**DATAfucking MINING PROJECT (aka Sphincter Destroyer):**

**Goal :**

To forecast the air quality in future (hourly) in Beijing and London :

* the concentration of PM2.5, PM10 and O3 for 35 stations in Beijing.
* predict the concentration of PM2.5 and PM10 for 13 stations in London.

**Technics to try (It is mandatory to use three or more diﬀerent algorithms in addition to a baseline):**

<https://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support#SVM_pour_la_r%C3%A9gression>

<https://en.wikipedia.org/wiki/Recurrent_neural_network>

<https://en.wikipedia.org/wiki/Gradient_boosting#Gradient_tree_boosting>

(Techniques mentionnées intéressantes par le prof pour ce projet)

🡪discuss the diﬀerent performances you have with your methods and explain why these work or not.

**Model comparison and selection (04) :**

Using cross-validation to select the best model based on statistical signiﬁcant test

**Error Estimation :**

* Cross-validation (+ simple et + stylé à mon avis qui compte pour que dalle)
* Bootstrap (méga casse-couilles, pas fini de lire le wiki)

**Some TODO’s :**

* Format the Beijing xlsx station data coordinates into csv (same format as London, easier to fetch – file pretty short)

**Questions :**

Use of London\_historical\_aqi… ?

Split the data set properly in training and validation sets taking into account that the **order matters**. -> what does it mean ?

Particles with a diameter of 2.5 m or less (called PM2.5) can -> mm right ? (sauf si on respire des trucs de bâtard, genre une voiture)

**Questions cons :**

**Apprendre le data mining pour les nuls :**

<https://fr.wikipedia.org/wiki/Validation_crois%C3%A9e> (rappel, je l’ai vu en IA au semestre 1 d’abord)

<https://fr.wikipedia.org/wiki/Bootstrap_(statistiques)>