**Digital Epidemiology and Surveillance, CCS2016 Satellite Workshop**

# Flu-Now : Nowcasting Flu based on Product Sales Time Series

**Ioanna Miliou[1], Salvatore Rinzivillo[2], Giulio Rossetti[2], Dino Pedreschi[1], Fosca Giannotti[2]**

[1]KDDLab, University of Pisa, {miliou, pedre}@di.unipi.it

[2]KDDLab, ISTI-CNR, {name.surname}@isti.cnr.it

Big Data offer nowadays the capability of creating a digital nervous system of our society, enabling the measurement, monitoring and prediction of various phenomena in quasi real time. But with that, comes the need of more timely forecast, in other words *nowcast* of changes and events in nearly real-time as well. The goal of *nowcasting* is to estimate up-to-date values for a time series whose actual observations are available only with a delay. Choi and Varian[1] introduced the term nowcasting to advocate the tendency of web searches to correlate with various indicators, which may reveal helpful for short term prediction. In the field of epidemiology, it was showed in various works that search data from Google Flu Trends, could help predict the incidence of influenza-like illnesses (ILI). But as Lazer and al. notice[2], in February 2013, Google Flu Trends predicted more than double the proportion of doctor visits for ILI than the Center for Disease Control.

In this work we are studying the flu time series, of cases from 2004/05 to 2014/2015 flu season, from physicians and pediatricians from all over Italy. We are interested to examine whether is possible to use retail market data as a proxy for flu prediction. Our dataset consists of economic transactions collected by COOP, a system of Italian consumers' cooperatives which operates the largest supermarket chain in Italy. The whole dataset contains retail market data in a time window that goes from January 1st, 2007 to April, 27th 2014. First, we identified the products that have adoption trend similar to the flu trend with the help of an 1-nearest neighbor classifier that uses dynamic time warping as the distance measure between time series. Based on these products, we identified the customers that buy them during the flu-peak, since those individuals would have higher possibility to be either infected or close to an infected individual. We extracted their most frequent baskets during the peak using the Apriori algorithm, an algorithm for frequent item set mining and association rule learning over transactional databases, and we use those baskets-sentinels as control set for the following year flu peak. Monitoring the behavior of these baskets-sentinels we are able to detect patterns similar to the ones of the previous year's flu peak, and as a result obtain an alarm for the appearance of the flu.

Many lines of research remain open for future work, such as studying whether the retail market data can manage to predict the flu peak even in particular cases such as the year 2009 non-seasonal H1N1 influenza (flu) pandemic that peaked in October and then declined quickly to below baseline levels by January.

## References

[1] H. Choi and H. Varian, "Predicting the present with google trends," *Technical Report*, 2009.

[2] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: Traps in big data analysis," *Science Magazine*, vol. 343, no. 6176, pp. 1203–1205, 2014.