1. **Health Wrap Up**

   **In-Sample and Out-of-Sample Fit of Prediction Model**

   ![Graph showing in-sample and out-of-sample fit of prediction model.](image)

   *Note: CDC official statistics in red; Google trends forecast in black*

   - Ginsberg et al. built a simple machine learning prediction model using Google search query data to forecast flu trends.
     - They correlated each of fifty million different search terms with the data from the CDC on actually flu data incidents using data from clinics.
     - Forty-five search queries made the optimal prediction model, based on an out of sample validation technique.
     - They showed how their forecast compared with the actual data. The Google data trends forecast is in black, the CDC data is in red, and then the true out of sample validation here, the test of their model, is shown to the right of the vertical dashed line.
     - Out of sample: in their original analysis, they only used data up to 2007. They showed that their forecast worked well up to 2007, then as a further test, they compared it to data they didn’t use at all in building the model. You can see their model still does very well.
   - The advantage of using the Google search data: you can get data more quickly and at finer geographies.
   - The graph below shows at each point in time what the Google forecast was two weeks in the future.
     - Top panel: when they had data available up to February 4, 2008, they forecasted an uptake in the coming two weeks before CDC actually put out that data.
     - Second panel: they were exactly right about that. Repeat for each panel.
     - As a leading indicator, the Google’s flu trends forecasts are actually working quite well.
This Google flu trend's predictive model broke down in late 2012 and became very inaccurate in forecasting outbreaks of the flu. Laser et al. document the model's failure by essentially extending the window used for out of sample validation a few years further.

2009 to 2011: the Google flu data and the CDC data match each other very closely.
• 2012: the Google estimates are more than double what CDC actually reports.
• The model broke down because of a change in the underlying technology of the platform.
  o Google search engine started to prompt users to search for additional diagnoses after entering a term like fever or cough. They put in the auto-fill feature, which changed what people were searching for because if you type cough, it now suggests other things that you might add to that.
  o If you want to be accurate you have to re-estimate the model after these changes to the platform.
• Ginsberg et al became the basis for Google Correlate, a publicly available tool to find searches that correlate with real world data.
  o Our research team is working with Google to try to construct leading indicators of economic and social variables: segregation, upward mobility, creating forecasts that cities and governments might be able to use to take action more quickly.
• Big data and machine learning still require careful and accurate measurements.
  o Very good that the CDC didn't abandon its program to collect data on flu incidents from clinics.
  o We're not at a point yet where machine learning is really a true substitute for traditional data collection method.
• Machine learning is not a black box to feed data into that will tell me an answer.
  o We need some technical skill to figure out how to fit these models but also careful judgment about whether the model really makes sense.
  o Models where the mechanisms that are underlying seem more sensible are more likely to yield stable predictions.
  o If you're getting predictions off of terms like Oscar nominations or high school basketball then you should start to worry that maybe some things are going to change, such that the basketball season no longer correlates with flu or something else changes about the nature of the searches.
  o That's where if the underlying predictions actually make sense in terms of being related to cough or fever, you understand the way the platform's set up, and the platform hasn't changed, you would have more confidence using these predictions going forward.
  o There's still an important role for human judgment, in addition to using these machine learning tools.