Core 109S IDWT: Leading Discussion

For each of the remaining classes, 4 students will be asked to provide discussion material and help lead discussion. Please sign up for a day by following the link on the course website `https://github.com/colgate-idwt/idwt-spring2017`.

Before Class

You must compose a list of **eight** discussion ideas. Each idea can be any of the type listed below and you are encouraged to have a few of each type. You do not have to coordinate with the people who are going the same day. The list is **due by 11:55pm the night before by email to me**. Please put your initials in front of each of your ideas. You do not need to categorize them as I have done below.

When there are supplemental readings, you are expected to browse them and bring up at least one point that you gleaned from the supplemental reading.

Here are categories of discussion ideas along with concrete examples for the Google Flu Trend (GFT) readings.

- Passages that you found interesting, controversial or inspirational
  - (MH) From *Science* article: "The most common explanation for GFT's error is a media-stoked panic last flu season (1, 15). Although this may have been a factor, it cannot explain why GFT has been missing high by wide margins for more than 2 years."
  - (MH) From *Science* article: "GFT bakes in an assumption that relative search volume for certain terms is statically related to external events, but search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider."
  - (MH) From *Science* article: "Instead of focusing on a 'big data revolution,' perhaps it is time we were focused on an 'all data revolution,' where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world."
- Questions that you find interesting or want to pose to the class
  - (MH) The second article is pretty critical of GFT and identifies some general problems with the GFT approach. What does their critique say about the "big data" approach more generally?
  - (MH) The second article claims that GFT was *overfitting* to the data. Do you find their critique accurate?
- Connections to things we have covered in class
  - (MH) There were a number of connections between the Nature article describing GFT and our readings from DSB on overfitting. In particular, the GFT designers used a validation set to determine which search terms to include. What did you think of their approach? Does it seem like an effective way of filtering our non-flu search terms? Or might some spurious terms still end up in their model?
  - (MH) In the Big Data book, we read a chapter about correlation vs. causation in the context of Big Data. That article essentially claimed that in many cases, correlation is good enough. In light of this discussion of GFT, what do you think of that argument now?
- Beyond the reading

  - (MH) Despite the criticism, GFT appears to provide some potential useful information about flu. This got me thinking... all this data is being collected about us all the time and while there is plenty of talk about how scary that can be in terms of privacy violations, etc., I wonder if there are ways we can use that data to benefit society (such as using searches to improve public health).

During Class

Whether leading the discussion or not, *everyone* is expected to participate. The leaders play a role in keeping the conversation going. For instance, if the conversation on one topic starts to die down, one of the discussion leaders can use their list of discussion ideas to initiate discussion on a new topic.