Seminar 4

The exam in January will contain essay questions based on the 2 papers on ELP. Based on the papers answer the following questions:

* How was data analysis used as part of trend forecasting

Data was initially captured through search queries relating to flu. They used a learning algorithm using previous data to help train the algorithm to build a more accurate predictor.

Data analysis of web searches pruned by google to specific influenza searches, limited by region and personal data pruned. This was then compiled into a national predictor, which allowed to predict trends in outbreaks of influenza in such a way as to be more readily available than official trend analysis.

(“Detecting influenza epidemics using search engine query data” by Jeremy Ginsberg)

The purpose of Google Flu Trends (GFT) is to use search keyword trends from Google.com to produce a daily estimate of the occurrence of flu two weeks in advance of publication of official surveillance data. The search results tied within influenza correlated highly with doctor visits. (GOOGLE DISEASE TRENDS: AN UPDATE by Patrick Copeland)

* What were the main factors that led to the failure to predict the flu pandemic in US

Predictive types of suggestions pushed results into more important search queries that may have made the algorithm more unreliable.

Media hype was a big factor in the failure to predict the flu pandemic, as this threw the results of the analytics that Google gathered from the web searches. This caused an uncharacteristically high amount of flu searches in a short time period, altering the predictive algorithm Google Flu Trends uses.

The algorithm also failed to consider long-term spikes caused by media hype and as such the initial system in place was not capable of dealing with this.

(“GOOGLE DISEASE TRENDS: AN UPDATE” by Patrick Copeland)

* Could the situation have been avoided and if so how?

The situation could have been avoided by cataloguing uncharacteristic spikes in the algorithm and removing them from the trend results, which hopefully would remove outlier data. Filtering out media hype within this algorithm would also help result in more consistent results.