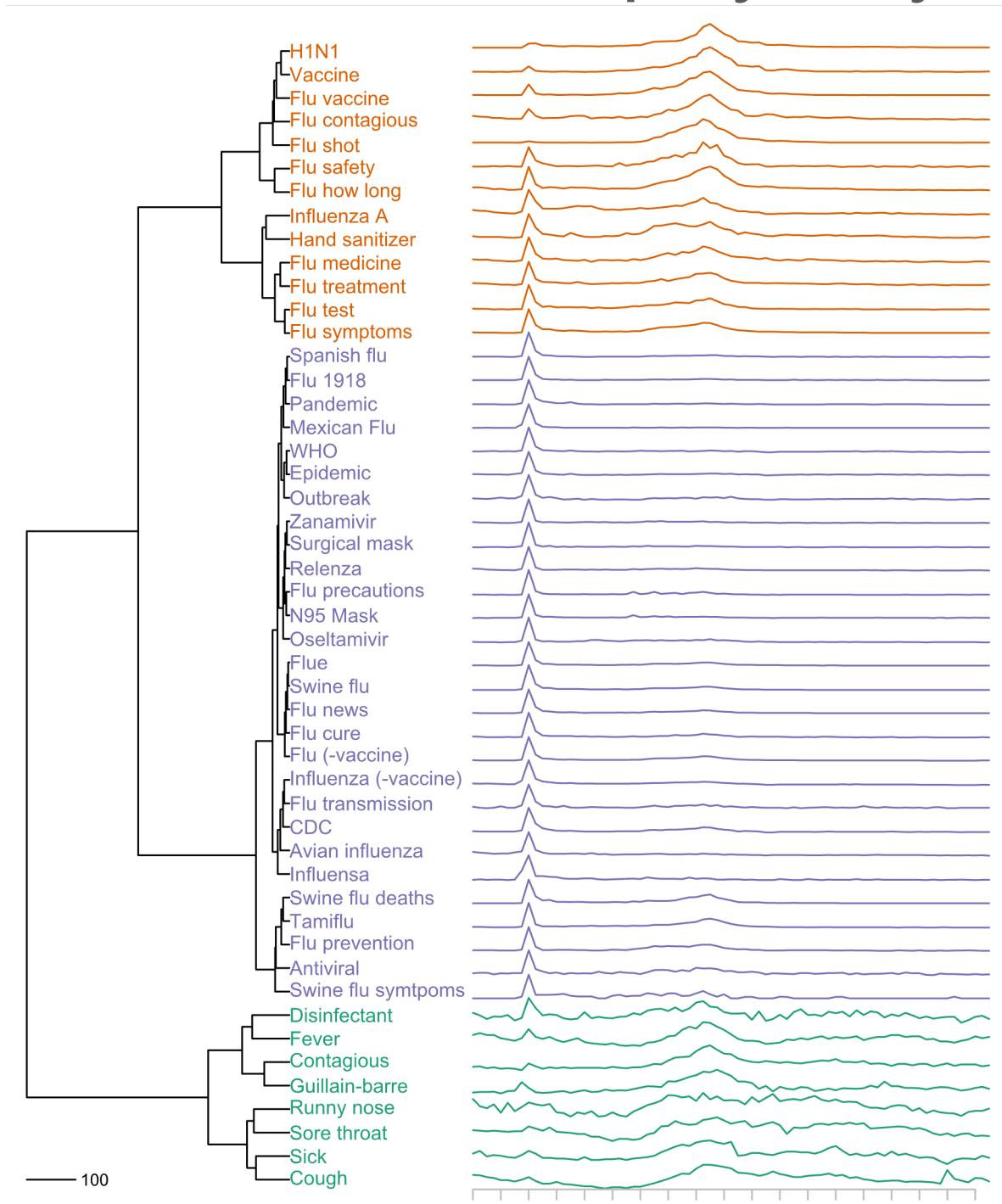


## COVID-19 Pandemic



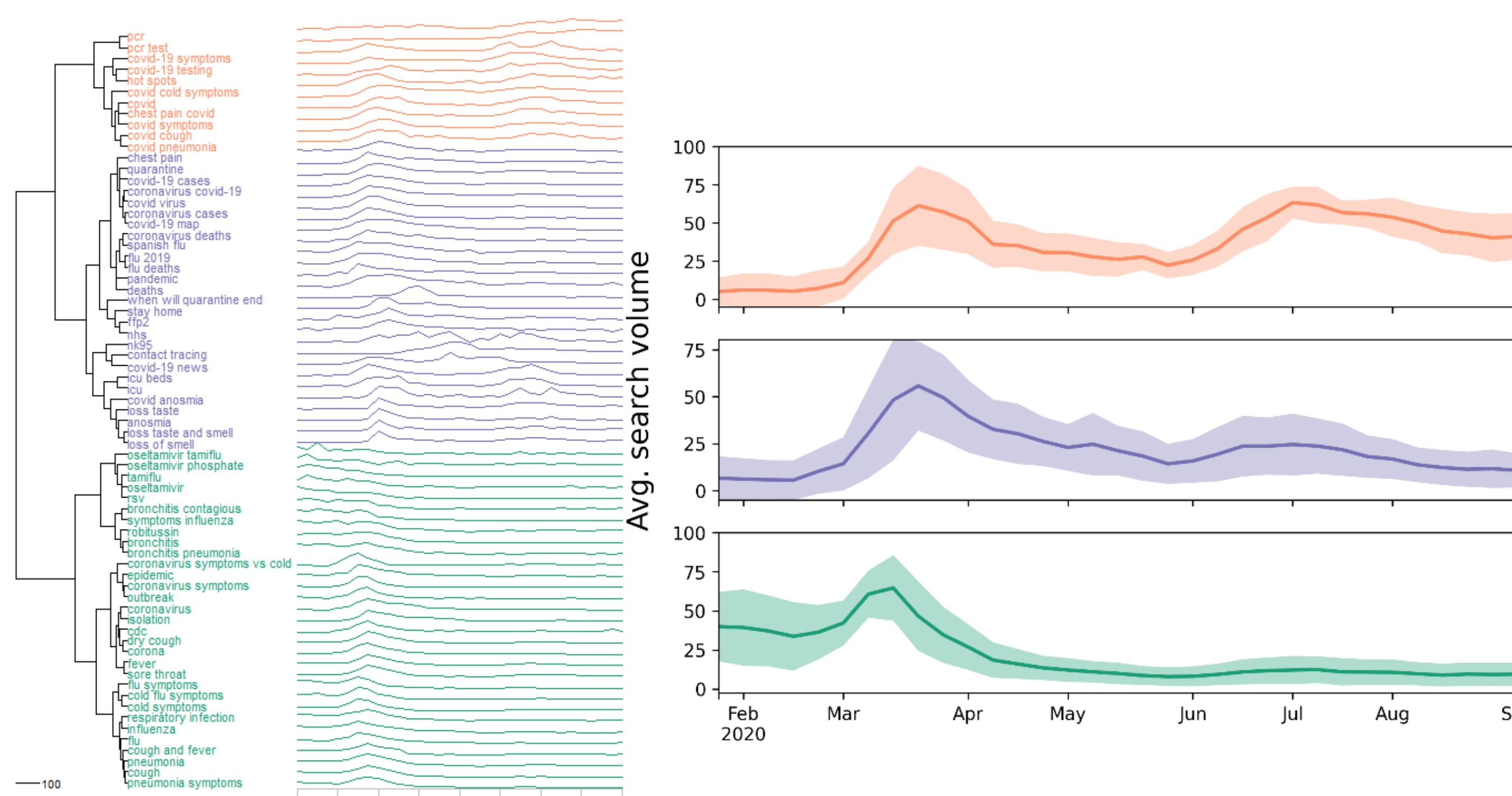
### Different temporal dynamics

similar terms display very different temporal dynamics



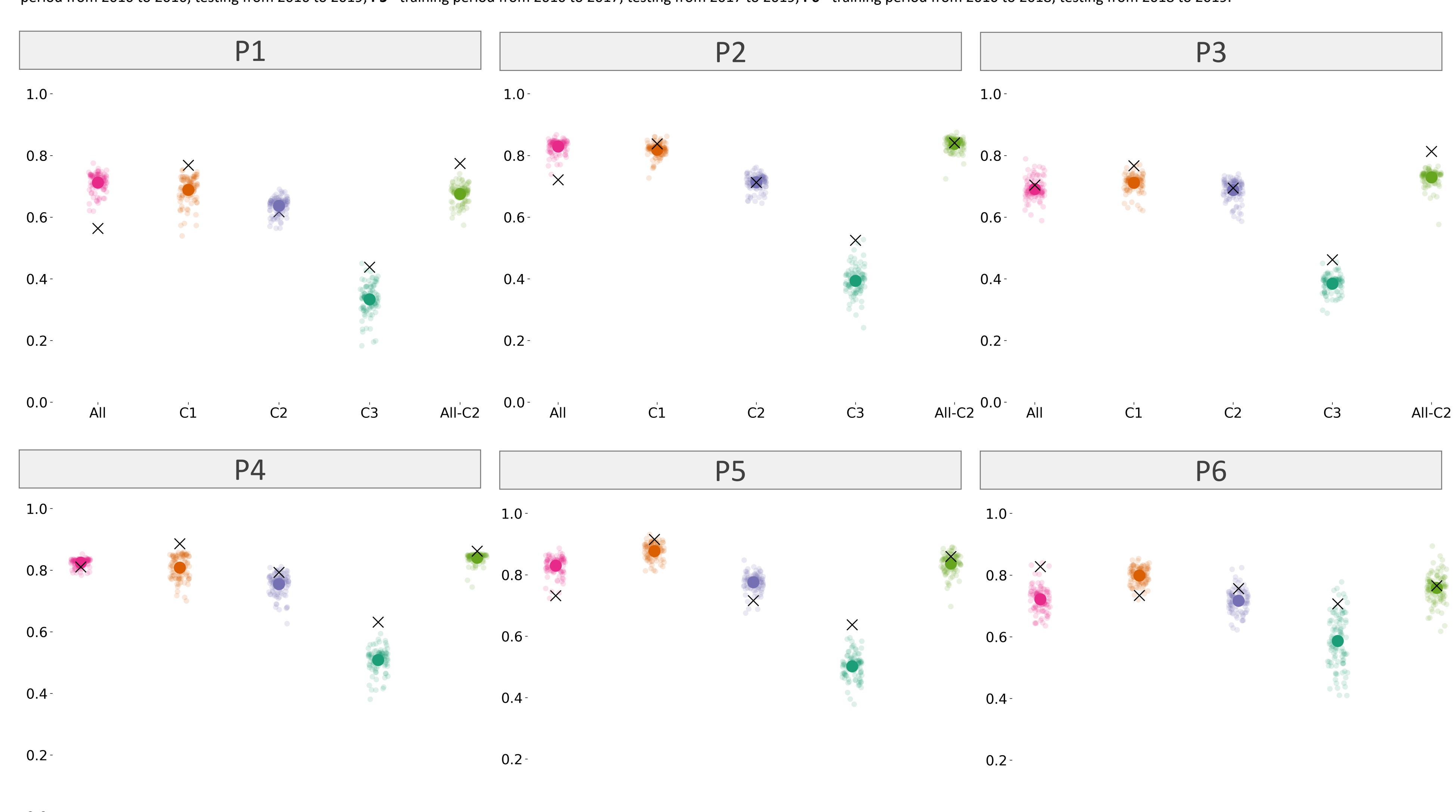
Hierarchical clustering:  
 3 main clusters identified

### Different temporal dynamics



## Models: Random Forest - RF (●) and Linear Regression - LR (x)

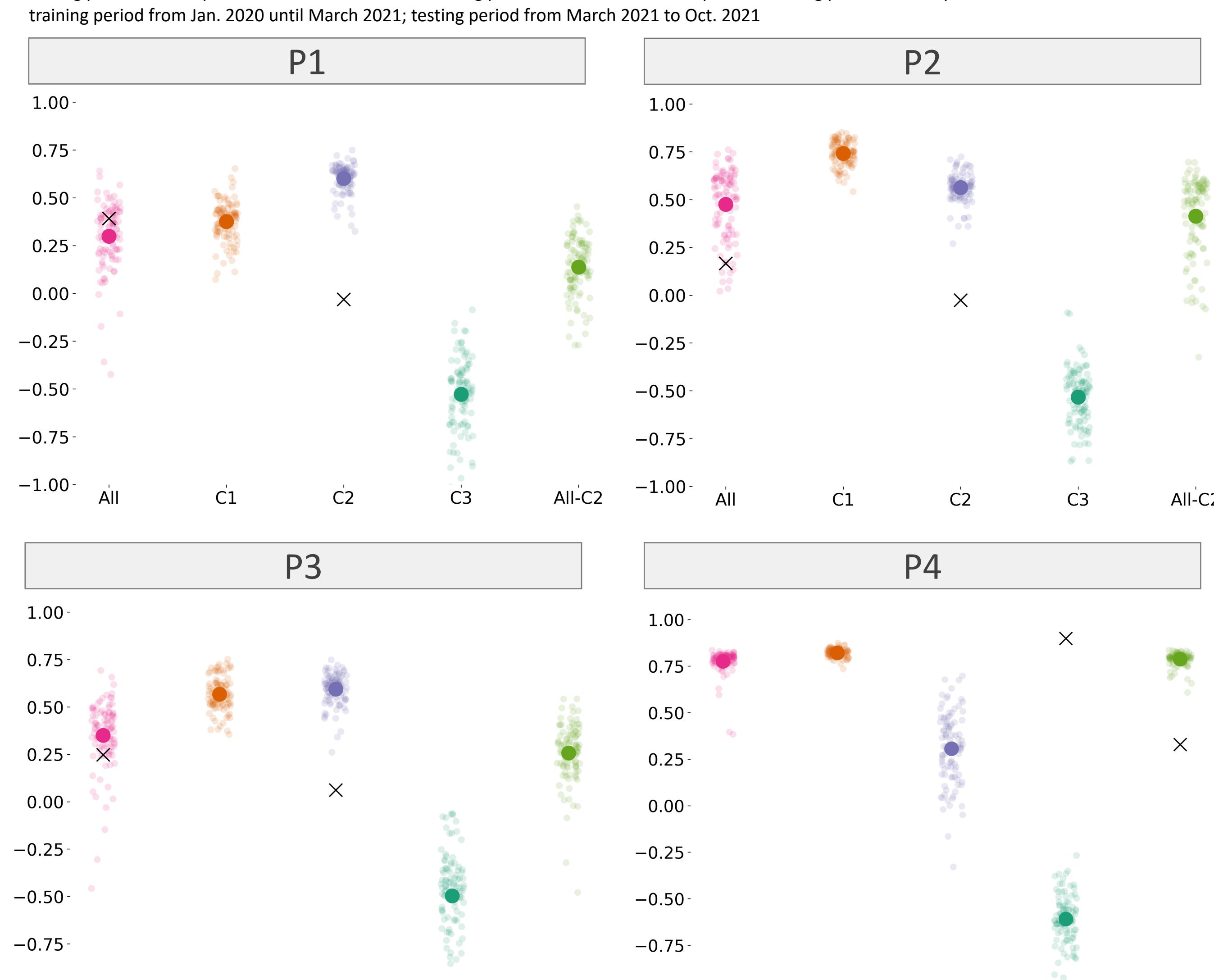
P1 - training period from 2010 to 2013; testing from 2013 to 2019; P2 - training period from 2010 to 2014; testing from 2014 to 2019; P3 - training period from 2010 to 2015; testing from 2015 to 2019; P4 - training period from 2010 to 2016; testing from 2016 to 2019; P5 - training period from 2010 to 2017; testing from 2017 to 2019; P6 - training period from 2010 to 2018; testing from 2018 to 2019.



C1 consistently better over different time periods (P)  
 than using all terms (AII)

## Models: RF(●) and LR(x)

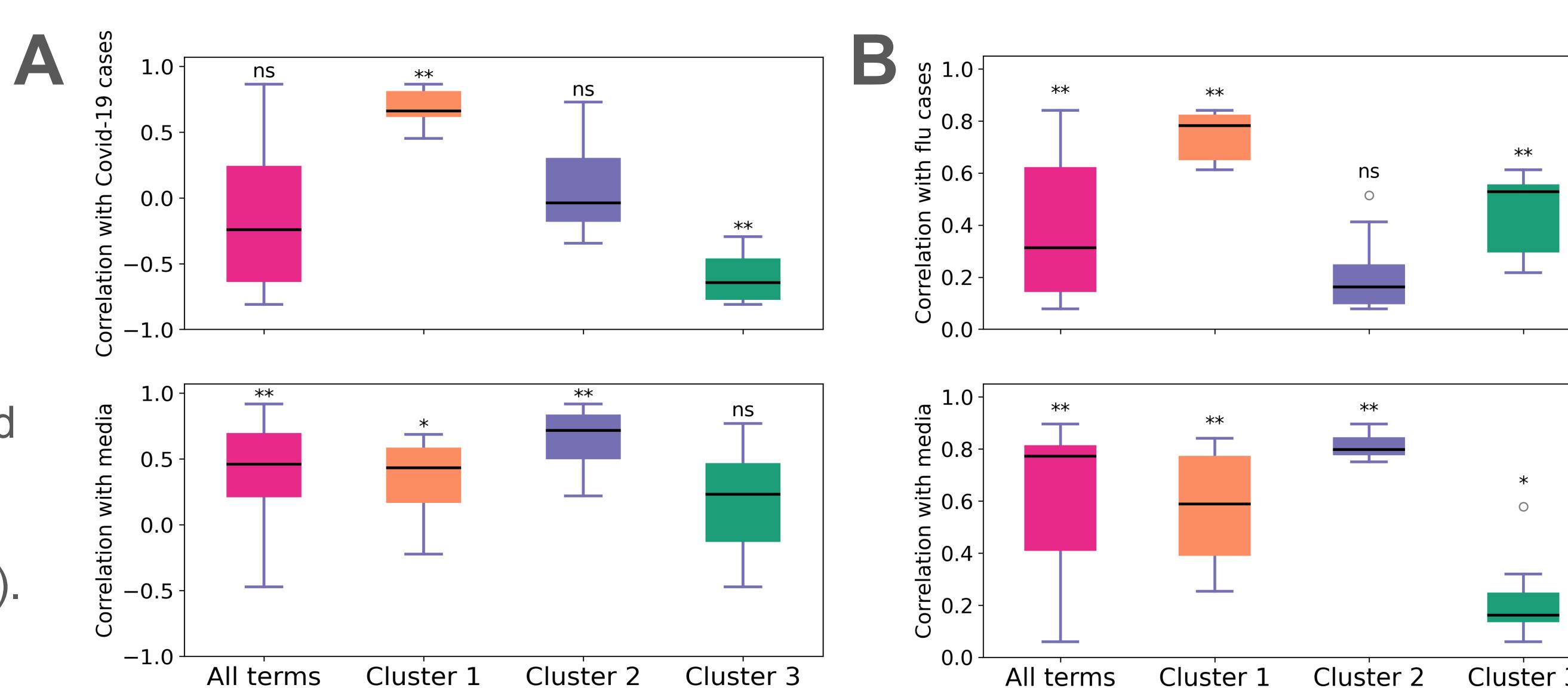
P1 - training period from Jan. 2020 to Sep. 2020; testing period from Sep 2020 to October 2021; P2 - training period from Jan. 2020 to Sep. 2020; testing period from Sep. 2020 to March 2021; P3 - training period from Jan. 2020 to Sep. 2020; testing period from Sep. 2020 to June 2021; P4 - training period from Jan. 2020 until March 2021; testing period from March 2021 to Oct. 2021



C1 consistently better over different time periods (P) than using all terms (AII)

### Correlation with cases vs media

Pearson correlation between the cluster centroids and:  
 (A) H1N1 infection cases (top) and media mentions (bottom),  
 (B) SARS-CoV-2 infection cases (top) and media mentions (bottom).



\* denotes  $p < 0.05$ , \*\* denotes  $p < 0.01$ , and ns a non-significant p-value.

### Take home message:

We can learn from pandemics to improve now-casting of future infection waves and including more data is not necessarily better.

### References

- [1] Ginsberg, J., et al., (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–4. 1
- [2] Lazer, D., et al., (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343, 1203–1205. 2
- [3] Tizzoni, M., et al., (2020). The impact of news exposure on collective attention in the united states during the 2016 zika epidemic. *Plos computational biology*, e1007633.
- [4] Kogan, Nicole E., et al. (2021) "An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time." *Science Advances* 7:10: eabd6989.
- [5] Mesquita, S., Vieira, C. H., Perfeito, L., & Gonçalves-Sá, J. (2021). Learning from pandemics: using extraordinary events can improve disease now-casting models. *arXiv preprint arXiv:2101.06774*.