Local-to-global analysis of influenza-like-illness data

J. Pita Costa, F. Fuart, L. Stopar
Quintelligence, Jozef Stefan Institute, Slovenia

D. Paolloti
ISI Foundation, Italy

M. Hirsch
German Research Center for AI (DFKI), Germany

R. Mexia
Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA), Portugal

P. Carlin
South Eastern Health and Social Care Trust, UK

J. Wallace
Ulster University, UK

ABSTRACT

The need for appropriate, robust and efficient epidemic intelligence tools is increasing in this age of a connected society. Global health initiatives, such as Influenzanet, potentially have a central role in the future of Public Health. This paper presents the contributions to the Influenzanet initiative, describing a new monitoring system for local hubs and their data sources, based on Elasticsearch.

It is often the case that the exploration of internally generated data is prioritised by national public health institutions, and therefore cannot be addressed in the global Influenzanet platform. This platform can be used by health professionals without programming expertise to encourage and enhance their independence from busy in-house IT departments and further contribute to the effectiveness of their own research.

The most meaningful data visualization modules can then be considered for integration into the full Influenzanet platform that will serve the complete network, thus collaborating at a global level. With this approach we also show the importance that an active hub in carrying out its own investigations towards its own priorities. In that regard and as an example, we also describe new results on the application of state-of-the-art approaches to a local data set, using the Portuguese ILI seasons between 2005 and 2013. This study is based on the application of the Streamstory approach. It aims to show the potential of this versatile approach in: (i) identifying data-driven ILI seasons; (ii) relating the ILI incidence to the dimensions of weather data; and (iii) comparing the incidence throughout four different ILI definitions.

CCS CONCEPTS

• Real-time systems • Data management systems • Life and medical science

KEYWORDS

Public health, Influenzanet, ILI, Elasticsearch, Streamstory

1 Introduction

With the recent worldwide threats to health being reported throughout the media, the need for efficient epidemic intelligence tools is paramount. It is important to note that the influenza virus is also part of these epidemic threats requiring monitorization, despite its less mediatic weight. The speed of mutation of the virus makes any epidemic unpredictable. Its socioeconomic impact is evident in the number of workplaces affected every year during the season and the associated mortality in particular demographic groups (very young, very old). Influenzanet is a participatory surveillance monitoring system based on volunteers, submitting an online symptom questionnaire on a weekly basis, this enables a real-time global view of the incidence of influenza-like illness (ILI) across Europe. Note that the confirmation of influenza virus would require biological evidence and, thus, (often expensive) sample collection. The data set is collected in real time by the Influenzanet system and each volunteer provides a profile survey (including important information such as being a smoker, usual transport, etc.) and the weekly questionnaire of symptoms (see an example of the latter in Figure 1). The latter gathers the information that permits the identification of the presence of ILI that can be defined in at least four different ways according to the symptoms considered [6].

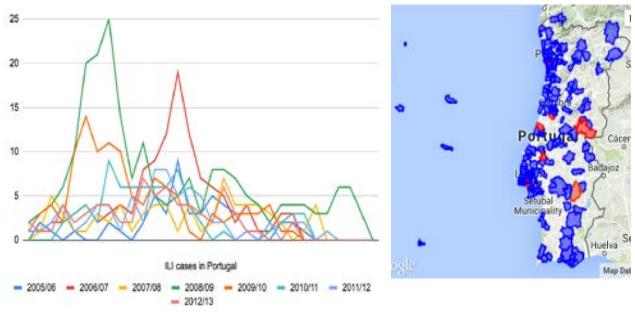

Figure 1: Incidence of influenza in Italy, between 2009 and 2013, collected by the local Influenzanet hub Gripenet.pt

2 Monitoring and exploring the local data

The Influenzanet system is deployed in more than ten European countries working in parallel under the same IT data collection framework, with some variety in the focus of the ILI monitoring [7]. This is usually aligned with the local public health priorities and ongoing studies, where most of the work is done by health experts, some of whom have some data science know-how or else are backed-up by in-house technical support.

To guarantee some independence to the less technical staff we have developed a data visualization dashboard that provides the user with real-time access to a local data set sourced on the national volunteer participants. This is based on Elastic Search technology, together with the Kibana open source data visualization plugin. Part of this work was developed in the context of the European Union research project MIDAS [3], by applying the know-how obtained

in building a similar system to monitor and manage the scientific knowledge open data set MEDLINE [5]. Note that the latter can be used to provide complementary information and be deployed in parallel to the Influenzanet dashboard.

The local Influenzanet data can be delivered to the dashboard through an API to the main platform. The update of the back-end system is driven by import scripts that appropriately load the new dataset into a new index in Elastic Search. This new Influenzanet-local index (comprised of one for surveys and the other for the symptom questionnaires) generates the database that serves the monitoring system. The dedicated dashboard based on Kibana has a native integration with Elastic Search and, therefore gets the index imported automatically to dynamically build the new visualization modules and dashboards. The public instance that can be derived from a dashboard is dependent of the choices in the definition of that dashboard.

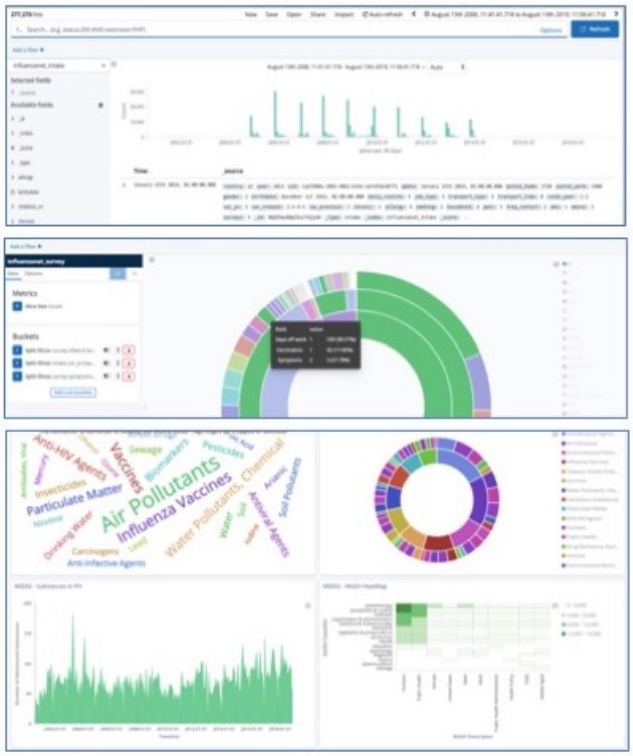

Figure 2: Dashboard of visual modules to monitor KPIs at each local Influenzanet hub, based on Elasticsearch and Kibana

It is often the case that specific interests in the local exploration of the Influenzanet data arise, directly related to the local public health priorities and ongoing studies. The possibility to explore their own local data through a technological tool offered as a service, can be of great value to the Influenzanet hubs (i.e., national Influenzanet-related institutions that collect the data). Although the data must be collected using a homogeneous approach to enable overall comparisons, the exploration of that data can target specific aims. The Elasticsearch-based system presented in this paper empowers the user with less technical expertise to build data visualization modules from subsets of the dataset that correspond to their own KPIs (Key Performance Indicators) they wish to monitor. This service will support an evidence-based policy-making by the national public health authority. The following are example queries made over the example data visualisation modules available in the Influenzanet dashboard:

- What are the most prominent symptoms per year?
- What is the coverage of Influenzanet surveys? (counts of questionnaires per country/year)
- When in the year the symptoms are more prevalent?
- What is the relationship between the incidence of ILI, the days at work and taking the Influenza vaccine?

The technical independence from the often busy IT departments enables health professionals to go further and faster in the exploration of their data through interactive visualization modules displayed through a dynamic dashboard (see Figure 2).

The architecture within the system relies on two useful tools provided by the Kibana technology (see Figure 3). The data collection is loaded by the local Influenzanet hub and is immediately made available at the Influenzanet data query dashboard, where the parameters of the data can be easily accessed and manipulated to subset the data or to produce powerful Lucene-based queries. With the saved subsets of data the user can create interactive visualization modules that will then integrate with the monitoring dashboard.

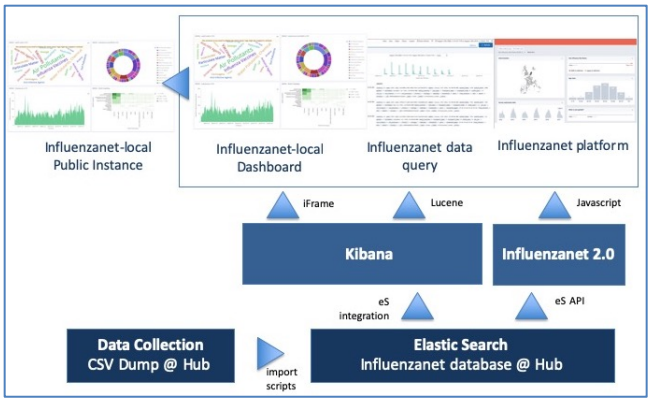

Figure 3: The architecture of the Elasticsearch-based system that enables the visualization of local own data at each Influenzanet hub

The Influenzanet network is preparing version 2.0 of its platform [6] that includes a modern architecture, making use of public APIs and storing the data locally, separating it by user data, survey data, and content data [4] [1]. The new Influenzanet platform will provide the Influenzanet hubs with a common set of data visualization modules. The plan is to augment "classical" Influenzanet data collection with additional sensor data from mobile phones [2]. Moreover, it includes a micro service architecture based system for better scalability and a more flexible development process. The backend of the new platform will be offered as Software as a Service (SaaS) but can also be downloaded as a self-hosted version. It will leverage this service in the

perspective of having access to the most meaningful data visualization modules throughout the Influenzanet hubs. The latter will be evaluated and can be integrated if they represent common value to other members of the Influenzanet network.

There are an ever increasing number of data sources that potentially could be used to gain new insights into areas such as disease prevention and policy formulation/evaluation, but these are not optimised for use within a data analytics type user interface. The **MIDAS** project was funded under a call for '*Big Data supporting Public Health policies*' to develop a big data platform that facilitates the utilisation of healthcare data beyond existing isolated systems, making that data amenable to enrichment with open and social data. This aligns closely with the efforts in Influenzanet, and the research work we have developed uses 5 year sample of this dataset. For this reason we made available a live demo page with videos and demos that can be shared with Influenzanet partners [8]. All of the tools and technologies presented in this paper are open source, available at the Quintelligence GitHub repository [11].

3 Using Streamstory to explore Influenzanet data

In the context of the visualization of complex data, the problem of visualization for the analysis and exploration of large multivariate time series is addressed by the Streamstory system [9][10]. This system computes and visualizes a hierarchical Markov chain model which captures the qualitative behaviour of the systems' dynamics. It provides us with a multi-scale representation of the data based on a hierarchical model which allows us to interactively find suitable scales for interpreting the data.

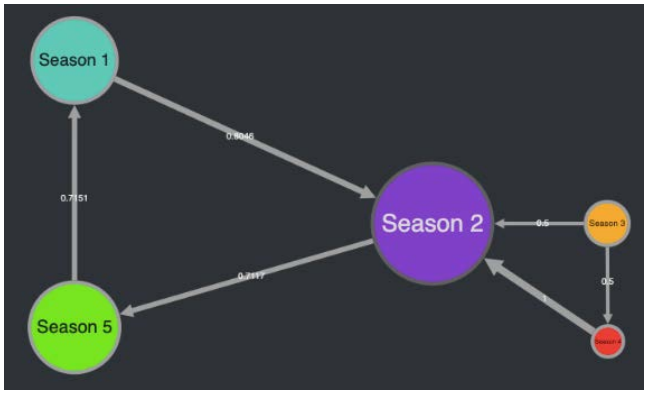

Figure 4: New data-driven seasons of the ILI incidence identified, subdividing the regular ILI season (marked as season 1, season 3, etc.), learning from historical Influenzanet data from Portugal during 2005-2013

We consider Streamstory in the context of the MIDAS project to look into the seasonality definition of ILI based on the sourced data from the Influenzanet platform. This system was developed by the AI Lab at the IJS and refocused by Quintelligence within the MIDAS project to visually analyse the Influenzanet dataset. It is Open Source. In this research, we consider the data across 8 seasons for Portugal, from 2005 to 2013 to try to identify time intervals during the ILI season where the dynamics of the time-series behaves similarly. In this first analysis we call data-season to each state and try to identify the most prominent ones throughout the ILI season. We can identify five seasons where most of the time is spent in the first and in the last data-seasons. Moreover, the data-seasons 3 and 4 seem to be skipped at times with a direct passage from data-season 2 to data-season 5 (see Figure 4). In a second analysis we used Streamstory to examine the relationship between the ILI incidence and two different dimensions of weather data: temperature and humidity (naturally correlated to rainfall). The diagram in Figure 5 shows that the second largest state is assigned to high humidity, eventually corresponding to the also high ILI incidence. The highlighted red coloured state of high ILI incidence is strongly related to high humidity but also low temperature which seems to be pointing to the weather in the end of winter.

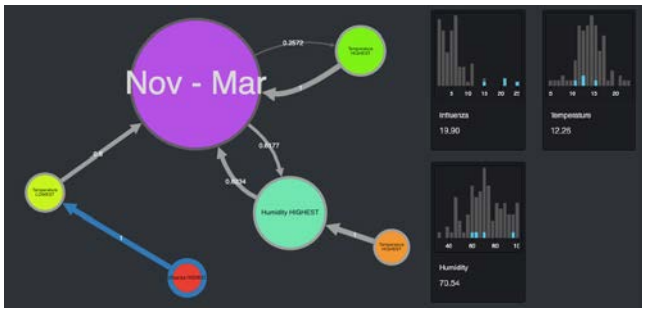

Figure 5: Correlation between the incidence of ILI and the different dimensions of weather data

A third analysis brings us to the comparison between the behaviour of ILI according to five coordinates, four of which corresponding to different ILI definitions: (i) historical, (ii) ECDC, (iii) including fever, and (iv) CDC. The diagram in Figure 6 shows the dynamics of the data when considering the 8 ILI seasons in Portugal altogether, highlighting the differences between the ILI definition used when counting incidence. The low incidence when considering ILI definition containing fever seems to have a strong expression in this analysis, followed by the high incidence of ILI defined using the historical and the ECDC definitions. The largest state is not related to any of the definitions in particular. When looking to its coordinates diagram, we can observe a higher influence of the ECDC definition, followed by the definition including fever. The CDC definition and the historical definition seem to have low weight in this largest state. A global analysis can be made through Streamstory to compare the ILI incidence in different countries. In the example of Figure 7 we compare five ILI seasons for Portugal and Italy. The close relation between the ILI behaviour in the two countries is usually similar between December and February, according to the diagram of states. Portugal and Italy tend to act distinctly in particular for the peaks of the epidemic, usually happening in Italy in November and February. Such a visualization might help us better understand the global behaviour of the epidemics throughout Europe, complementing the statistics provided by the Influenzanet platform itself.

Given the early stage of the Streamstory technology, it is difficult to have a clear global view on the obtained results and their meaning. Nevertheless, the approach seems to be promising, enabling a versatile analysis of the behaviour of the data through the diagrams of states, but also through the complementary coordinate diagram and the components histograms.

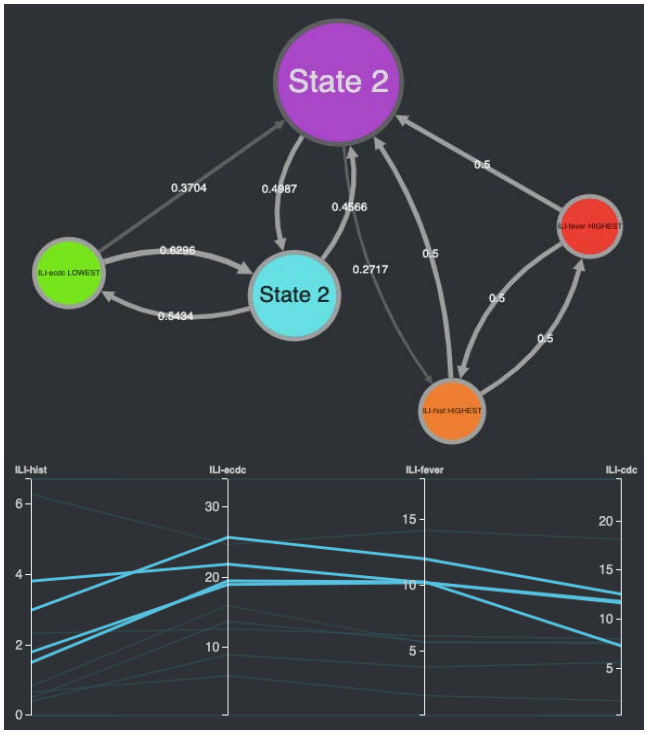

Figure 6: Comparison of the behaviour of ILI across four ILI definitions (with states generated by ILI cases per week)

4 Conclusions and further work

In this paper we discussed the research opportunities brought by the exploration of local data at each of the Influenzanet hubs. While the value of this data source is global, the local data explorations can target specific studies linked to priorities defined by the national public health institutes. The simplicity and ease of use of the discussed technology, offered as a service, permits the non-technical user to be much more independent of the in-house and overall technical support (often scarce in health organizations) to explore the local data in an almost real-time dimension. Moreover, some of the most meaningful data visualization models linked to those data explorations can be considered in the general platform to bring value to all the members of the Influenzanet network. Moreover, those visualization modules, common throughout the Influenzanet network, can provide means of comparison between countries and ILI seasons. Thus, the Kibana-based tool discussed in this paper can be of great value to digital epidemiology in general. Furthermore, the usage of advanced tools such as Streamstory might enable insights in the data that were otherwise unreachable. Versatile approaches such as this permit us to study the behaviour of the data through its dynamics over a diagram of related states. In that, we can identify new data-driven seasons, relate the ILI season to the several coordinates of the weather data, and look through the weight of the different ILI definitions. Nevertheless, the meaningfulness of this kind of general approach demands a large effort on the interpretation of the results in the public health context that they belong to. Although the obtained results are good indicators to the promising potential of the usage of this technology, clear interpretations of the relations between states must be tackled within public health experts to enhance the usability of the technological tool as a public health tool.

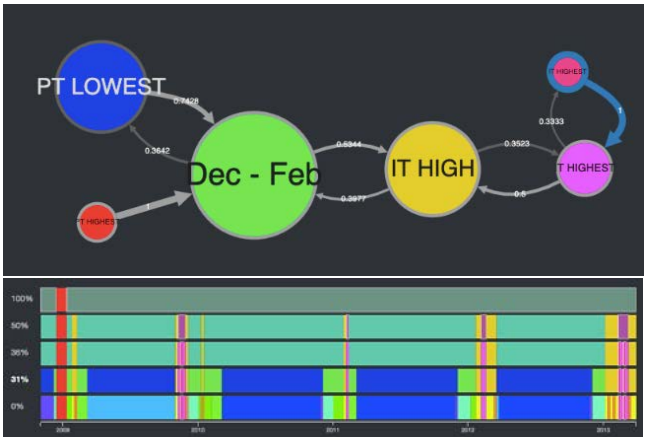

Figure 7: Comparison between the incidence of ILI in Portugal and Italy, and the corresponding state history

ACKNOWLEDGMENTS

We thank the support of the European Commission on the H2020 MIDAS project (GA nr. 727721).

REFERENCES

[1] C. Guerrisi et al (2016). Participatory Syndromic Surveillance of Influenza in Europe, J Infect Dis. 214 (supp 4): S386-S392
[2] M. Hirsch, O. Woolley-Meza, D. Paolotti, A. Flahault, and P. Lukowicz (2018). grippeNET App: Enhancing Participatory Influenza Monitoring Through Mobile Phone Sensors. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (pp. 833-841). ACM.
[3] MIDAS H2020 Project, "MIDAS Project Website," MIDAS Project, 2017. [Online]. Available: http://www.midasproject.eu/. [Accessed 8 8 2019]
[4] S. P. van Noort et al (2015). Ten-year performance of influenzanet: Ili time series, risks, vaccine effects, and care-seeking behaviour. Epidemics, 13:28–36.
[5] U.S. National Library of Medicine, "MEDLINE®: Description of the Database" [Online]. www.nlm.nih.gov/bsd/medline.html/. [Accessed 2019]
[6] D. Paolotti et al (2019). "Influenzanet 2.0. [Online] Available: influenzanet.github.io/docs/" [Accessed 20 8 2019]
[7] D. Paolotti et al. Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. Clinical Microbiology and Infection, 20(1):17–21, 2014.
[5] J. Pita Costa et al. (2018) "Influenzanet MIDAS toolset and demos" [Online]. midas.quintelligence.com/midas-influenzanet-demos [Accessed 20 8 2019]
[9] L. Stopar, "Streamstory," Institute Jozef Stefan, 2019. [Online]. Available: http://streamstory.ijs.si/. [Accessed 12 8 2019]
[10] L. Stopar, P. Škraba, M. Grobelnik, and D. Mladenić (2018). StreamStory: Exploring Multivariate Time Series on Multiple Scales. *IEEE transactions on visualization and computer graphics* 25.4: 1788-1802.
[11] Quintelligence, "GitHub repository" [Online]. Available: http:// github.com/quintelligence /. [Accessed 20 8 2019]