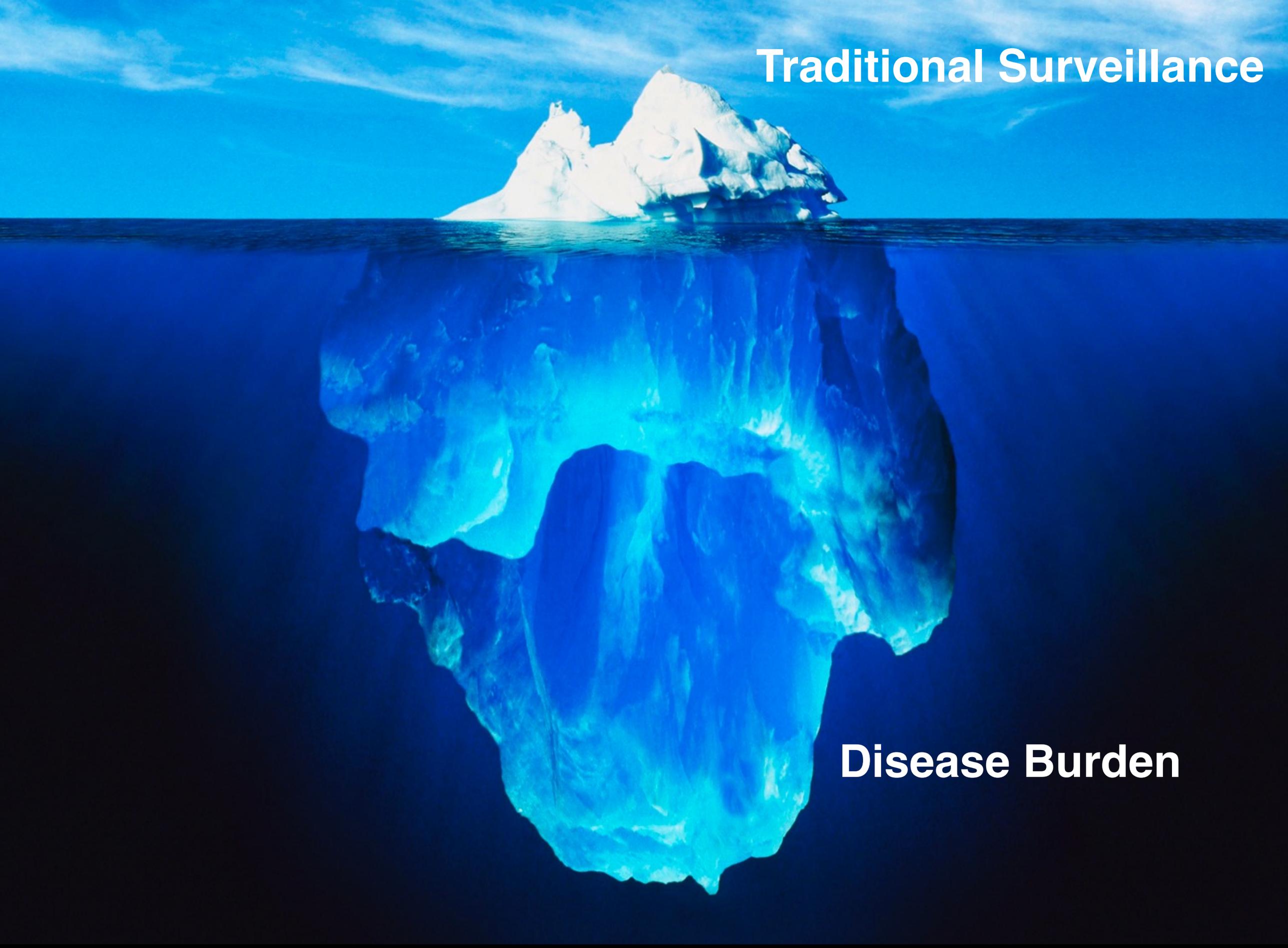
Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms

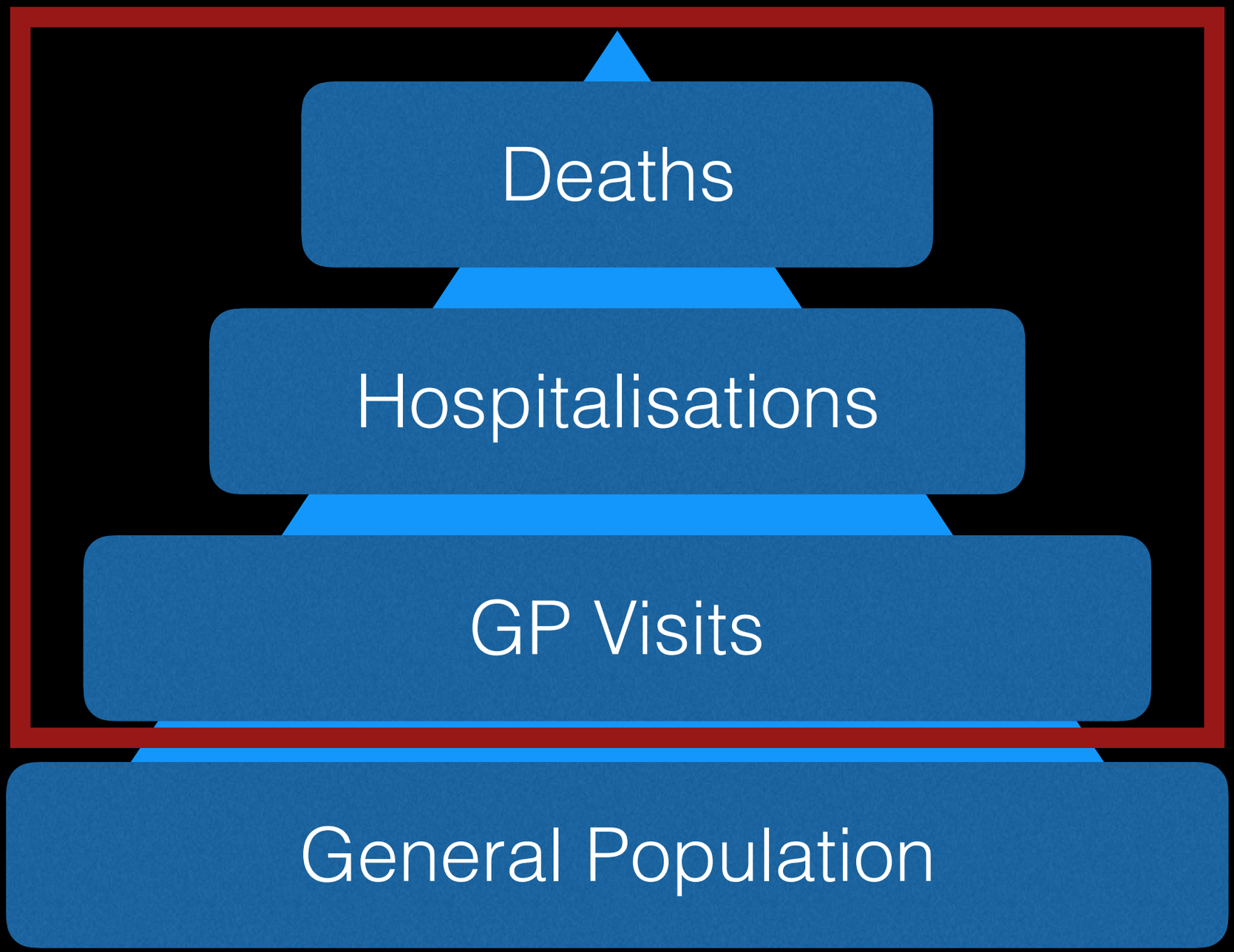
Daniela Paolotti
ISI Foundation
Turin, Italy

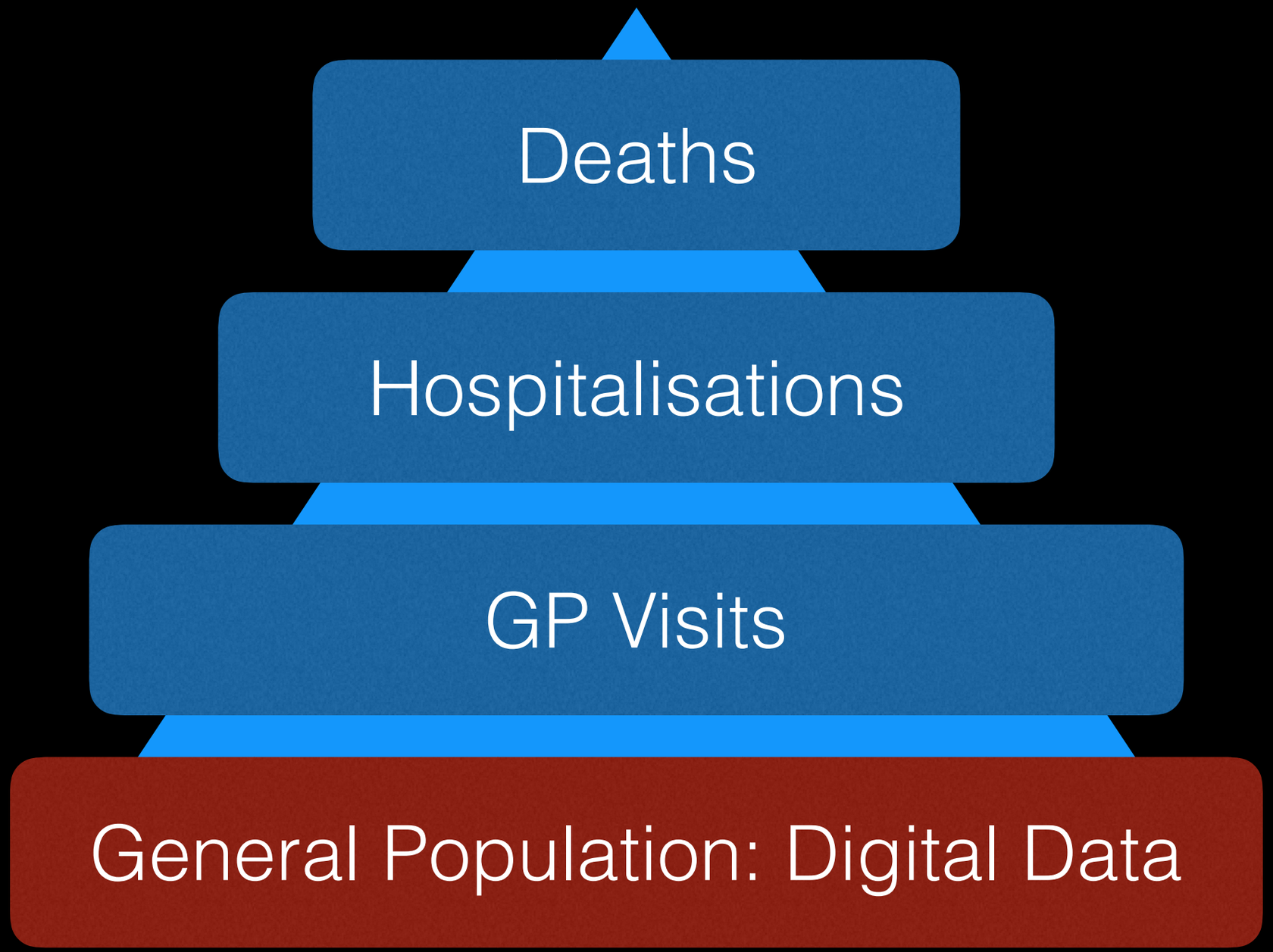
@danielapaolotti

Applied Machine Learning Days
Health Track
January 27th 2020


Traditional Surveillance
Disease Burden

Surveillance




Deaths
Hospitalisations
GP Visits
General Population: Digital Data

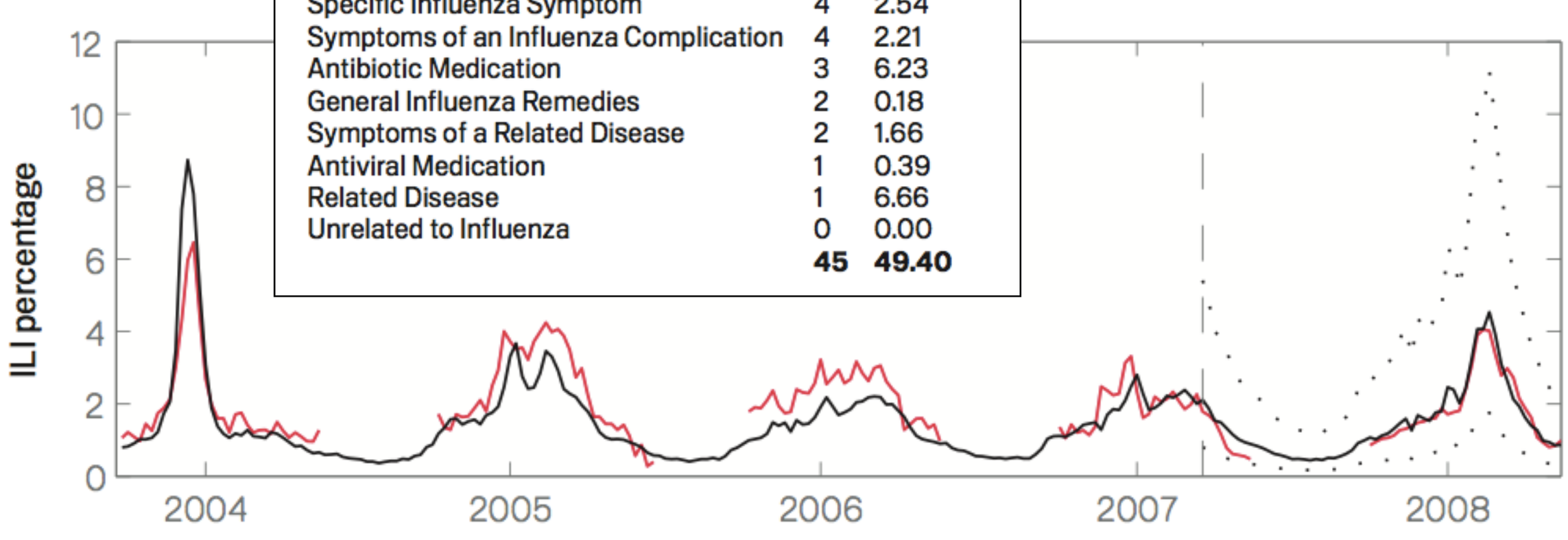

OPPORTUNITIES
DISTRIBUTED
RESEARCH
GENOMICS APPLIED
EXAMPLES
PETABYTES
CURRENTLY
USE
DESKTOP
TARGET
SHARED
DATABASES
GARTNER
MPP
DESCRIBING
INDEXING
CITATION
TYPES
RELATIONAL
BIG
USED
GROW
CAPACITY
DISK
TOLERABLE
NETWORKS
CREATED
SOFTWARE
SINCE
CAPTURE
SOCIAL
LARGER
ARCHIVES
TENS
ORGANIZATIONS
DATA
SAN
DEFINING
FORMS
NOW
NEEDED
HUNDREDS
TERABYTES
CONTINUES

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

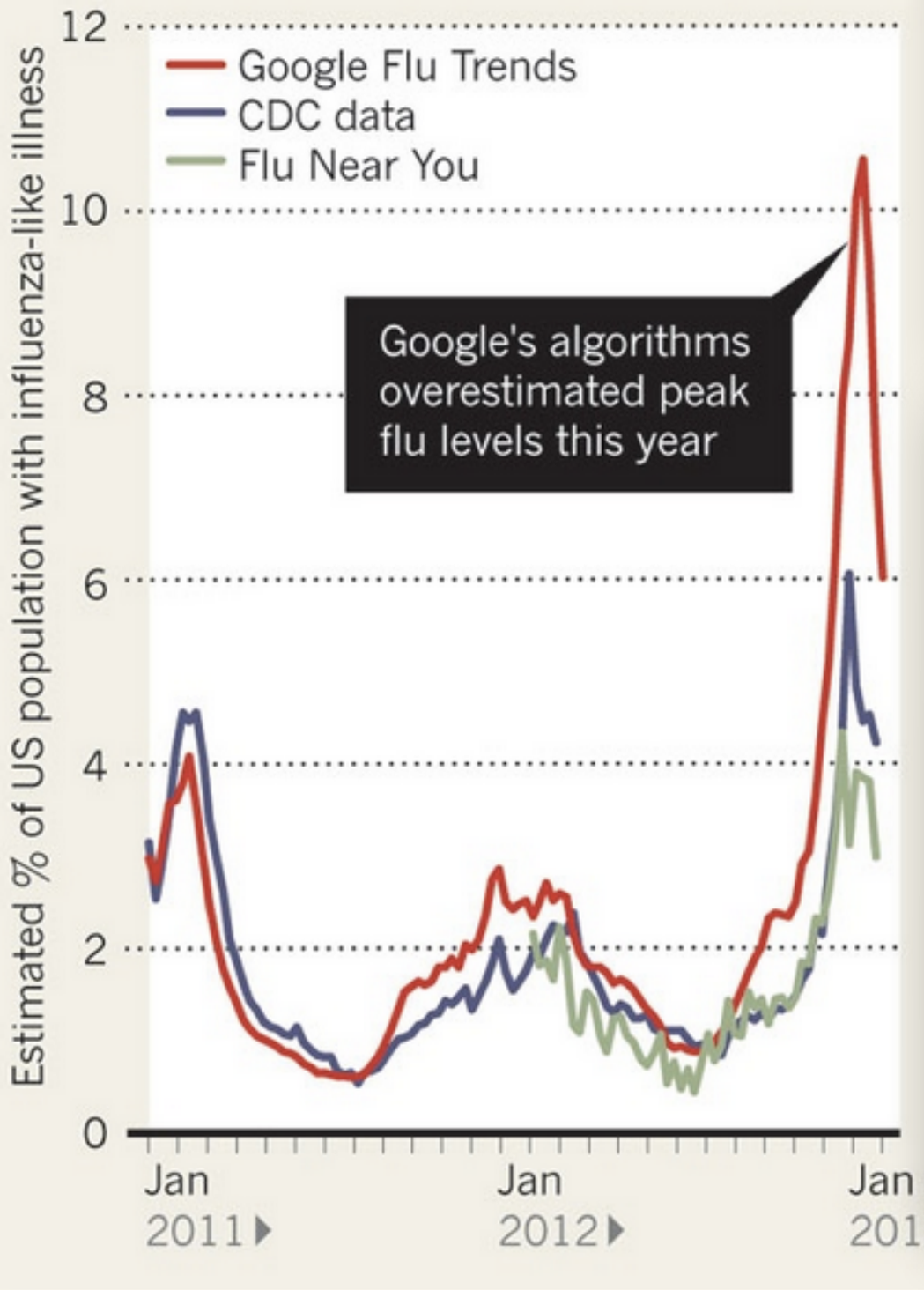
Search Query Topic	Top 45 Queries N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**



J. Ginsberg *et al.*, Nature 457, 1012 (2009)

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



nature
International weekly journal of science

NATURE | NEWS

عربي

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203–1205
DOI: 10.1126/science.1248506

POLICY FORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

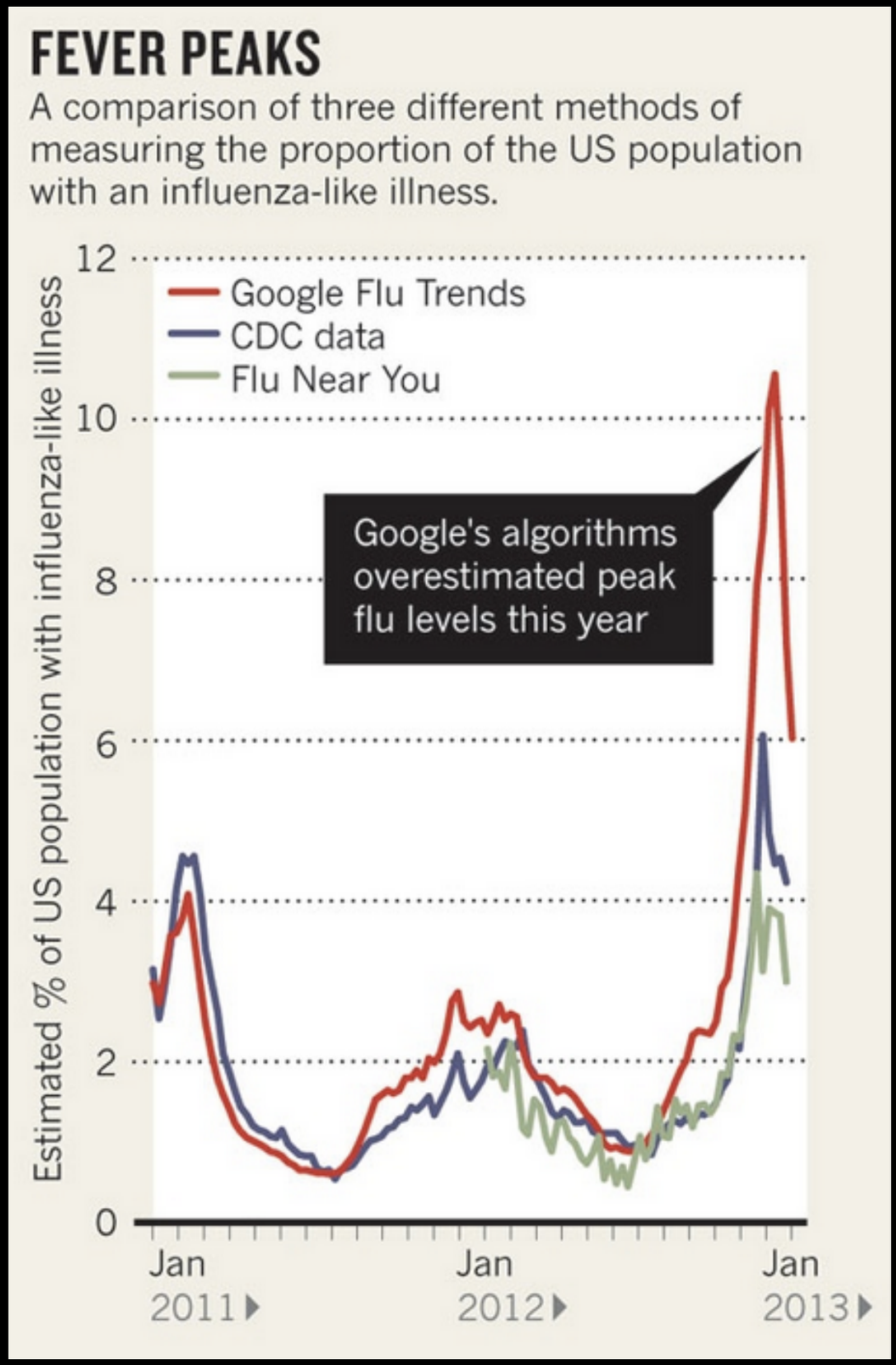
David Lazer[1,2,*], Ryan Kennedy[1,3,4], Gary King[3], Alessandro Vespignani[5,6,3]

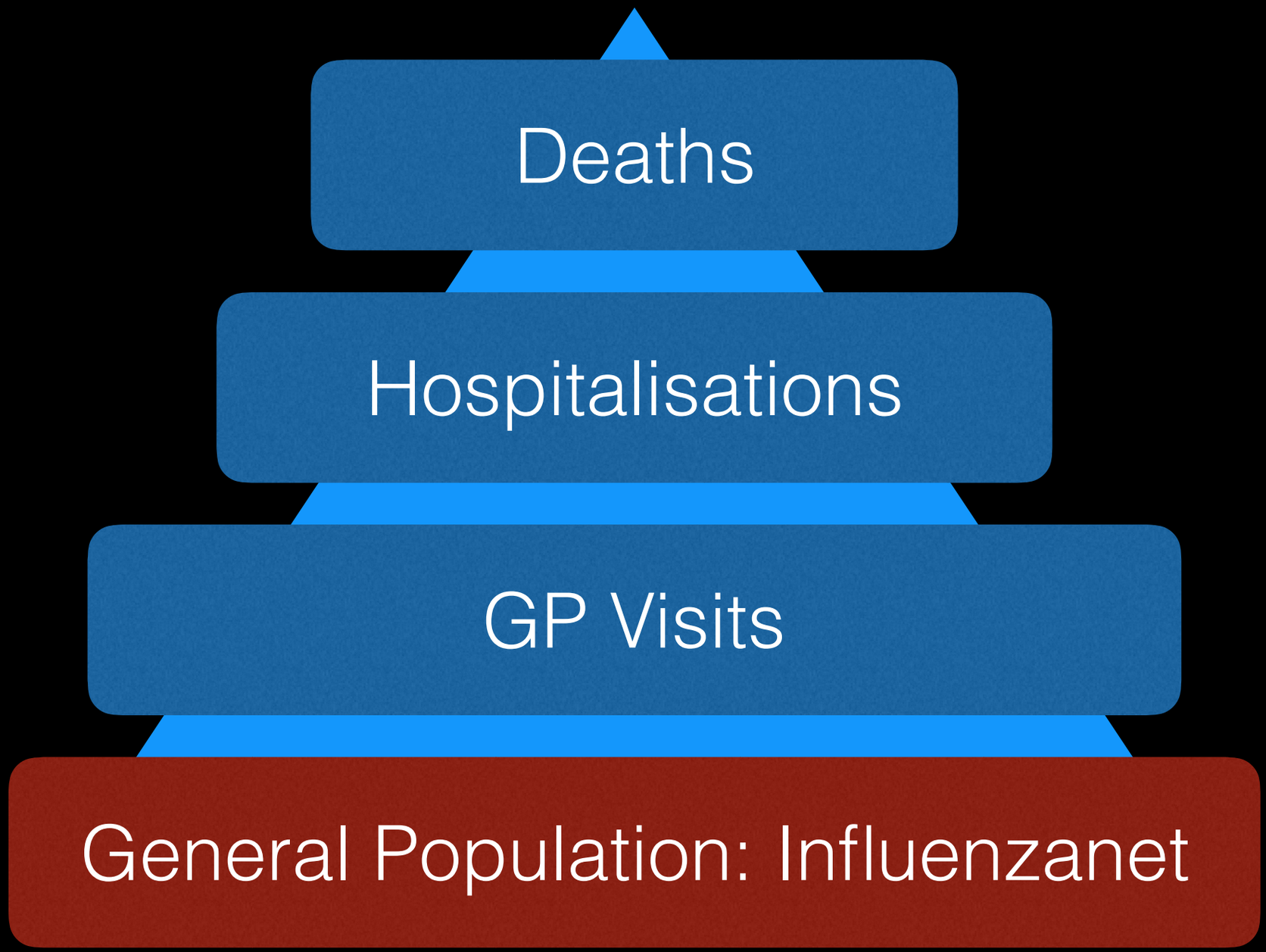
± Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

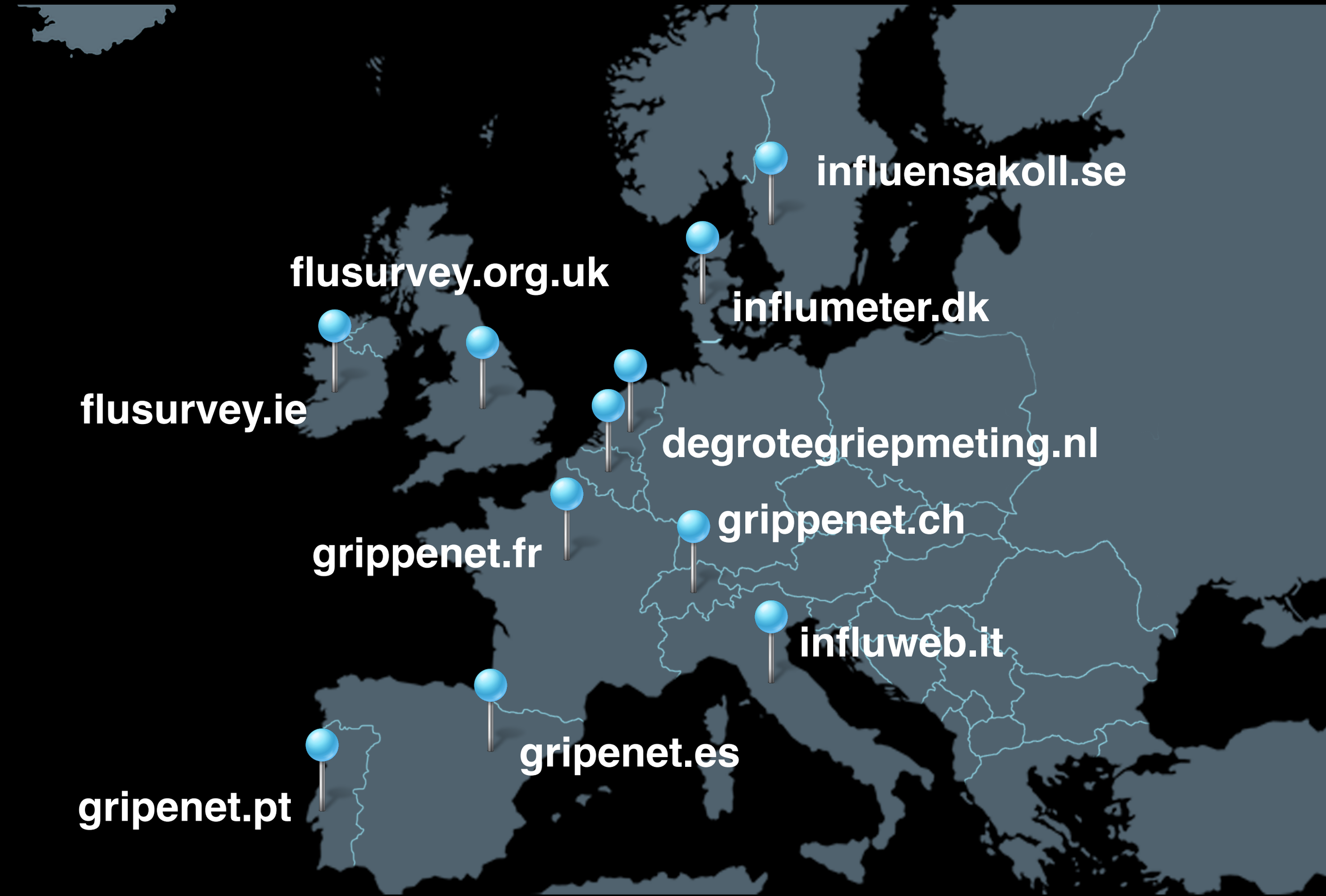
In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that G
executives or the creators of the flu tracking system would have hoped. *Nature* report
predicting more than double the proportion of doctor visits for influenza-like illness (
Centers for Disease Control and Prevention (CDC), which bases its estimates on survei
from laboratories across the United States (*1*, *2*). This happened despite the fact that
predict CDC reports. Given that GFT is often held up as an exemplary use of big data (
lessons can we draw from this error?

- Passive data sources don't describe who is well
- Low specificity

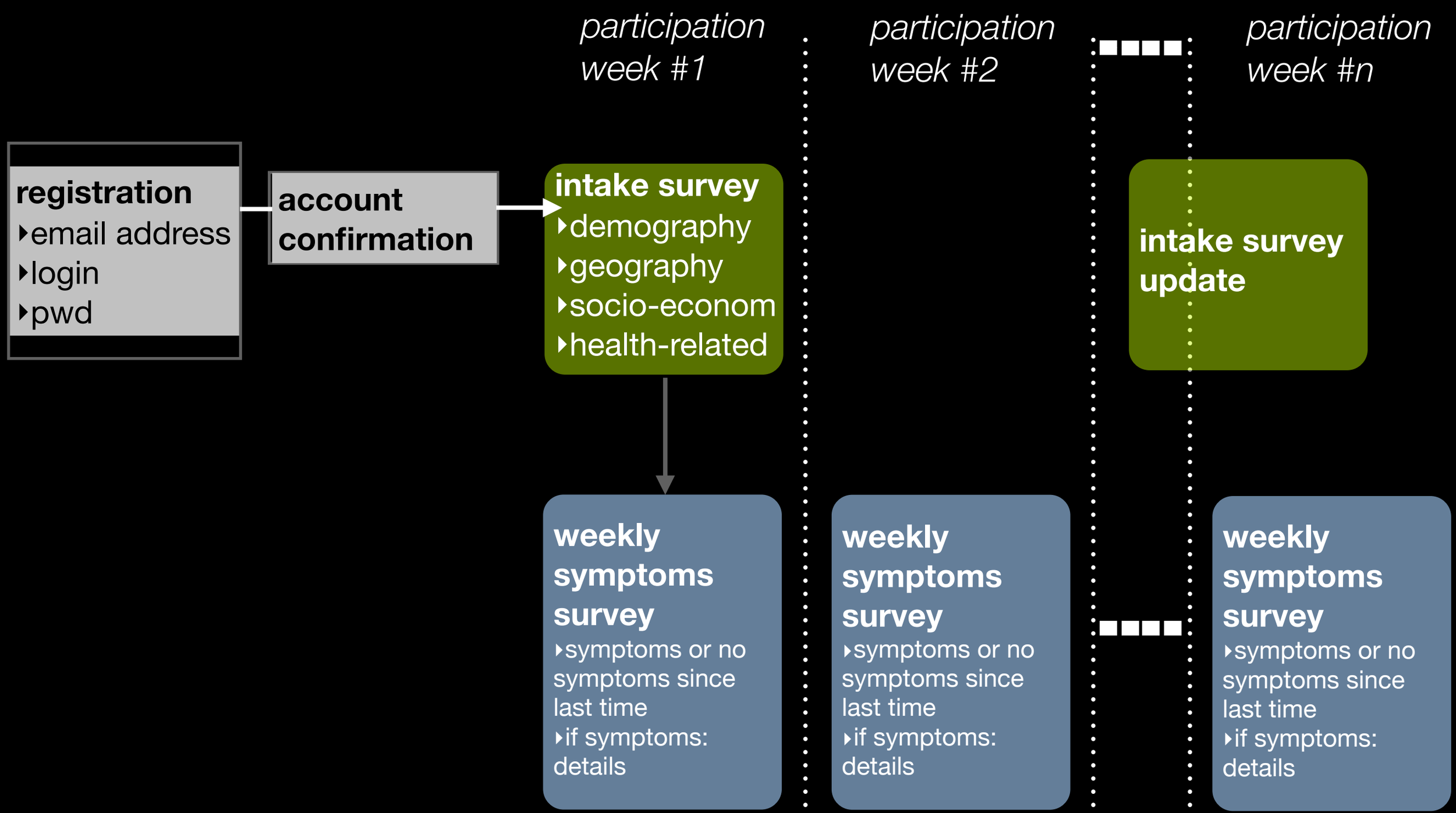



Deaths
Hospitalisations
GP Visits
General Population: Influenzanet

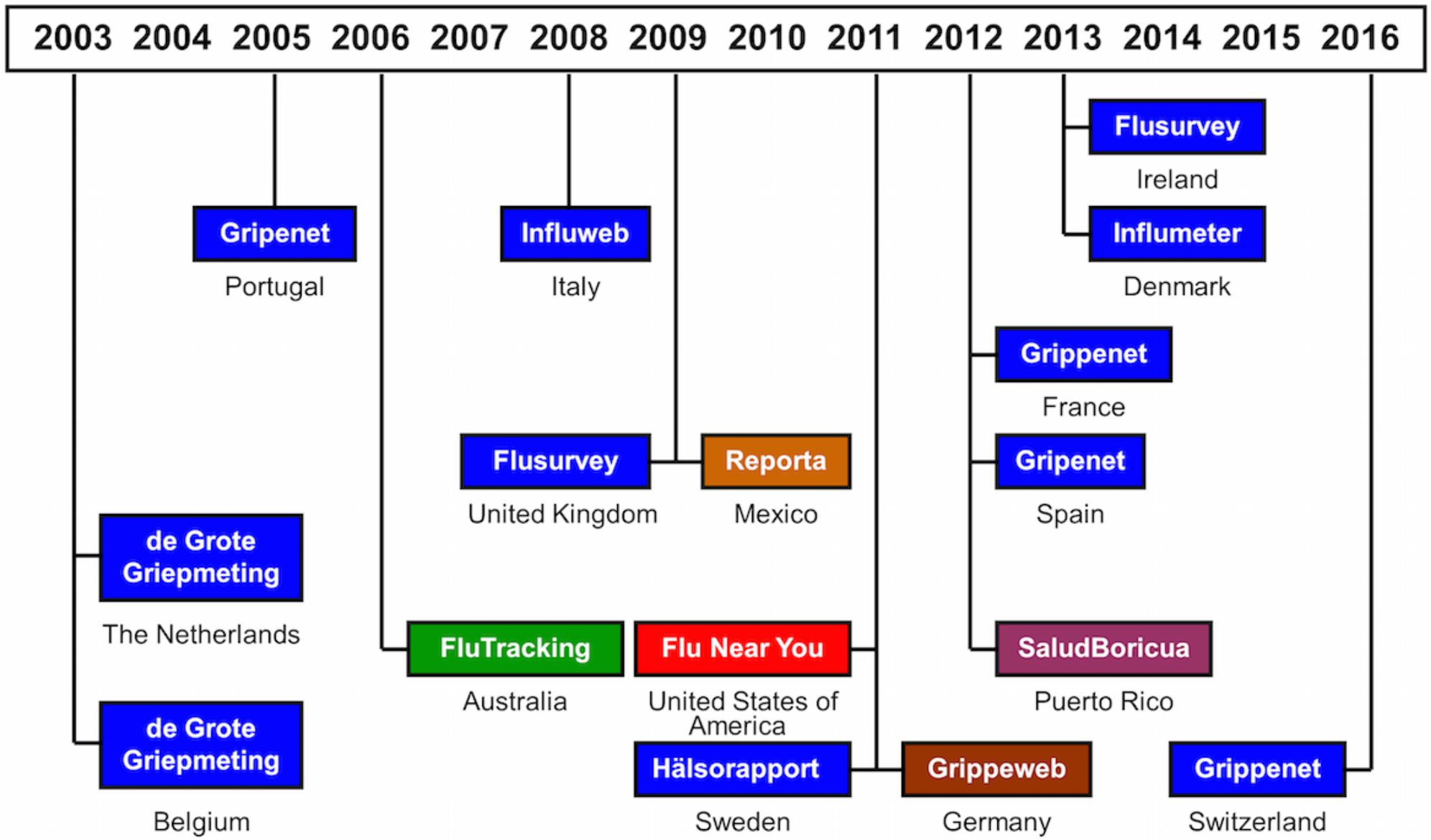
INFLUENZANET.INFO



INFLUENZANET study design

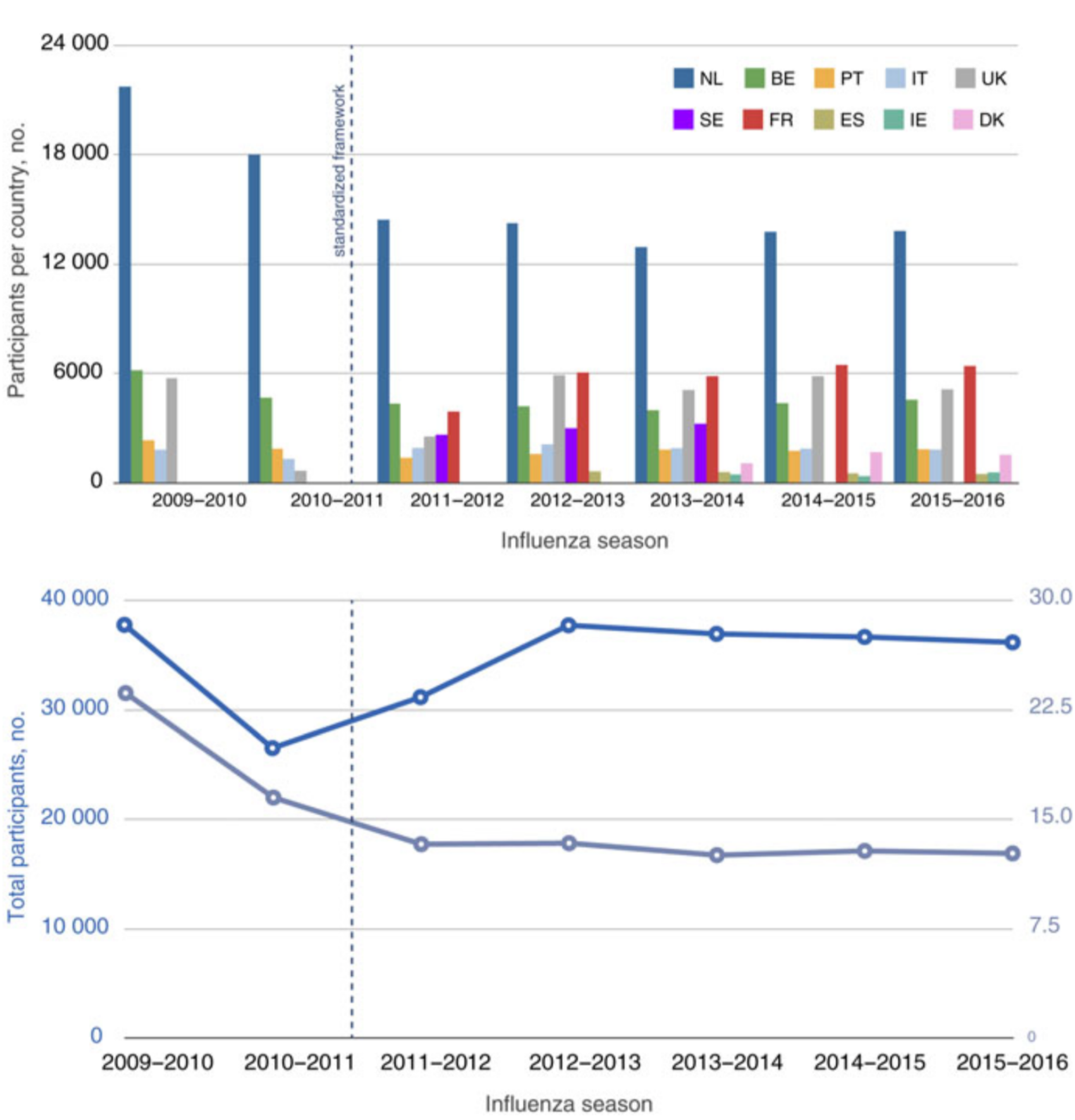


INFLUENZANET: a timeline



Koppeschaar CE, Colizza V, Guerrisi C, Turbelin C, Duggan J, Edmunds WJ, Kjelsø C, Mexia R, Moreno Y, Meloni S, Paolotti D, Perrotta D, van Straten E, Franco AO
Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe
JMIR Public Health Surveill 2017;3(3):e66

C. Guerrisi et al, *Participatory Syndromic Surveillance of Influenza in Europe* Journal of Infectious Diseases (2016) 214 (suppl 4): S386-S392.



INFLUENZANET.INFO

influweb.it - ISI Foundation & Istituto Superiore di Sanità, Italy

grippenet.fr - INSERM, France

gripenet.pt - Instituto Nacional de Saúde Doutor Ricardo Jorge, Portugal

flusurvey.net - Public Health England

influensakoll.se - Public Health Agency of Sweden

influmeter.dk - Staten Serum Institute, Denmark

grippenet.ch - Global Health Institute, Geneva

grippeweb.rki.de - Robert Koch Institute, Germany

influweb

grippenet.fr

gripenet

flusurvey

www.influensakoll.se
Kartlägger förekomst och spridning av influensa i Sverige

influmeter.dk
Kortlægger forekomst og spredning af influenza i Danmark

grippenet

GrippeWeb

Flu News Europe

Joint ECDC–WHO/Europe weekly influenza update

Week 02/2020 (6-12 January 2020)

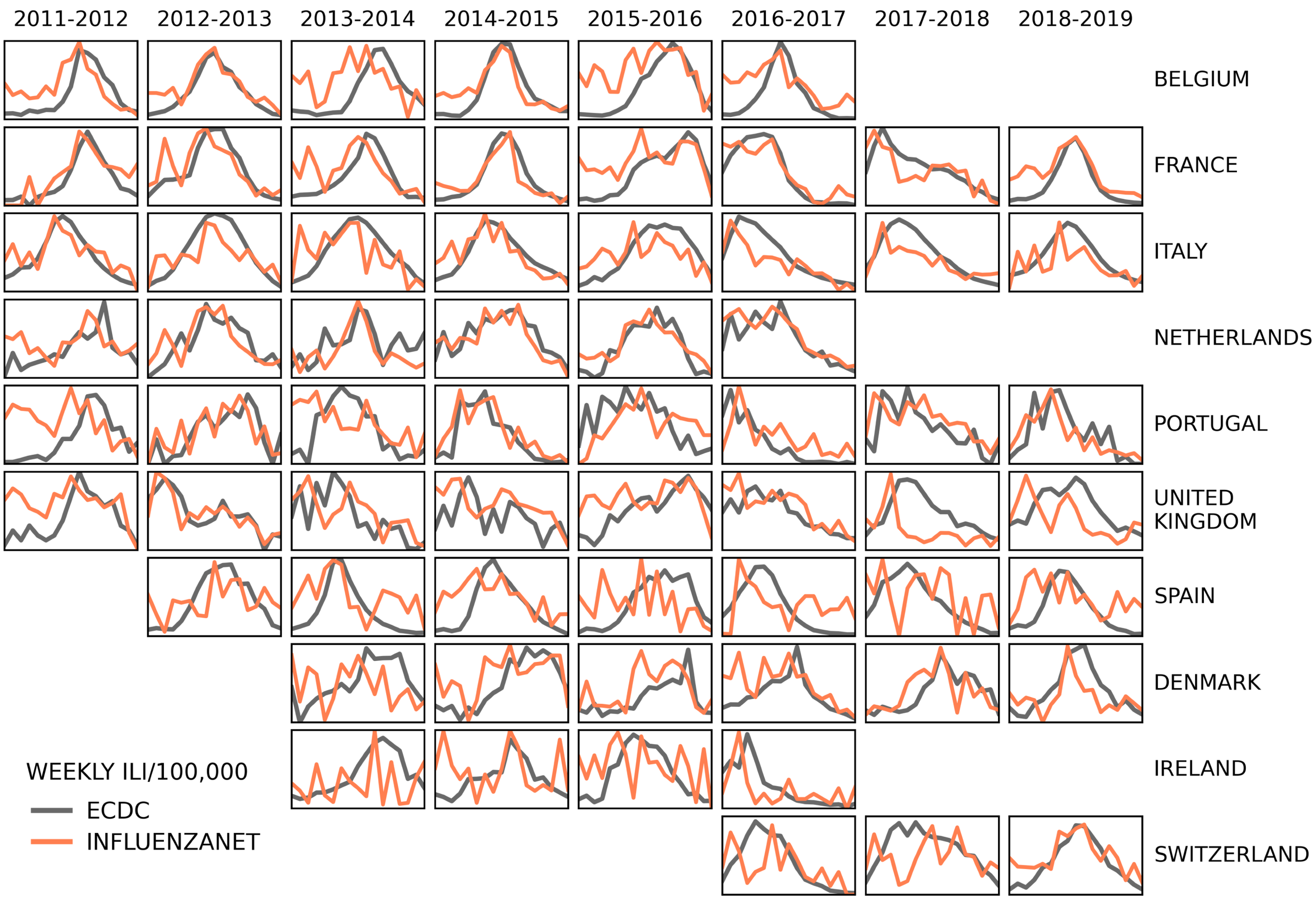
- Activity increased compared to week 01/2020, particularly in the southern part of the Region, with two Member States reporting high intensity and six reporting medium intensity. The remainder reported baseline or low intensity levels.
- The percentage of samples from sentinel ILI surveillance patients that tested positive for influenza virus increased from 27% in the previous week to 40% this week.
- The majority of reported influenza virus detections from sentinel ILI surveillance across the Region for week 02/2020 were type A (67%): this percentage has decreased from a high of 78% in week 49. The distribution of viruses detected varied between Member States and areas and within sub-regions.
- Data from the 22 countries or regions reporting to the EuroMOMO project indicated that all-cause mortality was at expected levels for this time of the year.
- ECDC published an Influenza virus characterization report, summarizing surveillance data in Europe through December 2019

2019/20 season overview

- For the Region as a whole, influenza activity commenced earlier than previous years.
- Influenza activity in the European Region, based on sentinel sampling, first exceeded a positivity rate of 10% in week 47/2019 and has remained over 10% for 8 weeks. There has been an overall increasing trend in the weekly positivity rate for influenza virus detections among sentinel ILI surveillance patients, following a dip in week 52.
- Type A viruses have dominated across the European Region, though several Member States and areas have reported influenza type B virus dominance or co-dominance of types A and B viruses.
- In sentinel sources, both influenza A subtypes, A(H3N2) and A(H1N1)pdm09, are co-circulating and of the influenza B viruses, the vast majority (98%) have been B/Victoria lineage.
- Influenzanet, which uses self-reported symptoms for ILI surveillance in the general population of European countries, is included in the bulletin as a pilot for the 2019/2020 season.
- ECDC and WHO Regional Office published a joint Regional Situation Assessment for the 2019–2020 influenza season up to week 49/2019, which focused on disease severity and impact on healthcare systems to assist forward planning in Member States.

Influenza intensity, spread and dominant virus type/subtype

Intensity | 2020-W02 | Map type EU Map | Export



D. Paolotti et al., Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience, Clinical Microbiology and Infection, 20(1), 17 (2014)

What is the definition of Influenza-like illness?

Sudden onset of symptoms

AND

at least one of the following four systemic symptoms:
Fever or feverishness, Malaise, Headache, Myalgia

AND

at least one of the following three respiratory symptoms:
Cough, Sore throat, Shortness of breath

ECDC case definition

“What if the observed symptoms are the result of a superposition of latent syndromes characterised by an unknown incidence and an unknown composition in terms of symptoms?”

Data

	NL	BE	IT	FR	UK	ES	PT	DK	IE
Number of seasons	6	6	6	6	6	5	6	4	4
Average participants per season	13,450	4,209	1,830	5,757	4,676	526	1,663	1,391	406
Average # surveys per season	206,987	67,420	17,807	68,567	45,543	5,894	17,852	22,782	3,220
Average % of surveys with symptoms	20%	16%	19%	20%	29%	22%	17%	18%	25%
Average of surveys per participant per season	15	16	9	12	9	11	10	16	8

Kalimeri et al, Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms, Plos Computational Biology 15(4): e1006173

Weekly Symptoms Survey

1. Fever
2. Chills
3. Runny/blocked nose
4. Sneezing
5. Sore throat
6. Cough
7. Shortness of breath
8. Headache
9. Muscle/joint pain
10. Chest pain
11. Feeling tired (malaise)
12. Loss of appetite
13. Phlegm
14. Watery, bloodshot eyes
15. Nausea
16. Vomiting
17. Diarrhoea
18. Stomachache
19. Sudden Onset

time series of daily symptoms counts

boolean variables

$$\mathbf{X} = [x_{ij}]$$

matrix whose elements contains the occurrence of symptom j on day i

Kalimeri et al, Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms, Plos Computational Biology 15(4): e1006173

Latent Syndromes detection

- it is reasonable to expect that a specific combination of symptoms reported by a user is the symptomatic expression of one or more illnesses, i.e. syndromes, experienced by the user.

- In accordance with this consideration, we postulate that the time series x_{ij} of observed symptoms counts are the result of a linear mixing process driven by K unknown sources, corresponding to the latent syndromes we want to detect.

$$x_{ij} = \sum_{k \in \{1,..,K\}} w_{ik}\, h_{kj} + e_{ij}.$$

Latent Syndromes detection

The mixing equations can be expressed in matrix notation:

$$\mathbf{X} = \mathbf{WH} + \mathbf{E}$$

$$\mathbf{W} = [w_{ik}], \mathbf{H} = [h_{kj}], \mathbf{E} = [e_{ij}]$$

In this notation, the problem of detecting the unknown K latent sources can then be formulated as a matrix decomposition problem

Non-negative matrix factorization

The specific factorization algorithm we used in this study is a non-negative matrix factorization (NMF) minimizing the Kullback-Leibler loss function:

$$argmin_{W,H} \sum_{i,j} x_{ij} \log\left(\frac{x_{ij}}{\hat{x}_{ij}}\right) - x_{ij} + \hat{x}_{ij}$$

where $\hat{x}_{ij} = \sum_k w_{ik} h_{kj}$

- this allows a probabilistic interpretation of the decomposition results and, as a consequence, a principled probabilistic way of choosing the intrinsic number *K* of latent sources or components, based on the model likelihood.

Non-negative matrix factorization

By leveraging on the same probabilistic framework, previously used in the context of semantic analysis of text corpora, we can then interpret the results of the decomposition of X as a mixture of multinomials.

From this probabilistic point of view, by decomposing the matrix X, we are effectively estimating the parameters of a probabilistic model containing a hidden variable which corresponds to the latent component we are looking for and approximating the observed daily proportions of symptoms:

$$\pi(i,j) = x_{ij}/N, \quad N = \sum_{i,j} x_{ij}$$

Non-negative matrix factorization

$$\pi(i,j) \approx p(i,j) = \sum_k p(k)\, p(i,j|k)$$

$$= \sum_k p(k)\, p(i|k)\, p(j|k)$$

mixture of conditionally independent multinomials

$$p(i|k) = \quad w_{ik} / \sum_i w_{ik}, \qquad where \sum_i p(i|k) = 1$$

$$p(j|k) = \quad h_{kj} / \sum_j h_{kj}, \qquad where \sum_j p(j|k) = 1,$$

$$p(k) = \quad N \sum_i w_{ik} \sum_j h_{jk} \qquad where \sum_k p(k) = 1.$$

Non-negative matrix factorization

- The total number of counts N will be proportionally split among K latent components according to p(k).
- These counts will in turn be distributed in each day i according to p(ilk) and finally contribute to the daily symptoms counts according to p(jlk), which describes each component in terms of the expected proportion of symptoms.
- According to this formulation, the total number of counts associated to a latent component k in day i will be given by:

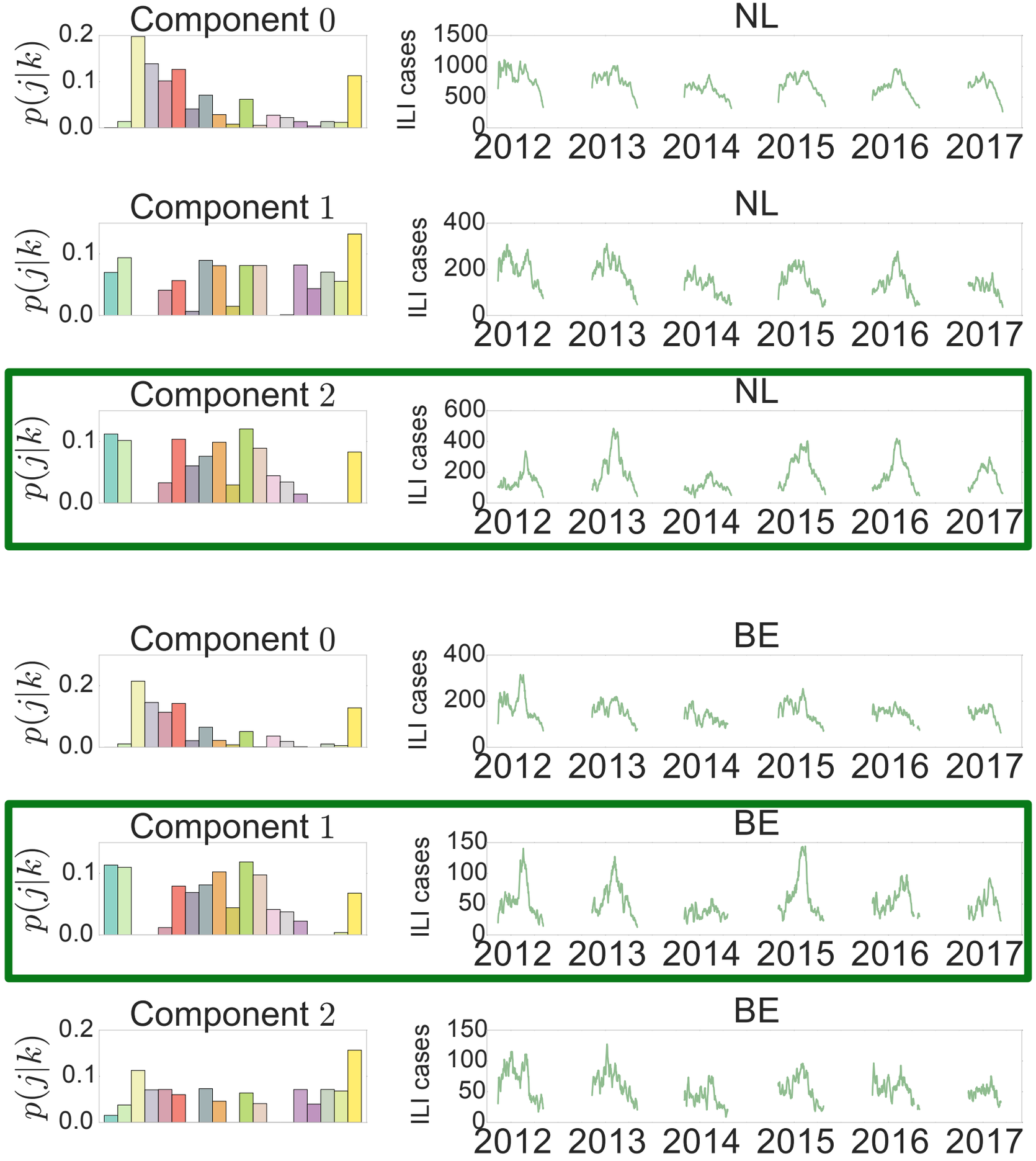
$$y_{ik} = N\,p(i,k) = N\,p(k)\,p(i|k)$$

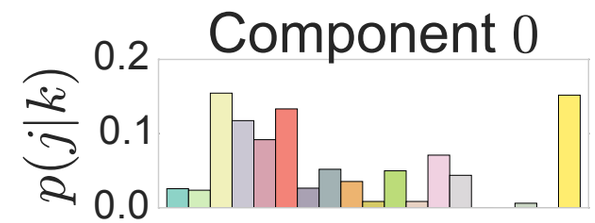
Model selection

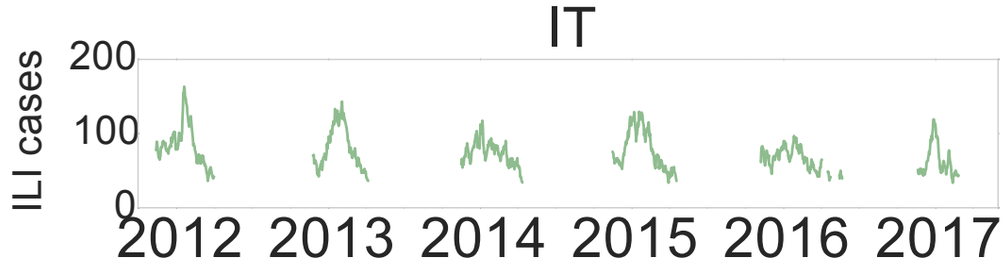
Given a set of candidate models obtained by minimizing the loss function by using an increasing number of hidden components *K*, we would like to select the best one in terms of its ability to correctly describe the observed phenomenon.The expected value of the Kullback-Liebler loss can be estimated in the asymptotic limit $N \rightarrow \infty$ leading to an approximated model selection criterion by means of the Akaike Information Criterion:

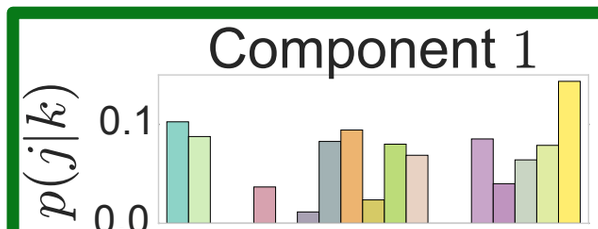
$$AICc = -2L(K) + 2P + 2\frac{P(P+1)}{N-P-1},$$

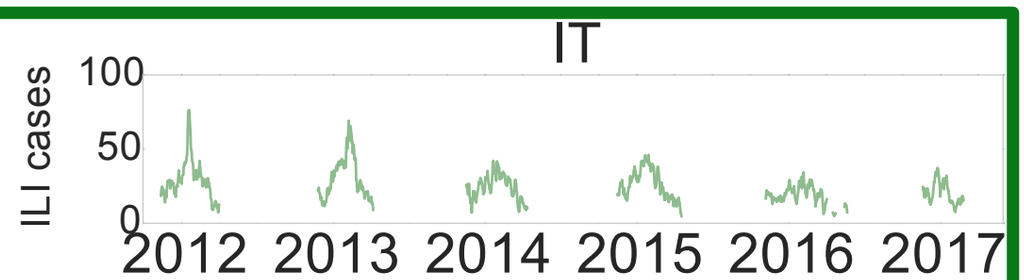
where L(K) is the log-likelihood of the model with K latent components, P is the number of effective parameters of the model: P = K \, (I + J - 2) - 1 and N is the the total number of counts. The best model in the set will be the one minimising AICc.

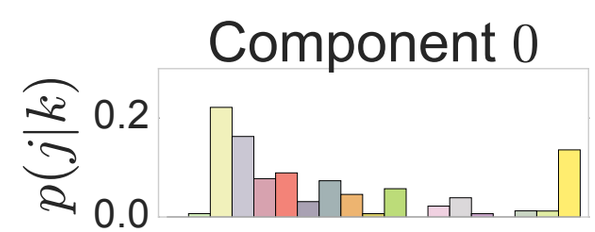

Component 0
$p(j|k)$
0.2
0.1
0.0
NL
ILI cases
1500
1000
500
0
2012 2013 2014 2015 2016 2017
Component 1
$p(j|k)$
0.1
0.0
NL
ILI cases
400
200
0
2012 2013 2014 2015 2016 2017
Component 2
$p(j|k)$
0.1
0.0
NL
ILI cases
600
400
200
0
2012 2013 2014 2015 2016 2017
Component 0
$p(j|k)$
0.2
0.0
BE
ILI cases
400
200
0
2012 2013 2014 2015 2016 2017
Component 1
$p(j|k)$
0.1
0.0
BE
ILI cases
150
100
50
0
2012 2013 2014 2015 2016 2017
Component 2
$p(j|k)$
0.2
0.1
0.0
BE
ILI cases
150
100
50
0
2012 2013 2014 2015 2016 2017

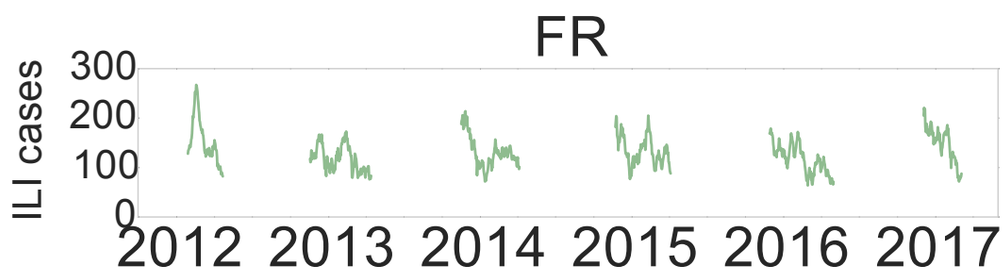

Component 0
$p(j|k)$
0.2
0.1
0.0

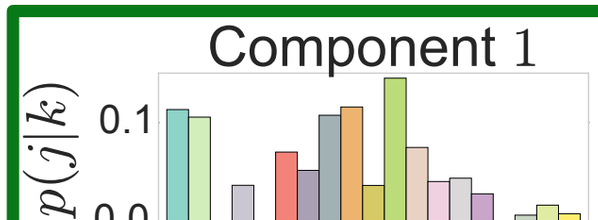

IT
ILI cases
200
100
0
2012
2013
2014
2015
2016
2017

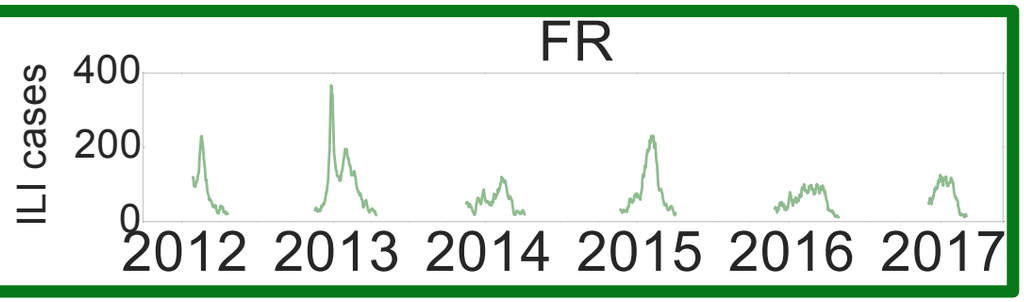

Component 1
$p(j|k)$
0.1
0.0

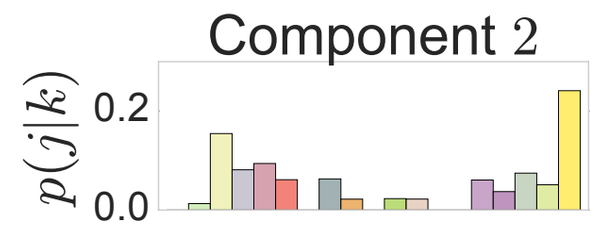

IT
ILI cases
100
50
0
2012
2013
2014
2015
2016
2017

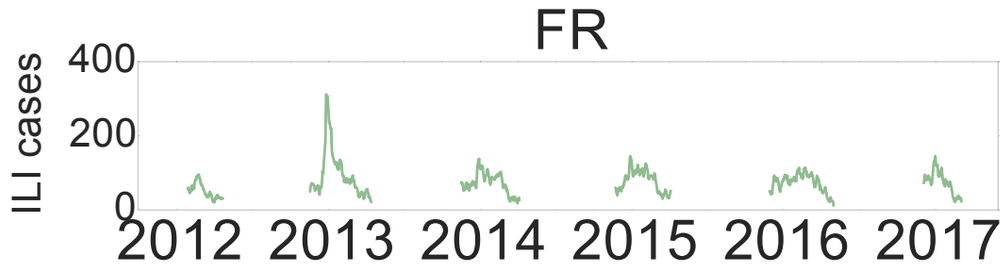

Component 0
$p(j|k)$
0.2
0.0

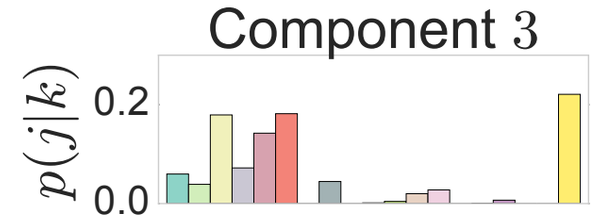

FR
ILI cases
300
200
100
0
2012
2013
2014
2015
2016
2017

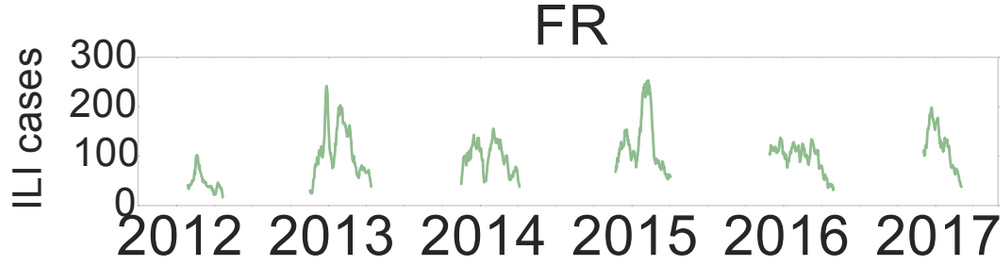

Component 1
$p(j|k)$
0.1
0.0

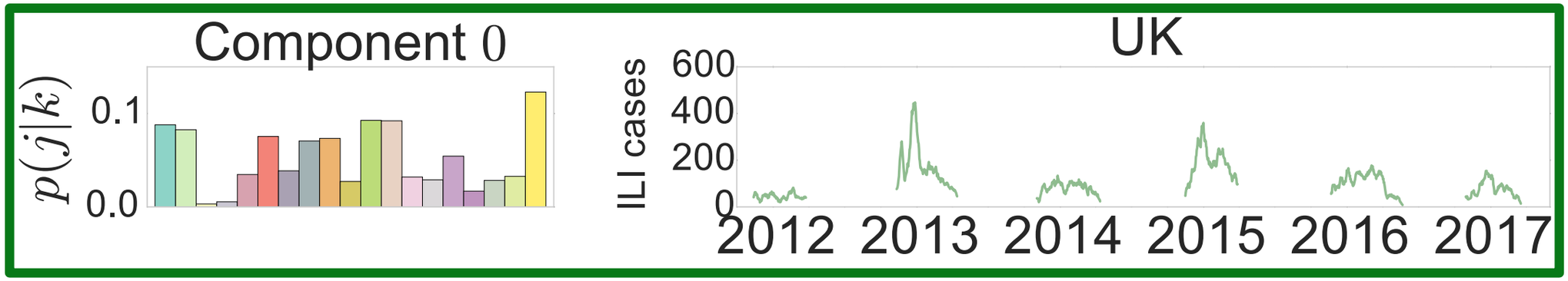

FR
ILI cases
400
200
0
2012
2013
2014
2015
2016
2017

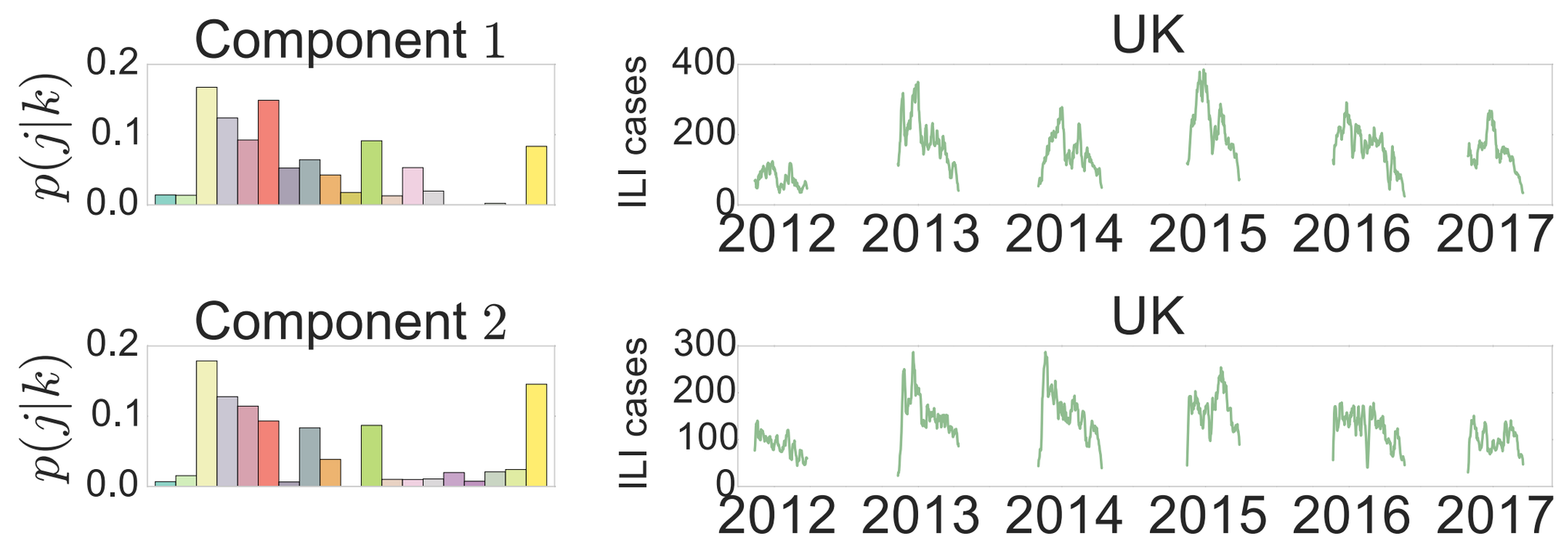

Component 2
$p(j|k)$
0.2
0.0


FR
ILI cases
400
200
0
2012
2013
2014
2015
2016
2017

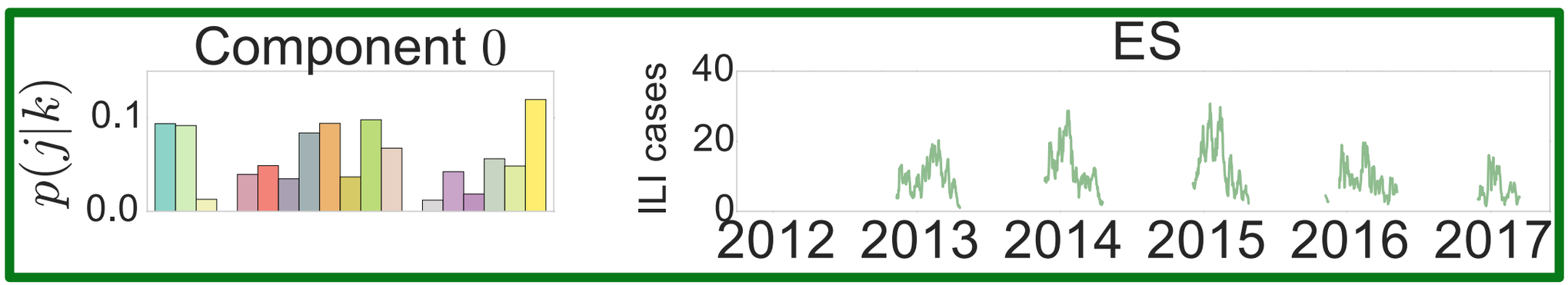

Component 3
$p(j|k)$
0.2
0.0

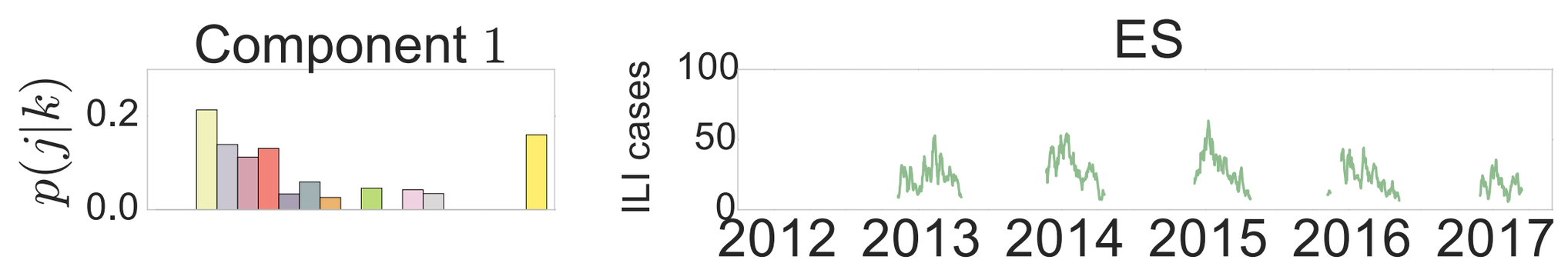

FR
ILI cases
300
200
100
0
2012
2013
2014
2015
2016
2017

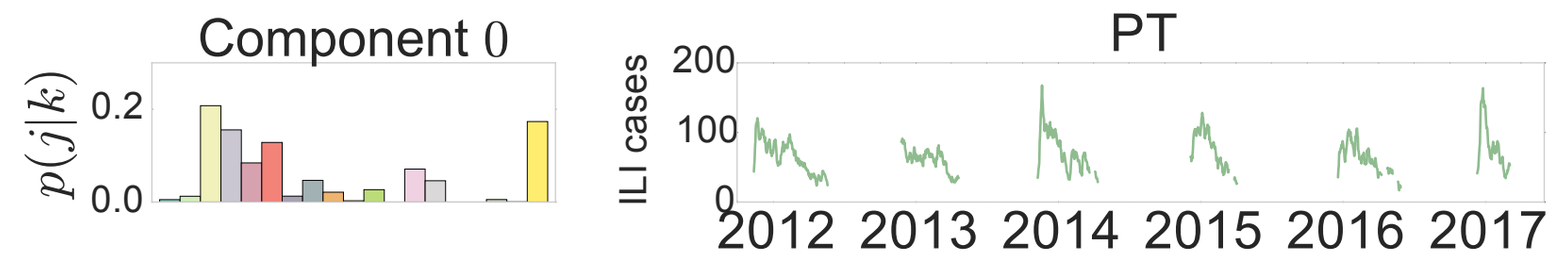

Component 0
$p(j|k)$
0.1
0.0
UK
ILI cases
600
400
200
0
2012 2013 2014 2015 2016 2017

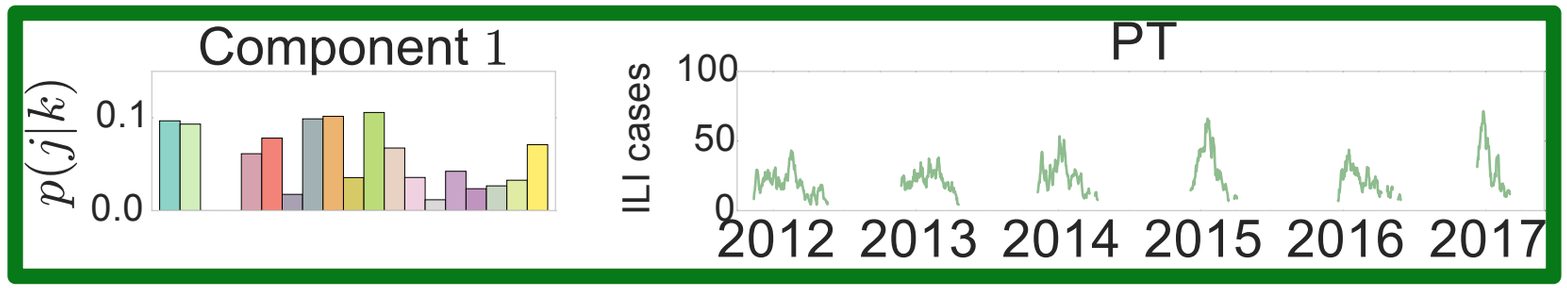

Component 1
$p(j|k)$
0.2
0.1
0.0
UK
ILI cases
400
200
0
2012 2013 2014 2015 2016 2017

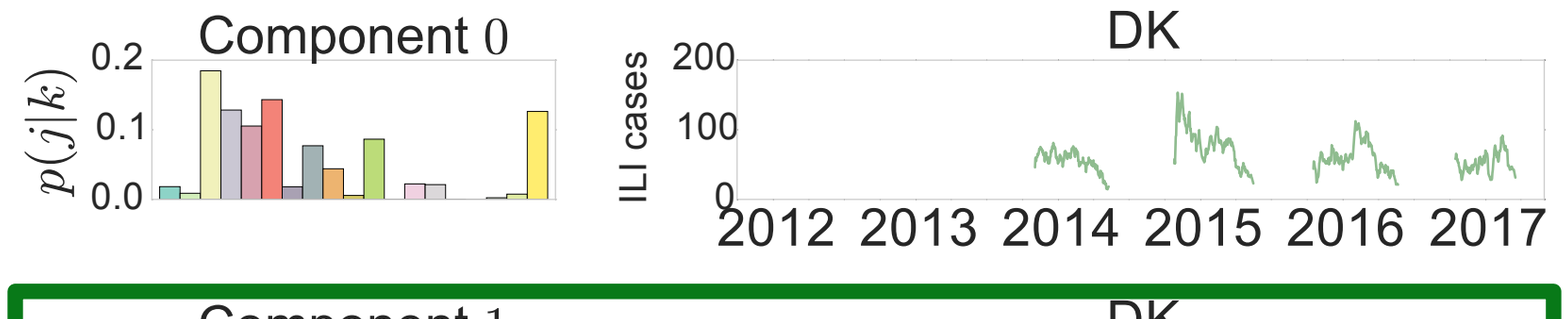
Component 2
$p(j|k)$
0.2
0.1
0.0
UK
ILI cases
300
200
100
0
2012 2013 2014 2015 2016 2017

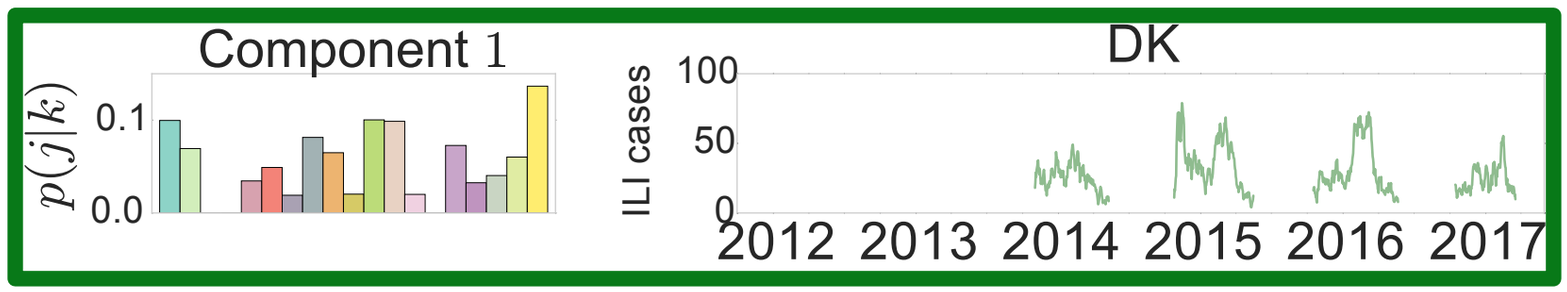

Component 0
$p(j|k)$
0.1
0.0
ES
ILI cases
40
20
0
2012 2013 2014 2015 2016 2017

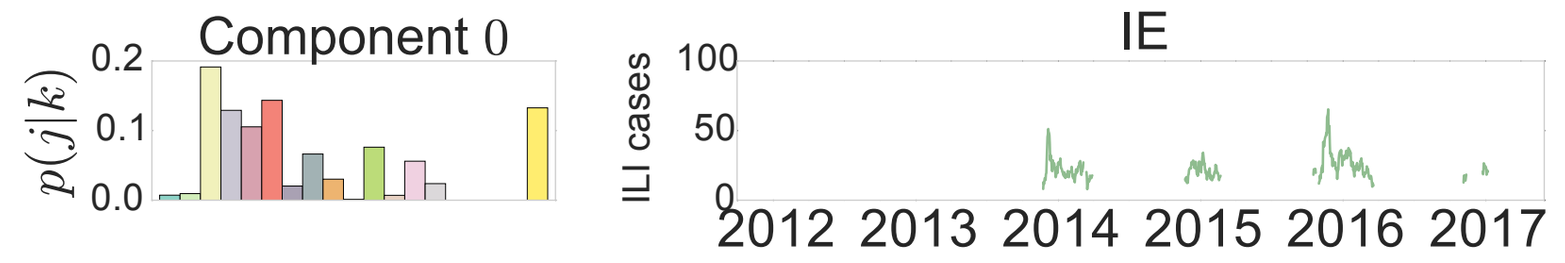

Component 1
$p(j|k)$
0.2
0.0
ES
ILI cases
100
50
0
2012 2013 2014 2015 2016 2017

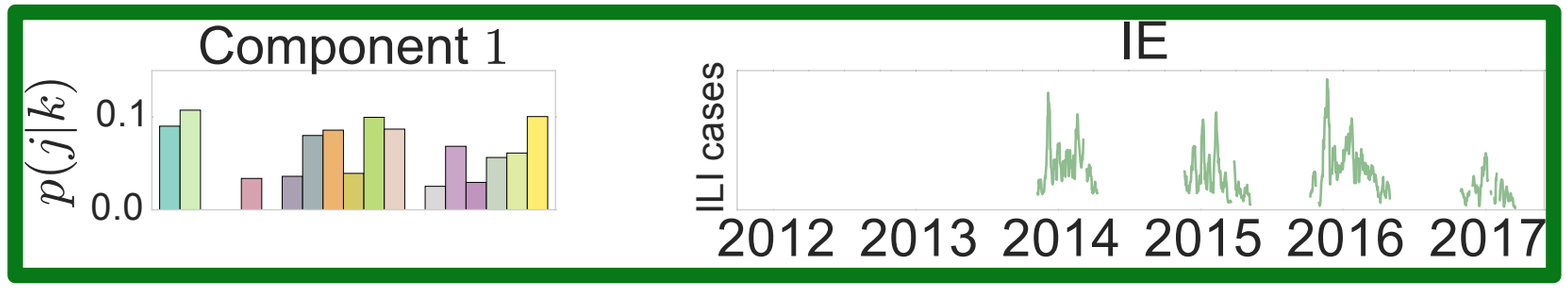

Component 0
$p(j|k)$
0.2
0.0
PT
ILI cases
200
100
0
2012 2013 2014 2015 2016 2017


Component 1
$p(j|k)$
0.1
0.0
PT
ILI cases
100
50
0
2012 2013 2014 2015 2016 2017


Component 0
$p(j|k)$
0.2
0.1
0.0
DK
ILI cases
200
100
0
2012 2013 2014 2015 2016 2017


Component 1
$p(j|k)$
0.1
0.0
DK
ILI cases
100
50
0
2012 2013 2014 2015 2016 2017


Component 0
$p(j|k)$
0.2
0.1
0.0
IE
ILI cases
100
50
0
2012 2013 2014 2015 2016 2017


Component 1
$p(j|k)$
0.1
0.0
IE
ILI cases
2012 2013 2014 2015 2016 2017

NL

BE

IT

FR

UK

ES

PT

DK

IE

GP incidence: ILI/100,000

IN_NMF GP

Fever
Chills
Runny nose
Sneezing
Sore thoat
Cough
Breath shortness
Headache
Muscle pain
Chest pain
Feeling tired
Loss of appetite
Coloured sputum
Watery eyes
Nausea
Vomiting
Diarrhoea
Stomach ache
Sudden Onset

$p(j|k)$

	NL	BE	IT	FR	UK	ES	PT	DK	IE
(i) Correlation between IN_ECDC and IN_NMF for the seasons 2011-2017									
	0.91	0.92	0.86	0.83	0.92	0.86	0.84	0.90	0.82
(ii) Correlation between IN_NMF and GP for the seasons 2011-2017									
	0.88	0.80	0.69	0.79	0.74	0.65	0.66	0.71	0.38
(iii) Correlation between IN_ECDC and GP for the seasons 2011-2017									
	0.79	0.72	0.80	0.86	0.75	0.67	0.63	0.68	0.23
(iv) Correlation between IN_NMF prediction for 2016-2017 and GP for the season 2016-2017									
	0.85	0.82	0.69	0.80	0.60	0.84	0.80	0.76	0.60
(v) Correlation between IN_ECDC and IN_NMF for the season 2016-2017									
	0.85	0.82	0.86	0.93	0.67	0.59	0.88	0.80	0.71

Kalimeri et al, Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms, Plos Computational Biology 15(4): e1006173

What's next - I?

- Virological confirmation is needed to estimate more accurately the scaling factor

- extension of the method to other countries and syndromes

- Assess the validity of the method for detection of new emerging diseases

Kalimeri et al, Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms, Plos Computational Biology 15(4): e1006173

INFLUENZANET.INFO

Daniela Paolotti, Daniela Perrotta (IT)

Vittoria Colizza, Clement Turbelin (FR)

Carl Koppeschaar, Ronald Smallenburg (NL, BE)

Yamir Moreno (ES)

Ricardo Mexia (PT)

Richard Pebody (UK)

Edward Van Straten (SE)

Charlotte Kjelsø (DK)

Jim Duggan (IE)

Antoine Flahaut (CH)

Udo Bucholz (DE)

- *D. Perrotta*
- *M. Delfino*
- *K. Kalimeri*
- *C. Cattuto*

ISI

ISI Foundation
& ISI Global Science
Foundation

Thank you!

–Daniela Paolotti
@danielapaolotti