**Big Data Challenge**

Data is increasingly easily accessible through the internet. A lot of data. The term ‘big’ is often used for data sets where normal analysis methods do not suffice anymore. This may be the case much faster than you think. For example, visualising correlation analyses on more than 20 variables becomes a bit of a challenge. And storing the output and checking it in your console quickly a mess. Let alone of you want to do this for hundreds or thousands of analyses.

The ‘Big Data Challenge’ is an assignment in which a big question will be tackled using online available data sets. The challenge has two levels:

1. ‘class’ level: as a class I want you to think of a general, overarching questions you want to tackle. The class will hand in one paper to do this
2. ‘group’ level: in sub groups of three or four you will work on a sub-question of the big overarching question. This part requires data analysis.

The learning goals of this challenge are three-fold:  
1) apply coding skills to a real life problem and challenge yourself solving new data problems.

2) work effectively as a group, both at the class level and within an among the different smaller groups. Help each other with coding, etc to answer questions.

3) define a big question, and start working on it using several smaller projects with sub-questions. Think about an analysis workflow.

The topic of your question is open and has to be decided and supported by the whole group. Examples are ‘effects of warming temperatures on humans’ (with potential sub-questions looking at correlations between temperature changes across the world, correlations between temperature and diseases, etc) and ‘how correlated are financial markets?’ (with potential sub-questions looking at correlations between different goods, spatial correlations, etc).

*Papers to hand in*

As a class you prepare a paper outlining the big question and sub-questions and a very general outline of the methods of each sub-project. Look-up references to back-up your question and rational why it is interesting to look at. This paper should be a nice an interesting read aimed at a general, educated audience. A minimum of one and a half page and a maximum of three pages (this includes references). Each student is on this paper and this determines 1/3 of your final assignment mark.

Each group hands in a paper for their own sub-questions. Each section has a suggested length in pages, you may deviate from this by 50% (so a 1 page, could be ½ page or 1 ½ page) but the final paper should not be longer than five pages. This paper should have a title, introduction (½ page), methods (2/3 page), results and discussion (3 pages) and references (<10 included in page count). The R file with all the code should be handed in separately. This is a very important part of the paper. Write the code clearly and it should be reproducible (i.e. I should be able to load it into RStudio follow some comments provided (e.g. which library to load) and run the code); test this!

**Some internet database resources**

Climate data:

Environment Canada data

<https://gist.github.com/gavinsimpson/8c13e3c5f905fd67cf85>

<https://www.r-bloggers.com/harvesting-canadian-climate-data/>

general

<https://ropensci.org/blog/blog/2013/08/18/sciordata>

Market data

[www.quandl.com](http://www.quandl.com)

Google search data:

<http://pmassicotte.github.io/2015-11-17-gtrendsR/> and <https://mran.revolutionanalytics.com/package/gtrendsR/#rtable>

overview

<https://mran.revolutionanalytics.com/documents/data/>

eBird data

<https://cran.r-project.org/web/packages/rebird/rebird.pdf>

<https://ropensci.org/tutorials/spocc_tutorial.html>