The **final project for EPS236** **(fall 2019)** is designed to be a capstone of the course. It is intended to be an original data analysis using real data, treating the problem in depth.

The projects will be due during reading period, and the presentations will occur during two dinner-time (vegan) sessions, for which attendance is mandatory. *The presentations will occur before 09 Dec 2019.*

1. The topic may be quite different from the student's thesis research (preferred), but can also be related. However, *it cannot be part of the student's thesis research currently or in the past.*

The topic and scope will be agreed between each student/team of 2 and the instructors. Several in-class work sessions will be scheduled.

A science (or data-science) question should be clearly articulated, and then addressed.

**2. Requirements**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* **a.** The bootstrap or MCMC should be applied in every case as \*

\* one method to assess confidence intervals. \*

\* **b.** Excellence in data visualization. \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

 **c.** In addition to the above, 3 methods from the list below should be applied:

 *i.* data cleaning and/or outlier removal, possibly including filling or interpolation

 *ii.* linear modeling

 *iii.* forecasting

 *iv.* trend assessment

 *v.* seasonal analysis

 *vi.* image processing and/or denoising

 *vii.* model optimization using MLE or similar, including regressions or more complex methods

 *viii.* cross validation

 *ix.* non-parametric models of the data

 *x.* other methods to model or statistically assess the data.

**3. Presentations and products.**

 **a.** The primary written product is a Powerpoint or pdf presentation-format product with sufficient text content to be self-explanatory.

 **b.** A full bibliography and citations to data sources.

 **c.** The actual data sets that were used in the work (even if the data set is downloadable, the data set itself is required). Formats may be NetCDF, flat ascii, or R-objects as .RData or RDS.

 **d.** The code used in the analysis. Instructors should be able to reproduce your results by running the code.

 **e.** A 15-minute presentation to the class, to be followed by a 5 minute discussion with the class and instructors.