digital epidemiology: combining big data and traditional methods

data researchers are using a combination of Google Flu Trends search data and Centers for Disease Control surveillance reports from hospital labs to estimate flu trends

impact errors in predicting flu patterns between 2011-13 was considerably reduced by using Google searches and surveillance reports together

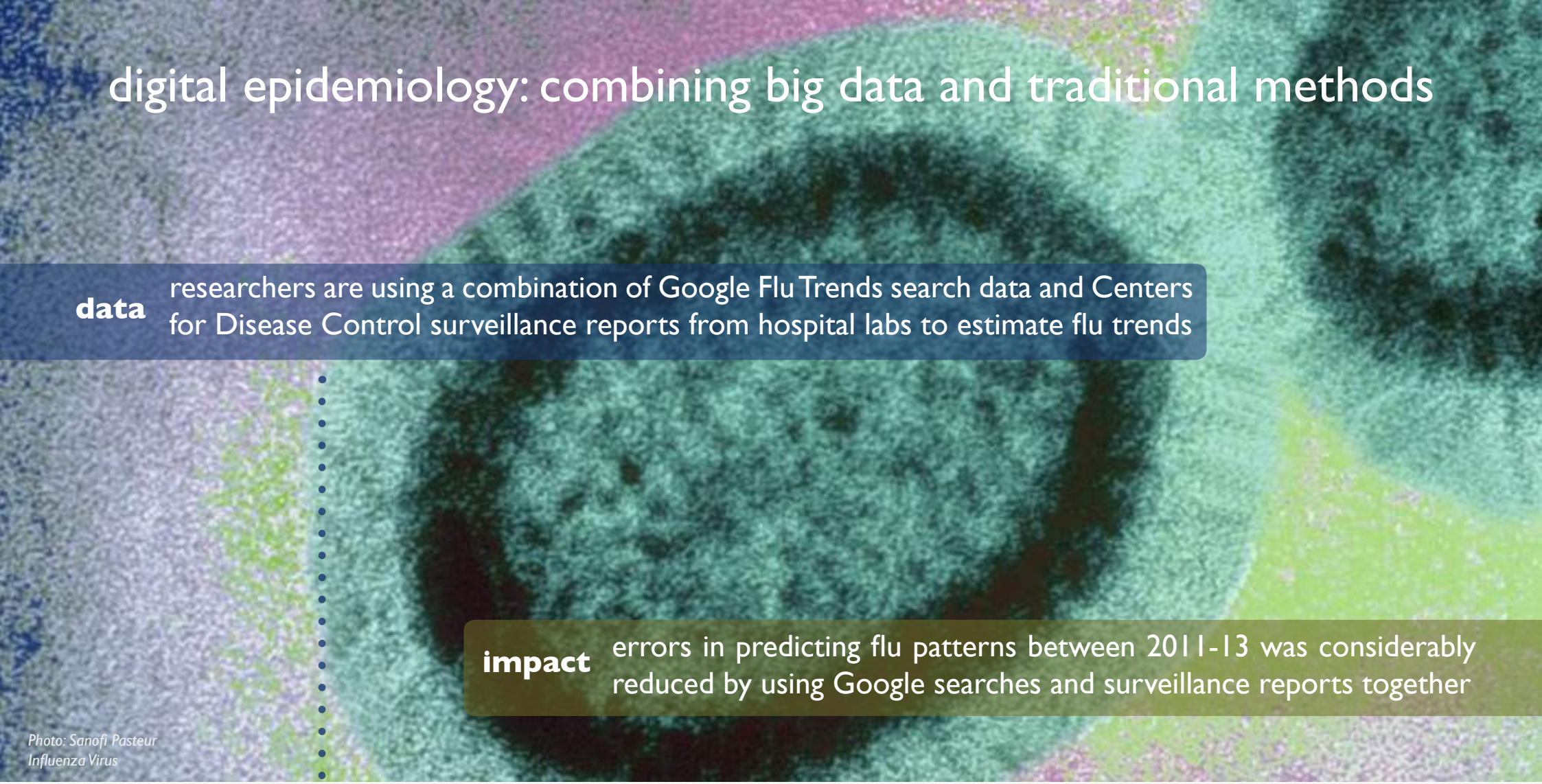
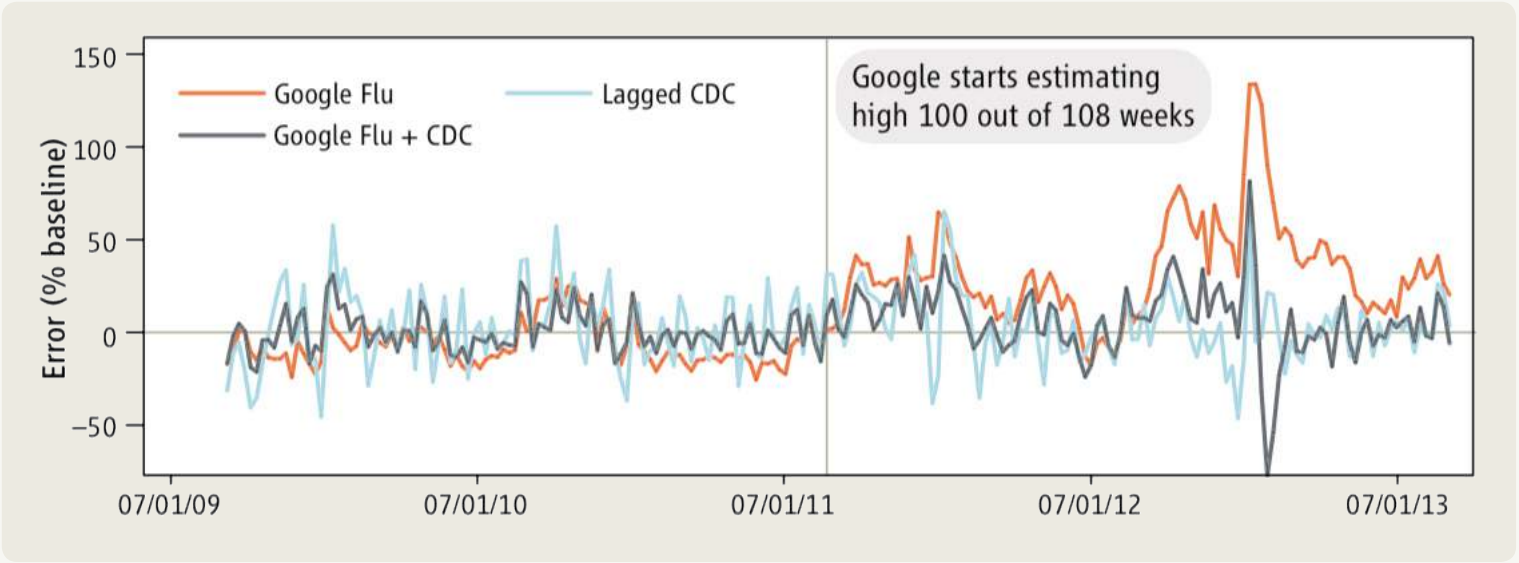


Photo: Sanofi Pasteur
Influenza Virus

After failing to detect the H1N1 (swine flu) epidemic in 2009, over-reporting the 2012 flu season by 50%, and overestimating flu prevalence in 100 out of 108 weeks from 2011 – 2013, Google Flu Trends (GFT) announced plans in 2014 to partner with the Centers for Disease Control (CDC) to improve its predictions.[1,2] One of the criticisms surrounding GFT is lack of transparency. By keeping the search algorithm and the relative weights attached to search terms hidden, it is impossible for outside researchers, tech experts, and health officials to help improve the predictive model.[3]



This graph shows the lack of accuracy of Google Flu Trends versus three week-lagged CDC estimates. From 2009 to 2013, the most accurate predictor was actually a combination of GFT with the lagged CDC estimates, suggesting the need for more collaboration between government agencies and private companies. Source: Lazer et al. 2014

the challenges of search-based prediction

GFT miscalculated flu levels for several reasons. One is its autocomplete feature, which gives a list of search predictions based on the search behavior of nearby users, causing some people to click on suggestions for which they might not otherwise have searched. A second factor is widespread media coverage of an issue, which can cause a spike in searches on a topic, even if not directly relevant to the searcher.[4] Google has tried to improve their algorithm to take these factors into consideration. A 2014 academic paper found that GFT is still overestimating flu prevalence about 74% of the time, although this is an improvement from 94% overestimation two years prior.[5]

data, big and small

Traditional surveillance reporting continues to predict more accurately than GFT, but the most powerful tool is a combination of both. Mean absolute error is a common way of comparing forecasts with actually observed outcomes; a zero score indicates perfect precision. In 2011-13, the mean absolute error for GFT was 0.49, 0.31 for CDC reports, but only 0.23 for a combination of GFT and lagged CDC combined.[6] Google and the CDC announced plans in 2014 to partner together to provide the most accurate information possible. This is a step in the right direction, but it is critical that Google search results also be presented separately from the CDC data so that the accuracy of signals from each independent source can be evaluated.[7]