Machine Learning Applied to Physics

Marcio Costa

Universidade Federal Fluminense

Project 2

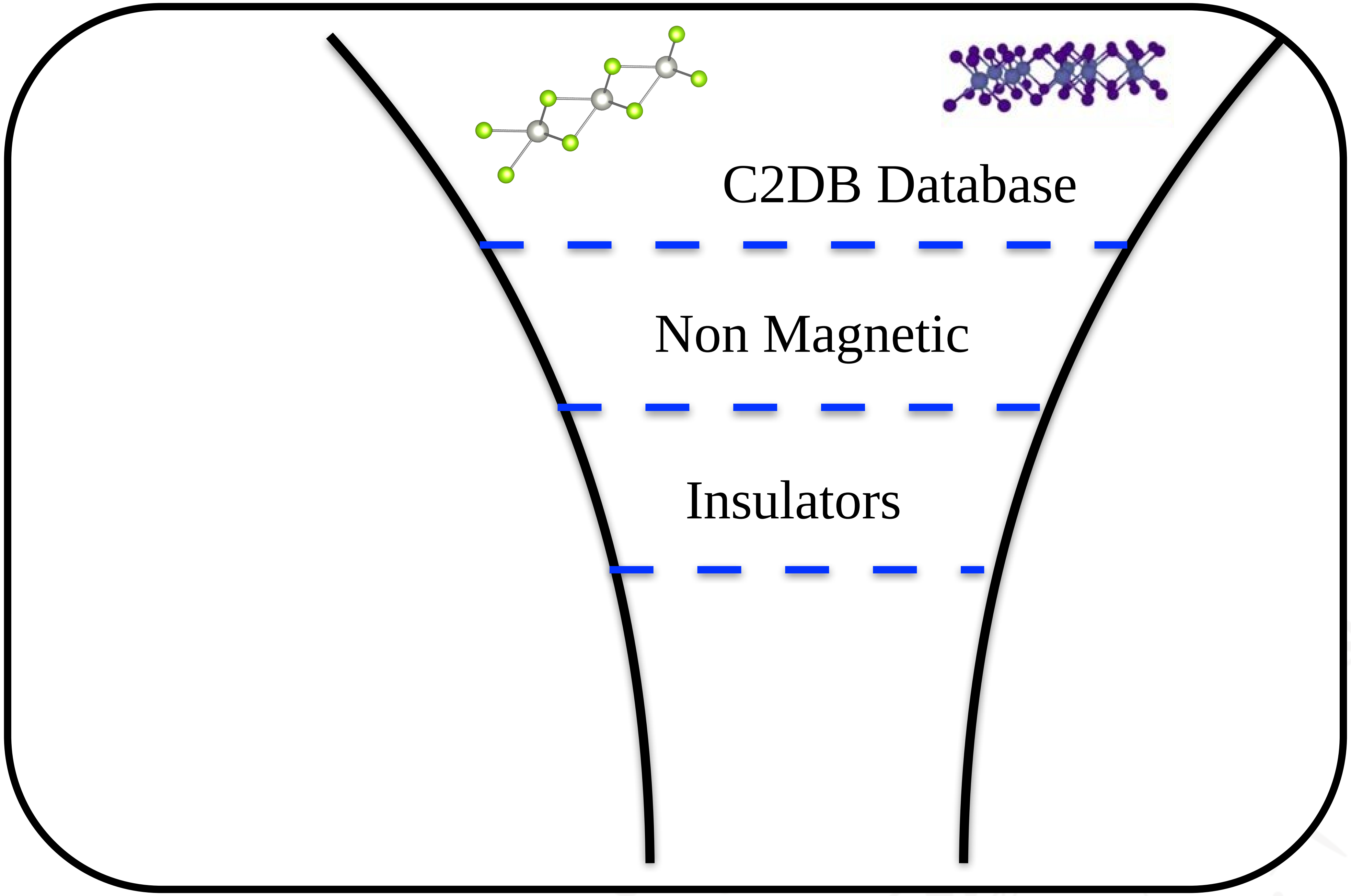
Build a **regression model** for the **band gap** of two dimensional non-magnetic materials. Including the prediction of novel two-dimensional insulators.

Step 1 - Construct the database, including all stoichiometry

- Remove all magnetic materials
- Remove all metals, i.e. zero band gap materials
- Create a data frame with : Formula, Band Gap (target), Space Group, Stoichiometry

Project 2

Step 1 - Construct the database, including all stoichiometry



C2DB Database

Non Magnetic

Insulators

Project 2

Step 2 - Construct the features space.

Now we are mixing AB2, ABC and others stoichiometry. In this case, the previous strategy of simply combining the atomic properties of atom A and atom B will not work. Since, for the AB2, AB, AB3 stoichiometry there will be 36 features. For the ABC stoichiometry there will be 54 features. Resulting in a feature space with non equal dimension. To circumvent this problem we will create new features based on averages of the original ones.

Step 2 - Construct the features space.

To circumvent this problem we will create new features based on averages of the original ones.

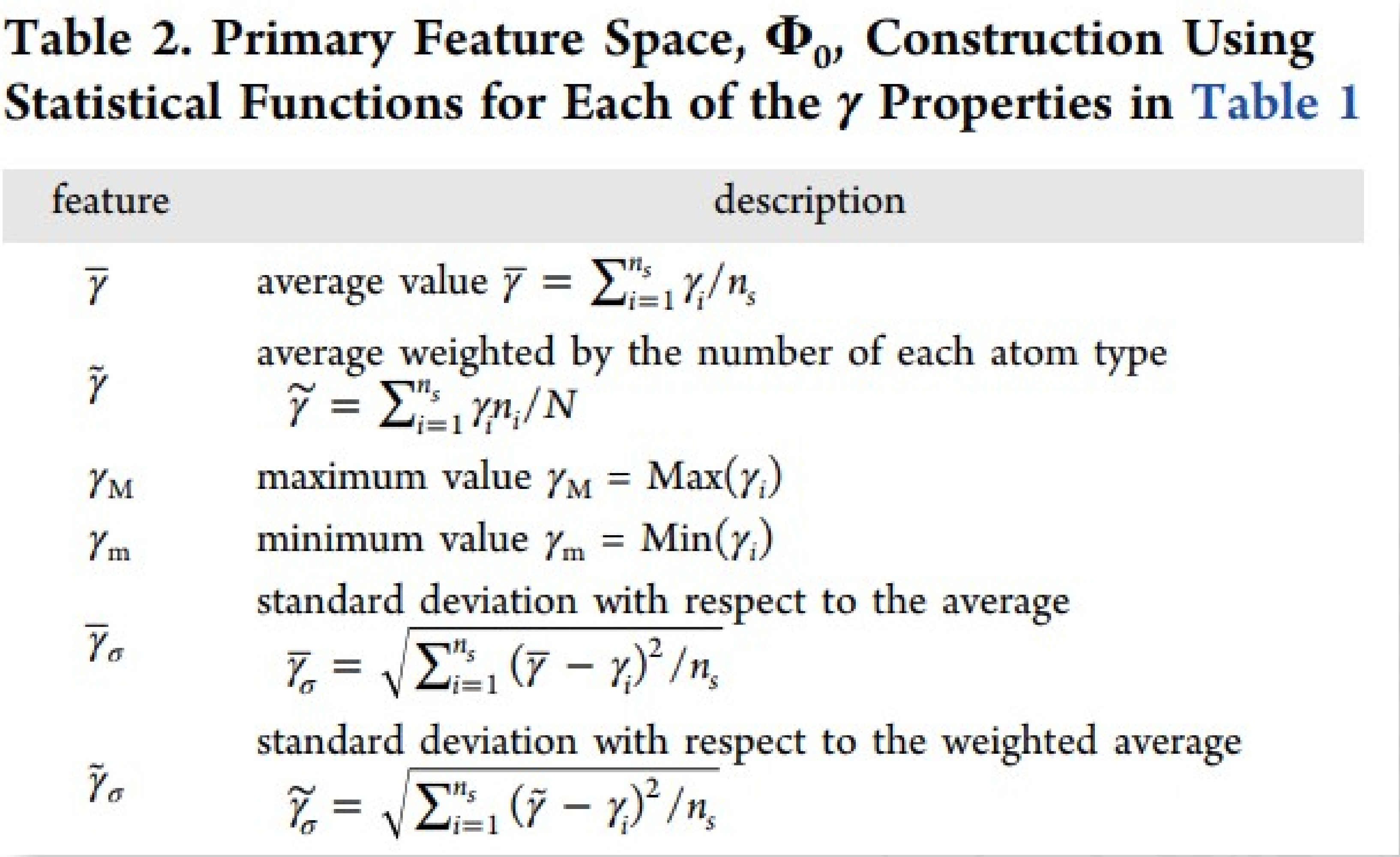


Table 2. Primary Feature Space, Φ_0, Construction Using Statistical Functions for Each of the γ Properties in Table 1

feature	description
$\bar{\gamma}$	average value $\bar{\gamma} = \sum_{i=1}^{n_s} \gamma_i / n_s$
$\tilde{\gamma}$	average weighted by the number of each atom type $\tilde{\gamma} = \sum_{i=1}^{n_s} \gamma_i n_i / N$
γ_{M}	maximum value $\gamma_{\mathrm{M}} = \mathrm{Max}(\gamma_i)$
γ_{m}	minimum value $\gamma_{\mathrm{m}} = \mathrm{Min}(\gamma_i)$
$\bar{\gamma}_\sigma$	standard deviation with respect to the average $\bar{\gamma}_\sigma = \sqrt{\sum_{i=1}^{n_s} (\bar{\gamma} - \gamma_i)^2 / n_s}$
$\tilde{\gamma}_\sigma$	standard deviation with respect to the weighted average $\tilde{\gamma}_\sigma = \sqrt{\sum_{i=1}^{n_s} (\tilde{\gamma} - \gamma_i)^2 / n_s}$

Step 3 – Train a model

In this step we will train different ML regression algorithms. So far we have been introduced to:

- Ridge Regression
- Lasso Regression
- Decision Trees
- Random Forrest
- Gradient Boosting

Nevertheless, you are free to use any other algorithm.

Step 3 – Train a model

You should test all methods we have used. Train/Test split, Cross Validation, Parameter Tuning, Bootstraping, Bagging, Boosting, Feature Engineering.

Step 4 – Deploy your model.

After, all the training/testing. You must select the best model and use it to predict novel materials.

- Use the output of the classification model selecting only the insulators.
- Use the best regression model to predict its band gap.
- Run a DFT calculation to confirm you prediction **(this part will be my responsibility)**.