BACKGROUND: GAUSSIAN PROCESS REGRESSION

According to the World Health Organization, air pollution is a major environmental health issue. Both short- and long-term exposure to polluted air increases the risk of heart and respiratory diseases. Hence, reducing the concentration of particulate matter (PM) in the air is an important task.

You are commissioned to help a city predict and audit the concentration of fine particulate matter ($PM_{2.5}$) per cubic meter of air. In an initial phase, the city has collected preliminary measurements using mobile measurement stations. The goal is now to develop a pollution model that can predict the air pollution concentration in locations without measurements. This model will then be used to determine particularly polluted areas where permanent measurement stations should be deployed

A pervasive class of models for weather and meteorology data are Gaussian Processes (GPs). In the following task, you will use Gaussian Process regression in order to model air pollution and try to predict the concentration of $PM_{2.5}$ at previously unmeasured locations.

CHALLENGES

We envisage that you need to overcome three challenges in order to solve this task - each requiring a specific strategy.

- **1. Model selection**: You will need to find the right kernel and its hyperparameters that model the data faithfully. With Bayesian models, a commonly used principle in choosing the right kernel or hyperparameters is to use the *data likelihood*, also known as the marginal likelihood. See more details here: Wikipedia (https://en.wikipedia.org/wiki/Marginal_likelihood).
- **2. Large scale learning**: GP inference grows computationally expensive for large datasets. In particular, posterior inference requires $\mathcal{O}(n^3)$ basic operations which becomes computationally infeasible for large datasets. Thus, you will have to mitigate computational issues. Practitioners often do so using a number of methods, among which you can opt for one of the following:
    - Undersampling: Sampling a subset from our initial dataset which is used for learning, either randomly or with clustering-based approaches.
    - Kernel low-rank approximations: Effectively avoid inverting the full kernel matrix. The most popular instances are the Nyström method and random Fourier features. The following excellent review on Wikipedia can serve as an introduction: Wikipedia (http://en.wikipedia.org/wiki/Low-rank_matrix_approximations). (Hint: for this kernel approximation method, you won't need sklearn.gaussian_process GP model)
    - Approximation of GP with multiple local GPs: For certain kernels, mutual dependence of distant samples diminish, so training a local GP for every region of our domain using the corresponding subset of local data samples from the original dataset can approximate the use of one global GP within that locallity.

  Of course, implementation of other methods or your unique solutions is highly encouraged.
- **3. Asymmetric cost**: We use a specialized cost function that weights different kinds of errors differently. As a result, the mean prediction might not be optimal. Here, the *mean prediction* refers to the optimal decision with respect to a general squared loss and some posterior distribution over the true value to be predicted. Thus, you might need to develop a custom decision rule for your model.