Phase 2: Innovation

- Prediction of earthquake methods generally consists of two methods.
    - Precursor method
    - Trend or Pattern

- Precursor method - identifying distinctive *precursors* to earthquakes.
    - **Dilatancy–diffusion**
    - **Changes in V_p/V_s**
    - **Radon emissions**
    - **Electromagnetic anomalies**

- Trend or Pattern - identifying some kind of geophysical *trend* or pattern in seismicity that might precede a large earthquake.
    - Date
    - Time
    - Latitude
    - Longitude
- Since we are using a kaggle dataset which consists of data that corresponds with the second method(location ,time) we are going to use that method.
- Using the location, time and date we are planning to create a machine learning model that will predict earthquakes.
- We are planning to use Ensemble model with multiple ML models that include KNN,SVM .

- **Steps involved in KNN:**
    - Choosing an appropriate value for k, which represents the number of nearest neighbors to consider. We also use cross-validation techniques to find the optimal k value.
    - Implementing the KNN algorithm using a machine learning library like scikit-learn in Python.
    - Training the KNN model using the training dataset, with the earthquake locations (latitude and longitude) as input features and potentially other relevant features.

- Evaluating the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC AUC on the testing dataset.

- **Steps involved in finding validity of data/ confidence of data**:
  - Certain features like Magnitude Type, Magnitude Error, Azimuthal Gap, Horizontal Distance, Horizontal Error, Root Mean Square, Source will help to detect the validity levels of the magnitude and depth information used in KNN.
  - Hence this can be found by detecting the outliers using **Support Vector Machine** or by providing confidence values that help the ensemble model decide.

- **Data Preprocessing:**
  - To guarantee the quality and appropriateness of the data for training the model, data preparation is an essential step. The tasks are broken out as follows:
    - Managing Missing Values:The dataset contains missing values like in magnitude error. In order to solve this, we will fill in the missing data points using the proper imputation techniques, such as mean or median imputation.
    - Normalization and Standardization: We will normalize the data to make sure that all features have comparable scales. This procedure stops the model from being dominated by one feature while it is being trained.
    - Data division: The dataset will be split into two sections: around 70% for training and 30% for testing. This allows us to train the model on one subset and evaluate its performance on another, ensuring that the model can generalize well to unseen data.

- **The ensemble model:**
  - We are planning to use models like AdaBoosting that can combine multiple weak learners and provide a result that predicts the magnitude of the earthquake along with the confidence.