# Homework (**Hand-in is 16th of May, 18:00)**

This homework should be done in groups of max 2 and it should use the notebooks provided in class (modified obviously). The goal here is to integrate text analysis, forecasting (and decision making). We want you to forecast a panel dataset using text features.

Grades will be given for creative ideas of what to forecast with what and a good explanation for why it matters. Bonus points for embedding in a decision framework. High forecast performance will be less important but the definition of onset or incidence and rolling forecast needs to be implemented correctly. We will pay close attention to correct use and arguments for things like the small data problem, onset, forwards, since variables etc. Please always explain your choices and interpret results.

1. Think of a (onset/incidence) variable to forecast. Optimally, this variable provides you with a large panel (long and many units), i.e. if you are using yearly data try to get many countries/units. Design this well with part 2 in mind. (10 pts)
2. Merge other time series to this outcome variable. One needs to be a text-based measure. Ideas: Google trends data, GDELT data, our topics from conflictforecast.org download section. You do not need to come up with your own text mining pipeline. (10 pts)
3. Define an onset or incidence variable based on your data in 1. Carefully explain why forecasting this is useful. Who could be the decision-makers that can make use of this forecast and how? Explain the prediction policy problem they face. (10 pts)
4. Do some feature engineering on the right-hand-side data if necessary. Explain the reasons for the feature engineering carefully. (10 pts)
5. Adapt the rolling forecast code that you should by now all be familiar with. Make sure you implement this correctly. Don't write your own code but adapt Ben's. (10 pts)
6. Plot ROC curve and precision recall curve of the merged fitted values. Discuss these curves in detail. (10 pts)

BONUS: Adapt the cost model to your purpose. Use it to find the optimal cut-off under assumptions on the costs on TP, FP, TN and FN. In order to argue for the reasonability of your assumptions it will help to think of what FP and FN in particular mean for your decision-maker from question 3). What decisions could be triggered by the forecast? (30 pts)