# Justify the methodology at each step.

## Step 1: Load and clean data

It was considere numerical columns for analysis and discarded expert demographics data. For missing values in numerical data, it was chosen median imputation, so that it may not disturb the data skewess. It was filled party names which had as party\_id equal to ‘2406’ as ‘NA’. Finally it was added the initial letters of each country based on the first 2 digits from the party\_id column. Finally it was dropped every row where the column had a complete null segment.

## Step 2: Dimensionality Reduction

It was taken into considerations two approaches, Principal Component Analysis and Rotation Factor Analysis. Both models perform a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. The 54 dimension data frame was downsized to a 2 dimensional and it was ploted for each approach. It was seen a cluster in terms of party lrgen.

## Step 3: Estimate the distribution of the 2D points from Step 2.

For this step it was implemented the Kernel Density Estimation, where it estimates the density distribution.

## Step 4: Find feature values from the high-dimensional space that map to the 2D points.

It was used random points to generate the higher dimensions. It was used standard scaler and inverse transform to try re-construct the higher dimension.

## Step 5: High dimensional bounds.

For this question I would create contour plot, which would show the limits and density for the reduced dataset, and plot the new sample points to check if the they are inside the limits. In step 4 we proved that each dot is inside the grey area, which represents the original data in 2D.

# What could you teach a politician about the European political landscape based on the dimensionality reduction and/or possible additional analyses you may produce?

In the bigger picture, a politician can see which political parties are similar to each other. In case there is a topic to be discussed, with dimensionality reduction you can check which parties maybe be a suitable target to the topic. The analysis made for the plot on step 2, you can see which parties are closer to each other taking into account the ideological stance.

# How would you deploy the model to a cloud environment so that it would be able to withstand 1 millon users per hour?

To answer this question, it was design a cloud solution in AWS. This model would be deployed taking into account the following steps:

1. Users input would have access to the architeture through an API.
2. The scalabilty is managed through the Load balancer which activates the zones, for better performance. At each zone there is a Amazon SageMaker responsible to all of the machine learning implemented with a docker container.
3. The input data would came from Amazon S3 bucket and the trained model would be saved on the same bucket.
4. At the end each SageMaker would trigger the lambda function so that an API could output the info to the client application.

